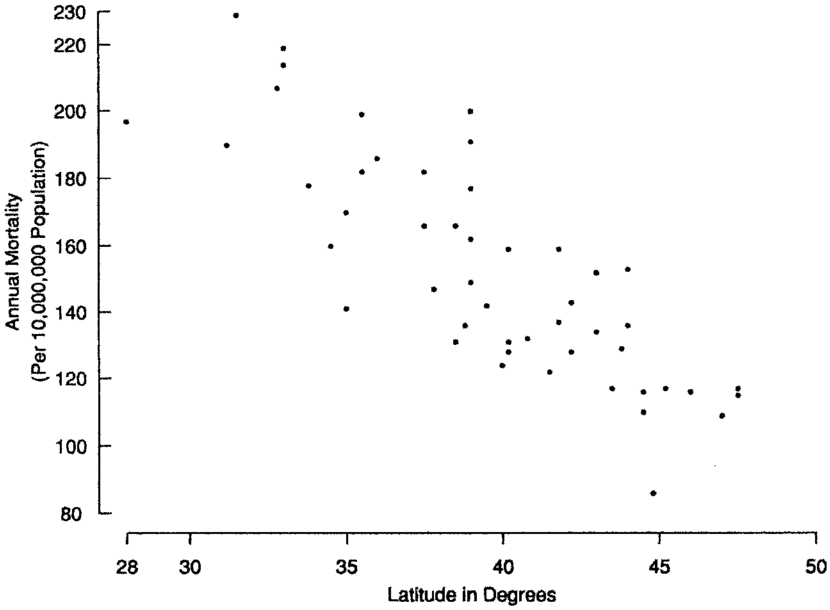

Figure 9.1 Annual mortality (per 10,000,000 population) due to malignant melanoma of the skin for white males by state and latitude of the center of the state for the period 1950–1959.

Example 9.2. To assess physical conditioning in normal subjects, it is useful to know how much energy they are capable of expending. Since the process of expending energy requires oxygen, one way to evaluate this is to look at the rate at which they use oxygen at peak physical activity. To examine the peak physical activity, tests have been designed where a person runs on a treadmill. At specified time intervals, the speed at which the treadmill moves and the grade of the treadmill both increase. The person is then run systematically to maximum physical capacity. The maximum capacity is determined by the person, who stops when unable to go further. Data from Bruce et al. [1973] are discussed.

The oxygen consumption was measured in the following way. The patient's nose was blocked off by a clip. Expired air was collected from a silicone rubber mouthpiece fitted with a very low resistance valve. The valve was connected by plastic tubes into a series of evacuated neoprene balloons. The inlet valve for each balloon was opened for 60 seconds to sample the expired air. Measurements were made of the volumes of expired air, and the oxygen content was obtained using a paramagnetic analyzer capable of measuring the oxygen. From this, the rate at which oxygen was used in mm/min was calculated. Physical conditioning, however, is relative to the size of the person involved. Smaller people need less oxygen to perform at the same speed. On the other hand, smaller people have smaller hearts, so relatively, the same level of effort may be exerted. For this reason, the maximum oxygen content is normalized by body weight; a quantity, VO_2 MAX, is computed by looking at the volume of oxygen used per minute per kilogram of body weight. Of course, the effort expended to go further on the treadmill increases with the duration of time on the treadmill, so there should be some relationship between VO_2 MAX and duration on the treadmill. This relationship is presented below.

Other pertinent variables that are used in the problems and in additional chapters are recorded in Table 9.2, including the maximum heart rate during exercise, the subject's age, height, and weight. The 44 subjects listed in Table 9.2 were all healthy. They were classified as active if they usually participated at least three times per week in activities vigorous enough to raise a sweat.

Table 9.2 Exercise Data for Healthy Active Males

Case	Duration (s)	$\text{VO}_{2\ \text{MAX}}$	Heart Rate (beats/min)	Age	Height (cm)	Weight (kg)
1	706	41.5	192	46	165	57
2	732	45.9	190	25	193	95
3	930	54.5	190	25	187	82
4	900	60.3	174	31	191	84
5	903	60.5	194	30	171	67
6	976	64.6	168	36	177	78
7	819	47.4	185	29	174	70
8	922	57.0	200	27	185	76
9	600	40.2	164	56	180	78
10	540	35.2	175	47	180	80
11	560	33.8	175	46	180	81
12	637	38.8	162	55	180	79
13	593	38.9	190	50	161	66
14	719	49.5	175	52	174	76
15	615	37.1	164	46	173	84
16	589	32.2	156	60	169	69
17	478	31.3	174	49	178	78
18	620	33.8	166	54	181	101
19	710	43.7	184	57	179	74
20	600	41.7	160	50	170	66
21	660	41.0	186	41	175	75
22	644	45.9	175	58	173	79
23	582	35.8	175	55	160	79
24	503	29.1	175	46	164	65
25	747	47.2	174	47	180	81
26	600	30.0	174	56	183	100
27	491	34.1	168	82	183	82
28	694	38.1	164	48	181	77
29	586	28.7	146	68	166	65
30	612	37.1	156	54	177	80
31	610	34.5	180	56	179	82
32	539	34.4	164	50	182	87
33	559	35.1	166	48	174	72
34	653	40.9	184	56	176	75
35	733	45.4	186	45	179	75
36	596	36.9	174	45	179	79
37	580	41.6	188	43	179	73
38	550	22.7	180	54	180	75
39	497	31.9	168	55	172	71
40	605	42.5	174	41	187	84
41	552	37.4	166	44	185	81
42	640	48.2	174	41	186	83
43	500	33.6	180	50	175	78
44	603	45.0	182	42	176	85

Source: Data from Bruce et al. [1973].

The duration of the treadmill exercise and $\text{VO}_{2\ \text{MAX}}$ data are presented in Figure 9.2. In this scattergram, we see that as the treadmill time increases, by and large, the $\text{VO}_{2\ \text{MAX}}$ increases. There is, however, some variability. The increase is not an infallible rule. There are subjects who run longer but have less oxygen consumption than someone else who has exercised for a shorter time period. Because of the expense and difficulty in collecting the expired air volumes,

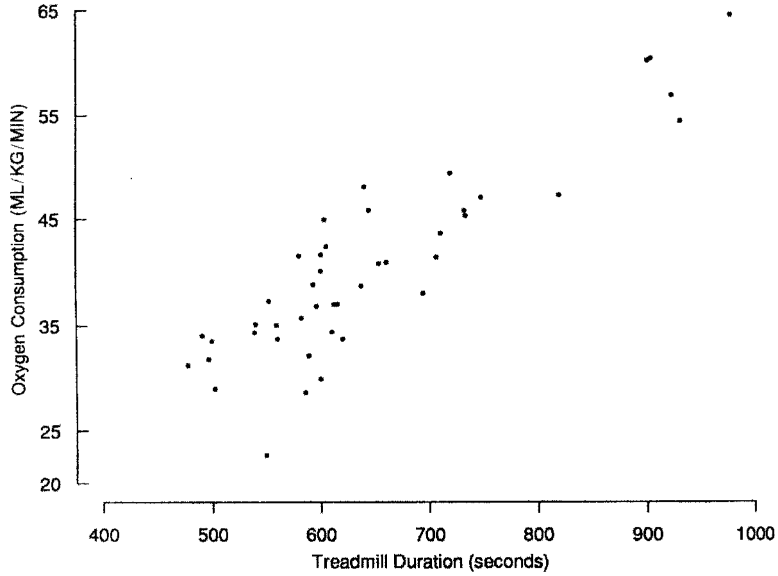

Figure 9.2 Oxygen consumption vs. treadmill duration.

it is useful to evaluate oxygen consumption and conditioning by having the subjects run on the treadmill and recording the duration. As we can see from Figure 9.2, this would not be a perfect solution to the problem. Duration would not totally determine the VO_2 MAX level. Nevertheless, it would give us considerable information. When we do this, how should we predict what the VO_2 MAX level would be from the duration? Clearly, such a predictive equation should be developed from the data at hand. When we do this, we want to characterize the accuracy of such predictions and succinctly summarize the relationship between the two variables.

Example 9.3. Dern and Wiorkowski [1969] collected data dealing with the erythrocyte adenosine triphosphate (ATP) levels in youngest and older sons in 17 families. The purpose of the study was to determine the effect of storage of the red blood cells on the ATP level. The level is important because it determines the ability of the blood to carry energy to the cells of the body. The study found considerable variation in the ATP levels, even before storage. Some of the variation could be explained on the basis of variation by family (genetic variation). The data for the oldest and youngest sons are extracted from the more complete data set in the paper. Table 9.3 presents the data for 17 pairs of brothers along with the ages of the brothers.

Figure 9.3 is a scattergram of the values in Table 9.3. Again, there appears to be some relationship between the two values, with both brothers tending to have high or low values at the same time. Again, we would like to consider whether or not such variability might occur by chance. If chance is not the explanation, how could we summarize the pattern of variation for the pairs of numbers?

The three scattergrams have certain features in common:

1. Each scattergram refers to a situation where two quantities are associated with each experimental unit. In the first example, the melanoma rate for the state and the latitude of the state are plotted. The state is the individual unit. In the second example, for each person studied on the treadmill, VO_2 MAX vs. the treadmill time in seconds was plotted. In the third example, the experimental unit was the family, and the ATP values of the youngest and oldest sons were plotted.

Table 9.3 Erythrocyte Adenosine Triphosphate (ATP) Levels[a] in Youngest and Oldest Sons in 17 Families Together with Age (Before Storage)

	Youngest		Oldest	
Family	Age	ATP Level	Age	ATP Level
1	24	4.18	41	4.81
2	25	5.16	26	4.98
3	19	4.85	27	4.48
4	28	3.43	32	4.19
5	22	4.53	25	4.27
6	7	5.13	23	4.87
7	21	4.10	24	4.74
8	17	4.77	25	4.53
9	25	4.12	26	3.72
10	24	4.65	25	4.62
11	12	6.03	25	5.83
12	16	5.94	24	4.40
13	9	5.99	22	4.87
14	18	5.43	24	5.44
15	14	5.00	26	4.70
16	24	4.82	26	4.14
17	20	5.25	24	5.30

Source: Data from Dern and Wiorkowski [1969].

[a] ATP levels expressed as micromoles per gram of hemoglobin.

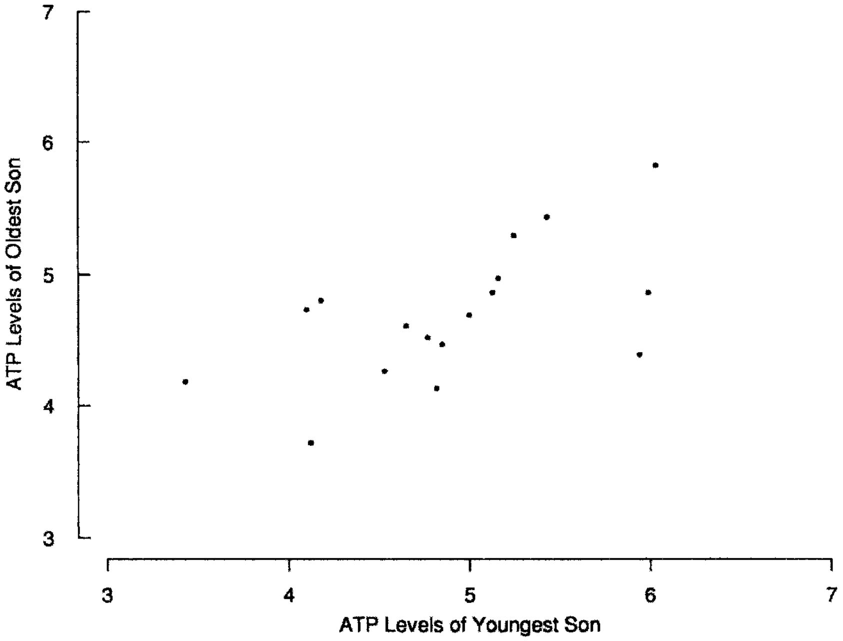

Figure 9.3 ATP levels (μmol/g of hemoglobin) of youngest and oldest sons in 17 families. (Data from Dern and Wiorkowski [1969].)

2. In each of the three diagrams, there appears to be a rough trend or association between the variables. In the melanoma rate date, as the latitude increases, the melanoma rate tends to decrease. In the treadmill data, as the duration on the treadmill increased, the $\mathrm{VO}_{2\ \mathrm{MAX}}$ also increased. In the ATP data, both brothers tended to have either a high or a low value for ATP.
3. Although increasing and decreasing trends were evident, there was not a one-to-one relationship between the two quantities. It was not true that every state with a higher latitude had a lower melanoma rate in comparison with a state at a lower latitude. It was not true that in each case when individual A ran on the treadmill a longer time than individual B that individual A had a higher $\mathrm{VO}_{2\ \mathrm{MAX}}$ value. There were some pairs of brothers for which one pair did not have the two highest values when compared to the other pair. This is in contrast to certain physical relationships. For example, if one plotted the volume of a cube as a function of the length of a side, there is the one-to-one relationship: the volume increases as the length of the side increases. In the data we are considering, there is a rough relationship, but there is still considerable variability or scatter.
4. To effectively use and summarize such scattergrams, there is a need for a method to quantitate how much of a change the trends represent. For example, if we consider two states where one has a latitude 5° south of the other, how much difference is expected in the melanoma rates? Suppose that we train a person to increase the duration of treadmill exercise by 70 seconds; how much of a change in $\mathrm{VO}_{2\ \mathrm{MAX}}$ capacity is likely to occur?
5. Suppose that we have some method of quantitating the overall relationship between the two variables in the scattergram. Since the relationship is not precisely one to one, there is a need to summarize how much of the variability the relationship explains. Another way of putting this is that we need a summary quantity which tells us how closely the two variables are related in the scattergram.
6. If we have methods of quantifying these things, we need to know whether or not any estimated relationships might occur by chance. If not, we still want to be able to quantify the uncertainty in our estimated relationships.

The remainder of this chapter deals with the issues we have just raised. In the next section we use a linear equation (a straight line) to summarize the relationship between two variables in a scattergram.

9.2 SIMPLE LINEAR REGRESSION MODEL

9.2.1 Summarizing the Data by a Linear Relationship

The three scattergrams above have a feature in common: the overall relationship is roughly linear; that is, a straight line that characterizes the relationships between the two variables could be placed through the data. In this and subsequent chapters, we look at linear relationships. A linear relationship is one expressed by a linear equation. For variables $U, V, W, \ldots,$ and constants $a, b, c, \ldots,$ a linear equation for Y is given by

$$Y = a + bU + cV + dW + \cdots$$

In the scattergrams for the melanoma data and the exercise data, let X denote the variable on the horizontal axis (*abscissa*) and Y be the notation for the variable on the vertical axis (*ordinate*). Let us summarize the data by fitting the straight-line equation $Y = a + bX$ to the data. In each case, let us think of the X variable as predicting a value for Y. In the first two examples, that would mean that given the latitude of the state, we would predict a value for the melanoma rate; given the duration of the exercise test, we would predict the VO_2 MAX value for each subject.

There is terminology associated with this procedure. The variable being predicted is called the *dependent variable* or *response variable*; the variable we are using to predict is called the *independent variable*, the *predictor variable*, or the *covariate*. For a particular value, say, X_i of the predictor variable, our value predicted for Y is given by

$$\widehat{Y_i} = a + bX_i \tag{1}$$

The fit of the values predicted to the values observed (X_i, Y_i) may be summarized by the difference between the value Y_i observed and the value $\widehat{Y_i}$ predicted. This difference is called a *residual value*:

$$\text{residual value} = y_i - \widehat{y_i} = \text{value observed} - \text{value predicted} \tag{2}$$

It is reasonable to fit the line by trying to make the residual values as small as possible. The *principle of least squares* chooses a and b to minimize the sum of squares of the residual values. This is given in the following definition:

Definition 9.2. Given data $(x_i, y_i), i = 1, 2, \ldots, n$, the *least squares fit* to the data chooses a and b to minimize

$$\sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

where $\widehat{y_i} = a + bx_i$.

The values a and b that minimize the sum of squares are described below. At this point, we introduce some notation similar to that of Section 7.3:

$$[y^2] = \sum_i (y_i - \overline{y})^2$$

$$[x^2] = \sum_i (x_i - \overline{x})^2$$

$$[xy] = \sum_i (x_i - \overline{x})(y_i - \overline{y})$$

We decided to choose values a and b so that the quantity

$$\sum_i (y_i - \widehat{y_i})^2 = \sum_i (y_i - a - bx_i)^2$$

is minimized. It can be shown that the values for a and b that minimize the quantity are given by

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{[xy]}{[x^2]}$$

and

$$a = \overline{y} - b\overline{x}$$

Note 9.4 gives another equivalent formula for b that emphasizes its role as a summary statistic of the slope of the X–Y relationship.

Table 9.4 Predicted Mortality Rates by Latitude for the Data of Table 9.1[a]

Latitude (x)	Predicted Mortality (y)	s_1	s_2	s_3
30	209.9	19.12	6.32	20.13
35	180.0	19.12	3.85	19.50
39.5 (mean)	152.9 (mean)	19.12	2.73	19.31
40	150.1	19.12	2.74	19.31
45	120.2	19.12	4.26	19.58
50	90.3	19.12	6.83	20.30

[a] For the quantities s_2 and s_3, see Section 9.2.3.

For the melanoma data, we have the following quantities:

$$\overline{x} = 39.533, \qquad \overline{y} = 152.878$$

$$\sum_i (x_i - \overline{x})(y_i - \overline{y}) = [xy] = -6100.171$$

$$\sum_i (x_i - \overline{x})^2 = [x^2] = 1020.499$$

$$\sum_i (y_i - \overline{y})^2 = [y^2] = 53,637.265$$

The least squares slope b is

$$b = \frac{-6100.171}{1020.499} = -5.9776$$

and the least squares intercept a is

$$a = 152.878 - (-5.9776 \times 39.533) = 389.190$$

Figure 9.4 presents the melanoma data with the line of least squares fit drawn in. Because of the method of selecting the line, the line goes through the data, of course. The least squares line always has the property that it goes through the point in the scattergram corresponding to the sample mean of the two variables. The sample means of the variables are located by the intersection of dotted lines. Further, the point for Tennessee is detailed in the box in the lower left-hand corner. The value predicted from the equation was 174, whereas the actual melanoma rate for this state was 186. Thus, the residual value is the difference, 12. We see that the value predicted, 174, is closer to the value observed than to the overall Y mean, which is 152.9.

For the melanoma data, the line of least squares fit is $Y = 389.19 - 5.9776X$. For each state's observed mortality rate, there is then a predicted mortality rate based on knowledge of the latitude. Some predicted values are listed in Table 9.4. The farther north the state, the lower the mortality due to malignant melanoma; but now we have quantified the change.

Note that the predicted mortality at the mean latitude (39.5°) is exactly the mean value of the mortalities *observed*; as noted above, the regression line goes through the point $(\overline{x}, \overline{y})$.

9.2.2 Linear Regression Models

With the line of least squares fit, we shall associate a mathematical model. This *linear regression model* takes the predictor or covariate observation as being fixed. Even if it is sampled at random,

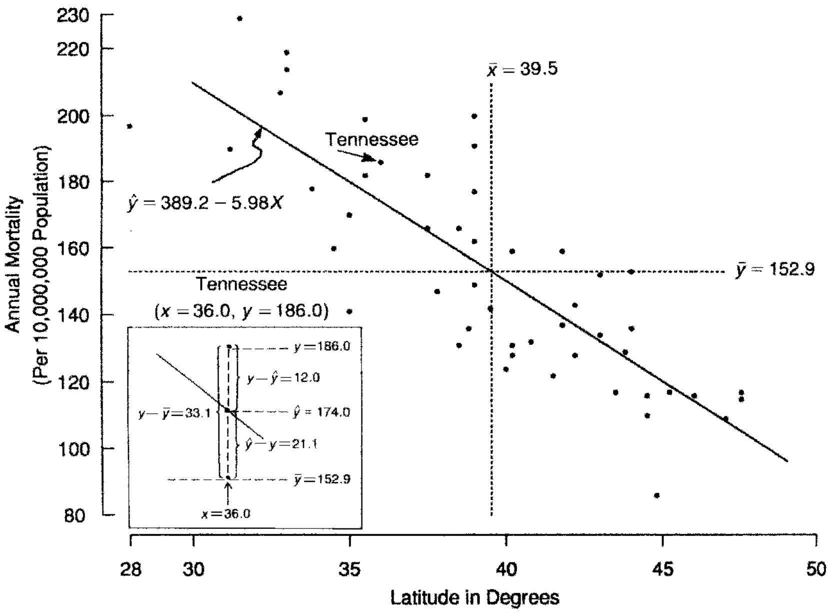

Figure 9.4 Annual mortality (per 10,000,000 population) due to malignant melanoma of the skin for white males by state and latitude of the center of the state for the period 1950–1959 (least squares regression line is given).

the analysis is conditional upon knowing the value of X. In the first example above, the latitude of each state is fixed. In the second example, the healthy people may be considered to be a representative—although not random—sample of a larger population; in this case, the duration may be considered a random quantity. In the linear regression analysis of this chapter, we know X and are interested in predicting the value of Y. The regression model assumes that for a fixed value of X, the expected value of Y is some function. In addition to this expected value, a random error term is added. It is assumed that the error has a mean value of zero. We shall restrict ourselves to situations where the expected value of Y for known X is a linear function. Thus, our linear regression model is the following:

$$\text{expected value of } Y \text{ knowing } X = E(Y|X) = \alpha + \beta X$$

$$Y = \alpha + \beta X + e, \qquad \text{where } e \text{ (error) has } E(e) = 0$$

The parameters α and β are population parameters. Given a sample of observations, the estimates a and b that we found above are estimates of the population parameters. In the mortality rates of the states, the random variability arises both because of the randomness of the rates in a given year and random factors associated with the state, other than latitude. These factors make the observations during a particular time period reasonably modeled as a random quantity. For the exercise test data, we may consider the normal subjects tested as a random sample from a population of active normal males who might have been tested.

Definition 9.3. The line $E(Y|X) = \alpha + \beta X$ is called the *population regression line*. Here, $E(Y|X)$ is the expected value of Y at X (assumed known). The coefficients α and β are called *population regression coefficients*. The line $Y = a + bX$ is called the *estimated regression line*, and a and b are called *estimated regression coefficients*. The term *estimated* is often dropped, and *regression line* and *regression coefficients* are used for these estimated quantities.

For each X, $E(Y|X)$ is the mean of a population of observations. On the left of Figure is shown a linear regression situation; on the right, the regression $E(Y|X)$ is not linear.

To simplify statistical inference, another assumption is often added: that the error term is normally distributed with mean zero and variance σ_1^2. As we saw with the t-test, the assumption of normality is important for testing and confidence interval estimation only in fairly small samples. In larger samples the central limit theorem replaces the need for distributional assumptions. Note that the variance of the error term is *not* the variance of the Y variable. It is the variance of the Y variable *when* the value of the X variable is known.

Given data, the variance σ_1^2 is estimated by the quantity $s_{y\cdot x}^2$, where this quantity is defined as

$$s_{y\cdot x}^2 = \sum \frac{(Y_i - \widehat{Y}_i)^2}{n-2}$$

Recall that the usual sample variance was divided by $n-1$. The $n-2$ occurs because two parameters, α and β, are estimated in fitting the data rather than one parameter, the sample mean, that was estimated before.

9.2.3 Inference

We have the model

$$Y = \alpha + \beta X + e, \qquad \text{where } e \sim N(0, \sigma_1^2)$$

On the basis of n pairs of observations we presented estimates a and b of α and β, respectively. To test hypotheses regarding α and β, we need to assume the normality of the term e.

The left panel of Figure 9.5 shows a situation where these assumptions are satisfied. Note that:

1. $E(Y|X)$ is linear.
2. For each X, the normal Y-distribution has the same variance.
3. For each X, the Y-distribution is normal (less important as the sample size is large).

The right panel of Figure 9.5 shows a situation where all these assumptions don't hold.

1. $E(Y|X)$ is not a straight line; it curves.
2. The variance of Y increases as X increases.
3. The distribution becomes more highly skewed as X increases.

It can be shown, under the correct normal model or in large samples, that

$$b \sim N\left(\beta, \frac{\sigma_1^2}{[x^2]}\right) \quad \text{and} \quad a \sim N\left(\alpha, \sigma_1^2\left[\frac{1}{n} + \frac{\overline{x}^2}{[x^2]}\right]\right)$$

Recall that σ_1^2 is estimated by $s_{y.x}^2 = \sum(Y_i - \widehat{Y}_i)^2/(n-2)$. Note that the divisor is $n-2$: the number of degrees of freedom. The reason, as just mentioned, is that now *two* parameters are estimated: α and β. Given these facts, we can now either construct confidence intervals or tests of hypotheses after constructing appropriate pivotal variables:

$$\frac{b-\beta}{\sigma_1/\sqrt{[x^2]}} \sim N(0,1), \qquad \frac{b-\beta}{s_{y.x}/\sqrt{[x^2]}} \sim t_{n-2}$$

and similar terms involving the intercept a are discussed below.

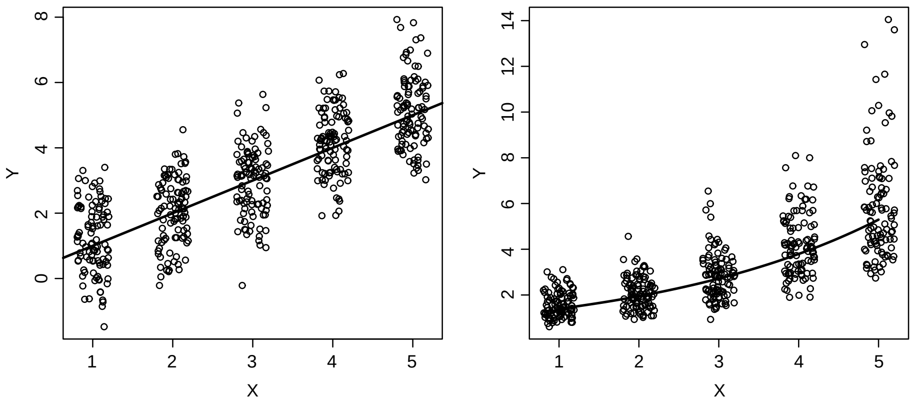

Figure 9.5 Linear regression assumptions and violations. On the left, the expected values of Y for given X values fall on a straight line. The variation about the line has the same variance at each X. On the right, the expected values fall on a curve, not a straight line. The distribution of Y is different for different X values, with variance and skewness increasing with X.

Returning to Example 9.1, the melanoma data by state, the following quantities are known or can be calculated:

$$a = 389.190, \qquad s_{y\cdot x}^2 = \sum_i \frac{(Y_i - \widehat{Y}_i)^2}{n-2} = \frac{17{,}173.1}{47} = 365.3844$$

$$b = -5.9776, \qquad [x^2] = 1020.499, \qquad s_{y.x} = 19.1150$$

On the assumption that there is no relationship between latitude and mortality, that is, $\beta = 0$, the variable b has mean zero. A t-test yields

$$t_{47} \doteq \frac{-5.9776}{19.1150/\sqrt{1020.499}} \doteq \frac{-5.9776}{0.59837} \doteq -9.99$$

From Table A.4, the critical value for a t-variable with 47 degrees of freedom, at the 0.0001 level (two-tailed) is approximately 4.25; hence, the hypothesis is rejected and we conclude that there is a relationship between latitude and mortality; the mortality increases about 6.0 persons per 10,000,000 for every degree farther south. This, of course, comes from the value of $b = -5.9776 \doteq -6.0$. Similarly, a 95% confidence interval for β can be constructed using the t-value of 2.01, and the standard error of the slope, $0.59837 = s_{y\cdot x}/\sqrt{[x^2]}$.

A 95% confidence interval is $-5.9776 \pm (2.01 \times 0.59837)$, producing lower and upper limits of -7.18 and -4.77, respectively. Again, the confidence interval does not include zero, and the same conclusion is reached as in the case of the hypothesis test.

The inference has been concerned with the slope β and intercept α up to now. We now want to consider two additional situations:

1. Inference about population means, $\alpha + \beta X$, for a fixed value of X
2. Inference about a future observation at a fixed value of X

To distinguish between the two cases, let $\widehat{\mu}_x$ and $\widehat{y}_x$ be the predicted mean and a new random single observation at the point x, respectively. It is important to note that for inference about a future observation the normality assumption is critical even in large samples. This is in contrast to inference about the predicted mean or about a and b, where normal distributions are required only in small samples and the central limit theorem substitutes in large samples.

First, then, inference about the population mean at a fixed X value: It is natural to estimate $\alpha + \beta X$ by $a + bx$; the predicted value of Y at the value of $X = x$. Rewrite this quantity as

$$\widehat{\mu}_x = \overline{y} + b(x - \overline{x})$$

It can be shown that $\overline{y}$ and b are statistically independent so that the variance of the quantity is

$$\begin{aligned}\text{var}[\overline{y} + b(x - \overline{x})] &= \text{var}(\overline{y}) + (x - \overline{x})^2 \text{ var}(b)\\ &= \frac{\sigma_1^2}{n} + (x - \overline{x})^2 \frac{\sigma_1^2}{[x^2]}\\ &= \sigma_1^2\left[\frac{1}{n} + \frac{(x - \overline{x})^2}{[x^2]}\right] = \sigma_2^2, \qquad \text{say}\end{aligned}$$

Tests and confidence intervals for $E(Y|X)$ at a fixed value of x may be based on the t-distribution.

The quantity σ_2^2 reduces to the variance for the intercept, a, at $X = 0$. It is useful to study this quantity carefully; there are important implications for design (see Note 9.3). The variance, σ_2^2, is not constant but depends on the value of x. The more x differs from $\overline{x}$, the greater the contribution of $(x - \overline{x})^2/[x^2]$ to the variance of $a + bx$. The contribution is zero at $x = \overline{x}$. At $x = \overline{x}$, $y = \overline{y}$ the slope is not used. Regardless of the slope the line goes through mean point $(\overline{X}, \overline{Y})$. Consider Example 9.1 again. We need the following information:

$$\begin{aligned} s_{y\cdot x} &= 19.1150\\ n &= 49\\ \overline{x} &= 39.533\\ [x^2] &= 1020.499\end{aligned}$$

Let

$$s_2^2 = s_{y\cdot x}^2\left[\frac{1}{n} + \frac{(x - \overline{x})^2}{[x^2]}\right]$$

That is, s_2^2 estimates σ_2^2. Values of s_2 as related to latitude are given in Table 9.4. Confidence interval bands for the mean, $\alpha + \beta X$ (at the 95% level), are given in Figure 9.6 by the narrower bands. The curvature is slight due to the large value of $[x^2]$ and the relatively narrow range of prediction.

We now turn to the second problem: predicting a future observation on the basis of the observed data. The variance is given by

$$s_3^2 = s_{y\cdot x}^2\left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{[x^2]}\right]$$

This is reasonable in view of the following argument: At the point $\alpha + \beta X$ an observation has variance σ_2^2 (estimated by $s_{y\cdot x}^2$). In addition, there is uncertainty in the true value $\alpha + \beta X$. This adds variability to the estimate. A future observation is assumed to be independent of past observations. Hence the variance can be added and the quantity s_3^2 results when σ_1^2 is estimated by $s_{y\cdot x}^2$. Confidence interval bands for future observations (95% level) are represented by outer lines in Figure 9.6. This band means that we are 95% certain that the next observation at the fixed point x will be within the given bands. Note that the curvature is not nearly as marked.

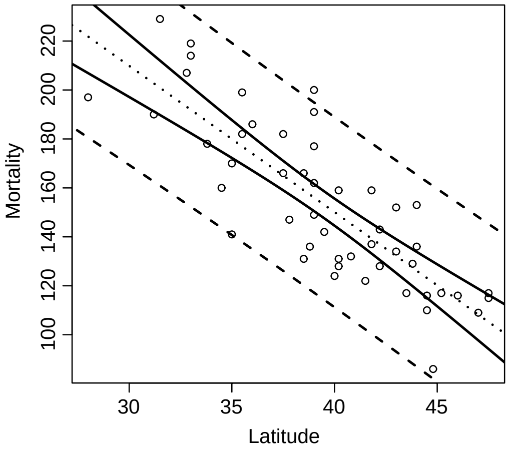

Figure 9.6 Data of Figure 9.1: 95% confidence bands for population means (solid) and 95% confidence bands for a future observation (dashed).

9.2.4 Analysis of Variance

Consider Example 9.1 and the data for Tennessee, as graphed in Figure 9.4. The basic data for this state are (omitting subscripts)

$$y = 186.0 = \text{observed mortality}$$

$$x = 36.0 = \text{latitude of center of state}$$

$$\widehat{y} = 174.0 = \text{predicted mortality using latitude of 36.0}$$

$$\overline{y} = 152.9 = \text{average mortality for United States}$$

Partition the data as follows:

$$\begin{aligned}(y - \overline{y}) &= (\widehat{y} - \overline{y}) + (y - \widehat{y})\\ \text{total variation} &= \text{attributable to regression} + \text{residual from regression}\\ 186.0 - 152.9 &= (174.0 - 152.9) \qquad + (186.0 - 174.0)\\ 33.1 &= 21.1 \qquad + 12.0\end{aligned}$$

Note that the quantity

$$\begin{aligned}\widehat{y} - \overline{y} &= a + bx - \overline{y}\\ &= \overline{y} - b\overline{x} + bx - \overline{y}\\ &= b(x - \overline{x})\end{aligned}$$

The quantity is zero if $b = 0$, that is, if there is no regression relationship between Y and X. In addition, it is zero if prediction is made at the point $x = \overline{x}$.

These quantities can be calculated for each state, as indicated in abbreviated form in Table 9.5. The sums of squares of these quantities are given at the bottom of the table. The remarkable fact is that

$$\begin{aligned}\sum(y_i - \overline{y})^2 &= \sum(\widehat{y}_i - \overline{y})^2 + \sum(y_i - \widehat{y}_i)^2\\ 53{,}637.3 &= 36{,}464.2 \qquad + 17{,}173.1\end{aligned}$$

Table 9.5 Deviations from Mean and Regression Based on Data of Table 9.1

					Variation				
		Observed							
		Mortality	Latitude	Predicted	Total	=	Regression	+	Residual
Case	State	(y)	(x)	Mortality[a]	$y - \overline{y}$	=	$\widehat{y} - \overline{y}$	+	$y - \widehat{y}$
1	Alabama	219.0	33.0	191.9	66.1	=	39.0	+	27.1
2	Arizona	160.0	34.5	183.0	7.1	=	30.1	+	−23.0
⋮	⋮		⋮				⋮		
41	Tennessee	186.0	36.0	174.0	33.1	=	21.1	+	12.0
⋮	⋮		⋮				⋮		
48	Wisconsin	110.0	44.5	123.2	−42.9	=	−29.7	+	−13.2
49	Wyoming	134.0	43.0	132.2	−18.9	=	−20.7	+	1.8
Total					0	=	0	+	0
Mean		152.9	39.5	152.9	0	=	0	+	0
Sum of squares					53,637.3	=	36,464.2	+	17,173.1

[a] Predicted mortality based on regression line $y = 389.19 - 5.9776x$, where x is the latitude at the center of the state.

that is, the total variation as measured by $\sum(y_i - \overline{y})^2$ has been partitioned additively into a part attributable to regression and the residual from regression. The quantity $\sum(\widehat{y}_i - \overline{y})^2 = \sum b^2(x_i - \overline{x})^2 = b^2[x^2]$. (But since $b = [xy]/[x^2]$, this becomes $\sum(\widehat{y}_i - \overline{y})^2 = [xy]^2/[x^2]$.) Associated with each sum of squares is a degree of freedom (d.f.) which can also be partitioned as follows:

$$\text{total variation} = \text{attributable to regression} + \text{residual variation}$$

$$\text{d.f.} = n - 1 = 1 + n - 2$$

$$49 = 1 + 48$$

The total variation has $n - 1$ d.f., not n, since we adjusted Y about the mean $\overline{Y}$. These quantities are commonly entered into an analysis of variance table as follows:

Source of Variation	**d.f.**	**SS**	**MS**	***F*-Ratio**
Regression	1	36,464.2	36,464.2	99.80
Residual	47	17,173.1	365.384	
Total	48	53,637.3		

The quantity 365.384 is precisely $s^2_{y\cdot x}$. The F-ratio is discussed below. The mean square is the sum of squares divided by the degrees of freedom. The analysis of variance table of any set of n pairs of observations (x_i, y_i), $i = 1, \ldots, n$, is

Source of Variation	**d.f.**	**SS**	**MS**	***F*-Ratio**
Regression	1	$[xy]^2/[x^2]$	$[xy]^2/[x^2]$	$\dfrac{[xy]^2/[x^2]}{s^2_{y\cdot x}}$
Residual	n - 2	By subtraction	$s^2_{y\cdot x}$	
Total	n - 1	$[y^2]$		

Several points should be noted about this table and the regression procedure:

1. Only five quantities need to be calculated from the raw data to completely determine the regression line and sums of squares: $\sum x_i$, $\sum y_i$, $\sum x_i^2$, $\sum y_i^2$, and $\sum x_i y_i$. From these quantities one can calculate

$$[x^2] = \sum (x_i - \overline{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$[y^2] = \sum (y_i - \overline{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

$$[xy] = \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{\sum y_i \sum x_i}{n}.$$

2. The greater the slope, the greater the SS due to regression. That is,

$$\text{SS(regression)} = b^2 \sum (x_i - \overline{x})^2 = \frac{[xy]^2}{[x^2]}$$

If the slope is "negligible," SS(regression) will tend to be "small."

3. The proportion of the total variation attributable to regression is usually denoted by r^2; that is,

$$\begin{aligned} r^2 &= \frac{\text{variation attributable to regression}}{\text{total variation}} \\ &= \frac{[xy]^2/[x^2]}{[y^2]} \\ &= \frac{[xy]^2}{[x^2][y^2]} \end{aligned}$$

It is clear that $0 \leq r^2 \leq 1$ (why?). If $b = 0$, then $[xy]^2/[x^2] = 0$ and the variation attributable to regression is zero. If $[xy]^2/[x^2]$ is equal to $[y^2]$, all of the variation can be attributed to regression; to be more precise, to *linear* regression; that is, all the observations fall on the line $a + bx$. Thus, r^2 measures the degree of *linear* relationship between X and Y. The *correlation coefficient*, r, is studied in Section 9.3. For the data in Table 9.4,

$$r^2 = \frac{36{,}464.2}{53{,}637.3} = 0.67983$$

That is, approximately 68% of the variation in mortality can be attributed to variation in latitude. Equivalently, the variation in mortality can be reduced 68% knowing the latitude.

4. Now consider the ratio

$$F = \frac{[xy]^2/[x^2]}{s_{y\cdot x}^2}$$

Under the assumption of the model [i.e., $y \sim N(\alpha + \beta X, \sigma_1^2)$], the ratio F tends to be near 1 if $\beta = 0$ and tends to be larger than 1 if $\beta \neq 0$ (either positively or negatively). F has the F-distribution, as introduced in Chapter 5. In the example $F_{1,47} = 99.80$, the critical value at the 0.05 level is $F_{1,47} = 4.03$ (by interpolation). The critical value at the 0.001 level is $F_{1,47} = 12.4$ (by interpolation). Hence, we reject the hypotheses that $\beta = 0$. We tested the significance of the slope using a t-test given the value

$$t_{47} = -9.9898$$

The F-value we obtained was

$$F_{1,47} = 99.80$$

In fact,

$$(-9.9898)^2 = 99.80$$

That is,

$$t_{47}^2 = F_{1,47}$$

Recall that

$$t_v^2 = F_{1,v}$$

Thus, the t-test and the F-test for the significance of the slope are equivalent.

9.2.5 Appropriateness of the Model

In Chapter 5 we considered the appropriateness of the model $y \sim N(\mu, \sigma^2)$ for a set of data and discussed briefly some ways of verifying the appropriateness of this model. In this section we have the model

$$y \sim N(\alpha + \beta X, \sigma_1^2)$$

and want to consider its validity. At least three questions can be asked:

1. Is the relationship between Y and X linear?
2. The variance σ_1^2 is assumed to be constant for all values of X (homogeneity of variable). Is this so?
3. Does the normal model hold?

Two very simple graphical procedures, both utilizing the residuals from regression $y_i - \widehat{y}_i$, can be used to verify the assumptions above. Also, one computation on the residuals is useful. The two graphical procedures are considered first.

To Check for:	Graphical Procedure
1. Linearity of regression and homogeneity of variance	Plot $(y_i - \widehat{y}_i)$ against $\widehat{y}_i$, $i = 1, \dots, n$
2. Normality	Normal probability plot of $y_i - \widehat{y}_i, i = 1, \dots, n$

We illustrate these with data created by Anscombe [1973]. As we noted above, just five summaries of the data specify everything about the linear regression model. Anscombe created four data sets in which these five summaries, and thus the fitted model, were identical, but where the data were very different. Only one of these sets of data is appropriate for linear regression.

Linearity of Regression and Homogeneity of Variance

Given only one predictor variable, X, the graph of Y vs. X will suggest nonlinearity or heterogeneity of variance, see the top row of regression patterns in Figure 9.7. But if there is more than one predictor variable, as in Chapter 11, the simple two-dimensional graph is not possible. But there is a way of detecting such patterns by considering residual plots $y - \widehat{y}$ against a variety of variables. A common practice is to plot $y - \widehat{y}$ against $\widehat{y}$; this graph is usually referred to as a *residual plot*. The advantage is, of course, that no matter how many predictor variables are used,

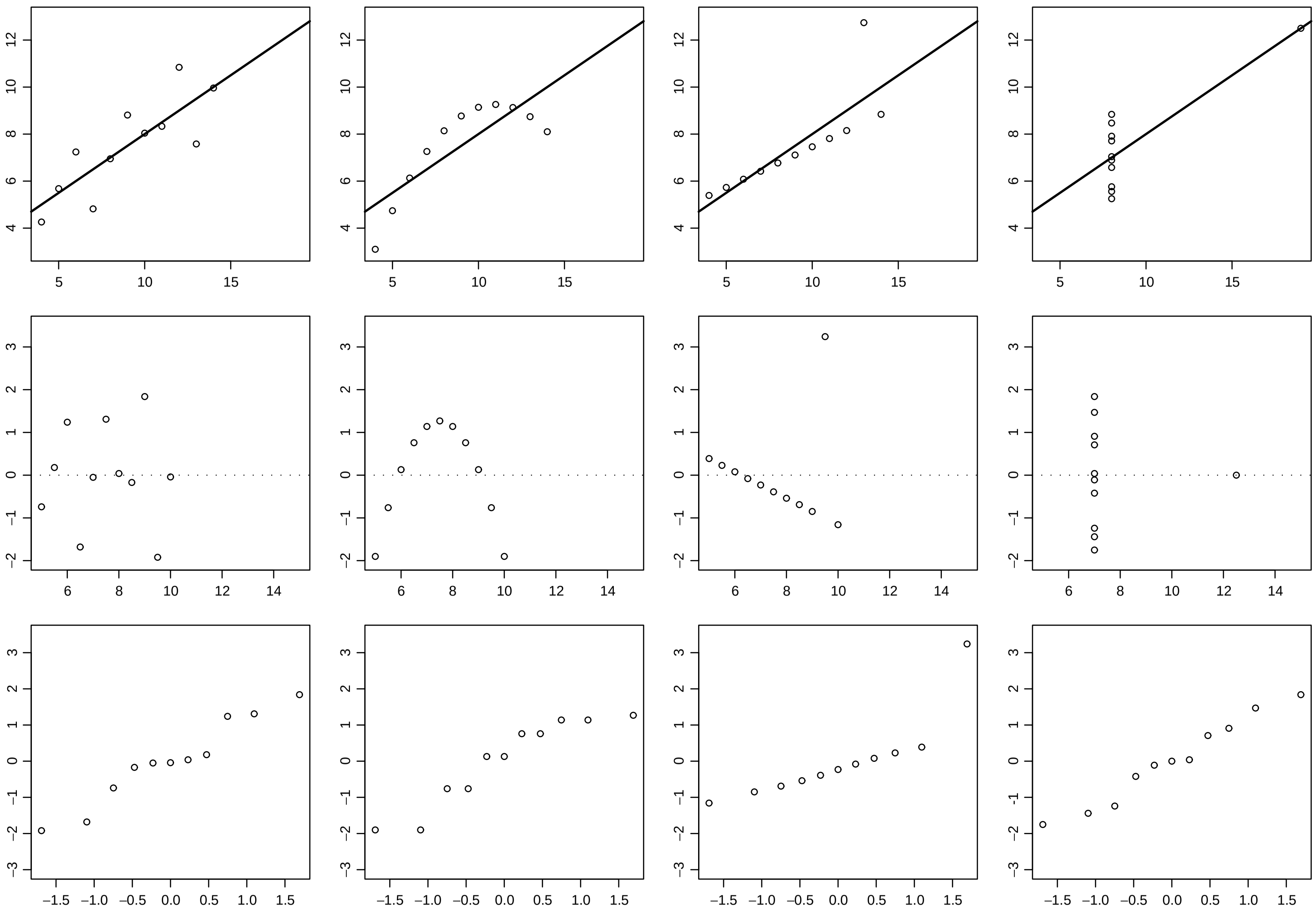

Figure 9.7 Regression patterns (Y vs. X), residuals patterns ($y - \widehat{y}$ vs. $\widehat{y}$) and normal probability plot of residuals ($y - \widehat{y}$).

it is always possible to plot $y - \widehat{y}$. The second row of graphs in Figure 9.7 indicate the residual patterns associated with the regression patterns of the top row. Pattern 1 indicates a reasonable linear trend, pattern 2 shows a very strong pattern in the residuals. Pattern 3 has a single very large residual, and in pattern 4 it is the distribution of X rather than Y that is suspicious.

Before turning to the questions of normality of the data, consider the same kind of analysis carried out on the melanoma data. The residuals are plotted in the left panel of Figure 9.8. There is no evidence that there is nonlinearity or heterogeneity of variance.

Normality

One way of detecting gross deviations from normality is to graph the residuals from regression against the expected quantiles of a normal distribution as introduced in Chapter 4. The last row of patterns in Figure 9.6 are the normal probability plots of the deviations from linear regression. The last row in Figure 9.6 indicates that a normal probability plot indicates outliers clearly but is not useful in detecting heterogeneity of variance or curvilinearity.

Of particular concern are points not fit closely by the data. The upper right and lower left points often tail in toward the center in least squares plot. Points on the top far to the right and on the bottom far to the left (as in pattern 2) are of particular concern.

The normal probability plot associated with the residuals of the melanoma are plotted in the right panel of Figure 9.8. There is no evidence against the normality assumption.

9.2.6 Two-Sample t-Test as a Regression Problem

In this section we show the usefulness of the linear model approach by illustrating how the two sample t-test can be considered a special kind of linear model. For an example, we again return to the data on mortality rates due to melanoma contained in Table 9.1. This time we consider the rates in relationship to contiguity to an ocean; there are two groups of states: those that border on an ocean and those that do not. The question is whether the average mortality rate for the first group differs from that of the second group. The t-test and analysis are contained in Table 9.6.

The mean difference, $\overline{y}_1 - \overline{y}_2 = 31.486$, has a standard error of 8.5468 so that the calculated t-value is $t = 3.684$ with 47 degrees of freedom, which exceeds the largest value in the t-table at 40 or 60 degrees of freedom and consequently, $p < 0.001$. The conclusion then is that the mortality rate due to malignant melanoma is appreciably higher in states contiguous to an ocean as compared to "inland" states, the difference being approximately 31 deaths per 10^7 population per year.

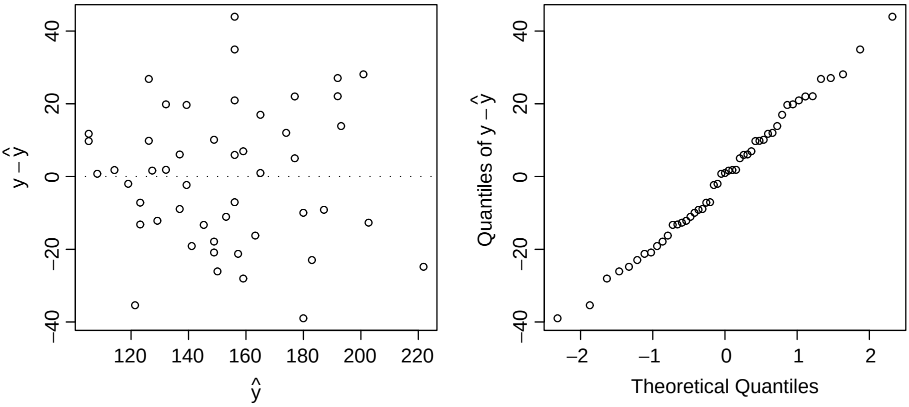

Figure 9.8 Melanoma data (left) residuals $(y - \widehat{y})$ from regression lines $Y = 389.19 - 589.8X$ plotted against $\widehat{y}$ and (right) normal quantile plot of residuals, $y - \widehat{y}$.

Table 9.6 Comparison by Two-Sample t-Test of Mortality Rates Due to Melanoma (Y) by Contiguity to Ocean

Contiguity to ocean	No = 0		Yes = 1
Number of states	$n_1 = 27$		$n_2 = 22$
Mean mortality	$\overline{y}_1 = 138.741$		$\overline{y}_2 = 170.227$
Variance[a]	$s_1^2 = 697.97$		$s_2^2 = 1117.70$
Pooled variance	$s_p^2 = 885.51$		
Standard error of difference		$s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 8.5468$	
Mean difference		$\overline{y}_2 - \overline{y}_1 = 31.487$	
t-Value		$t = 3.684$	
Degrees of freedom		d.f. = 47	
p-Value		$p < 0.001$	

[a] Subscripts on variances denote group membership in this table.

Now consider the following (equivalent) regression problem. Let Y be the mortality rate and X the predictor variable; "X = contiguity to ocean" and X takes on only two values, 0, 1. (For simplicity, we again label all the variables and parameters, Y, X, α, β, and σ_1^2, but except for Y, they obviously are different from the way they were defined in earlier sections.) The model is

$$Y \sim N(\alpha + \beta X, \sigma_1^2)$$

The data are graphed in Figure 9.9. The calculations for the regression line are as follows:

$$n = 49, \qquad b = \frac{[xy]}{[x^2]} = 31.487$$

$$[y^2] = 53637.265, \qquad a = 138.741$$

$$[xy] = 381.6939, \qquad \text{Regression line}$$

$$[x^2] = 12.12245, \qquad Y = 138.741 + 31.487\text{X}$$

$$\overline{y} = 152.8776, \qquad (n-2)s_{y\cdot x}^2 = [y^2] - \frac{[xy]^2}{[x^2]}$$

$$\overline{x} = 0.44898, \qquad = 41{,}619.0488$$

$$s_{y\cdot x}^2 = 885.51$$

The similarity to the t-test becomes obvious, the intercept $a = 138.741$ is precisely the mean mortality for the "inland" states. The "slope," $b = 31.487$, is the mean difference between the two groups of states, and $s_{y\cdot x}^2$, the residual variance, is the pooled variance. The t-test for the slope is equivalent to the t-test for the difference in the two means.

$$\begin{aligned}\text{variance of slope} = s_b^2 &= \frac{s_{y\cdot x}^2}{[x^2]}\\ &= \frac{885.51}{12.12245}\\ &= 73.0471\end{aligned}$$

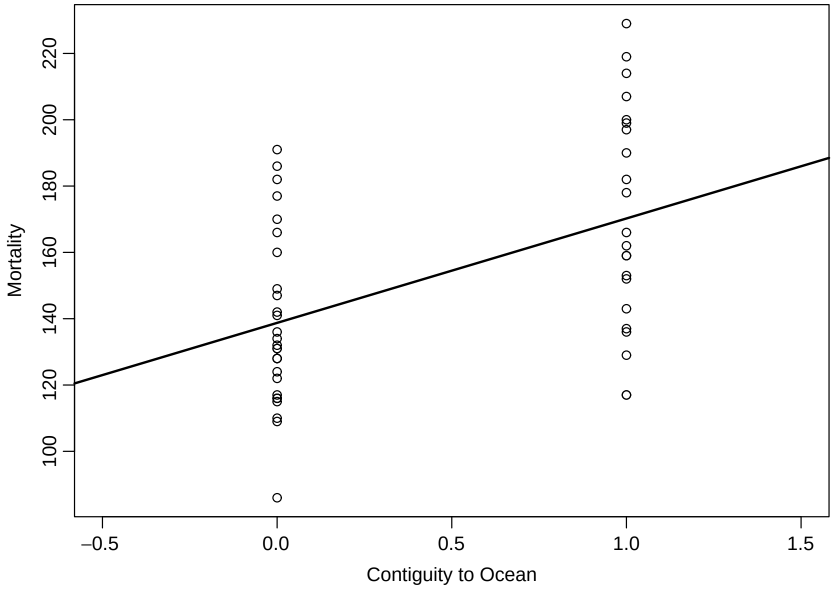

Figure 9.9 Melanoma data: regression of mortality rate on contiguity to ocean, coded 0 if not contiguous to ocean, 1 if contiguous to ocean.

$$\begin{aligned} s_b &= 8.5468 \\ t &= \frac{31.487}{8.5468} \\ &= 3.684 \end{aligned}$$

The t-test for the slope has 47 degrees of freedom, as does the two-sample t-test. Note also that s_b is the standard error of the differences in the two-sample t-test.

Finally, the regression analysis can be put into analysis of variance form as displayed in Table 9.7:

$$\begin{aligned} \text{SS(regression)} &= \frac{[xy]^2}{[x^2]} \\ &= \frac{(381.6939)^2}{12.12245} \\ &= 12{,}018.22 \\ \text{SS(residual)} &= [y^2] - \frac{[xy]^2}{[x^2]} \\ &= 53{,}637.26 - 12{,}018.22 \\ &= 41{,}619.04 \end{aligned}$$

We note that the proportion of variation in mortality rates attributable to "contiguity to ocean" is

$$\begin{aligned} r^2 &= \frac{[xy]^2/[x^2]}{[y^2]} \\ &= \frac{12{,}018.22}{53{,}637.06} \\ &= 0.2241 \end{aligned}$$

Table 9.7 Regression Analysis of Mortality and Contiguity to Ocean

Source of Variation	d.f.	SS	MS	F-Ratio
Regression	1	12,018.22	12,018.22	13.57[a]
Residual	47	41,619.04	885.51	
Total	48	53,637.26		

[a] Significant at the 0.001 level.

Approximately 22% of the variation in mortality can be attributed to the predictor variable: "contiguity to ocean."

In Chapter 11 we deal with the relationships among the three variables: mortality, latitude, and contiguity to an ocean. The predictor variable "contiguity to ocean," which takes on only two values, 0 and 1 in this case, is called a *dummy variable* or *indicator variable*. In Chapter 11 more use is made of such variables.

9.3 CORRELATION AND COVARIANCE

In Section 9.2 the method of least squares was used to find a line for predicting one variable from the other. The response variable Y, or dependent variable Y, was random for given X. Even if X and Y were jointly distributed so that X was a random variable, the model only had assumptions about the distribution of Y given the value of X. There are cases, however, where both variables vary jointly, and there is a considerable amount of symmetry. In particular, there does not seem to be a reason to predict one variable from the other. Example 9.3 is of that type. As another example, we may want to characterize the length and weight relationship of newborn infants. The basic sampling unit is an infant, and two measurements are made, both of which vary. There is a certain symmetry in this situation: There is no "causal direction"—length does not cause weight, or vice versa. Both variables vary together in some way and are probably related to each other through several other underlying variables which determine (cause) length and weight. In this section we provide a quantitative measure of the strength of the relationship between the two variables and discuss some of the properties of this measure. The measure (the correlation coefficient) is a measure of the strength of the linear relationship between two variables.

9.3.1 Correlation and Covariance

We would like to develop a measure (preferable one number) that summarizes the strength of any linear relationship between two variables X and Y. Consider Example 9.2, the exercise test data. The X variable is measured in seconds and the Y variable is measured in milliliters per minute per kilogram. When totally different units are used on the two axes, one can change the units for one of the variables, and the picture seems to change. For example, if we went from seconds to minutes where 1 minute was graphed over the interval of 1 second in Figure 9.2, the data of Figure 9.2 would go almost straight up in the air. Whatever measure we use should not depend on the choice of units for the two variables. We already have one technique of adjusting for or removing the units involved: to standardize the variables. We have done this for the t-test, and we often had to do it for the construction of test statistics in earlier chapters. Further, since we are just concerned with how closely the family of points is related, if we shift our picture (i.e., change the means of the X and Y variables), the strength of the relationship between the two variables should not change. For that reason, we subtract the mean of each variable, so that the pictures will be centered about zero. In order that we have a solution that does not depend

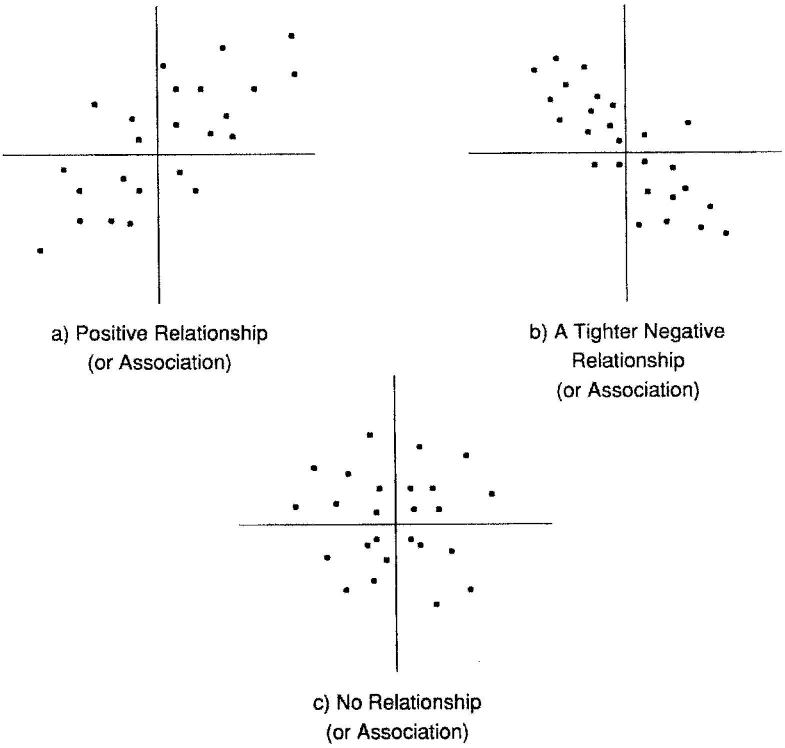

Figure 9.10 Scatter diagrams for the standardized variables.

on units, we standardize each variable by dividing by the standard deviation. Thus, we are now working with two new variables, say U and V, which are related to X and Y as follows:

$$U_i = \frac{X_i - \overline{X}}{s_x}, \qquad V_i = \frac{Y_i - \overline{Y}}{s_y}$$

where

$$s_x^2 = \sum \frac{(X_i - \overline{X})^2}{n-1} \quad \text{and} \quad s_y^2 = \sum \frac{(Y_i - \overline{Y})^2}{n-1}$$

Let us consider how the variables U_i and V_i vary together. In Figure 9.10 we see three possible types of association. Part (a) presents a positive relationship, or association between, the variables. As one increases, the other tends to increase. Part (b) represents a tighter, negative relationship. As one decreases, the other tends to increase, and vice versa. By the word *tighter*, we mean that the variability about a fitted regression line would not be as large. Part (c) represents little or no association, with a somewhat circular distribution of points.

One mathematical function that would capture these aspects of the data results from multiplying U_i and V_i. If the variables tend to be positive or negative together, the product will always be positive. If we add up those multiples, we would get a positive number. On the other hand, if one variable tends to be negative when the other is positive, and vice versa, when we multiply the U_i and V_i together, the product will be negative; when we add them, we will get a negative number of substantial absolute value.

On the other hand, if there is no relationship between U and V, when we multiply them, half the time the product will be positive and half the time the product will be negative; if we sum them, the positive and negative terms will tend to cancel out and we will get something close to zero. Thus, adding the products of the standardized variables seems to be a reasonable method of characterizing the association between the variables. This gives us our definition of the correlation coefficient.

Definition 9.4. The *sample Pearson product moment correlation coefficient*, denoted by r, or r_{XY}, is defined to be

$$r = \frac{[xy]}{\sqrt{[x^2][y^2]}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{1}{n-1}\sum u_i v_i$$

This quantity is usually called the *correlation coefficient.*

Note that the denominator looks like the product of the sample standard deviations of X and Y except for a factor of $n - 1$. If we define the sample covariance by the following equation, we could define the correlation coefficient according to the second alternative definition.

Definition 9.5. The *sample covariance*, s_{xy}, is defined by

$$s_{xy} = \sum_i \frac{(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Alternative Definition 9.4. The *sample Pearson product moment correlation coefficient* is defined by

$$r = \frac{[xy]}{\sqrt{[x^2][y^2]}} = \frac{s_{xy}}{s_x s_y}$$

The prefix co- is a prefix meaning “with,” “together,” and “in association,” occurring in words derived from Latin: thus, the co-talks about the two variables varying together or in association. The term *covariance* has the same meaning as the variance of one variable: how spread out or variable things are. It is hard to interpret the value of the covariance alone because it is composed of two parts; the variability of the individual variables and their linear association. A small covariance can occur because X and/or Y has small variability. It can also occur because the two variables are not associated. Thus, in interpreting the covariance, one usually needs to have some idea of the variability in both variables. A large covariance, however, does imply that at least one of the two variables has a large variance.

The correlation coefficient is a rescaling of the covariance by the standard deviations of X and Y. The motivation for the construction of the covariance and correlation coefficient is the following: s_{xy} is the average of the product of the deviations about the means of X and Y. If X tends to be large when Y is large, both deviations will be positive and the product will be positive. Similarly, if X is small when Y is small, both deviations will be negative but their products will still be positive. Hence, the average of the products for all the cases will tend to be positive. If there is no relationship between X and Y, a positive deviation in X may be paired with a positive or negative deviation in Y and the product will either be positive or negative, and on the average will tend to center around zero. In the first case X and Y are said to be positively correlated, in the second case there is no correlation between X and Y. A third case results when large values of X tend to be associated with small values of Y, and vice versa. In this situation, the product of deviations will tend to be negative and the variables are said to be negatively correlated. The statistic r rescales the average of the product of the deviations about the means by the standard deviations of X and Y.

The statistic r has the following properties:

1. r has value between -1 and 1.
2. $r = 1$ if and only if all the observations are on a straight line with positive slope.
3. $r = -1$ if and only if all observations are on a straight line with negative slope.
4. r takes on the same value if X, or Y, changes units or has a constant added or subtracted.
5. r measures the extent of *linear* association between two variables.
6. r tends to be close to zero if there is no linear association between X and Y.

Some typical scattergrams and associated values of r are given in Figure 9.11. Figure 9.11(*a*) and (*b*) indicate perfect linear relationships between two variables. Figure 9.11(*c*) indicates no correlation. Figure 9.11(*d*) and (*e*) indicate typical patterns representing less than perfect correlation. Figure 9.11(*f*) to (*j*) portray various pathological situations. Figure 9.11(*f*) indicates that although there is an explicit relationship between X and Y, the linear relationship is zero; thus $r = 0$ does not imply that there is no relationship between X and Y. In statistical terminology, $r = 0$ does not imply that the variables are statistically independent. There is one important exception to this statement that is discussed in Section 9.3.3. Figure 9.11(*g*) indicates that except for the one extreme point there is no correlation. The coefficient of correlation is very sensitive to such outliers, and in Section 9.3.7 we discuss correlations that are not as sensitive, that is, more robust. Figure 9.11(*h*) indicates that an explicit relationship between X and Y is not identified by the correlation coefficient if the relationship is not linear. Finally, Figure 9.11(*j*)

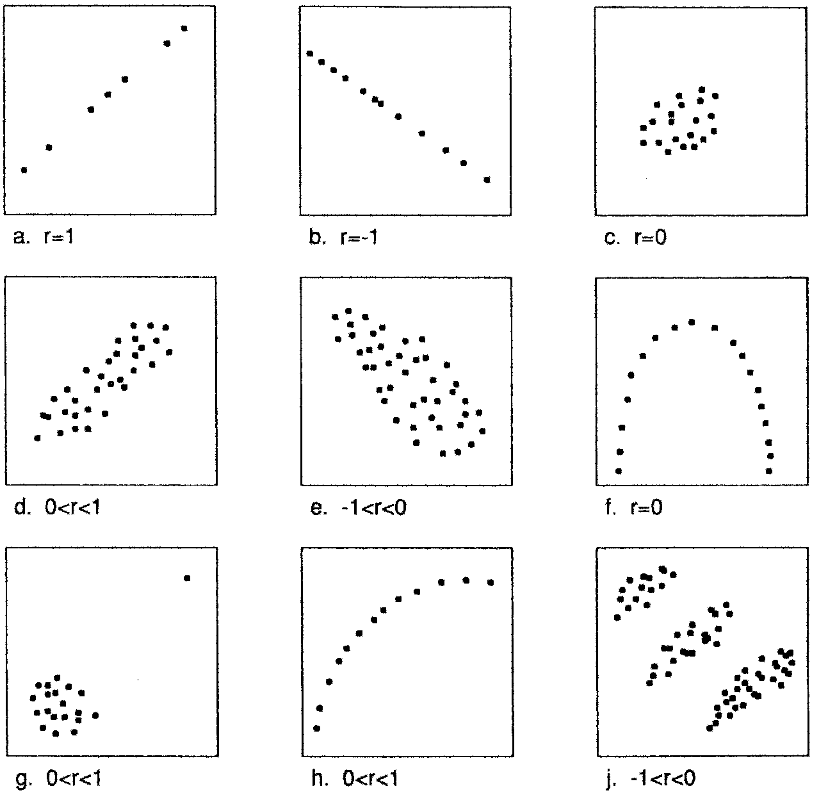

Figure 9.11 Some patterns of association.

suggests that there are three subgroups of cases; within each subgroup there is a positive correlation, but the correlation is negative when the subgroups are combined. The reason is that the subgroups have different means and care must be taken when combining data. For example, natural subgroups defined by gender or race may differ in their means in a direction opposite to the correlation within each subgroup.

Now consider Example 9.3. The scattergram in Figure 9.3 suggests a positive association between the ATP level of the youngest son (X) and that of the oldest son (Y). The data for this example produce the following summary statistics (the subscripts on the values of X and Y have been suppressed: for example, $\sum x_i = \sum x$).

$$\begin{array}{lll}
\text{n} = 17, & & \\
\sum x = 83.38, & \overline{x} = 4.90, & \\
\sum y = 79.89, & \overline{y} = 4.70, & \\
\sum x^2 = 417.1874, & \sum (x-\overline{x})^2 = 8.233024, & s_x = 0.717331 \\
\sum y^2 = 379.6631, & \sum (y-\overline{y})^2 = 4.227094, & s_y = 0.513997 \\
\sum xy = 395.3612, & \sum (x-\overline{x})(y-\overline{y}) = 3.524247, & s_{xy} = 0.220265
\end{array}$$

$$r = \frac{0.220265}{(0.717331)(0.513997)} = 0.597$$

In practice, r will simply be calculated from the equivalent formula

$$r = \frac{[xy]}{\sqrt{[x^2][y^2]}} = \frac{3.524247}{\sqrt{(8.233024)(4.227094)}} = \frac{3.524247}{5.899302} = 0.597$$

The sample correlation coefficient and covariance estimate the population parameters. The expected value of the covariance is

$$\begin{aligned} E(S_{xy}) &= E((X-\mu_x)(Y-\mu_y)) \\ &= \sigma_{xy} \end{aligned}$$

where

$$\mu_x = E(X) \quad \text{and} \quad \mu_y = E(Y)$$

The population covariance is the average of the product of X about its mean times Y about its mean.

The sample correlation coefficient estimates the population correlation coefficient ρ, defined as follows:

Definition 9.6. Let (X, Y) be two jointly distributed random variables. The (*population*) *correlation coefficient* is

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{E((X-\mu_x)(Y-\mu_y))}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

where σ_{xy} is the covariance of X and Y, σ_x the standard deviation of X, and σ_y the standard deviation of Y. ρ is zero if X and Y are statistically independent variables.

There is now a question about the statistical "significance" of a value r. In practical terms, suppose that we have sampled 17 families and calculated the correlation coefficient in ATP levels between the youngest son and the oldest son. How much variation could we have expected relative to the value observed for this set? Could the population correlation coefficient $\rho = 0$? In the next two sections we deal with this question.

9.3.2 Relationship between Correlation and Regression

In Section 9.2.4, r^2 was presented, indicating a close connection between correlation and regression. In this section, the connection will be made explicit in several ways. Formally, one of the variables X, Y could be considered the dependent variable and the other the predictor variable and the techniques of Section 9.2 applied. It is easy to see that in most cases the slope of the regression of Y on X will not be the same as that of X on Y. To keep the distinction clear, the following notation will be used:

b_{yx} = slope of the regression of the "dependent" variable Y on the "predictor" variable X

a_y = intercept of the regression of Y on X

Similarly,

b_{xy} = slope of the regression of the "dependent" variable X on the "predictor" variable Y

a_x = intercept of the regression of X on Y

These quantities are calculated as follows:

	Regress Y on X	**Regress X on Y**
Slope	$b_{yx} = \dfrac{[xy]}{[x^2]}$	$b_{xy} = \dfrac{[xy]}{[y^2]}$
Intercept	$a_y = \overline{y} - b_{yx}\overline{x}$	$a_x = \overline{x} - b_{xy}\overline{y}$
Residual variance	$S^2_{y\cdot x} = \dfrac{[y^2] - [xy]^2/[x^2]}{n-2}$	$S^2_{x.y} = \dfrac{[x^2] - [xy]^2/[y^2]}{n-2}$

From these quantities, the following relationships can be derived:

1. Consider the product

$$\begin{aligned} b_{yx}b_{xy} &= \frac{[xy]^2}{[x^2][y^2]} \\ &= r^2 \end{aligned}$$

Hence

$$r = \pm\sqrt{b_{yx}b_{xy}}$$

In words, r is the geometric mean of the slope of the regression of Y on X and the slope of the regression of X on Y.

2.

$$b_{yx} = r\frac{S_y}{S_x}, \qquad b_{xy} = r\frac{S_x}{S_y}$$

where S_x and S_y are the sample standard deviations of X and Y, respectively.

3. Using the relationships in (2), the regression line of Y on X,

$$\widehat{Y} = a_y + b_{yx}X$$

can be transformed to

$$\begin{aligned}\widehat{Y} &= a_y + \frac{rS_y}{S_x}X \\ &= \overline{y} + \frac{rS_y}{S_x}(X - \overline{x})\end{aligned}$$

4. Finally, the t-test for the slope, in the regression of Y on X,

$$\begin{aligned}t_{n-2} &= \frac{b_{yx}}{S_{b_{yx}}} \\ &= \frac{b_{yx}}{S_{y\cdot x}/\sqrt{[x^2]}}\end{aligned}$$

is algebraically equivalent to

$$r\bigg/\sqrt{\frac{1-r^2}{n-2}}$$

Hence, testing the significance of the slope is equivalent to testing the significance of the correlation.

Consider Example 9.3 again. The data are summarized in Table 9.8. This table indicates that the two regression lines are not the same but that the t-tests for testing the significance of the slopes produce the same observed value, and this value is identical to the test of significance of the correlation coefficient. If the corresponding analyses of variance are carried out, it will be found that the F-ratio in the two analyses are identical and give an equivalent statistical test.

9.3.3 Bivariate Normal Distribution

The statement that a random variable Y has a normal distribution with mean μ and variance σ^2 is a statement about the distribution of the values of Y and is written in a shorthand way as

$$Y \sim N(\mu, \sigma^2)$$

Such a distribution is called a *univariate distribution.*

Definition 9.7. A specification of the distribution of two (or more) variables is called a *bivariate* (or *multivariate*) *distribution.*

The definition of such a distribution will require the specification of the numerical characteristics of each of the variables separately as well as the relationships among the variables. The most common bivariate distribution is the normal distribution. The equation for the density of this distribution as well as additional properties are given in Note 9.6.

We write that (X, Y) have a bivariate normal distribution as

$$(X, Y) \sim N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$$

Table 9.8 Regression Analyses of ATP Levels of Oldest and Youngest Sons

Dependent variable	Y^a		X^a
Predictor variable	X^b		Y^b
Slope	$b_{yx} = 0.42806$		$b_{xy} = 0.83373$
Intercept	$a_y = 2.59989$		$a_x = 0.98668$
Regression line	$\widehat{Y} = 2.600 + 0.428X$		$\widehat{X} = 0.987 + 0.834Y$
Variance about mean	$s_y^2 = 0.26419$		$s_x^2 = 0.51456$
Residual variance	$s_{y\cdot x}^2 = 0.18123$		$s_{x.y}^2 = 0.35298$
Standard error of slope	$s_{b_{y\cdot x}} = 0.14837$		$s_{b_{x.y}} = 0.28897$
Test of significance	$t_{15} = \dfrac{0.42806}{0.14837} = 2.885$		$t_{15} = \dfrac{0.83373}{0.28897} = 2.885$
Correlation		$r_{xy} = r_{yx} = r = 0.597401$	
Test of significance		$t_{15} = \dfrac{0.597401}{\sqrt{\dfrac{1-(0.597401)^2}{17-2}}}$ $= \dfrac{0.597401}{0.20706}$ $= 2.885$	

Source: Data from Dern and Wiorkowski [1969].

[a] ATP level of oldest son.

[b] ATP level of youngest son.

Here μ_x, μ_y, σ_x^2, and σ_y^2 are the means and variances of X and Y, respectively. The quantity ρ is the (population) correlation coefficient. If we assume this model, it is this quantity, ρ, that is estimated by the sample correlation, r.

The following considerations may help to give you some feeling for the bivariate normal distribution. A continuous distribution of two variables, X and Y, may be modeled as follows. Pour 1 pound of sand on a floor (the X–Y plane). The probability that a pair (X, Y) falls into an area, say A, on the floor is the weight of the sand on the area A. For a bivariate normal distribution, the sand forms one mountain, or pile, sloping down from its peak at (μ_x, μ_y), the mean of (X, Y). Cross sections of the sand at constant heights are all ellipses. Figure 9.12 shows a bivariate normal distribution. On the left is shown a view of the sand pile; on the right, a topographical map of the terrain.

The bivariate normal distribution has the property that at every fixed value of X (or Y) the variable Y (or X) has a univariate normal distribution. In particular, write

$$Y_x = \text{random variable } Y \text{ at a fixed value of } X = x$$

It can be shown that at this fixed value of $X = x$,

$$Y_x \sim N\left(\alpha_y + \frac{\sigma_y}{\sigma_x}\rho x, \quad \sigma_y^2(1-\rho^2)\right)$$

This is the regression model discussed previously:

$$Y_x \sim N(\alpha + \beta x, \sigma_1^2)$$

where

$$\alpha = \mu_y - \beta\mu_x, \qquad \beta = \frac{\sigma_y}{\sigma_x}\rho, \qquad \sigma_1^2 = \sigma_y^2(1-\rho^2)$$

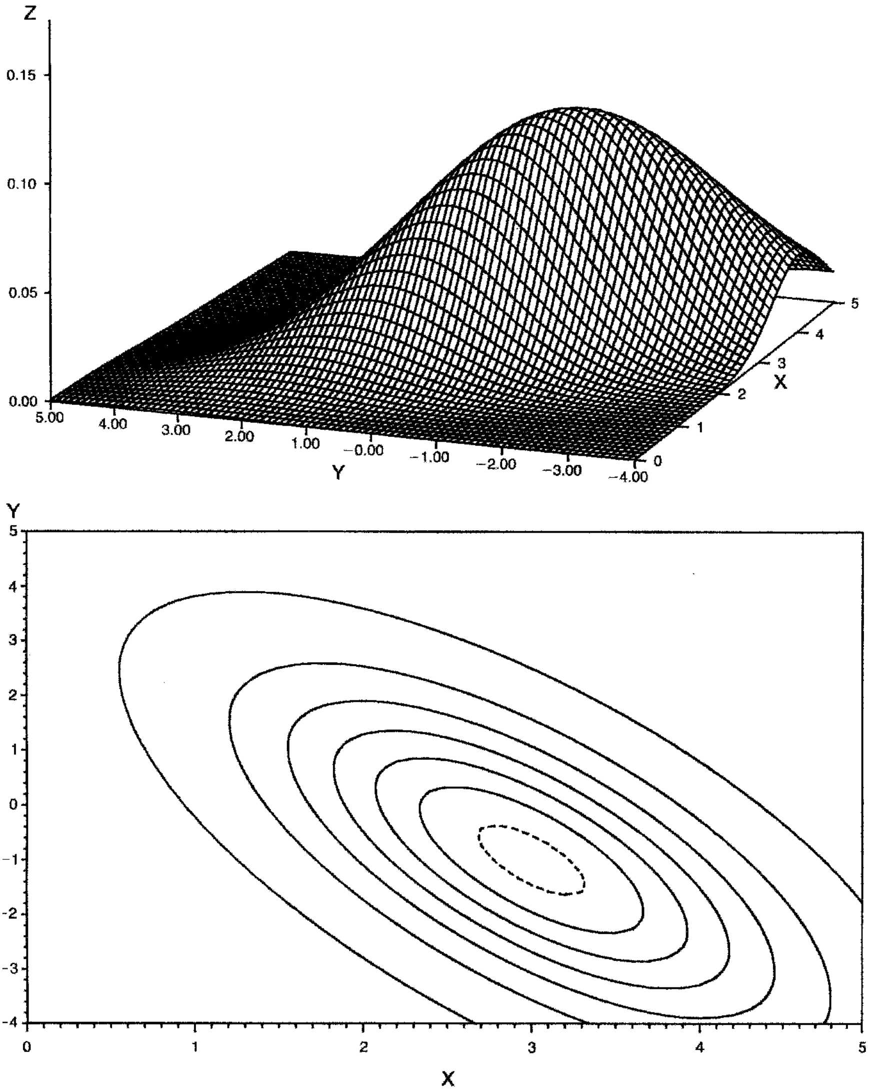

Figure 9.12 Bivariate normal distribution.

Similarly, for

$$X_y = \text{random variable } X \text{ at a fixed value of } Y = y$$

it can be shown that

$$X_y \sim N\left(\alpha_x + \frac{\sigma_x}{\sigma_y}\rho y, \sigma_x^2(1-\rho^2)\right)$$

The null hypothesis $\beta_{yx} = 0$ (or, $\beta_{xy} = 0$) is equivalent then to the hypothesis $\rho = 0$, and the t-test for $\beta = 0$ can be applied.

Suppose now that the null hypothesis is

$$\rho = \rho_0$$

where ρ_0 is an arbitrary but specified value. The sample correlation coefficient r does not have a normal distribution and the usual normal theory cannot be applied. However, R. A. Fisher showed that the quantity

$$Z_r = \frac{1}{2}\log_e \frac{1+r}{1-r}$$

has the following approximate normal distribution:

$$Z_r \sim N\left(\frac{1}{2}\log_e \frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right)$$

where n is the number of pairs of values of X and Y from which r is computed. Not only does Z_r have approximately a normal distribution, but the variance of this normal distribution does not depend on the true value ρ; that is, $Z_r - Z_\rho$ is a pivotal quantity (5.2). This is illustrated graphically in Figure 9.13, which shows the distribution of 1000 simulated values of r and Z_r from distributions with $\rho = 0$ and $\rho = 1/\sqrt{2} \approx 0.71$. The distribution of r has a different variance and different shape for the two values of ρ, but the distribution of Z_r has the same shape and same variance, differing only in location.

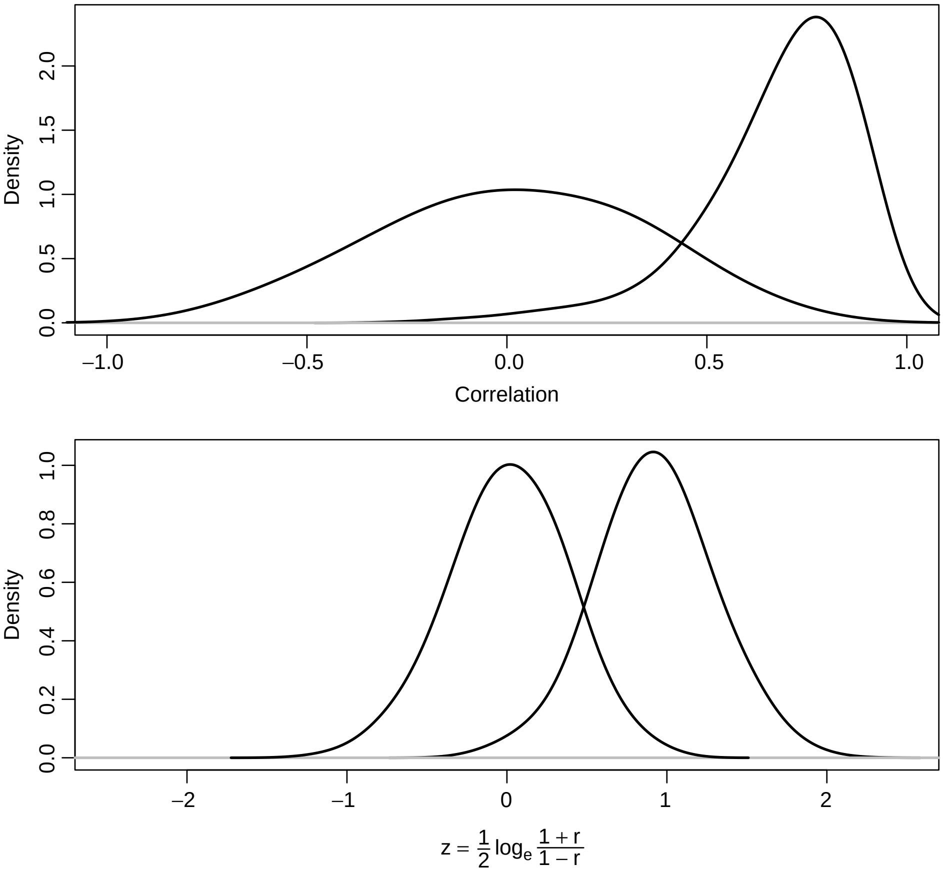

Figure 9.13 Sampling distribution of correlation coefficient, r, before and after transformation, for $\rho = 0, 1/\sqrt{2}$. Estimated from 1000 samples of size 10.

Although the approximate distribution of Z_r was derived under the assumption of a bivariate normal distribution for X and Y, it is not very sensitive to this assumption and is useful quite broadly. Z_r may be used to test hypotheses about ρ and to construct a confidence interval for ρ. This is illustrated below. The inverse, or reverse, function to r is $(e^{2Z} - 1)/(e^{2Z} + 1)$. Z_r is also the inverse of the hyperbolic tangent, tanh. To "undo" the operation, tanh is used.

Consider again Example 9.3 involving the ATP levels of youngest and oldest sons in the 17 families. The correlation coefficient was calculated to be

$$r = 0.5974$$

This value was significantly different from zero; that is, the null hypothesis $\rho = 0$ was rejected. However, the authors show in the paper that genetic theory predicts the correlation to be $\rho = 0.5$. Does the observed value differ significantly from this value? To test this hypothesis we use the Fisher Z_r transformation. Under the genetic theory, the null hypothesis stated in terms of Z_r is

$$\begin{aligned} Z_r &\sim N\left(\frac{1}{2}\log_e\left(\frac{1+0.5}{1-0.5}\right), \frac{1}{17-3}\right) \\ &\sim N(0.5493, 0.07143) \end{aligned}$$

The value observed is

$$Z_r = \frac{1}{2}\log_e\left(\frac{1+0.5974}{1-0.5974}\right) = 0.6891$$

The corresponding standard normal deviate is

$$z = \frac{0.6891 - 0.5493}{\sqrt{0.07143}} = \frac{0.1398}{0.2673} = 0.5231$$

This value does not exceed the critical values at, say, the 0.05 level, and there is no evidence to reject this null hypothesis.

Confidence intervals for ρ may be formed by first using Z_r to find a confidence interval for $1/2\log_e[(1+\rho)/(1-\rho)]$. We then transform back to find the confidence interval for ρ. To illustrate: a $100(1-\alpha)\%$ confidence interval for $1/2\log_e[(1+\rho)/(1-\rho)]$ is given by

$$Z_r \pm z_{1-\alpha/2}\sqrt{\frac{1}{n-3}}$$

For a 90% confidence interval with these data, the interval is $(0 : 6891 - 1.645\sqrt{1/14}, 0.6891 + 1.645\sqrt{1/14}) = (0.249, 1.13)$. When $Z_r = 0.249, r = 0.244$, and when $Z_r = 0.811$. Thus the 90% confidence interval for ρ is (0.244, 0.811). This value straddles 0.5.

9.3.4 Critical Values and Sample Size

We discussed the t-test for testing the hypothesis $\rho = 0$. The formula was

$$t_{n-2} = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

This formula is very simple and can be used for finding critical values and for sample size calculations: Given that the number of observation pairs is specified, the critical value for t with

$n-2$ degrees of freedom is determined, and hence the r critical value can be calculated. For simplicity, write $t_{n-1}=t$; solving the equation above for r^2 yields

$$r^2=\frac{t^2}{t^2+n-2}$$

For example, suppose that $n=20$, the corresponding t-value with 18 degrees of freedom at the 0.05 level is $t_{18}=2.101$. Hence,

$$r^2=\frac{(2.101)^2}{(2.101)^2+18}=0.1969$$

and the corresponding value for r is $r \pm 0.444$; that is, with 20 observations the value of r must exceed 0.444 or be less than -0.444 to be significant at the 0.05 level. Table A.11 lists critical values for r, as a function of sample size.

Another approach is to determine the sample size needed to "make" an observed value of r significant. Algebraic manipulation of the formula gives

$$n=\frac{t^2}{r^2}-t^2+2$$

A useful approximation can be derived if it is assumed that we are interested in reasonably small values of r, say $r<0.5$; in this case, $t \doteq 2$ at the 0.05 level and the formula becomes

$$n=\left(\frac{2}{r}\right)^2-2$$

For example, suppose that $r=0.3$; the sample size needed to make this value significant is

$$n=\left(\frac{2}{0.3}\right)^2-2=44-2=42$$

A somewhat more refined calculation yields $n=43$, so the approximation works reasonably well.

9.3.5 Using the Correlation Coefficient as a Measure of Agreement for Two Methods of Measuring the Same Quantity

We have seen that for X and Y jointly distributed random variables, the correlation coefficient ρ is a population parameter value: ρ is a measure of how closely X and Y have a linear association, ρ^2 is the proportion of the Y variance that can be explained by linear prediction from X, and vice versa.

Suppose that the regression holds and we may choose X. Figure 9.14 shows data from a regression model with three different patterns of X variables chosen. The same errors were added in each figure. The X values were spread out over larger and larger intervals. Since the spread *about* the regression line remains the same and the range of Y increases as the X range increases, the proportion of Y variability explained by X increases: 0.50 to 0.68 to 0.79. For the same random errors and population regression line, r can be anywhere between 0 and 1, depending on which X values are used! In this case the correlation coefficient depends not only on the model, but also on experimental design, where the X's are taken. For this reason some authors say that the r should never be used unless one has a *bi*variate sample: Otherwise, we do not know what r means; another experimenter with the same regression model could choose different X values and obtain a radically different result.

We discuss these ideas in the context of the exercise data of Example 9.2. Suppose that we were strong supporters of maximal treadmill stress testing and wanted to show how closely treadmill duration and VO2 MAX are related. Our strategy for obtaining a large correlation coefficient will be to obtain a large spread of X values, duration. We may know that some of the largest duration and VO2 MAX values were obtained by world-class cross-country skiers; so we would recruit some. For low values we might search for elderly overweight and deconditioned persons. Taking a combined group of these two types of subjects should result in a large value of r. If the same experiment is run using only very old, very overweight, and very deconditioned subjects, the small range will produce a small, statistically insignificant r value.

Since the same treadmill test procedure is associated with large and small r values, what does r mean? A preferable summary indicator is the estimate, $s_{y \cdot x}$ of the residual standard deviation σ_1. If the linear regression model holds, this would be estimated to be the same in each case.

Is it wrong to calculate or present r when a bivariate sample is not obtained? Our answer is a qualified no; that is, it is all right to present r in regression situations provided that:

1. The limitations are kept in mind and discussed. Possible comments on the situation for other sorts of X values might be appropriate.
2. The standard deviation of the residuals should be estimated and presented.

In Chapter 7, the kappa statistic was presented. This was a measure of the amount of agreement when two categorical measurements of the same objects were available. If the two measurements were continuous, the correlation coefficient r is often used as a measure of the agreement of the two techniques. Such use of r is subject to the comments above.

9.3.6 Errors in Both Variables

An assumption in the linear regression model has been that the predictor variable could be measured without error and that the variation in the dependent variable was of one kind only

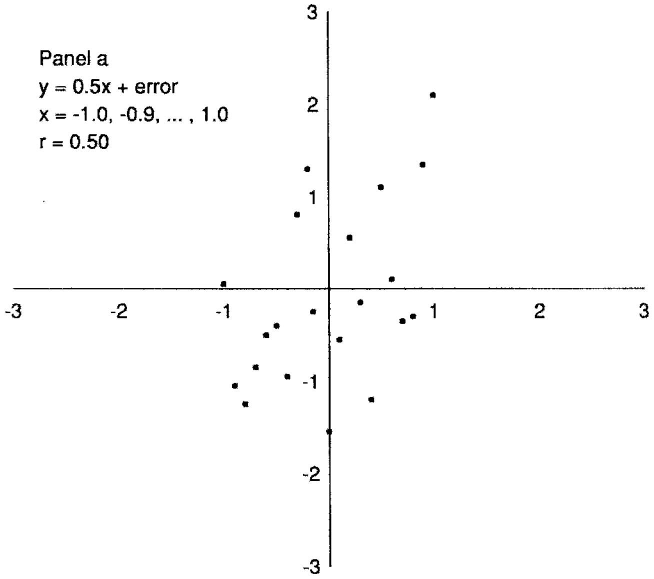

Figure 9.14 The regression model $Y = 0.5X + e$ was used. Twenty-one random $N(0, 1)$ errors were generated by computer. The same errors were used in each panel.

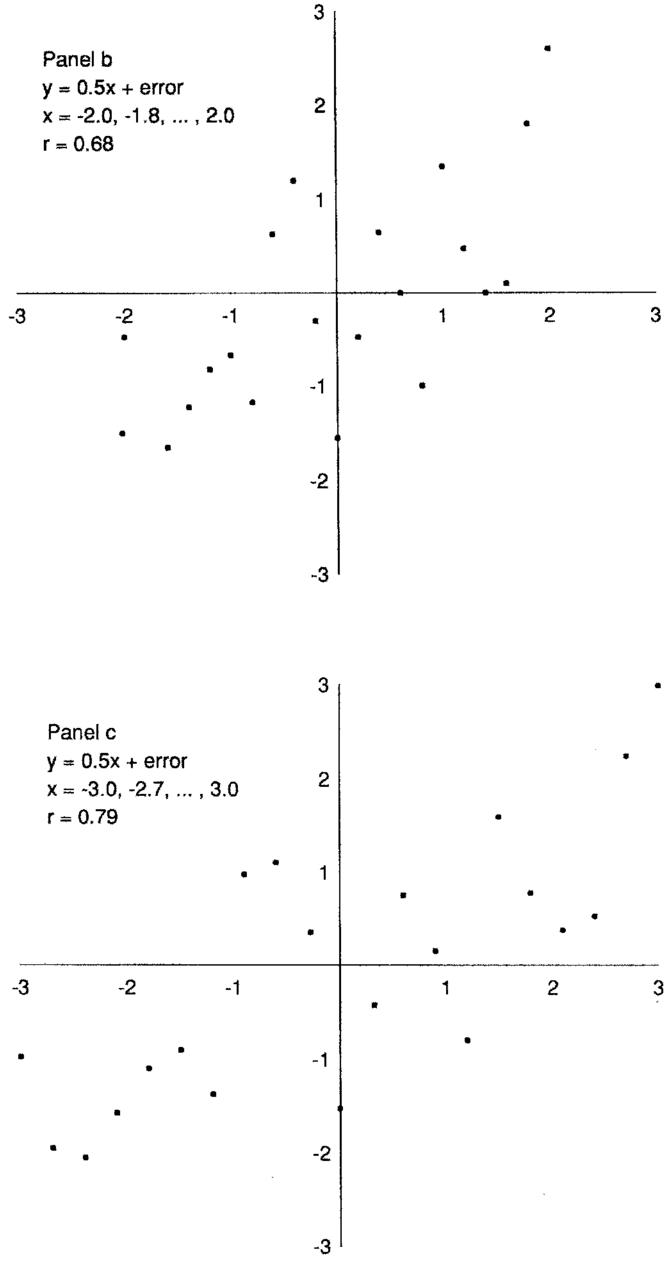

Figure 9.14 (*continued*)

and could be modeled completely if the value of the predictor variable was fixed. In almost all cases, these assumptions do not hold. For example, in measuring psychological characteristics of individuals, there is (1) variation in the characteristics from person to person; and (2) error in the measurement of these psychological characteristics. It is almost certainly true that this problem is present in all scientific work. However, it may be that the measurement error is "small" relative to the variation of the individuals, and hence the former can be neglected.

Another context where the error is unimportant is where the scientific interest is in the variable as measured, not some underlying quantity. For example, in examining how well blood pressure predicts stroke, we are interested in practical prediction, not in what might hypothetically be possible with perfect measurements.

The problem is difficult and we will not discuss it beyond the effect of errors on the correlation coefficient. For a more complete treatment, consult Acton [1984] or Kendall and Stuart [1967, Vol. 2], and for a discussion of measurement error in more complex models, see Carrol et al. [1995].

Suppose that we are interested in the correlation between two random variables X and Y which are measured with errors so that instead of X and Y, we observe that

$$W = X + d, \qquad V = Y + e$$

where d and e are errors. The sampling we have in mind is the following: a "case" is selected at random from the population of interest. The characteristics X and Y are measured but with random independent errors d and e. It is assumed that these errors have mean zero and variances σ_1^2 and σ_2^2, respectively. Another "case" is then selected and the measurement process is repeated with error. Of interest is the correlation ρ_{XY} between X and Y, but the correlation ρ_{VW} is estimated. What is the relationship between these two correlations? The correlation ρ_{XY} can be written

$$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

The reason for writing the correlation this way can be understood when the correlation between V and W is considered:

$$\begin{aligned}\rho_{VW} &= \frac{\sigma_{XY}}{\sqrt{(\sigma_X^2 + \sigma_1^2)(\sigma_Y^2 + \sigma_2^2)}}\\ &= \frac{\sigma_{XY}}{\sigma_X \sigma_Y \sqrt{\left(1 + \sigma_1^2/\sigma_X^2\right)\left(1 + \sigma_2^2/\sigma_Y^2\right)}}\\ &= \frac{\rho_{XY}}{\sqrt{\left(1 + \sigma_1^2/\sigma_X^2\right)\left(1 + \sigma_2^2/\sigma_Y^2\right)}}\end{aligned}$$

The last two formulas indicate that the correlation between V and W is smaller in absolute value than the correlation between X and Y by an amount determined by the ratio of the measurement errors to the variance in the population. Table 9.9 gives the effect on ρ_{XY} as related to the ratios of σ_1^2/σ_X^2 and σ_2^2/σ_Y^2.

A 10% error of measurement in the variables X and Y produces a 9% reduction in the correlation coefficient. The conclusion is that errors of measurement reduce the correlation between two variables; this phenomenon is called *attenuation*.

Table 9.9 Effect of Errors of Measurement on the Correlation between Two Random Variables

$\frac{\sigma_1^2}{\sigma_X^2}$	$\frac{\sigma_2^2}{\sigma_Y^2}$	ρ_{VW}	$\frac{\sigma_1^2}{\sigma_X^2}$	$\frac{\sigma_2^2}{\sigma_Y^2}$	ρ_{VW}
0	0	1 ρ_{XY}	0.20	0.10	$0.87\rho_{XY}$
0.05	0.05	$0.95\rho_{XY}$	0.20	0.20	$0.83\rho_{XY}$
0.10	0.10	$0.91\rho_{XY}$	0.30	0.30	$0.77\rho_{XY}$

Table 9.10 Schema for Spearman Rank Correlation

Case	X	Rank(X)	Y	Rank(Y)	d = Rank(X) − Rank(Y)
1	x_1	R_{x_1}	y_1	R_{y_1}	$d_1 = R_{x_1} - R_{y_1}$
2	x_2	R_{x_2}	y_2	R_{y_2}	$d_2 = R_{x_2} - R_{y_2}$
3	x_3	R_{x_3}	y_3	R_{y_3}	$d_3 = R_{x_3} - R_{y_3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	R_{x_n}	y_n	R_{y_n}	$d_n = R_{x_n} - R_{y_n}$

9.3.7 Nonparametric Estimates of Correlation

As indicated earlier, the correlation coefficient is quite sensitive to outliers. There are many ways of getting estimates of correlation that are more robust; the paper by Devlin et al. [1975] contains a description of some of these methods. In this section we want to discuss two methods of testing correlations derived from the ranks of observations.

The procedure leading to the Spearman rank correlation is as follows: Given a set of n observations on the variables X, Y, the values for X are replaced by their ranks, and similarly, the values for Y. Ties are simply assigned the average of the ranks associated with the tied observations. The scheme shown in Table 9.10 illustrates the procedure.

The correlation is then calculated between R_x and R_y. In practice, the *Spearman rank correlation formula* is used:

$$r_s = r_{R_x R_y} = 1 - \frac{6\sum d_i^2}{n^3 - n}$$

It can be shown that the usual Pearson product-moment correlation formula reduces to this formula when the calculations are made on the ranks, if there are no ties. *Note:* For one or two ties, the results are virtually the same. It is possible to correct the Spearman formula for ties, but a simpler procedure is to calculate r_s by application of the usual product-moment formula to the ranks. Table A.12 gives percentile points for testing the hypothesis that X and Y are independent.

Example 9.4. Consider again the data in Table 9.3 dealing with the ATP levels of the oldest and youngest sons. These data are reproduced in Table 9.11 together with the ranks, the ATP levels being ranked from lowest to highest.

Note that the oldest sons in families 6 and 13 had the same ATP levels; they would have been assigned ranks 12 and 13 if the values had been recorded more accurately; consequently, they are both assigned a rank of 12.5. For this example,

$$n = 17$$

$$\sum d_i^2 = 298.5$$

$$r_s = 1 - \frac{(6)(298.5)}{17^3 - 17} = 1 - 0.3658 = 0.6342$$

This value compares reasonably well with the value $r_{xy} = 0.597$ calculated on the actual data. If the usual Pearson product-moment formula is applied to the ranks, the value $r_s = 0.6340$ is obtained. The reader may verify that this is the case. The reason for the slight difference is due to the tie in values for two of the oldest sons. Table A.12 shows the statistical significance at the two-sided 0.05 level since $r_s = 0.6342 > 0.490$.

The second nonparametric correlation coefficient is the *Kendall rank correlation coefficient*. Recall our motivation for the correlation coefficient r. If there is positive association, increase in

Table 9.11 Rank Correlation Analysis of ATP Levels in Youngest and Oldest Sons in 17 Families

	Youngest		Oldest		
Family	ATP Level	Rank (X)	ATP Level	Rank (Y)	d^a
1	4.18	4	4.81	11	−7
2	5.16	12	4.98	14	−2
3	4.85	9	4.48	6	3
4	3.43	1	4.19	3	−2
5	4.53	5	4.27	4	1
6	5.13	11	4.87	12.5	−1.51
7	4.10	2	4.74	10	−8
8	4.77	7	4.53	7	0
9	4.12	3	3.72	1	2
10	4.65	6	4.62	8	−2
11	6.03	17	5.83	17	0
12	5.94	15	4.40	5	10
13	5.99	16	4.87	12.5	3.5
14	5.43	14	5.44	16	−2
15	5.00	10	4.70	9	1
16	4.82	8	4.14	2	6
17	5.25	13	5.30	15	−2
					$\sum d = 0$
					$\sum d^2 = 298.5$

[a] Rank(X) − rank(Y).

X will tend to correspond to increase in Y. That is, given two data points (X_1, Y_1) and (X_2, Y_2), if $X_1 - X_2$ is positive, $Y_1 - Y_2$ is positive. In this case, $(X_1 - X_2)(Y_1 - Y_2)$ is usually positive. If there is negative association, $(X_1 - X_2)(Y_1 - Y_2)$ will usually be negative. If X and Y are independent, the expected value is zero. Kendall's rank correlation coefficient is based on this observation.

Definition 9.8. Consider a bivariate sample of size $n, (X_1, Y_1), \ldots, (X_n, Y_n)$. For each pair, count 1 if $(X_i - X_j)(Y_i - Y_j) > 0$. Count −1 if $(X_i - X_j)(Y_i - Y_j) < 0$. Count zero if $(X_i - X_j)(Y_i - Y_j) = 0$. Let κ be the sum of these $n(n-1)/2$ counts. (Note that this κ is not related to the kappa of Chapter 7.) Kendall's τ is

$$\tau = \frac{\kappa}{n(n-1)/2}$$

1. The value of τ is between −1 and 1. Under the null hypothesis of independence, τ is symmetric about zero.
2. Note that $(R_{X_i} - R_{X_j})(R_{Y_i} - R_{Y_j})$ has the same sign as $(X_i - X_j)(Y_i - Y_j)$. That is, both are positive or both are negative or both are zero. If we calculated τ from the ranks of the (X_i, Y_i), we get the same number. Thus, τ is a nonparametric quantity based on ranks; it does not depend on the distributions of X and Y.
3. The expected value of τ is

$$P[(X_i - X_j)(Y_i - Y_j) > 0] - P[(X_i - X_j)(Y_i - Y_j) < 0]$$

Table 9.12 Data for Example 9.4[a]

	j															
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	1															
3	−1	1														
4	1	1	1													
5	−1	1	1	1												
6	1	1	1	1	1											
7	1	1	−1	1	−1	1										
8	−1	1	−1	1	1	1	−1									
9	1	1	1	−1	1	1	−1	1								
10	−1	1	−1	1	1	1	−1	−1	1							
11	1	1	1	1	1	1	1	1	1	1						
12	−1	−1	−1	1	1	−1	−1	−1	1	−1	1					
13	1	−1	1	1	1	0	1	1	1	1	1	1				
14	1	1	1	1	1	1	1	1	1	1	1	−1	−1			
15	−1	1	1	1	1	1	−1	1	1	1	1	−1	1	1		
16	−1	1	1	−1	−1	1	−1	−1	1	−1	1	1	1	1	1	
17	1	1	1	1	1	1	1	1	1	1	1	−1	−1	1	1	1

[a] Consider $(X_i - X_j)(Y_i - Y_j)$: the entries are 1 if this is positive, 0 if this equals 0, and −1 if this is negative.

4. For moderate to large n and no or few ties, an approximate standard normal test statistic is

$$Z = \frac{\kappa}{\sqrt{n(n-1)(2n+5)/18}}$$

More information where there are ties is given in Note 9.7.

5. If $(X_i - X_j)(Y_i - Y_j) > 0$, the pairs are said to be concordant. If $(X_i - X_j)(Y_i - Y_j) < 0$, the pairs are discordant.

Return to the ATP data of Table 9.11. $(X_1 - X_2)(Y_1 - Y_2) = (4.18 - 5.16)(4.81 - 4.98) > 0$, so we count +1. Comparing each of the $17 \times 16/2 = 136$ pairs gives the +1's, 0's and −1's in Table 9.12. Adding these numbers, $\kappa = 67$, and $\tau = 67/(17 \times 16/2) = 0.493$. The asymptotic Z-value is

$$Z = \frac{67}{\sqrt{17 \times 16 \times 39/18}} = 2.67$$

with $p = 0.0076$ (two-sided).

9.3.8 Change and Association

Consider two continuous measurements of the same quantity on the same subjects at different times or under different circumstances. The two times might be before and after some treatment. They might be for a person taking a drug and not taking a drug. If we want to see if there is a difference in the means at the two times or under the two circumstances, we have several statistical tests: the paired t-test, the signed rank test, and the sign test. Note that we have observed pairs of numbers on each subject.

We now have new methods when pairs of numbers are observed: linear regression and correlation. Which technique should be used in a given circumstance? The first set of techniques looks for *changes between the two measurements*. The second set of techniques look for association and sometimes *the ability to predict*. The two concepts are different ideas:

1. Consider two independent length measurements from the same x-rays of a sample of patients. Presumably there is a "true" length. The measurements should fluctuate about the true length. Since the true length will fluctuate from patient to patient, the two readings should be associated, hopefully highly correlated. Since both measurements are of the same quantity, there should be little or no change. This would be a case where one expects association, but no change.
2. Consider cardiac measurements on patients before and after a heart transplant. The initial measurements refer to a failing heart. After heart transplant the measurements refer to the donor heart. There will be little or no association because the measurements of output, and so on, refer to different (somewhat randomly paired) hearts.

There are situations where both change and prediction or association are relevant. After observing a change, one might like to investigate how the new changed values relate to the original values.

9.4 COMMON MISAPPLICATION OF REGRESSION AND CORRELATION METHODS

In this section we discuss some of the pitfalls of regression and correlation methods.

9.4.1 Regression to the Mean

Consider Figure 9.15, which has data points with approximately zero correlation or association, considered as measurements before and after some intervention. On the left we see that the before and after measurements have no association. The solid line indicates before = 0, and the dashed line indicates before = after. On the right we plot the change against the value before intervention. Again, the two lines are before = 0 and before = after (i.e., change = 0), and we can see how selecting based on the value of the measurement before intervention distorts the average change.

Cases with low initial values (circles on the graph) tend to have positive changes; those with high initial values (triangles) have negative changes. If we admitted to our study only the subjects with low values, it would appear that the intervention led to an increase. In fact, the change would be due to random variability and the case selection. This phenomenon is called *regression to the mean*.

As another example, consider subjects in a quantitative measurement of the amount of rash due to an allergy. Persons will have considerable variability due to biology and environment. Over time, in a random fashion, perhaps related to the season, the severity of rash will ebb and flow. Such people will naturally tend to seek medical help when things are in a particularly bad state. Following the soliciting of help, biological variability will give improvement with or without treatment. Thus, if the treatment is evaluated (using before and after values), there would be a natural drop in the amount of rash simply because medical help was solicited during particularly bad times. This phenomenon again is *regression to the mean*. The phenomenon of regression to the mean is one reason that control groups are used in clinical studies. Some approaches to addressing it are given by Yanez et al. [1998].

9.4.2 Spurious Correlation

Consider a series of population units, for example, states. Suppose that we wish to relate the occurrence of death from two distinct causes, for example, cancer at two different sites on the body. If we take all the states and plot a scatter diagram of the number of deaths from the two causes, there will be a relationship simply because states with many more people, such as

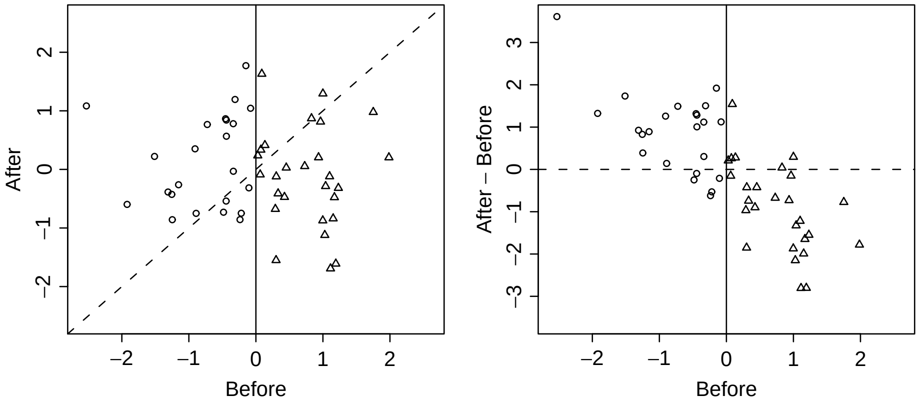

Figure 9.15 Regression to the mean in variables with no association: Before vs. after and before vs. change.

California or New York, will have a large number of deaths, compared to a smaller state such as Wyoming or New Hampshire.

It is necessary to somehow adjust for or take the population into account. The most natural thing to do is to take the death rate from certain causes, that is, to divide the number of deaths by the population of the state. This would appear to be a good solution to the problem. This introduces another problem, however. If we have two variables, X and Y, which are *not related* and we divide them by a third variable, Z, which is random, the two ratios X/Z and *Y/Z will be related*. Suppose that Z is the true denominator measured with error. The reason for the relationship is that when Z is on the low side, since we are dividing by Z, we will increase both numbers at the same time; when Z is larger than it should be and we divide X and Y by Z, we decrease both numbers. The introduction of correlation due to computing rates using the same denominator is called *spurious correlation*. For further discussion on this, see Neyman [1952] and Kronmal [1993], who gives a superb, readable review. A preferable way to adjust for population size is to use the techniques of multiple regression, which is discussed in Chapter 11.

9.4.3 Extrapolation beyond the Range of the Data

For many data sets, including the three of this chapter, the linear relationship does a reasonable job of summarizing the association between two variables. In other situations, the relationship may be reasonably well modeled as linear over a part of the range of X but not over the entire range of X. Suppose, however, that data had been collected on only a small range of X. Then a linear model might fit the accumulated data quite well. If one takes the regression line and uses it as an indication of what would happen for data values *outside the range covered by the actual data*, trouble can result. To have confidence in such extrapolation, one needs to know that indeed the linear relationship holds over a broader range than the range associated with the actual data. Sometimes this assumption is valid, but often, it is quite wrong. There is no way of knowing in general to what extent extrapolation beyond the data gives problems. Some of the possibilities are indicated graphically in Figure 9.16. Note that virtually any of these patterns of curves, when data are observed over a short range, can reasonably be approximated by a linear function. Over a wider range, a linear approximation is not adequate. But if one does not have data over the wide range, this cannot be seen.

Sometimes it is necessary to extrapolate beyond the range of the data. For example, there is substantial concern in Britain over the scale of transmission of “mad cow disease” to humans, causing variant Creutzfeld–Jakob disease (vCJD). Forecasting the number of future cases is

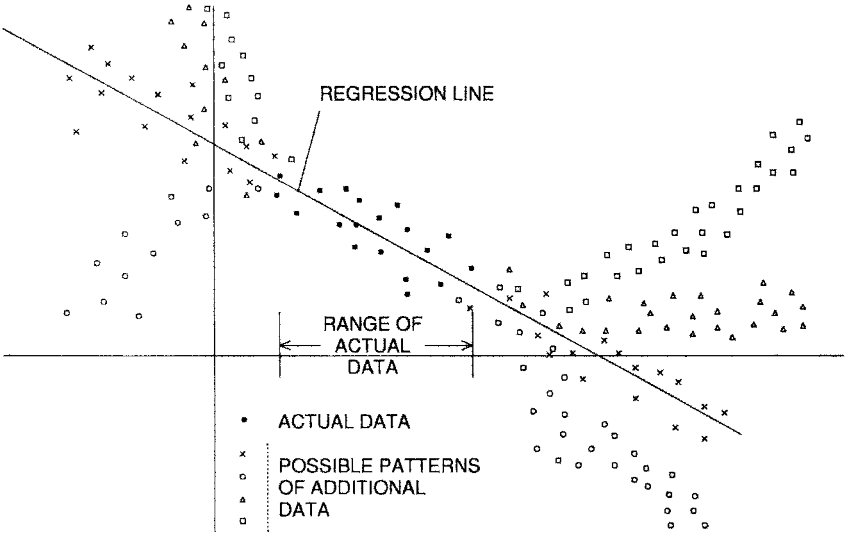

Figure 9.16 Danger of extrapolating beyond observed data.

important for public health, and intrinsically, requires extrapolation. A responsible approach to this type of problem is to consider carefully what models (linear or otherwise) are consistent with the data available and more important, with other existing knowledge. The result is a range of predictions that acknowledge both the statistical uncertainty within each model and the (often much greater) uncertainty about which model to use.

9.4.4 Inferring Causality from Correlation

Because two variables are associated does not necessarily mean that there is any causal connection between them. For example, if one recorded two numbers for each year—the numbers of hospital beds and the total attendance at major league baseball games—there would be a positive association because both of these variables have increased as the population increased. The direct connection is undoubtedly slight at best. Thus, regression and correlation approaches show observed relationships, which may or may not represent a causal relationship. In general, the strongest inference for causality comes from experimental data; in this case, factors are changed by the experimenter to observe change in a response. Regression and correlation from observational data may be very suggestive but do not definitively establish causal relationships.

9.4.5 Interpretation of the Slope of the Regression Line

During the discussion, we have noted that the regression equation implies that if the predictor or independent variable X is higher by an amount ΔX, then on the average, Y is higher by an amount $\Delta Y = b\,\Delta X$. This is sometimes interpreted to mean that if we can modify a situation such that the X variable is changed by ΔX, the Y variable will change correspondingly; this may or may not be the case. For example, if we look at a scatter diagram of adults' height and weight, it does not follow if we induce a change in a person's weight, either by dieting or by excess calories that the person's height will change correspondingly. Nevertheless, there is an association between height and weight. Thus, the particular inference depends on the science involved. Earlier in this chapter, it was noted that from the relation between $\text{VO}_2\ _{\text{MAX}}$ and the duration of the exercise test that if a person is trained to have an increased duration, the $\text{VO}_2\ _{\text{MAX}}$ will also increase. This particular inference is correct and has been documented by serial studies recording both variables. It follows from other data and scientific understanding. It is *not* a logical consequence of the observed association.

9.4.6 Outlying Observations

As noted above, outlying observations can have a large effect on the actual regression line (see Figure 9.7, for example). If one examines these scattergrams or residual plots, the problem should be recognized. In many situations, however, people look at large numbers of correlations and do not have the time, the wherewithal, or possibly the knowledge to examine all of the necessary visual presentations. In such a case, an outlier can be missed and data may be interpreted inappropriately.

9.4.7 Robust Regression Models

The least squares regression coefficients result from minimizing

$$\sum_{i=1}^{n} g(Y_i - a - bX_i)$$

where the function $g(z) = z^2$. For large z (large residuals) this term is very large. In the second column of figures in Figure 9.7 we saw that one outlying value could heavily modify an otherwise nice fit.

One way to give less importance to large residuals is to choose the function g to put less weight on outlying values. Many robust regression techniques take this approach. We can choose g so that for most z, $g(z) = z^2$, as in the least squares estimates, but for very large $|z|$, $g(z)$ is less than z^2, even zero for extreme z! See Draper and Smith [1998, Chap. 25] and Huber [2003, Chap. 7]. These *resistant M–estimators* protect against outlying Y but not against outlying X, for which even more complex estimators are needed. It is also important to note that protection against outliers is not always desirable. Consider the situation of a managed care organization trying to determine if exercise reduces medical costs. A resistant regression estimator would effectively ignore information on occasional very expensive subjects, who may be precisely the most important in managing costs. See Chapter 8 and Lumley et al. [2002] for more discussion of these issues.

NOTES

9.1 Origin of the Term Regression

Sir Francis Galton first used the term in 1885. He studied the heights of parents and offspring. He found (on the average) that children of tall parents were closer to the average height (were shorter); children of short parents were taller and closer to the average height. The children's height *regressed* to the average.

9.2 Maximum Likelihood Estimation of Regression and Correlation Parameters

For a data set from a continuous probability density, the probability of observing the data is proportional to the probability density function. It makes sense to *estimate the parameters by choosing parameters to make the probability of the observed data as large as possible*. Such estimates are called *maximum likelihood estimates* (MLEs). Knowing $X_1, \dots, X_n$ in the regression problem, the likelihood function for the observed $Y_1, \ldots, Y_n$ is (assuming normality)

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}[Y_i - (\alpha + \beta X_i)]^2\right\}$$

The maximum likelihood estimates of α and β are the least squares estimates a and b. For the bivariate normal distribution, the MLE of ρ is r.

9.3 Notes on the Variance of a, Variance of a + bx, and Choice of x for Small Variance (Experimental Design)

1. The variance of a in the regression equation $y = a + bx$ can be derived as follows: $a = \overline{y} + b\overline{x}$; it is true that $\overline{y}$ and b are statistically independent; hence,

$$\begin{aligned} \text{var}(a) &= \text{var}(\overline{y} + b\overline{x}) \\ &= \text{var}(\overline{y}) + \overline{x}^2\text{var}(b) \\ &= \frac{\sigma_1^2}{n} + \overline{x}^2 \frac{\sigma_1^2}{[x^2]} \\ &= \sigma_1^2\left(\frac{1}{n} + \frac{\overline{x}^2}{[x^2]}\right) \end{aligned}$$

2. Consider the variance of the estimate of the mean of y at some arbitrary fixed point X:

$$\sigma_1^2\left(\frac{1}{n} + \frac{(x - \overline{x})^2}{[x^2]}\right)$$

a. Given a choice of x, the quantity is minimized at $x = \overline{x}$.

b. For values of x close to $\overline{x}$ the contribution to the variance is minimal.

c. The contribution increase as the *square* of the distance the predictor variable x is from $\overline{x}$.

d. If there was a choice in the selection of the predictor variables, the quantity $[x^2] = \sum(x_i - \overline{x})^2$ is maximized if the predictor variables are spaced as far apart as possible. If X can have a range of values, say, X_{min} to X_{max}, the quantity $[x^2]$ is maximized if half the observations are placed at X_{min} and the other half at X_{max}. The quantity $(x - \overline{x})^2/[x^2]$ will then be as small as possible. Of course, a price is paid for this design: it is not possible to check the linearity of the relationship between Y and X.

9.4 Average-Slope Formula for b

An alternative formula for the slope estimate b emphasizes the interpretation as an average difference in Y for each unit difference in X. Suppose that we had just two points (X_1, Y_1) and (X_2, Y_2). The obvious estimate of the slope comes from simply joining the points with a line:

$$b_{21} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

With more than two points we could calculate all the pairwise slope estimates

$$b_{ij} = \frac{Y_i - Y_j}{X_i - X_j}$$

and then take some summary of these as the overall slope. More weight should be give to estimates b_{ij} where $X_i - X_j$ is larger, as the expected difference in Y, $\beta(X_i - X_j)$ is larger relative to the residual error in Y_i and Y_j. If we assign weights $w_{ij} = (X_i - X_j)^2$, a little algebra shows that an alternative formula for the least squares estimate b is

$$b = \frac{\sum_{i,j} w_{ij} b_{ij}}{\sum_{i,j} w_{ij}}$$

a weighted average of the pairwise slopes.

This formulation makes it clear that b estimates the average slope of Y with respect to X under essentially no assumptions. Of course, if the relationship is not at least roughly linear, the average slope may be of little practical interest, and in any case some further assumptions are needed for statistical inference.

9.5 Regression Lines through the Origin

Suppose that we want to fit the model $Y \sim N(\beta X, \sigma^2)$, that is, the line goes through the origin. In many situations this is an appropriate model (e.g., in relating body weight to height, it is reasonable to assume that the regression line must go through the origin). However, the regression relationship may not be linear over the entire range, and often, the interval of interest is quite far removed from the origin.

Given n pairs of observation (x_i, y_i), $i = 1, \dots, n$, and a regression line through the origin is desired, it can be shown that the least squares estimate, b, of β is

$$b = \frac{\sum x_i y_i}{\sum x_i^2}$$

The residual sum of squares is based on the quantity

$$\sum (y_i - \widehat{y}_i)^2 = \sum (y_i - bx_i)^2$$

and has associated with it, $n - 1$ degrees of freedom, since only one parameter, β, is estimated.

9.6 Bivariate Normal Density Function

The formula for the density of the bivariate normal distribution is

$$f_{X,Y}(x, y) = \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1 - p^2}} \exp\left[-\frac{1}{2(1-\rho^2)} (Z_X^2 - 2\rho Z_X Z_Y + Z_Y^2) \right]$$

where

$$Z_X = \frac{x - \mu_X}{\sigma_X} \quad \text{and} \quad Z_Y = \frac{y - \mu_Y}{\sigma_Y}$$

The quantities μ_X, μ_Y, σ_X, and σ_Y are, as usual, the means and standard deviations of X and Y, respectively. Several characteristics of this distribution can be deduced from this formula:

1. If $\rho = 0$, the equation becomes

$$f_{X,Y}(x, y) = \frac{1}{2\pi \sigma_X \sigma_Y} \exp\left[-\frac{1}{2} (Z_X^2 + Z_Y^2) \right]$$

and can be written as

$$= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{1}{2}Z_X^2\right) \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left(-\frac{1}{2}Z_Y^2\right)$$

$$= f_X(x) f_Y(y)$$

Thus in the case of the bivariate normal distribution, $\rho = 0$ (i.e., the correlation is zero), implies that the random variables X and Y are statistically independent.

2. Suppose that $f_{X,Y}(x, y)$ is fixed at some specified value; this implies that the expression in the exponent of the density $f_{X,Y}(x, y)$ has a fixed value, say, K:

$$K = \frac{-1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]$$

This is the equation of an ellipse centered at (μ_X, μ_Y).

9.7 Ties in Kendall's Tau

When there are ties in the X_i and/or Y_i values for Kendall's tau, the variability is reduced. The asymptotic formula needs to be adjusted accordingly [Hollander and Wolfe, 1999]. Let the X_i values have g distinct values with ties with t_j tied observations at the jth tied value. Let the Y_i values have h distinct tied values with u_k tied observations at the kth tied value. Under the null hypothesis of independence between the X and Y values, the variance of K is

$$\begin{aligned}\operatorname{var}(K) = {} & \frac{n(n-1)(2n+5)}{18} \\ & - \sum_{j=1}^{g} \frac{t_j(t_j-1)(2t_j+5)}{18} \\ & - \sum_{j=1}^{h} \frac{u_k(u_k-1)(2u_k+5)}{18} \\ & + \frac{\left[\sum_{j=1}^{g} t_j(t_j-1)(t_j-2)\right]\left[\sum_{k=1}^{h} u_k(u_k-1)(u_k-2)\right]}{9n(n-1)(n-2)} \\ & + \frac{\left[\sum_{j=1}^{g} t_j(t_j-1)\right]\left[\sum_{k=1}^{h} u_k(u_k-1)\right]}{2n(n-1)}\end{aligned}$$

The asymptotic normal Z value is

$$Z = \frac{K}{\sqrt{\operatorname{var}(K)}}$$

Note that the null hypothesis is independence, not $\tau = 0$. If the data are not independent but nevertheless have $\tau = 0$ (e.g., a U-shaped relationship), the test will be incorrect.

9.8 Weighted Regression Analysis

In certain cases the assumption of homogeneity of variance of the dependent variable, Y, at all levels of X is not tenable. Suppose that the precision of value $Y = y$ is proportional to a value

W, the weight. Usually, the precision is the reciprocal of the variance at X_i. The data can then be modeled as follows:

Case	X	Y	W
1	x_1	y_1	w_1
2	x_2	y_2	w_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i	w_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	w_n

Define $\sum w_i(x_i - \overline{x}_i)^2 = [wx^2]$, $\sum w(x_i - \overline{x})(y_i - \overline{y}) = [wxy]$. It can be shown that the *weighted* least squares line has slope and intercept,

$$b = \frac{[wxy]}{[wx^2]} \quad \text{and} \quad a = \overline{y} - b\overline{x}$$

where

$$\overline{y} = \frac{\sum w_i y_i}{\sum w_i} \quad \text{and} \quad \overline{x} = \frac{\sum w_i x_i}{\sum w_i}$$

It is a weighted least squares solution in that the quantity $\sum w_i(y_i - \widehat{y}_i)^2$ is minimized. If all the weights are the same, say equal to 1, the ordinary least squares solutions are obtained.

9.9 Model-Robust Standard Error Estimates

We showed that Student's t-test can be formulated as a regression problem. This raises the question of whether we can also find a regression formulation of the Z-test or the unequal-variance approximate t-test of Note 5.2. The answer is in the affirmative. Standard error estimates are available that remove subsidiary assumptions such as equality of variance for a wide range of statistical estimators. These model-robust or "sandwich" standard errors were discovered independently in different fields of statistics and are typically attributed to Huber in biostatistics and to White in econometrics. The Huber–White standard error estimates are available for linear models in SAS and for nearly all regression models in State. In the case of linear regression with a binary X variable, they are equivalent to the unequal-variance t-test except that there is not complete agreement on whether n or $n - 1$ should be used as a denominator in computing variances. See Huber [2003] for further discussion.

PROBLEMS

In most of the problems below, you are asked to perform some subset of the following tasks:

(a) Plot the scatter diagram for the data.

(b) Compute for $\overline{X}$, $\overline{Y}$, $[x^2]$, $[y^2]$, and $[\,xy]$ those quantities not given.

(c) Find the regression coefficients a and b.

(d) Place the regression line on the scatter diagram.

(e) Give $s^2_{y \cdot x}$ and $s_{y \cdot x}$.

(f) Compute the missing predicted values, residuals, and normal deviates for the given portion of the table.

(g) Plot the residual plot.

(h) Interpret the residual plot.

(i) Plot the residual normal probability plot.

(j) Interpret the residual normal probability plot.

(k) **i.** Construct the 90% confidence interval for β.

ii. Construct the 95% confidence interval for β.

iii. Construct the 99% confidence interval for β.

iv. Compute the t-statistic for testing $\beta = 0$. What can you say about its p-value?

(l) **i.** Construct the 90% confidence interval for α.

ii. Construct the 95% confidence interval for α.

iii. Construct the 99% confidence interval for α.

(m) Construct the ANOVA table and use Table A.7 to give information about the p-value.

(n) Construct the 95% confidence interval for $\alpha + \beta X$ at the X value(s) specified.

(o) Construct the interval such that one is 95% certain that a new observation at the specified X value(s) will fall into the interval.

(p) Compute the correlation coefficient r.

(q) **i.** Construct the 90% confidence interval for ρ.

ii. Construct the 95% confidence interval for ρ.

iii. Construct the 99% confidence interval for ρ.

(r) Test the independence of X and Y using Spearman's rank correlation coefficient. Compute the coefficient.

(s) Test the independence of X and Y using Kendall's rank correlation coefficient. Compute the value of the coefficient.

(t) Compute Student's paired t-test for the data, if not given; in any case, interpret.

(u) Compute the signed rank statistic, if not given; in any case, interpret.

The first set of problems, 9.1 to 9.4, come from the exercise data in Example 9.2.

9.1 Suppose that we use duration, X, to predict VO2 MAX, Y. The scatter diagram is shown in Figure 9.2. $\overline{X} = 647.4$, $\overline{Y} = 40.57$, $[x^2] = 673{,}496.4$, $[y^2] = 3506.2$, and $[xy] = 43{,}352.5$. Do tasks (c), (e), (f), (h), (k-ii), (k-iv), (l-ii), (m), (n) at $x = 650$, (p), and (q-ii) (the residual plot is Figure 9.17). All the data are listed in Table 9.13. What proportion of the Y variance is explained by X? (In practice, duration is used as a reasonable approximation to VO2 MAX.)

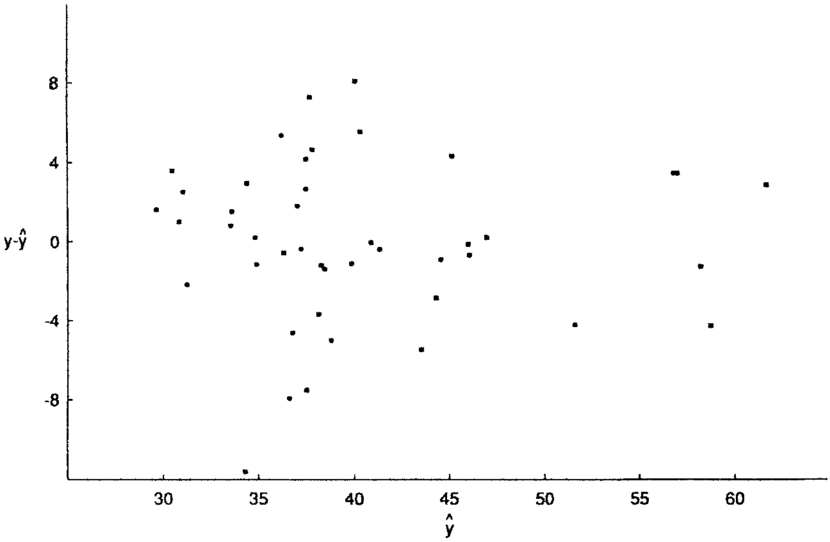

Figure 9.17 Residual plot for the data of Example 9.2; VO2 MAX predicted from duration.

Table 9.13 Oxygen Data for Problem 9.1

X	Y	$\widehat{Y}$	$Y - \widehat{Y}$	Normal Deviate
706	41.5	44.5	−3.0	−0.80
732	45.9	46.13	−0.23	−0.06
930	54.5	?	?	?
900	60.3	?	3.59	0.96
903	60.5	56.90	3.60	0.97
976	64.6	61.50	3.10	0.83
819	47.4	?	−4.21	−1.13
922	57.0	58.10	−1.10	−0.29
600	40.2	37.82	?	0.64
540	35.2	?	1.16	0.31
560	33.8	35.30	−1.50	?
637	38.8	40.15	−1.35	−0.36
593	38.9	?	1.52	0.41
719	49.5	45.31	?	1.23
615	37.1	38.77	−1.67	−0.45
589	32.2	37.13	?	−1.32
478	31.3	30.14	1.16	0.31
620	33.8	39.08	−5.28	?
710	43.7	44.75	−1.05	−0.28
600	41.7	37.82	3.88	1.04
660	41.0	41.60	−0.60	−0.16

9.2 One expects exercise performance to reduce with age. In this problem, X = age and Y = duration. $\overline{X} = 47.2, \overline{Y} = 647.4, [x^2] = 4303.2, [y^2] = 673,496.4$, and $[xy] = -36,538.5$. Do tasks (c), (e), (k-i), (l-i), (p), and (q-i).

9.3 To see if maximum heart rate changes with age, the following numbers are found where X = age and Y = maximum heart rate. $\overline{X} = 47.2, \overline{Y} = 174.8, [x^2] = 4303.2, [y^2] = 5608.5$, and $[xy] = -2915.4$. Do tasks (c), (e), (k-iii), (k-iv), (m), (p), and (p-iii).

9.4 The relationship between height and weight was examined in these active healthy males. X = height, Y = weight, $\overline{X} = 177.7, \overline{Y} = 77.8, [x^2] = 1985.2, [y^2] = 3154.5$, and $[xy] = 1845.6$. Do tasks (c), (e), (m), (p), and (q-i). How do the p-values for the F-test [in part (m)] and for the transformed Z for r compare? There were two normal deviates of values 3.44 and 2.95. If these two people were removed from the calculation, $\overline{X} = 177.5, \overline{Y} = 76.7, [x^2] = 1944.5, [y^2] = 2076.12$, and $[xy] = 1642.5$. How much do the regression coefficients a and b, and correlation coefficient r, change?

Problems 9.5 to 9.8 also refer to the Bruce et al. [1973] paper, as did Example 9.2 and Problems 9.1 to 9.4. The data for 43 active females are given in Table 9.14.

9.5 The duration and VO_2 $_{\text{MAX}}$ relationship for the active females is studied in this problem. $\overline{X} = 514.9, \overline{Y} = 29.1, [x^2] = 251,260.4, [y^2] = 1028.7$, and $[xy] = 12,636.5$. Do tasks (c), (e), (f), (g), (h), (i), (j), (k-iv), (m), (p), and (q-ii). Table 9.15 contains the residuals. If the data are rerun with the sixth case omitted, the values of $\overline{X}, \overline{Y}, [x^2], [y^2]$, and $[xy]$ are changed to 512.9, 29.2, 243,843.1, 1001.5, and 13,085.6, respectively. Find the new estimates a, b, and r. By what percent are they changed?

9.6 With X = age and Y = duration, $\overline{X} = 45.1, \overline{Y} = 514.9, [x^2] = 4399.2, [y^2] = 251,260.4$, and $[xy] = -22,911.3$. For each 10-year increase in age, how much does duration tend to change? What proportion of the variability in VO_2 $_{\text{MAX}}$ is accounted for by age? Do tasks (m) and (q-ii).

9.7 With X = age and Y = maximum heart rate, $\overline{X} = 45.1, \overline{Y} = 180.6, [x^2] = 4399.2, [y^2] = 5474.6$, and $[xy] = -2017.3$. Do tasks (c), (e), (k-i), (k-iv), (l-i), (m), (n) at $X = 30$ and $X = 50$, (o) at $X = 45$, (p), and (q-ii).

9.8 X = height and Y = weight, $\overline{X} = 164.7, \overline{Y} = 61.3, [x^2] = 1667.1, [y^2] = 2607.4$, and $[xy] = 1006.2$. Do tasks (c), (e), (h), (k-iv), (m), and (p). Check that $t^2 = F$. The residual plot is shown in Figure 9.18.

For Problems 9.9 to 9.12, additional Bruce et al. [1973] data are used. Table 9.16 presents the data for 94 sedentary males.

9.9 The duration, X, and VO_2 $_{\text{MAX}}, Y$, give $\overline{X} = 577.1, \overline{Y} = 35.6, [x^2] = 1,425,990.9, [y^2] = 5245.3$, and $[xy] = 78,280.1$. Do tasks (c), (e), (j), (k-i), (k-iv), (l-i), (m), and (p). The normal probability plot is shown in Figure 9.19.

9.10 X = age is related to Y = duration. $\overline{X} = 49.8, \overline{Y} = 577.1, [x^2] = 11,395.7, [y^2] = 1,425,990.9$, and $[xy] = -87,611.9$. Do tasks (c), (e), (m), (p), and (q-ii).

9.11 The prediction of age by maximal heart rate for sedentary males is considered here. $\overline{X} = 49.8, \overline{Y} = 18.6, [x^2] = 11,395.7, [y^2] = 32,146.4$, and $[xy] = -12,064.1$. Do tasks (c), (m), and (p). Verify (to accuracy given) that $(\overline{X}, \overline{Y})$ lies on the regression line.

9.12 The height and weight data give $\overline{X} = 177.3, \overline{Y} = 79.0, [x^2] = 4030.1, [y^2] = 7060.0$, and $[xy] = 2857.0$. Do tasks (c), (e), (k-iv), (n) at $X = 160$, 170, and 180, and (p).

Table 9.14 Exercise Data for Healthy Active Females

Duration	VO2 MAX	Heart Rate	Age	Height	Weight
660	38.1	184	23	177	83
628	38.4	183	21	163	52
637	41.7	200	21	174	61
575	33.5	170	42	160	50
590	28.6	188	34	170	68
600	23.9	190	43	171	68
562	29.6	190	30	172	63
495	27.3	180	49	157	53
540	33.2	184	30	178	63
470	26.6	162	57	161	63
408	23.6	188	58	159	54
387	23.1	170	51	162	55
564	36.6	184	32	165	57
603	35.8	175	42	170	53
420	28.0	180	51	158	47
573	33.8	200	46	161	60
602	33.6	190	37	173	56
430	21.0	170	50	161	62
508	31.2	158	65	165	58
565	31.2	186	40	154	69
464	23.7	166	52	166	67
495	24.5	170	40	160	58
461	30.5	188	52	162	64
540	25.9	190	47	161	72
588	32.7	194	43	164	56
498	26.9	190	48	176	82
483	24.6	190	43	165	61
554	28.8	188	45	166	62
521	25.9	184	52	167	62
436	24.4	170	52	168	62
398	26.3	168	56	162	66
366	23.2	175	56	159	56
439	24.6	156	51	161	61
549	28.8	184	44	154	56
360	19.6	180	56	167	79
566	31.4	184	40	165	56
407	26.6	156	53	157	52
602	30.6	194	52	161	65
488	27.5	190	40	178	64
526	30.9	188	55	162	61
524	33.9	164	39	166	59
562	32.3	185	57	168	68
496	26.9	178	46	156	53

Source: Data from Bruce et al. [1973].

Mehta et al. [1981] studied the effect of the drug dipyridamole on blood platelet function in eight patients with at least 50% narrowing of one or more coronary arteries. Active platelets are sequestered in the coronary arteries, giving reduced platelet function in the coronary venous blood, that is, in blood leaving the heart muscle after delivering oxygen and nutrients. More active platelets in the coronary arteries can lead to thrombosis, blood clots, and a heart attack. Drugs lessening the chance of thrombosis may be useful in treatment.

Table 9.15 Data for Problem 9.5

X	Y	$\widehat{Y}$	Residual	Normal Deviate
660	38.1	36.35	1.75	0.56
628	38.4	34.74	3.66	1.18
637	41.7	35.19	6.51	2.10
575	33.5	32.08	1.42	0.46
590	28.6	32.83	−4.23	−1.37
600	23.9	?	?	?
562	29.6	31.42	−1.82	−0.59
495	27.3	28.05	−0.75	−0.24
540	33.2	?	2.88	0.93
470	26.6	26.80	−0.20	−0.06
408	23.6	23.68	−0.07	−0.02
387	23.1	22.62	0.48	0.15
564	36.6	31.52	5.08	1.64
603	35.8	33.49	2.21	0.75
420	28.0	24.28	3.72	1.20
573	33.8	?	?	0.59
602	33.6	33.43	0.17	0.05
430	21.0	24.78	−3.78	?
508	31.2	28.71	2.49	?
565	31.2	31.57	−0.37	−0.12
464	23.7	26.49	−2.79	−0.90
495	24.5	28.05	−3.55	−1.10
461	30.5	26.34	4.16	1.34
540	25.9	30.32	−4.42	−1.43
588	32.7	?	−0.03	−0.00
498	26.9	?	−1.30	−0.42
483	24.6	27.45	−2.85	−0.92
554	28.8	31.02	−2.22	−0.72
521	25.9	29.36	−3.46	−1.12
436	24.4	25.09	−0.69	−0.22
398	26.3	23.18	3.12	1.01
366	23.2	21.57	1.63	0.53
439	24.6	25.24	−0.64	−0.21
549	28.8	30.77	−1.97	−0.64
360	19.6	21.26	−1.66	−0.54
566	31.4	31.62	−0.22	−0.07
407	26.6	23.63	2.97	0.96
602	30.6	33.43	−2.83	−0.92
488	27.5	27.70	−0.20	−0.06
526	30.9	29.61	1.29	0.42
524	33.9	29.51	4.39	1.42
562	32.3	31.42	0.88	0.28
496	26.9	28.10	−1.20	−0.39

Platelet aggregation measures the extent to which platelets aggregate or cluster together in the presence of a chemical that stimulates clustering or aggregation. The measure used was the percent increase in light transmission after an aggregating agent was added to plasma. (The clustering of the cells make more "holes" in the plasma to let light through.) Two aggregating agents, adenosine diphosphate (ADP) and epinephrine (EPI), were used in this experiment. A second measure taken from the blood count was the count of platelets.

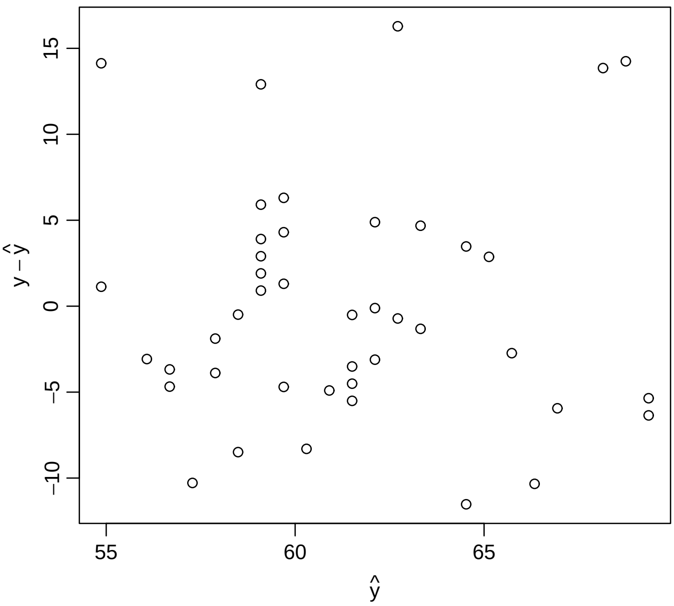

Figure 9.18 Residual plot for Problem 9.8.

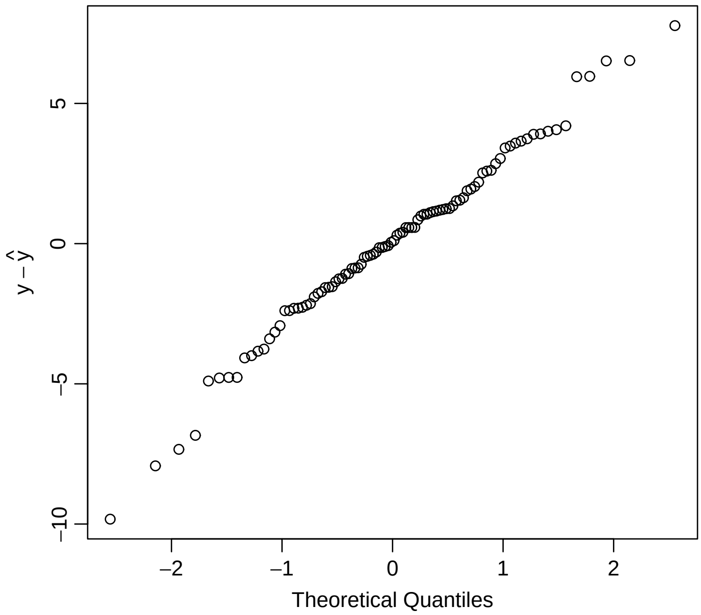

Figure 9.19 Normal probability plot for Problem 9.9.

Blood was sampled from two sites, the aorta (blood being pumped from the heart) and the coronary sinus (blood returning from nourishing the heart muscle). Control samples as well as samples after intravenous infusion of 100 mg of dipyridamole were taken. The data are given in Table 9.17 and 9.18. Problems 9.13 to 9.22 refer to these data.

Table 9.16 Exercise Data for Sedentary Males

Duration	$VO_{2\,\text{MAX}}$	Heart Rate	Age	Height	Weight
360	24.7	168	40	175	96
770	46.8	190	25	168	68
663	41.2	175	41	187	82
679	31.4	190	37	176	82
780	45.7	200	26	179	73
727	47.6	210	28	185	84
647	38.6	208	26	177	77
675	43.2	200	42	162	72
735	48.2	196	30	188	85
827	50.9	184	21	178	69
760	47.2	184	33	182	87
814	41.8	208	31	182	82
778	42.9	184	29	174	73
590	35.1	174	42	188	93
567	37.6	176	40	184	86
648	47.3	200	40	168	80
730	44.4	204	44	183	78
660	46.7	190	44	176	81
663	41.6	184	40	174	78
589	40.2	200	43	193	92
600	35.8	190	41	176	68
480	30.2	174	44	172	84
630	38.4	164	39	181	72
646	41.3	190	39	187	90
630	31.2	190	42	173	69
630	42.6	190	53	181	53
624	39.4	172	57	172	57
572	35.4	164	58	181	58
622	35.9	190	61	168	61
209	16.0	104	74	171	74
536	29.3	175	57	181	57
602	36.7	175	49	175	49
727	43.0	168	53	172	53
260	15.3	112	75	170	75
622	42.3	175	47	185	47
705	43.7	174	51	169	51
669	40.3	174	65	170	65
425	28.5	170	56	167	56
645	38.0	175	50	177	50
576	30.8	184	48	188	48
605	40.2	156	46	187	46
458	29.5	148	61	185	61
551	32.3	188	49	182	49
607	35.5	179	53	179	53
599	35.3	166	55	182	55
453	32.3	160	69	182	69
337	23.8	204	68	176	68
663	41.4	182	47	171	47
603	39.0	180	48	180	48
610	38.6	190	55	180	55
472	31.5	175	53	192	85

Table 9.16 ***(continued)***

Duration	$VO_{2\ MAX}$	Heart Rate	Age	Height	Weight
458	25.7	166	58	178	81
446	24.6	160	50	178	77
532	30.0	160	51	175	82
656	42.0	186	52	176	73
583	34.4	175	52	172	77
595	34.9	180	48	179	78
552	35.5	156	45	167	89
675	38.7	162	58	183	85
622	38.4	186	45	175	76
591	32.4	170	62	175	79
582	33.6	156	63	171	69
518	30.0	166	57	174	75
444	28.9	170	48	180	105
473	29.5	175	52	177	77
490	30.4	168	59	173	74
596	34.4	192	46	190	92
529	37.0	175	54	168	82
652	43.4	156	54	180	85
714	46.0	175	46	174	77
646	43.0	184	45	178	80
551	29.3	160	54	172	86
601	36.8	184	48	169	82
579	35.0	170	54	180	80
325	21.9	140	61	175	76
392	25.4	168	60	180	89
659	40.7	178	45	181	81
631	33.8	184	48	173	74
405	28.8	170	63	168	79
560	35.8	180	60	181	82
615	40.3	190	47	178	78
580	33.4	180	66	173	68
530	39.0	174	47	169	64
495	23.2	145	69	171	84
330	20.5	138	60	185	87
600	36.4	200	50	182	81
443	23.5	166	50	175	84
508	29.7	188	61	188	80
596	43.2	168	57	174	66
461	30.4	170	47	171	65
583	34.7	164	46	187	83
620	37.1	174	61	165	71
620	41.4	190	45	171	79
180	19.8	125	71	185	80

Source: Data from Bruce et al. [1973]

9.13 Relate the control platelet counts in the aorta, X, and coronary sinus, Y. Do tasks (a), (b), (c), (d), (e), compute the (X, Y, $\widehat{Y}$, residual, normal deviate) table, (g), (h), (i), (j), (k-i), (k-iv), (l), (m), (p), (r), and (s).

9.14 Look at the association between the platelet counts in the aorta, X, and coronary sinus, Y, when being treated with dipyridamole. Do tasks (a), (b), (c), (d), (m), (r), and (s).

Table 9.17 Platelet Aggregation Data for Problem 9.12

	Platelet Aggregation (%)							
	Control				Dipyridamole			
	Aorta		Coronary Sinus		Aorta		Coronary Sinus	
Case	EPI	ADP	EPI	ADP	EPI	ADP	EPI	ADP
1	87	75	89	23	89	75	89	35
2	70	23	42	14	45	16	47	18
3	96	75	96	31	96	83	96	84
4	65	51	70	33	70	55	70	57
5	85	16	79	4	69	13	53	22
6	98	83	98	80	83	70	94	88
7	77	14	97	13	84	35	73	67
8	98	50	99	40	85	50	91	48
Mean	85	48	84	30	78	50	77	52
±SEM	5	10	7	8	6	9	7	9

Source: Data from Mehta et al. [1981].

9.15 Examine the control platelet aggregation percent for EPI, X, and ADP, Y, in the aorta. Do tasks (a), (b), (c), (d), (e), and (m).

9.16 Examine the association between the EPI, X, and ADP, Y, in the control situation at the coronary sinus. Do tasks (a), (b), (c), (d), (e), (m), (p), (r), and (s).

9.17 Interpret at the 5% significance level. Look at the platelet aggregation % for epinephrine in the aorta and coronary sinus under the control data. Do tasks (m), (p) and (t), (u). Explain in words how there can be association but no (statistical) difference between the values at the two locations.

9.18 Under dipyridamole treatment, study the platelet aggregation percent for EPI in the aorta, X, and coronary sinus, Y. Do tasks (a), (b), (c), (d), (e), (g), (h), (m), (p), (r), (s), (t), and (u).

9.19 The control aggregation percent for ADP is compared in the aorta, X, and coronary sinus, Y, in this problem. Do tasks (a), (b), (c), (d), (e), (f), (g), (h), (i), (j), (m), (p), and (q-ii).

9.20 Under dipyridamole, the aggregation percent for ADP in the aorta, X, and coronary sinus, Y, is studied here. Do tasks (b), (c), (e), (k-ii), (k-iv), (l-ii), (m), (p), (q-ii), (r), and (s).

9.21 The aortic platelet counts under the control, X, and dipyridamole, Y, are compared in this problem. Do tasks (b), (c), (e), (m), (p), (q-ii), (t), and (u). Do the platelet counts differ under the two treatments? (Use $\alpha = 0.05$.) Are the platelet counts associated under the two treatments? ($\alpha = 0.05$.)

9.22 The coronary sinus ADP aggregation percent was studied during the control period, the X variable, and on dipyridamole, the Y variable. Do tasks (b), (c), (d), (e), (m), and (t). At the 5% significance level, is there a change between the treatment and control periods? Can you show association between the two values? How do you reconcile these findings?

Table 9.18 Platelet Count Data for Problem 9.12

	Platelet Counts ($\times 1000/mm^3$)[a]			
	Control		Dipyridamole	
Case	Aorta	Coronary Sinus	Aorta	Coronary Sinus
1	390	355	455	445
2	240	190	248	205
3	135	125	150	145
4	305	268	285	290
5	255	195	230	220
6	283	307	291	312
7	435	350	457	374
8	290	250	301	284
Mean	292	255	302	284
±SEM	32	29	38	34

Source: Data from Mehta et al. [1981].

Problems 9.23 to 9.29 deal with the data in Tables 9.19 and 9.20. Jensen et al. [1980] studied 19 patients with coronary artery disease. Thirteen had a prior myocardial infarction (heart attack); three had coronary bypass surgery. The patients were evaluated before and after three months or more on a structured supervised training program.

The cardiac performance was evaluated using radionuclide studies while the patients were at rest and also exercising with bicycle pedals (while lying supine). Variables measured included (1) ejection fraction (EF), the fraction of the blood in the left ventricle ejected during a heart beat, (2) heart rate (HR) at maximum exercise in beats per minute, (3) systolic blood pressure (SBP) in millimeters of mercury, (4) the rate pressure product (RPP) maximum heart rate times the maximum systolic blood pressure divided by 100, and (5) the estimated maximum oxygen consumption in cubic centimeters of oxygen per kilogram of body weight per minute.

9.23 The resting ejection fraction is measured before, X, and after, Y, training. $\overline{X} = 0.574$, $\overline{Y} = 0.553$, $[x^2] = 0.29886$, $[y^2] = 0.32541$, $[xy] = 0.23385$, and paired $t = -0.984$. Do tasks (c), (e), (k-iv), (m), and (p). Is there a change in resting ejection fraction demonstrated with six months of exercise training? Are the two ejection fractions associated?

9.24 The ejection fraction at maximal exercise was measured before, X, and after, Y, training. $\overline{X} = 0.556$, $\overline{Y} = 0.564$, $[x^2] = 0.30284$, $[y^2] = 0.46706$, and $[xy] = 0.2809$. Is there association ($\alpha = 0.05$) between the two ejection fractions? If yes, do tasks (c), (k-iii), (l-iii), (p), and (q-ii). Is there a change ($\alpha = 0.05$) between the two ejection fractions? If yes, find a 95% confidence interval for the average difference.

9.25 The maximum systolic blood pressure was measured before, X, and after, Y, training. $\overline{X} = 173.8$, $\overline{Y} = 184.2$, $[x^2] = 11{,}488.5$, $[y^2] = 10{,}458.5$, $[xy] = 7419.5$, and paired $t = 2.263$. Do tasks (a), (b), (c), (d), (e), (m), (p), and (t). Does the exercise training produce a change? How much? Can we predict individually the maximum SBP after training from that before? How much of the variability in maximum SBP after exercise is accounted for by knowing the value before exercise?

9.26 The before, X, and after, Y, rate pressure product give $\overline{X} = 223.0$, $\overline{Y} = 245.7$, $[x^2] = 58{,}476$, $[y^2] = 85{,}038$, $[xy] = 54{,}465$, and paired $t = 2.256$ (Table 9.21). Do tasks (c), (e), (f), (g), (h), and (m). Find the large-sample p-value for Kendall's tau for association.

Table 9.19 Resting and Maximal Ejection Fraction Measured by Radionuclide Ventriculography, and Maximal Heart Rate

	Resting EF		Maximal EF		Maximal HR	
Case	Pre	Post	Pre	Post	Pre	Post
1	0.39	0.48	0.46	0.48	110	119
2	0.57	0.49	0.51	0.57	120	125
3	0.77	0.63	0.70	0.82	108	105
4	0.48	0.50	0.51	0.51	85	88
5	0.55	0.46	0.45	0.55	107	103
6	0.60	0.50	0.52	0.54	125	115
7	0.63	0.61	0.75	0.68	170	166
8	0.73	0.61	0.53	0.71	160	142
9	0.70	0.68	0.80	0.79	125	114
10	0.66	0.68	0.54	0.43	131	150
11	0.40	0.31	0.42	0.30	135	174
12	0.48	0.46	0.48	0.30	97	94
13	0.63	0.78	0.60	0.75	135	132
14	0.41	0.37	0.41	0.44	127	162
15	0.75	0.54	0.76	0.57	126	148
16	0.58	0.64	0.62	0.72	102	112
17	0.50	0.58	0.54	0.65	145	140
18	0.71	0.81	0.65	0.60	152	145
19	0.37	0.38	0.32	0.31	155	170
Mean	0.57	0.55	0.56	0.56	127	132
±SD	0.13	0.13	0.13	0.16	23	26

Table 9.20 Systolic Blood Pressure, Rate Pressure Product and Estimate $VO_{2\ MAX}$ before (Pre) and after (Post) Training

	Maximal SBP		Maximal RPP		Est. $VO_{2\ MAX}$ ($cm^3/kg \cdot min$)	
Case	Pre	Post	Pre	Post	Pre	Post
1	148	156	163	186	24	30
2	180	196	216	245	28	44
3	185	200	200	210	28	28
4	150	148	128	130	34	38
5	150	156	161	161	20	28
6	164	172	205	198	30	36
7	180	210	306	349	64	54
8	182	176	291	250	44	40
9	186	170	233	194	30	28
10	220	230	288	345	30	30
11	188	205	254	357	28	44
12	120	165	116	155	22	20
13	175	160	236	211	20	36
14	190	180	241	292	36	38
15	140	170	176	252	36	44
16	200	230	204	258	28	36
17	215	185	312	259	44	44
18	165	190	251	276	28	34
19	165	200	256	340	44	52
Mean	174	184	223	246	31	37
±SD	25	24	57	69	8	9

Table 9.21 Blood Pressure Data for Problem 9.26

Maximal SBP				
Pre	Post			
X	Y	$\widehat{Y}$	$Y - \widehat{Y}$	Normal Deviate
163	186	189.90	−3.80	−0.08
216	245	239.16	?	?
200	210	224.26	−14.26	−0.32
128	130	157.20	−27.20	−0.61
161	161	?	−26.94	?
205	198	228.92	−30.92	−0.69
306	349	322.99	26.01	?
291	250	309.02	−59.02	−1.31
233	194	255.00	−61.00	−1.36
288	345	306.22	38.77	0.86
254	357	?	?	?
116	155	146.02	8.98	0.20
236	211	257.79	−46.79	−1.04
241	292	262.45	29.55	0.66
176	252	201.91	50.09	1.12
204	258	227.99	30.01	0.67
312	259	328.58	−69.58	−1.55
251	276	271.76	4.24	0.09
256	340	276.42	63.58	1.42

9.27 The maximum oxygen consumption, $\text{VO}_{2\ \text{MAX}}$, is measured before, X, and after, Y. Here $\overline{X} = 32.53, \overline{Y} = 37.05, [x^2] = 2030.7, [y^2] = 1362.9, [xy] = 54465$, and paired $t = 2.811$. Do tasks (c), (k-ii), (m), (n), at $x = 30$, 35, and 40, (p), (q-ii), and (t).

9.28 The ejection fractions at rest, X, and at maximum exercise, Y, before training is used in this problem. $\overline{X} = 0.574, \overline{Y} = 0.556, [x^2] = 0.29886, [y^2] = 0.30284, [xy] = 0.24379$, and paired $t = -0.980$. Analyze these data, including a scatter diagram, and write a short paragraph describing the change and/or association seen.

9.29 The ejection fractions at rest, X, and after exercises, Y, for the subjects after training: (1) are associated, (2) do not change on the average, (3) explain about 52% of the variability in each other. Justify statements (1)–(3). $\overline{X} = 0.553, \overline{Y} = 0.564, [x^2] = 0.32541, [y^2] = 0.4671, [xy] = 0.28014$, and paired $t = 0.424$.

Problems 9.30 to 9.33 refer to the following study. Boucher et al. [1981] studied patients before and after surgery for isolated aortic regurgitation and isolated mitral regurgitation. The aortic valve is in the heart valve between the left ventricle, where blood is pumped from the heart, and the aorta, the large artery beginning the arterial system. When the valve is not functioning and closing properly, some of the blood pumped from the heart returns (or regurgitates) as the heart relaxes before its next pumping action. To compensate for this, the heart volume increases to pump more blood out (since some of it returns). To correct for this, open heart surgery is performed and an artificial valve is sewn into the heart. Data on 20 patients with aortic regurgitation and corrective surgery are given in Tables 9.22 and 9.23.

"NYHA Class" measures the amount of impairment in daily activities that the patient suffers: I is least impairment, II is mild impairment, III is moderate impairment, and IV is severe impairment; HR, heart rate; SBP, the systolic (pumping or maximum) blood pressure; EF, the ejection fraction, the fraction of blood in the left ventricle pumped out during a beat; EDVI,

Table 9.22 Preoperative Data for 20 Patients with Aortic Regurgitation

Case	Age (yr) and Gender	NYHA Class	HR (beats/min)	SBP (mmHG)	EF	EDVI (mL/m^2)	SVI (mL/m^2)	ESVI (mL/m^2)
1	33M	I	75	150	0.54	225	121	104
2	36M	I	110	150	0.64	82	52	30
3	37M	I	75	140	0.50	267	134	134
4	38M	I	70	150	0.41	225	92	133
5	38M	I	68	215	0.53	186	99	87
6	54M	I	76	160	0.56	116	65	51
7	56F	I	60	140	0.81	79	64	15
8	70M	I	70	160	0.67	85	37	28
9	22M	II	68	140	0.57	132	95	57
10	28F	II	75	180	0.58	141	82	59
11	40M	II	65	110	0.62	190	118	72
12	48F	II	70	120	0.36	232	84	148
13	42F	III	70	120	0.64	142	91	51
14	57M	III	85	150	0.60	179	107	30
15	61M	III	66	140	0.56	214	120	94
16	64M	III	54	150	0.60	145	87	58
17	61M	IV	110	126	0.55	83	46	37
18	62M	IV	75	132	0.56	119	67	52
19	64M	IV	80	120	0.39	226	88	138
20	65M	IV	80	110	0.29	195	57	138
Mean	49		75	143	0.55	162	85	77
±SD	14		14	25	0.12	60	26	43

Table 9.23 Postoperative Data for 20 Patients with Aortic Regurgitation

Case	Age (yr) and Gender	NYHA Class	HR (beats/min)	SBP (mmHG)	EF	EDVI (mL/m^2)	SVI (mL/m^2)	ESVI (mL/m^2)
1	33M	I	80	115	0.38	113	43	43
2	36M	I	100	125	0.58	56	32	24
3	37M	I	100	130	0.27	93	25	68
4	38M	I	85	110	0.17	160	27	133
5	38M	I	94	130	0.47	111	52	59
6	54M	I	74	110	0.50	83	42	42
7	56F	I	85	120	0.56	59	33	26
8	70M	I	85	130	0.59	68	40	28
9	22M	II	120	136	0.33	119	39	80
10	28F	II	92	160	0.32	71	23	48
11	40M	II	85	110	0.47	70	33	37
12	48F	II	84	120	0.24	149	36	113
13	42F	III	84	100	0.63	55	35	20
14	57M	III	86	135	0.33	91	72	61
15	61M	III	100	138	0.34	92	31	61
16	64M	III	60	130	0.30	118	35	83
17	61M	IV	88	130	0.62	63	39	24
18	62M	IV	75	126	0.29	100	29	71
19	64M	IV	78	110	0.26	198	52	147
20	65M	IV	75	90	0.26	176	46	130
Mean	49		87	123	0.40	102	38	65
±SD	14		13	15	0.14	41	11	39

Table 9.24 Preoperative Data for 20 Patients with Mitral Regurgitation

Case	Age (yr) and Gender	NYHA Class	HR (beats/min)	SBP (mmHG)	EF	EDVI (mL/m^2)	SVI (mL/m^2)	ESVI (mL/m^2)
1	23M	II	75	95	0.69	71	49	22
2	31M	II	70	150	0.77	184	142	42
3	40F	II	86	90	0.68	84	57	30
4	47M	II	120	150	0.51	135	67	66
5	54F	II	85	120	0.73	127	93	34
6	57M	II	80	130	0.74	149	110	39
7	61M	II	55	120	0.67	196	131	65
8	37M	III	72	120	0.70	214	150	64
9	52M	III	108	105	0.66	126	83	43
10	52F	III	80	115	0.52	167	70	97
11	52M	III	80	105	0.76	130	99	31
12	56M	III	80	115	0.60	136	82	54
13	58F	III	65	110	0.62	146	91	56
14	59M	III	102	90	0.63	82	52	30
15	66M	III	60	100	0.62	76	47	29
16	67F	III	75	140	0.71	94	67	27
17	71F	III	88	140	0.65	111	72	39
18	55M	IV	80	125	0.66	136	90	46
19	59F	IV	115	130	0.72	96	69	27
20	60M	IV	64	140	0.60	161	97	64
Mean	53		81	121	0.66	131	86	45
±SD	12		17	17	0.09	40	30	19

Table 9.25 Postoperative Data for 20 Patients with Mitral Regurgitation

Case	Age (yr) and Gender	NYHA Class	HR (beats/min)	SBP (mmHG)	EF	EDVI (mL/m^2)	SVI (mL/m^2)	ESVI (mL/m^2)
1	23M	II	90	100	0.60	67	40	27
2	31M	II	95	110	0.64	64	41	23
3	40F	II	80	110	0.77	59	45	14
4	47M	II	90	120	0.36	96	35	61
5	54F	II	100	110	0.41	59	24	35
6	57M	II	75	115	0.54	71	38	33
7	61M	II	140	120	0.41	165	68	97
8	37M	III	95	120	0.25	84	21	63
9	52M	III	100	125	0.43	67	29	38
10	52F	III	90	90	0.44	124	55	69
11	52M	III	98	116	0.55	68	37	31
12	56M	III	61	108	0.56	112	63	49
13	58F	III	88	120	0.50	76	38	38
14	59M	III	100	100	0.48	40	19	21
15	66M	III	85	124	0.51	31	16	15
16	67F	III	84	120	0.39	81	32	49
17	71F	III	100	100	0.44	76	33	43
18	55M	IV	108	124	0.43	63	27	36
19	59F	IV	100	110	0.49	62	30	32
20	60M	IV	90	110	0.36	93	34	60
Mean	53		93	113	0.48	78	36	42
±SD	12		15	9	0.11	30	14	21

the volume of the left ventricle after the heart relaxes (adjusted for physical size, to divide by an estimate of the patient's body surface area (BSA); SVI, the volume of the left ventricle after the blood is pumped out, adjusted for BSA; ESVI, the volume of the left ventricle pumped out during one cycle, adjusted for BSA; ESVI = EDVI − SVI. These values were measured before and after valve replacement surgery. The patients in this study were selected to have left ventricular volume overload; that is, expanded EDVI.

Another group of 20 patients with mitral valve disease and left ventricular volume overload were studied. The mitral valve is the valve allowing oxygenated blood from the lungs into the left ventricle for pumping to the body. Mitral regurgitation allows blood to be pumped "backward" and to be mixed with "new" blood coming from the lungs. The data for these patients are given in Tables 9.24 and 9.25.

9.30 **(a)** The preoperative, X, and postoperative, Y, ejection fraction in the patients with aortic valve replacement gave $\overline{X} = 0.549, \overline{Y} = 0.396, [x^2] = 0.26158, [y^2] = 0.39170, [xy] = 0.21981$, and paired $t = -6.474$. Do tasks (a), (c), (d), (e), (m), (p), and (t). Is there a change? Are ejection fractions before and after surgery related?

(b) The mitral valve cases had $\overline{X} = 0.662, \overline{Y} = 0.478, [x^2] = 0.09592, [y^2] = 0.24812, [xy] = 0.04458$, and paired $t = -7.105$. Perform the same tasks as in part (a).

(c) When the emphasis is on the change, rather than possible association and predictive value, a figure like Figure 9.20 may be preferred to a scatter diagram. Plot the scatter diagram for the aortic regurgitation data and comment on the relative merits of the two graphics.

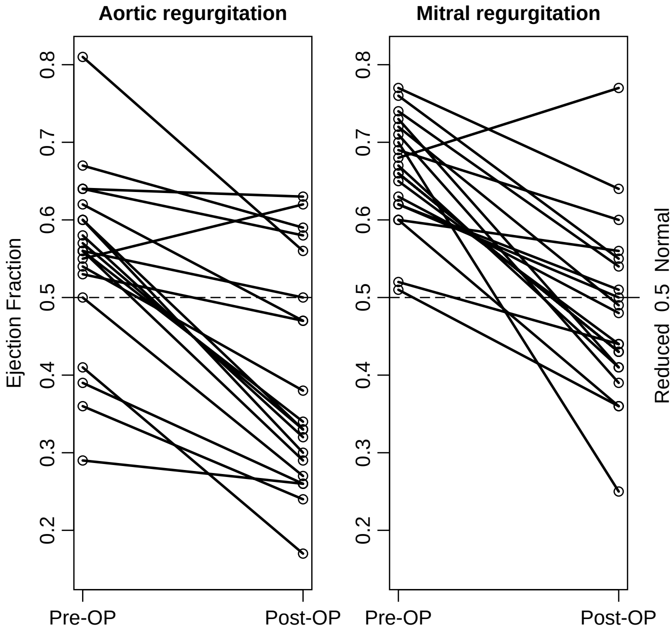

Figure 9.20 Figure for Problem 9.30(c). Individual values for ejection fraction before (pre-OP) and early after (post-OP) surgery are plotted; preoperatively, only four patients with aortic regurgitation had an ejection fraction below normal. After operation, 13 patients with aortic regurgitation and 9 with mitral regurgitation had an ejection fraction below normal. The lower limit of normal (0.50) is represented by a dashed line. (From Boucher et al. [1981].).

Table 9.26 Data for Problem 9.31

X	Y	$\widehat{Y}$	Residuals	Normal Deviate
22	67	51.26	15.74	0.75
42	64	74.18	−10.18	−0.48
30	59	60.42	−1.42	−0.06
66	96	101.68	−5.68	−0.27
34	59	65.01	−6.01	−0.28
39	71	70.74	0.26	0.01
65	165	?	?	?
64	84	99.39	15.29	−0.73
43	67	75.32	?	−0.39
97	124	137.20	−13.20	?
31	68	61.57	?	?
54	112	87.93	24.07	1.14
56	76	?	?	−0.67
30	40	?	−20.42	−0.97
29	31	?	?	?
27	81	56.99	24.01	1.14
39	76	70.74	5.26	0.25
46	63	78.76	−15.76	−0.75
27	62	56.99	5.01	0.24
64	93	99.39	−6.39	−0.30

9.31 **(a)** For the mitral valve cases, we use the end systolic volume index (ESVI) before surgery to try to predict the end diastolic volume index (EDVI) after surgery. $\overline{X} = 45.25, \overline{Y} = 77.9, [x^2] = 6753.8, [y^2] = 16,885.5$, and $[xy] = 7739.5$. Do tasks (c), (d), (e), (f), (h), (j), (k-iv), (m), and (p). Data are given in Table 9.26. The residual plot and normal probability plot are given in Figures 9.21 and 9.22.

(b) If subject 7 is omitted, $\overline{X} = 44.2, \overline{Y} = 73.3, [x^2] = 6343.2, [y^2] = 8900.1$, and $[xy] = 5928.7$. Do tasks (c), (m), and (p). What are the changes in tasks (a), (b), and (r) from part (a)?

(c) For the aortic cases; $\overline{X} = 75.8$, $\overline{Y} = 102.3$, $[x^2] = 35{,}307.2$, $[y^2] = 32{,}513.8$, $[xy] = 27,076$. Do tasks (c), (k-iv), (p), and (q-ii).

9.32 We want to investigate the predictive value of the preoperative ESVI to predict the postoperative ejection fraction, EF. For each part, do tasks (a), (c), (d), (k-i), (k-iv), (m), and (p).

(a) The aortic cases have $\overline{X} = 75.8, \overline{Y} = 0.396, [x^2] = 35307.2, [y^2] = 0.39170$, and $[xy] = 84.338$.

(b) The mitral cases have $\overline{X} = 45.3, \overline{Y} = 0.478, [x^2] = 6753.8, [y^2] = 0.24812$, and $[xy] = -18.610$.

9.33 Investigate the relationship between the preoperative heart rate and the postoperative heart rate. If there are outliers, eliminate (their) effect. Specifically address these questions: (1) Is there an overall change from preop to postop HR? (2) Are the preop and postop HRs associated? If there is an association, summarize it (Tables 9.27 and 9.28).

(a) For the aortic cases, $\sum X = 1502, \sum Y = 17.30, \sum X^2 = 116,446, \sum Y^2 = 152,662$, and $\sum XY = 130,556$. Data are given in Table 9.27.

(b) For the mitral cases: $\sum X = 1640, \sum Y = 1869, \sum X^2 = 140,338, \sum Y^2 = 179,089$, and $\sum XY = 152,860$. Data are given in Table 9.28.

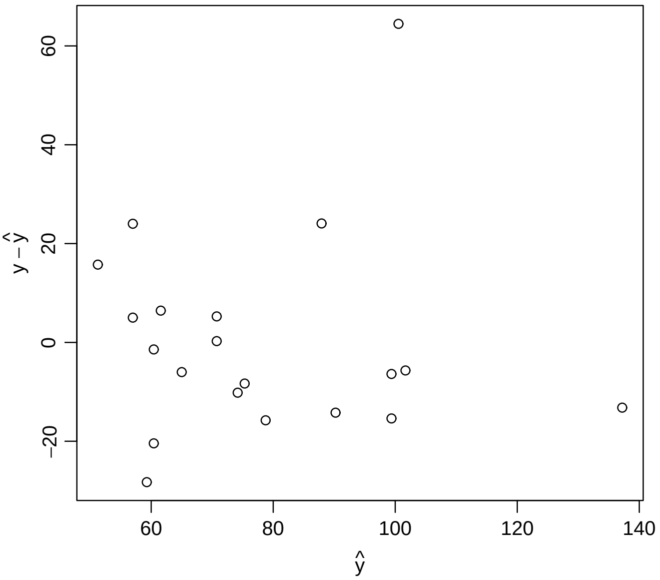

Figure 9.21 Residual plot for Problem 9.31(a).

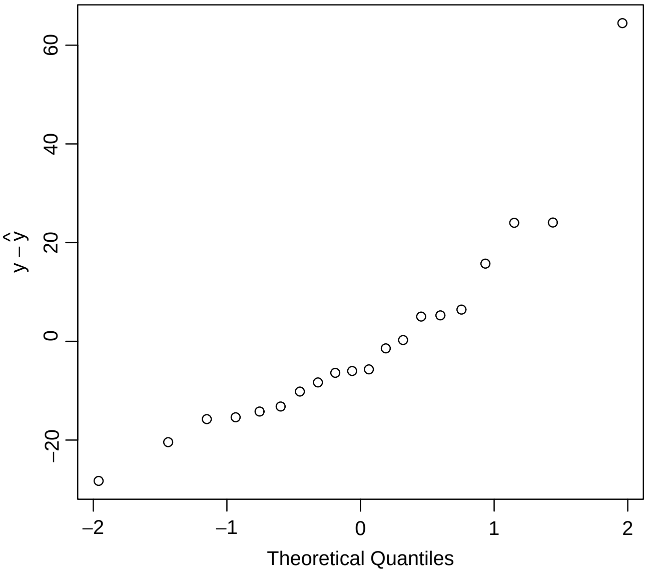

Figure 9.22 Normal probability plot for Problem 9.31(a).

9.34 The Web appendix to this chapter contains county-by-county electoral data for the state of Florida for the 2000 elections for president and for governor of Florida. The major Democratic and Republican parties each had a candidate for both positions, and there were two minor party candidates for president and one for governor. In Palm Beach County a poorly designed ballot was used, and it was suggested that this led to some voters who intended to vote for Gore in fact voting for Buchanan.

Table 9.27 Data for Problem 9.33(a)

X	Y	$\widehat{Y}$	Residuals	Normal Deviate
75	80	86.48	−6.48	−0.51
110	100	92.56	7.44	0.59
75	100	86.48	13.52	1.06
70	85	85.61	0.61	−0.04
68	94	85.27	8.73	0.69
76	74	86.66	−12.66	−1.00
60	85	83.88	1.12	0.08
70	85	85.61	0.61	−0.04
68	120	85.27	34.73	2.73
75	92	86.48	5.52	0.43
65	85	84.75	0.25	0.02
70	84	85.61	−1.61	−0.13
70	84	85.61	−1.61	−0.13
85	86	88.22	−2.22	−0.17
66	100	84.92	15.08	1.19
54	60	82.84	−22.84	−1.80
110	88	92.56	−4.56	0.36
75	75	86.48	−11.48	−0.90
80	78	87.35	−9.35	−0.74
80	75	87.35	−12.35	−0.97

Table 9.28 Data for Problem 9.33(b)

X	Y	$\widehat{Y}$	Residuals	Normal Deviate
75	90	93.93	−3.93	−0.25
70	95	94.27	0.73	0.04
86	80	93.18	−13.18	−0.84
120	90	90.87	−0.87	−0.05
85	100	93.25	6.75	0.43
80	75	93.59	−18.59	−1.19
55	140	95.28	44.72	2.86
72	95	94.13	0.87	0.05
108	100	91.68	8.32	0.53
80	90	93.59	−3.59	−0.23
80	98	93.59	4.41	0.28
80	61	93.95	−32.59	−2.08
65	88	94.61	−6.61	0.42
102	100	92.09	7.91	0.51
60	85	94.94	−9.94	−0.64
75	84	93.93	−9.93	−0.63
88	100	93.04	6.96	0.44
80	108	93.59	14.41	0.92
115	100	91.21	8.79	0.56
64	90	94.67	−4.67	−0.30

(a) Using simple linear regression and graphs, examine whether the data support this claim.

(b) Read the analyses linked from the Web appendix and critically evaluate their claims.

REFERENCES

Acton, F. S. [1984]. *Analysis of Straight-Line Data*. Dover Publications, New York.

Anscombe, F. J. [1973]. Graphs in statistical analysis. *American Statistician*, **27**: 17–21.

Boucher, C. A., Bingham, J. B., Osbakken, M. D., Okada, R. D., Strauss, H. W., Block, P. C., Levine, F. H., Phillips, H. R., and Phost, G. M. [1981]. Early changes in left ventricular volume overload. *American Journal of Cardiology*, **47**: 991–1004.

Bruce, R. A., Kusumi, F., and Hosmer, D. [1973]. Maximal oxygen intake and nomographic assessment of functional aerobic impairment in cardiovascular disease. *American Heart Journal*, **65**: 546–562.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. [1995]. *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

Dern, R. J., and Wiorkowski, J. J. [1969]. Studies on the preservation of human blood: IV. The hereditary component of pre- and post storage erythrocyte adenosine triphosphate levels. *Journal of Laboratory and Clinical Medicine*, **73**: 1019–1029.

Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. [1975]. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, **62**: 531–545.

Draper, N. R., and Smith, H. [1998]. *Applied Regression Analysis*, 3rd ed. Wiley, New York.

Hollander, M., and Wolfe, D. A. [1999]. *Nonparametric Statistical Methods*. 2nd ed. Wiley, New York.

Huber, P. J. [2003]. *Robust Statistics*. Wiley, New York.

Jensen, D., Atwood, J. E., Frolicher, V., McKirnan, M. D., Battler, A., Ashburn, W., and Ross, J., Jr., [1980]. Improvement in ventricular function during exercise studied with radionuclide ventriculography after cardiac rehabilitation. *American Journal of Cardiology*, **46**: 770–777.

Kendall, M. G., and Stuart, A. [1967]. *The Advanced Theory of Statistics*, Vol. 2, *Inference and Relationships*, 2nd ed. Hafner, New York.

Kronmal, R. A. [1993]. Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society, Series A*, **60**: 489–498.

Lumley, T., Diehr, P., Emerson, S., and Chen, L. [2002]. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, **23**: 151–169.

Mehta, J., Mehta, P., Pepine, C. J., and Conti, C. R. [1981]. Platelet function studies in coronary artery disease: X. Effects of dipyridamole. *American Journal of Cardiology*, **47**: 1111–1114.

Neyman, J. [1952]. On a most powerful method of discovering statistical regularities. *Lectures and Conferences on Mathematical Statistics and Probability*. U.S. Department of Agriculture, Washington, DC, pp. 143–154.

U.S. Department of Health, Education, and Welfare [1974].

U.S. Cancer Mortality by County: 1950–59. DHEW Publication (NIH) 74–615. U.S. Government Printing Office, Washington, DC.

Yanez, N. D., Kronmal, R. A., and Shemanski, L. R. [1998]. The effects of measurement error in response variables and test of association of explanatory variables in change models. *Statistics in Medicine* **17**(22): 2597–2606.

CHAPTER 10

Analysis of Variance

10.1 INTRODUCTION

The phrase *analysis of variance* was coined by Fisher [1950], who defined it as "the separation of variance ascribable to one group of causes from the variance ascribable to other groups." Another way of stating this is to consider it as a partitioning of total variance into component parts. One illustration of this procedure is contained in Chapter 9, where the total variability of the dependent variable was partitioned into two components: one associated with regression and the other associated with (residual) variation about the regression line. Analysis of variance models are a special class of linear models.

Definition 10.1. An *analysis of variance model* is a linear regression model in which the predictor variables are classification variables. The categories of a variable are called the *levels* of the variable.

The meaning of this definition will become clearer as you read this chapter.

The topics of analysis of variance and design of experiments are closely related, which has been evident in earlier chapters. For example, use of a paired t-test implies that the data are paired and thus may indicate a certain type of experiment. Similarly, a partitioning of total variation in a regression situation implies that two variables measured are linearly related. A general principle is involved: The analysis of a set of data should be appropriate for the design. We indicate the close relationship between design and analysis throughout this chapter.

The chapter begins with the one-way analysis of variance. Total variability is partitioned into a variance between groups and a variance within groups. The groups could consist of different treatments or different classifications. In Section 10.2 we develop the construction of an analysis of variance from group means and standard deviations, and consider the analysis of variance using ranks. In Section 10.3 we discuss the two-way analysis of variance: A special two-way analysis involving randomized blocks and the corresponding rank analysis are discussed, and then two kinds of classification variables (random and fixed) are covered. Special but common designs are presented in Sections 10.4 and 10.5. Finally, in Section 10.6 we discuss the testing of the assumptions of the analysis of variance, including ways of transforming the data to make the assumptions valid. Notes and specialized topics conclude our discussion.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

A few comments about notation and computations: The formulas for the analysis of variance look formidable but follow a logical pattern. The following rules are followed or held (we remind you on occasion):

1. Indices for groups follow a mnemonic pattern. For example, the subscript i runs from $1, \dots, I$; the subscript j from $1, \dots, J$; k from $1, \dots, K$, and so on.

2. Sums of values of the random variables are indicated by replacing the subscript by a dot. For example,

$$Y_{i\cdot} = \sum_{j=1}^{J} Y_{ij}, \qquad Y_{\cdot jk} = \sum_{i=1}^{I} Y_{ijk}, \qquad Y_{\cdot j \cdot} = \sum_{i=1}^{I}\sum_{k=1}^{K} Y_{ijk}$$

3. It is expensive to print subscripts and superscripts on $\sum$ signs. A very simple rule is that summations are always over the given subscripts. For example,

$$\sum Y_i = \sum_{i=1}^{I} Y_i, \qquad \sum Y_{ijk} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}$$

We may write expressions initially with the subscripts and superscripts, but after the patterns have been established, we omit them. See Table 10.6 for an example.

4. The symbol n_{ij} denotes the number of Y_{ijk} observations, and so on. The total sample size is denoted by n rather than $n_{\dots}$; it will be obvious from the context that the total sample size is meant.

5. The means are indicated by $\overline{Y}_{ij\cdot}$, $\overline{Y}_{\cdot j \cdot}$, and so on. The number of observations associated with a mean is always n with the same subscript (e.g., $\overline{Y}_{ij\cdot} = Y_{ij\cdot}/n_{ij}$ or $\overline{Y}_{\cdot j \cdot} = Y_{\cdot j \cdot}/n_{\cdot j}$).

6. The analysis of variance is an analysis of variability associated with a single observation. This implies that sums of squares of subtotals or totals must always be divided by the number of observations making up the total; for example, $\sum Y_{i\cdot}^2/n_i$ if $Y_{i\cdot}$ is the sum of n_i observations. The rule is then that the divisor is always the number of observations represented by the dotted subscripts. Another example: $Y_{\cdot\cdot}^2/n_{\cdot\cdot}$, since $Y_{\cdot\cdot}$ is the sum of $n_{\cdot\cdot}$ observations.

7. Similar to rules 5 and 6, a sum of squares involving means always have as weighting factor the number of observations on which the mean is based. For example,

$$\sum_{i=1}^{I} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$$

because the mean $\overline{Y}_{i\cdot}$ is based on n_i observations.

8. The ANOVA models are best expressed in terms of means and deviations from means. The computations are best carried out in terms of totals to avoid unnecessary calculations and prevent rounding error. (This is similar to the definition and calculation of the sample standard deviation.) For example,

$$\sum n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 = \sum \frac{Y_{i\cdot}^2}{n_i} - \frac{Y_{\cdot\cdot}^2}{n_{\cdot\cdot}}$$

See Problem 10.25.

10.2 ONE-WAY ANALYSIS OF VARIANCE

10.2.1 Motivating Example

Example 10.1. To motivate the one-way analysis of variance, we return to the data of Zelazo et al. [1972], which deal with the age at which children first walked (see Chapter 5). The experiment involved reinforcement of the walking and placing reflexes in newborns. The walking and placing reflexes disappear by about 8 weeks of age. In this experiment, newborn children were randomly assigned to one of four treatment groups: active exercise; passive exercise; no exercise; or an 8-week control group. Infants in the active-exercise group received walking and placing stimulation four times a day for eight weeks, infants in the passive-exercise group received an equal amount of gross motor stimulation, infants in the no-exercise group were tested along with the first two groups at weekly intervals, and the eight-week control group consisted of infants observed only at 8 weeks of age to control for possible effects of repeated examination. The response variable was age (in months) at which the infant first walked. The data are presented in Table 10.1. For purposes of this example we have added the mean of the fourth group to that group to make the sample sizes equal; this will not change the mean of the fourth group. Equal sample sizes are not required for the one-way analysis of variance.

Assume that the age at which an infant first walks alone is normally distributed with variance σ^2. For the four treatment groups, let the means be $\mu_1, \mu_2, \mu_3,$ and μ_4. Since σ^2 is unknown, we could calculate the sample variance for each of the four groups and come up with a pooled estimate, s_p^2, of σ^2. For this example, since the sample sizes per group are assumed to be equal, this is

$$s_p^2 = \frac{1}{4}(2.0938 + 3.5938 + 2.3104 + 0.7400) = 2.1845$$

But we have one more estimate of σ^2. If the four treatments do not differ ($H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$), the sample means are normally distributed with variance $\sigma^2/6$. The quantity $\sigma^2/6$ can be estimated by $s_{\overline{Y}}^2$, the variance of the sample means. For this example it is

$$s_{\overline{Y}}^2 = 0.87439$$

Table 10.1 Distribution of Ages (in Months) at which Infants First Walked Alone

	Active Group	Passive Group	No-Exercise Group	Eight-Week Control Group
	9.00	11.00	11.50	13.25
	9.50	10.00	12.00	11.50
	9.75	10.00	9.00	12.00
	10.00	11.75	11.50	13.50
	13.00	10.50	13.25	11.50
	9.50	15.00	13.00	12.35[a]
Mean	10.125	11.375	11.708	12.350
Variance	2.0938	3.5938	2.3104	0.7400
$Y_{i\cdot}$	60.75	68.25	70.25	74.10

Source: Data from Zelazo et al. [1972].

[a] This observation is missing from the original data set. For purposes of this illustration, it is estimated by the sample mean. See the text for further discussion.

Hence, $6s_{\bar{Y}}^2 = 5.2463$ is also an estimate of σ^2. Under the null hypothesis, $6s_{\bar{Y}}^2/s_p^2$ will follow an F-distribution. How many degrees of freedom are involved? The quantity $s_{\bar{Y}}^2$ has three degrees of freedom associated with it (since it is a variance based on four observations). The quantity s_p^2 has 20 degrees of freedom (since each of its four component variances has five degrees of freedom). So the quantity $6s_{\bar{Y}}^2/s_p^2$ under the null hypothesis has an F-distribution with 3 and 20 degrees of freedom. What if the null hypothesis is not true (i.e., the μ_1, μ_2, μ_3, and μ_4 are not all equal)? It can be shown that $6s_{\bar{Y}}^2$ then estimates σ^2+ *positive constant*, so that the ratio $6s_{\bar{Y}}^2/s_p^2$ tends to be larger than 1. The usual hypothesis-testing approach is to reject the null hypothesis if the ratio is "too large," with the critical value selected from an F-table. The analysis is summarized in an *analysis of variance table* (ANOVA), as in Table 10.2.

The variances $6s_{\bar{Y}}^2/s_p^2$ and s_p^2 are called *mean squares* for reasons to be explained later. It is clear that the first variance measures the variability between groups, and the second measures the variability within groups. The F-ratio of 2.40 is referred to an F-table. The critical value at the 0.05 level is $F_{3,20,0.95} = 3.10$, the observed value 2.40 is smaller, and we do not reject the null hypothesis at the 0.05 level. The data are displayed in Figure 10.1. From the graph it can be seen that the active group had the lowest mean value. The nonsignificance of the F-test suggests that the active group mean is not significantly lower than that of the other three groups.

Table 10.2 Simplified ANOVA Table of Data of Table 10.1

Source of Variation	d.f.	MS	F-Ratio
Between groups	3	$6s_{\bar{Y}}^2 = 5.2463$	$\dfrac{6s_{\bar{Y}}^2}{s_p^2} = \dfrac{5.2463}{2.1845} = 2.40$
Within groups	20	$s_p^2 = 2.1845$	

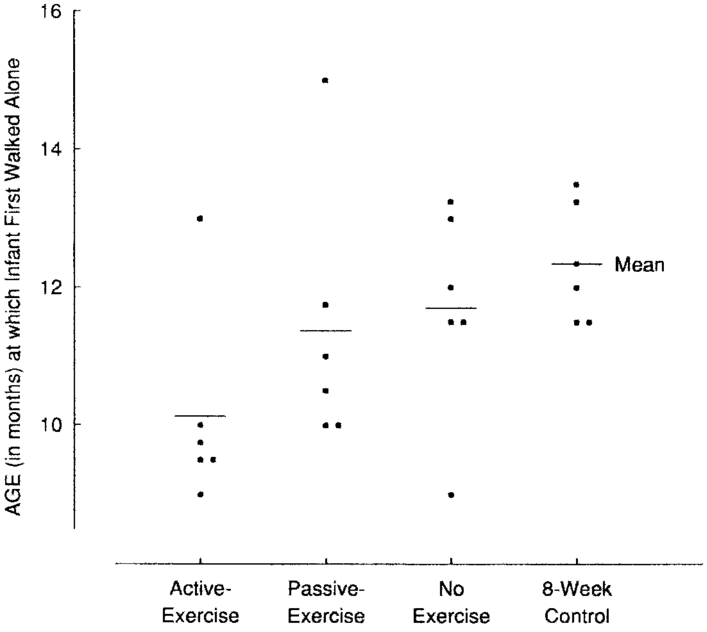

Figure 10.1 Distribution of ages at which infants first walked alone. (Data from Zelazo et al. [1972]; see Table 10.1.)

10.2.2 Using the Normal Distribution Model

Basic Approach

The one-way analysis of variance is a generalization of the t-test. As in the motivating example above, it can be used to examine the age at which groups of infants first walk alone, each group receiving a different treatment; or we may compare patient costs (in dollars per day) in a sample of hospitals from a metropolitan area. (There is a subtle distinction between the two examples; see Section 10.3.4 for a further discussion.)

Definition 10.2. An analysis of variance of observations, each of which belongs to one of I disjoint groups, is a *one-way analysis of variance of I groups*.

Suppose that samples are taken from I normal populations that differ at most in their means; the observations can be modeled by

$$Y_{ij} = \mu_i + \epsilon_{ij}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, n_i \tag{1}$$

The mean for normal population i is μ_i; we assume that there are n_i observations from this population. Also, by assumption, the ϵ_{ij} are independent $N(0, \sigma^2)$ variables. In words: Y_{ij} denotes the jth sample from a population with mean μ_i and variance σ^2. If $I = 2$, you can see that this is precisely the model for the two-sample t-test.

The only difference between the situation now and that of Section 10.2.1 is that we allow the number of observations to vary from group to group. The within-group estimate of the variance σ^2 now becomes a weighted sum of sample variances. Let s_i^2 be the sample variance from group i, where $i = 1, \ldots, I$. The within-group estimate σ^2 is

$$\frac{\sum (n_i - 1)s_i^2}{\sum (n_i - 1)} = \frac{\sum (n_i - 1)s_i^2}{n - I}$$

where $n = n_1 + n_2 + \cdots + n_I$ is the total number of observations.

Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$, the variability among the group of sample means also estimates σ^2. We will show below that the proper expression is

$$\frac{\sum n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}{I - 1}$$

where

$$\overline{Y}_{i\cdot} = \sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}$$

is the sample mean for group i, and

$$\overline{Y}_{\cdot\cdot} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \frac{Y_{ij}}{n} = \sum \frac{n_i \overline{Y}_{i\cdot}}{n}$$

is the grand mean. These quantities can again be arranged in an ANOVA table, as displayed in Table 10.3. Under the null hypothesis, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$, the quantity A/B in Table 10.3 follows an F-distribution with $(I - 1)$ and $(n - I)$ degrees of freedom.

We now reanalyze our first example in Section 10.2.1, deleting the sixth observation, 12.35, in the eight-week control group. The means and variances for the four groups are now:

Table 10.3 One-Way ANOVA Table for I Groups and n_i Observations per Group ($i = 1, \dots, I$)

Source of Variation	d.f.	MS	F-Ratio
Between groups	$I-1$	$A = \dfrac{\sum n_i(\overline{Y}_i. - \overline{Y}..)^2}{I-1}$	A/B
Within groups	$n-I$	$B = \sum \dfrac{(n_i-1)s_i^2}{n-I}$	

Table 10.4 ANOVA of Data from Example 10.1, Omitting the Last Observation

Source of Variation	d.f.	MS	F-Ratio
Between groups	3	4.9253	2.14
Within groups	19	2.2994	

	Active	Passive	No Exercise	Control	Overall
Mean ($\overline{Y}_i.$)	10.125	11.375	11.708	12.350	11.348
Variance (s_i^2)	2.0938	3.5938	2.3104	0.925	—
n_i	6	6	6	5	23

Therefore,

$$\begin{aligned}\sum n_i(\overline{Y}_i. - \overline{Y}..)^2 &= 6(10.125 - 11.348)^2 + 6(11.375 - 11.348)^2 \\ &\quad + 6(11.708 - 11.348)^2 + 5(12.350 - 11.348)^2 \\ &= 14.776\end{aligned}$$

The between-group mean square is $14.776/(4-1) = 4.9253$. The within-group mean square is

$$\frac{1}{23-4}[5(2.0938) + 5(3.5938) + 5(2.3104) + 4(0.925)] = 2.2994$$

The ANOVA table is displayed in Table 10.4.

The critical value $F_{3,19,0.95} = 3.13$, so again, the four groups do not differ significantly.

Linear Model Approach

In this section we approach the analysis of variance using linear models. The model $Y_{ij} = \mu_i + \epsilon_{ij}$ is usually written as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, \dots, I, \quad j = 1, \dots, n_i \tag{2}$$

The quantity μ is defined as

$$\mu = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \frac{\mu_i}{n}$$

where $n = \sum n_i$ (the total number of observations). The quantity α_i is defined as $\alpha_i = \mu - \mu_i$. This implies that

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i} \alpha_i = \sum n_i\alpha_i = 0 \tag{3}$$

Definition 10.3. The quantity $\alpha_i = \mu - \mu_i$ is the *main effect* of the ith population.

Comments:

1. The symbol α with a subscript will denote an element of the analysis of variance model, not the type I error. The context will make it clear which meaning is intended.
2. The equation $\sum n_i\alpha_i = 0$ is a constraint. It implies that fixing any $(I-1)$ of the main effects determines the remaining value.

If we hypothesize that the I populations have the same means,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$$

then an equivalent statement is

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0 \quad \text{or} \quad H_0 : \alpha_i = 0, \qquad i = 1, \ldots, I$$

How are the quantities $\mu_i, i = 1, \ldots, I$ and σ^2 to be estimated from the data? (Or, equivalently, $\mu, \alpha_i, i = 1, \ldots, I$ and σ^2.) Basically, we follow the same strategy as in Section 10.2.1. The variances within the I groups are pooled to provide an estimate of σ^2, and the variability between groups provides a second estimate under the null hypothesis. The data can be displayed as shown in Table 10.5. For this set of data, a partitioning can be set up that mimics the model defined by equation (2):

$$\left.\begin{array}{ll}\text{Model}: & Y_{ij} = \mu + \alpha_i + \epsilon_{ij}\\ \text{Data}: & Y_{ij} = \overline{Y}.. + a_i + e_{ij}\end{array}\right\} i = 1, \ldots, I, \quad j = 1, \ldots, n_i \tag{4}$$

where $a_i = \overline{Y}_i. - \overline{Y}..$ and $e_{ij} = Y_{ij} - \overline{Y}_i.$ for $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$. It is easy to verify that the condition $\sum n_i\alpha_i = 0$ is mimicked by $\sum n_i a_i = 0$. Each data point is partitioned into three component estimates:

$$Y_{ij} = \overline{Y}.. + (\overline{Y}_i - \overline{Y}..) + (Y_{ij} - \overline{Y}_i.) = \text{mean} + i\text{th main effect} + \text{error}$$

Table 10.5 Pooled Variances of I Groups

	Sample				
	1	2	3	$\cdots$	I
	Y_{11}	Y_{21}	Y_{31}	$\cdots$	Y_{I1}
	Y_{12}	Y_{22}	Y_{32}	$\cdots$	Y_{I2}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Y_{1n_1}	Y_{2n_2}	Y_{3n_3}	$\cdots$	Y_{In_I}
Observations	n_1	n_2	n_3	$\cdots$	n_I
Means	$\overline{Y}_1.$	$\overline{Y}_2.$	$\overline{Y}_3.$	$\cdots$	$\overline{Y}_I.$
Totals	$Y_1.$	$Y_2.$	$Y_3.$	$\cdots$	$Y_I.$

The expression on the right side of Y_{ij} is an algebraic identity. It is a remarkable property of this partitioning that the sum of squares of the Y_{ij} is equal to the sum of the three sums of squares of the elements on the right side:

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}^2 &= \sum_{i=1}^{I}\sum_{j=1}^{n_i} \overline{Y}_{..}^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\overline{Y}_{i\cdot} - \overline{Y}_{..})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2 \\ &= n\overline{Y}_{..}^2 + \sum_{i=1}^{I} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{..})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2 \qquad (5)\end{aligned}$$

and the degrees of freedom can also be partitioned: $n = 1+(I-1)+(n-I)$. You will recognize the terms on the right side as the ingredients needed for setting up the analysis of variance table as discussed in the preceding section. It should also be noted that the quantities on the right side are random variables (since they are based on statistics). It can be shown that their expected values are

$$E\left(\sum n_i(\overline{Y}_{i\cdot} - \overline{Y}_{..})^2\right) = \sum n_i\alpha_i^2 + (I-1)\sigma^2 \qquad (6)$$

and

$$E\left(\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2\right) = (n-I)\sigma^2 \qquad (7)$$

If the null hypothesis $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$ is true (i.e., $\mu_1 = \mu_2 = \cdots = \mu_I = \mu$), then $\sum n_i\alpha_i^2 = 0$, and both of the terms above provide an estimate of σ^2 [after division by $(I-1)$ and $(n-I)$, respectively]. This layout and analysis is summarized in Table 10.6.

The quantities making up the component parts of equation (5) are called *sums of squares* (SS). "Grand mean" is usually omitted; it is used to test the null hypothesis that $\mu = 0$. This is rarely of very much interest, particularly if the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ is rejected (but see Example 10.7). "Between groups" is used to test the latter null hypothesis, or the equivalent hypothesis, $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$.

Before returning to Example 10.1, we give a few computational notes.

Computational Notes

As in the case of calculating standard deviations, the computations usually are not based on the means but rather, on the group totals. Only three quantities have to be calculated for the one-way ANOVA. Let

$$Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij} = \text{total in the } i\text{th treatment group} \qquad (8)$$

and

$$Y_{..} = \sum Y_{i\cdot} = \text{grand total} \qquad (9)$$

The three quantities that have to be calculated are

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}^2 = \sum\sum Y_{ij}^2, \qquad \sum_{i=1}^{I}\frac{Y_{i\cdot}^2}{n_i} = \sum\frac{Y_{i\cdot}^2}{n_i}, \qquad \frac{Y_{..}^2}{n}$$

Table 10.6 Layout for the One-Way ANOVA

Source of Variation	d.f.	SS[a]	MS	F-Ratio	d.f. of F-Ratio	E(MS)	Hypothesis Tested
Grand mean	1	$\text{SS}_\mu = n\overline{Y}^2_{..}$	$\text{MS}_\mu = \text{SS}_\mu$	$\dfrac{\text{MS}_\mu}{\text{MS}_\epsilon}$	$(1, n-1)$	$n\mu^2 + \sigma^2$	$\mu = 0$
Between groups (main effects)	$I-1$	$\text{SS}_\alpha = \sum n_i(\overline{Y}_{i\cdot} - \overline{Y}_{..})^2$	$\text{MS}_\alpha = \dfrac{\text{SS}_\alpha}{I-1}$	$\dfrac{\text{MS}_\alpha}{\text{MS}_\epsilon}$	$(I-1, n-I)$	$\dfrac{\sum n_i\alpha_i^2}{I-1} + \sigma^2$	$\alpha_1 = \cdots = \alpha_I$ or $\mu_1 = \cdots = \mu_I$
Within groups (residuals)	$n-I$	$\text{SS}_\epsilon = \sum\sum(\overline{Y}_{ij} - \overline{Y}_{i\cdot})^2$	$\text{MS}_\epsilon = \dfrac{\text{SS}_\epsilon}{n-I}$	—	—	σ^2	σ^2
Total	n	$\sum\sum Y_{ij}^2$					

[a] Summation is over all displayed subscripts.

$$
\begin{aligned}
\text{Model}: \quad Y_{ij} &= \mu_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \text{iid } N(0, \sigma^2) \\
&= \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, n_i \\
\text{Data}: \quad Y_{ij} &= \overline{Y}_{..} + (\overline{Y}_{i\cdot} - \overline{Y}_{..}) + (Y_{ij} - Y_{i\cdot})
\end{aligned}
$$

(iid = independent and identically distributed). An equivalent model is

$$Y_{ij} \sim N(\mu_i, \sigma^2), \qquad \text{where } Y_{ij}\text{'s are independent}$$

where $n = \sum n_i =$ total observations. It is easy to establish the following relationships:

$$\text{SS}_\mu = \frac{Y_{\cdot\cdot}^2}{n} \tag{10}$$

$$\text{SS}_\alpha = \sum \frac{Y_{i\cdot}^2}{n_i} - \frac{Y_{\cdot\cdot}^2}{n} \tag{11}$$

$$\text{SS}_\epsilon = \sum\sum Y_{ij}^2 - \sum \frac{Y_{i\cdot}^2}{n_i} \tag{12}$$

The subscripts are omitted.

We have an algebraic identity in $\sum\sum Y_{ij}^2 = \text{SS}_\mu + \text{SS}_\alpha + \text{SS}_\epsilon$. Defining SS_{TOTAL} as $\text{SS}_{\text{TOTAL}} = \sum\sum Y_{ij}^2 - \text{SS}_\mu$, we get $\text{SS}_{\text{TOTAL}} = \text{SS}_\alpha + \text{SS}_\epsilon$ and degrees of freedom $(n-1) = (i-1)+(n-I)$. This formulation is a simplified version of equation (5). Note that the original data are needed only for $\sum\sum Y_{ij}^2$; all other sums of squares can be calculated from group or overall totals.

Continuing Example 10.1, omitting again the last observation (12.35):

$$\sum\sum Y_{ij}^2 = 9.00^2 + 9.50^2 + \cdots + 11.50^2 = 3020.2500$$

$$\sum \frac{Y_{i\cdot}^2}{n_i} = \frac{60.75^2}{6} + \frac{68.25^2}{6} + \frac{70.25^2}{6} + \frac{61.75^2}{5} = 2976.5604$$

$$\frac{Y_{\cdot\cdot}^2}{n} = \frac{261.00^2}{23} = 2961.7826$$

The ANOVA table omitting rows for SS_μ and SS_{TOTAL} becomes

Source of Variation	d.f.	SS	MS	F-Ratio
Between groups	3	14.7778	4.9259	2.14
Within groups	19	43.6896	2.2995	

The numbers in this table are not subject to rounding error and differ slightly from those in Table 10.4.

Estimates of the components of the expected mean squares of Table 10.6 can now be obtained. The estimate of σ^2 is $\widehat{\sigma}^2 = 2.2995$, and the estimate of $\sum n_i\alpha_i^2/(I-1)$ is

$$\frac{\sum n_i\widehat{\alpha}_i^2}{I-1} = 4.9259 - 2.2995 = 2.6264$$

How is this quantity to be interpreted in view of the nonrejection of the null hypothesis? Theoretically, the quantity can never be less than zero (all the terms are positive). The best interpretation looks back to MS_α, which is a random variable which (under the null hypothesis) estimates σ^2. Under the null hypothesis, MS_α and MS_ϵ both estimate σ^2, and $\sum n_i\alpha_i^2/(I-1)$ is zero.

10.2.3 One-Way ANOVA from Group Means and Standard Deviation

In many research papers, the raw data are not presented but rather, the means and standard deviations (or variances) for each of the, say, I treatment groups under consideration. It is instructive to construct an analysis of variance from these data and see how the assumption of the equality of the population variances for each of the groups enters in. Advantages of constructing the ANOVA table are:

1. Pooling the sample standard deviations (variances) of the groups produces a more precise estimate of the population standard deviation. This becomes very important if the sample sizes are small.
2. A simultaneous comparison of all group means can be made by means of the F-test rather than by a series of two-sample t-tests. The analysis can be modeled on the layout in Table 10.3.

Suppose that for each of I groups the following quantities are available:

Group	Sample Size	Sample Mean	Sample Variance
i	n_i	$\overline{Y}_{i\cdot}$	s_i^2

The quantities $n = \sum n_i$, $Y_{i\cdot} = n_i\overline{Y}_{i\cdot}$, and $Y_{\cdot\cdot} = \sum Y_{i\cdot}$ can be calculated. The "within groups" SS is the quantity B in Table 10.3 times $n - I$, and the "between groups" SS can be calculated as

$$\text{SS}_\alpha = \sum \frac{Y_{i\cdot}^2}{n_i} - \frac{Y_{\cdot\cdot}^2}{n}$$

Example 10.2. Barboriak et al. [1972] studied risk factors in patients undergoing coronary bypass surgery for coronary artery disease. The authors looked for an association between cholesterol level (a putative risk factor) and the number of diseased blood vessels. The data are:

Diseased Vessels (i)	Sample Size (n_i)	Mean Cholesterol Level ($\overline{Y}_{i\cdot}$)	Standard Deviation (s_i)
1	29	260	56.0
2	49	289	87.5
3	76	295	72.4

Using equations (8)–(12), we get $n = 29 + 49 + 76 = 154$,

$$Y_{1\cdot} = n_1\overline{Y}_{1\cdot} = 29(260) = 7540, \qquad Y_{3\cdot} = n_3\overline{Y}_{3\cdot} = 76(295) = 22{,}420$$

$$Y_{2\cdot} = n_2\overline{Y}_{2\cdot} = 49(289) = 14{,}161, \qquad Y_{\cdot\cdot} = \sum n_i\overline{Y}_{i\cdot} = \sum Y_{i\cdot} = 44,121$$

$$\begin{aligned}\text{SS}_\alpha &= \frac{7540^2}{29} + \frac{14{,}161^2}{49} + \frac{22{,}420^2}{76} - \frac{44{,}121^2}{154}\\ &= 12{,}666{,}829.0 - 12{,}640{,}666.5 = 26{,}162.5\\ \text{SS}_\epsilon &= \sum(n_i - 1)s_i^2 = 28 \times 56.0^2 + 48 \times 87.5^2 + 75 \times 72.4^2 = 848,440\end{aligned}$$

The ANOVA table (Table 10.7) can now be constructed. (There is no need to calculate the total SS.)

The critical value for F at the 0.05 level with 2 and 120 degrees of freedom is 3.07; the observed F-value does not exceed this critical value, and the conclusion is that the average cholesterol levels do not differ significantly.

Table 10.7 ANOVA of Data of Example 10.2

Source	d.f.	SS	MS	F-Ratio
Main effects (disease status)	2	26,162.50	13,081.2	2.33
Residual (error)	151	848,440.0	5,618.5	—

10.2.4 One-Way ANOVA Using Ranks

In this section the rank procedures discussed in Chapter 8 are extended to the one-way analysis of variance. For three or more groups, Kruskal and Wallis [1952] have given a one-way ANOVA based on ranks. The model is

$$Y_{ij} = \mu_i + \epsilon_{ij}, \qquad i = 1, \dots, I, \quad j = 1, \dots, n_i$$

The only assumption about the ϵ_{ij} is that they are independently and identically distributed, not necessarily normal. It is assumed that there are no ties among the observations. For a small number of ties in the data, the average of the ranks for the tied observations is usually assigned (see Note 10.1). The test procedure will be conservative in the presence of ties (i.e., the p-value will be smaller when adjustment for ties is made).

The null hypothesis of interest is

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$$

The procedure for obtaining the ranks is similar to that for the two-sample Wilcoxon rank-sum procedure: The $n_1 + n_2 + \cdots + n_I = n$ observations are ranked without regard to which group they belong. Let $R_{ij} =$ rank of observation j in group i.

$$T_{\text{KW}} = \frac{12 \sum n_i (\overline{R}_i. - \overline{R}..)^2}{n(n+1)} \tag{13}$$

where $\overline{R}_i.$ is the average of the ranks of the observations in group i:

$$\overline{R}_i. = \sum_{j=1}^{n_i} \frac{R_{ij}}{n_i}$$

and $\overline{R}..$ is the grand mean of the ranks. The value of the mean $(\overline{R}..)$ must be $(n+1)/2$ (why?) and this provides a partial check on the arithmetic. Large values of T_{KW} imply that the average ranks for the group differ, so that the null hypothesis is rejected for large values of this statistic. If the null hypothesis is true and all the n_i become large, the distribution of the statistic T_{KW} approaches a χ^2-distribution with $I - 1$ degrees of freedom. Thus, for large sample sizes, critical values for T_{KW} can be read from a χ^2-table. For small values of n_i, say, in the range 2 to 5, exact critical values have been tabulated (see, e.g., CRC Table X.9 [Beyer, 1968]). Such tables are available for three or four groups.

An equivalent formula for T_{KW} as defined by equation (13) is

$$T_{\text{KW}} = \frac{12 \sum R_{i.}^2 / n_i}{n(n+1)} - 3(n+1) \tag{14}$$

where $R_i.$ is the total of the ranks for the ith group.

Example 10.3. Chikos et al. [1977] studied errors in the reading of chest x-rays. The opinion of 10 radiologists about the status of the left ventricle of the heart ("normal" vs. "abnormal") was compared to data obtained by ventriculography (which consists of the insertion of a catheter into the left ventricle, injection of a radiopague fluid, and the taking of a series of x-rays). The ventriculography data were used to classify a subject's left ventricle as "normal" or "abnormal." Using this gold standard, the percentage of errors for each radiologist was computed. The authors were interested in the effect of experience, and for this purpose the radiologists were classified into one of three groups: senior staff, junior staff, and residents. The data for these three groups are shown in Table 10.8.

To compute the Kruskal–Wallis statistic T_{KW}, the data are ranked disregarding groups:

Observation	7.3	7.4	10.6	13.3	14.7	15.0	20.7	22.7	23.0	26.6
Rank	1	2	3	4	5	6	7	8	9	10
Group	1	1	2	2	3	2	2	3	3	3

The sums and means of the ranks for each group are calculated to be

$$\begin{aligned} R_{1\cdot} &= 1 + 2 = 3, & \overline{R}_{1\cdot} &= 1.5 \\ R_{2\cdot} &= 3 + 4 + 6 + 7 = 20, & \overline{R}_{2\cdot} &= 5.0 \\ R_{3\cdot} &= 5 + 8 + 9 + 10 = 32, & \overline{R}_{3\cdot} &= 8.0 \end{aligned}$$

[The sum of the ranks is $R_1 + R_2 + R_3 = 55 = (10 \times 11)/2$, providing a partial check of the ranking procedure.]

Using equation (14), the T_{KW} statistic has a value of

$$T_{\text{KW}} = \frac{12(3^2/2 + 20^2/4 + 32^2/4)}{10(10 + 1)} - 3(10 + 1) = 6.33$$

This value can be referred to as a χ^2-table with two degrees of freedom. The p-value is $0.025 < p < 0.05$. The exact p-value can be obtained from, for example, Table X.9 of the CRC tables [Beyer, 1968]. (This table does not list the critical values of T_{KW} for $n_1 = 2$, $n_2 = 4$, $n_3 = 4$; however, the order in which the groups are labeled does not matter, so that the values $n_1 = 4, n_2 = 4$, and $n_3 = 2$ may be used.) From this table it is seen that $0.011 < p < 0.046$, indicating that the chi-square approximation is satisfactory even for these small sample sizes. The conclusion from both analyses is that among staff levels there are significant differences in the accuracy of reading left ventricular abnormality from a chest x-ray.

Table 10.8 Data for Three Radiologist Groups

	Senior Staff	Junior Staff	Residents
i	1	2	3
n_i	2	4	4
Y_{ij}	7.3	13.3	14.7
	7.4	10.6	23.0
(Percent error)		15.0	22.7
		20.7	26.6

10.3 TWO-WAY ANALYSIS OF VARIANCE

10.3.1 Using the Normal Distribution Model

In this section we consider data that arise when a response variable can be classified in two ways. For example, the response variable may be blood pressure and the classification variables type of drug treatment and gender of the subject. Another example arises from classifying people by type of health insurance and race; the response variable could be number of physician contacts per year.

Definition 10.4. An analysis of variance of observations, each of which can be classified in two ways is called a *two-way analysis of variance.*

The data are usually displayed in "cells," with the row categories the values of one classification variable and the columns representing values of the second classification variable.

A completely general two-way ANOVA model with each cell mean any value could be

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \tag{15}$$

where $i = 1, \ldots, I, j = 1, \ldots, J$, and $k = 1, \ldots, n_{ij}$. By assumption, the ϵ_{ijk} are iid $N(0, \sigma^2)$: independently and identically distributed $N(0, \sigma^2)$. This model could be treated as a one-way ANOVA with IJ groups with a test of the hypothesis that all μ_{ij} are the same, implying that the classification variables are not related to the response variable. However, if there is a significant difference among the IJ group means, we want to know whether these differences can be attributed to:

1. One of the classification variables,
2. Both of the classification variables acting separately (no interaction), or
3. Both of the classification variables acting separately and jointly (interaction).

In many situations involving classification variables, the mean μ_{ij} may be modeled as the sum of two terms, an effect of variable 1 plus an effect of variable 2:

$$\mu_{ij} = u_i + v_j, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J \tag{16}$$

Here μ_{ij} depends, in an additive fashion, on the ith level of the first variable and the jth level of the second variable. One problem is that u_i and v_j are not defined uniquely; for any constant C, if $\mu_i^* = u_i + C$ and $v_j^* = v_j - C$, then $\mu_{ij} = u_i^* + v_j^*$. Thus, the values of u_i and v_j can be pinned down to within a constant. The constant is specified by convention and is associated with the experimental setup. Suppose that there are n_{ij} observations at the ith level of variable 1 and the jth level of variable 2. The frequencies of observations can be laid out in a contingency table as shown in Table 10.9.

The experiment has a total of $n_{\cdot\cdot}$ observations. The notation is identical to that used in a two-way contingency table layout. (A major difference is that all the frequencies are usually chosen by the experimenter; we shall return to this point when talking about a balanced ANOVA design.) Using the model of equation (16), the value of μ_{ij} is defined as

$$\mu_{ij} = \mu + \alpha_i + \beta_j \tag{17}$$

where $\mu = \sum\sum n_{ij}\mu_{ij}/n_{\cdot\cdot}$, $\sum n_{i\cdot}\alpha_i = 0$, and $\sum n_{\cdot j}\beta_j = 0$. This is similar to the constraints put on the one-way ANOVA model; see equations (2) and (10.3), and Problem 10.25(d).

Example 10.4. An experimental setup involves two explanatory variables, each at three levels. There are 24 observations distributed as shown in Table 10.10. The effects of the first

Table 10.9 Contingency Table for Variables

Levels of	Levels of Variable 2						
Variable 1	1	2	$\cdots$	j	$\cdots$	J	Total
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1j}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot J}$	$n_{\cdot\cdot}$

Table 10.10 Observation Data

Levels of	Levels of Variable 2			
Variable 1	1	2	3	Total
1	2	2	2	6
2	3	3	3	9
3	3	3	3	9
Total	8	8	8	24

Table 10.11 Data for Variable Effects

Effects of the	Effects of the Second Variable			
First Variable	$\beta_1 = 1$	$\beta_2 = -3$	$\beta_3 = 2$	Total
$\alpha_1 = 3$	$\mu_{11} = 24$	$\mu_{12} = 20$	$\mu_{13} = 25$	$\mu_{1\cdot} = 23$
$\alpha_2 = 6$	$\mu_{21} = 27$	$\mu_{22} = 23$	$\mu_{23} = 28$	$\mu_{2\cdot} = 26$
$\alpha_3 = -8$	$\mu_{31} = 13$	$\mu_{32} = 9$	$\mu_{33} = 14$	$\mu_{3\cdot} = 12$
Total	$\mu_{\cdot 1} = 21$	$\mu_{\cdot 2} = 17$	$\mu_{\cdot 3} = 22$	$\mu = 20$

variable are assumed to be $\alpha_1 = 3, \alpha_2 = 6$, and $\alpha_3 = -8$; the effects of the second variable are $\beta_1 = 1, \beta_2 = -3$, and $\beta_3 = 2$. The overall level is $\mu = 20$. If the model defined by equation (17) holds, the cell means μ_{ij} are specified completely as shown in Table 10.11.

For example, $\mu_{11} = 20 + 3 + 1 = 24$ and $\mu_{33} = 20 - 8 + 2 = 14$. Note that $\sum n_{i\cdot}\alpha_i = 6.3 + 9.6 + 9(-8) = 0$ and, similarly, $\sum n_{\cdot j}\beta_j = 0$. Note also that $\mu_{1\cdot} = \sum n_{1j}\mu_{1j}/\sum n_{ij} = \mu + \alpha_1 = 20 + 3 = 23$; that is, a marginal mean is just the overall mean plus the effect of the variable associated with that margin. The means are graphed in Figure 10.2. The points have been joined by dashed lines to make the pattern clear; there need not be any continuity between the levels. A similar graph could be made with the level of the second variable plotted on the abscissa and the lines indexed by the levels of the first variable.

Definition 10.5. A two-way ANOVA model satisfying equation (17) is called an *additive model.*

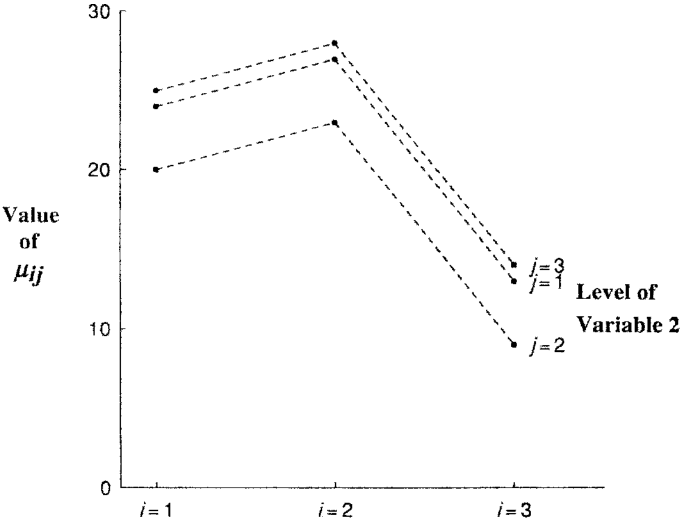

Figure 10.2 Graph of additive ANOVA model (see Example 10.4).

Some implications of this model are discussed. You will find it helpful to refer to Example 10.4 and Figure 10.2 in understanding the following:

1. The statement of equation (17) is equivalent to saying that "changing the level of variable 1 while the level of the second variable remains fixed changes the value of the mean by the same amount regardless of the (fixed) level of the second variable."
2. Statement 1 holds with variables 1 and 2 interchanged.
3. If the values of $\mu_{ij}(i = 1, \dots , I)$ are plotted for the various levels of the second variable, the curves are parallel (see Figure 10.2).
4. Statement 3 holds with the roles of variables 1 and 2 interchanged.
5. The model defined by equation (17) imposes $1 + (I - 1) + (J - 1)$ constraints on the IJ means μ_{ij}, leaving $(I - 1)(J - 1)$ degrees of freedom.

We now want to define a nonadditive model, but before doing so, we must introduce one other concept.

Definition 10.6. A two-way ANOVA has a *balanced (orthogonal) design* if for every i and j,

$$n_{ij} = \frac{n_i.n._j}{n..}$$

That is, the cell frequencies are functions of the product of the marginal totals. The reason this characteristic is needed is that only for balanced designs can the total variability be partitioned in an additive fashion. In Section 10.5 we introduce a discussion of unbalanced or nonorthogonal designs; the topic is treated in terms of multiple regression models in Chapter 11.

Definition 10.7. A *balanced two-way* ANOVA *model with interaction* (a nonadditive model) is defined by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \qquad \begin{array}{l} i = 1, \dots , I \\ j = 1, \dots , J \\ k = 1, \dots , n_{ij} \end{array} \tag{18}$$

subject to the following conditions:

1. $n_{ij} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$ for every i and j.
2. $\sum n_{i\cdot}\alpha_i = \sum n_{\cdot j}\beta_j = 0$.
3. $\sum n_{i\cdot}\gamma_{ij} = 0$ for all $j = 1, \ldots, J$, $\sum n_{\cdot j}\gamma_{ij} = 0$ for all $i = 1, \ldots, I$.
4. The ϵ_{ijk} are iid $N(0, \sigma^2)$. This assumption implies homogeneity of variances among the IJ cells.

If the γ_{ij} are zero, the model is equivalent to the one defined by equation (17), there is no interaction, and the model is additive.

As in Section 10.2, equations (4) and (5), a set of data as defined at the beginning of this section can be partitioned into parts, each of which estimates the component part of the model:

$$Y_{ijk} = \overline{Y}_{\ldots} + a_i + b_j + g_{ij} + e_{ijk} \tag{19}$$

where

$$\begin{aligned}
\overline{Y}_{\ldots} &= \text{grand mean}\\
a_i &= \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\ldots} = \text{main effect of } i\text{th level of variable 1}\\
b_j &= \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\ldots} = \text{main effect of } j\text{th level of variable 2}\\
g_{ij} &= \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\ldots} = \text{interaction of } i\text{th and } j\text{th levels of variables 1 and 2}\\
e_{ijk} &= \overline{Y}_{ijk} - \overline{Y}_{ij\cdot} = \text{residual effect (error)}
\end{aligned}$$

The quantities $\overline{Y}_{i\cdot\cdot}$ and $\overline{Y}_{\cdot j\cdot}$ are the means of the ith level of variable 1 and the jth level of variable 2. In symbols,

$$\overline{Y}_{i\cdot\cdot} = \sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} \frac{Y_{ijk}}{n_{i\cdot}} \quad \text{and} \quad \overline{Y}_{\cdot j\cdot} = \sum_{i=1}^{I}\sum_{k=1}^{n_{ij}} \frac{Y_{ijk}}{n_{\cdot j}}$$

The interaction term, g_{ij}, can be rewritten as

$$g_{ij} = (\overline{Y}_{ij\cdot} - \overline{Y}_{\ldots}) - (\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\ldots}) - (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\ldots})$$

which is the overall deviation of the mean of the ijth cell from the grand mean minus the main effects of variables 1 and 2. If the data can be fully explained by main effects, the term g_{ij} will be zero. Hence, g_{ij} measures the extent to which the data deviate from an additive model.

For a balanced design the total sum of squares, $\text{SS}_{\text{TOTAL}} = \sum\sum\sum(Y_{ijk} - \overline{Y}_{\ldots})^2$ and degrees of freedom can be partitioned additively into four parts:

$$\begin{aligned}
\text{SS}_{\text{TOTAL}} &= \text{SS}_\alpha + \text{SS}_\beta + \text{SS}_\gamma + \text{SS}_\epsilon\\
n_{\cdot\cdot} - 1 &= (I-1) + (J-1) + (I-1)(J-1) + (n_{\cdot\cdot} - IJ)
\end{aligned} \tag{20}$$

Let

$$Y_{ij\cdot} = \sum_{k=1}^{n_{ij}} Y_{ijk} = \text{total for cell } ij$$

$$Y_{i\cdot\cdot} = \sum_{j=1}^{J} Y_{ij\cdot} = \text{total for row } i$$

$$Y_{\cdot j\cdot} = \sum_{i=1}^{I} Y_{ij\cdot} = \text{total for column } j$$

Then the equations for the sums of squares together with computationally simpler formulas are

$$\begin{aligned}
\mathrm{SS}_\alpha &= \sum n_{i\cdot}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2 = \sum \frac{Y_{i\cdot\cdot}^2}{n_{i\cdot}} - \frac{Y_{\cdots}^2}{n_{\cdot\cdot}} \\
\mathrm{SS}_\beta &= \sum n_{\cdot j}(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})^2 = \sum \frac{Y_{\cdot j\cdot}^2}{n_{\cdot j}} - \frac{Y_{\cdots}^2}{n_{\cdot\cdot}} \\
\mathrm{SS}_\gamma &= \sum\sum n_{ij}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2 = \sum\sum \frac{Y_{ij\cdot}^2}{n_{ij}} - \frac{Y_{\cdots}^2}{n} - \mathrm{SS}_\alpha - \mathrm{SS}_\beta \\
\mathrm{SS}_\epsilon &= \sum\sum\sum (Y_{ijk} - \overline{Y}_{ij\cdot})^2 = \sum\sum\sum Y_{ijk}^2 - \sum\sum \frac{Y_{ij\cdot}^2}{n_{ij}}
\end{aligned} \tag{21}$$

The partition of the sum of squares, the mean squares, and the expected mean squares are given in Table 10.12.

A series of F-tests can be carried out to test the significance of the components of the model specified by equation (18). The first test carried out is usually the test for interaction: $\mathrm{MS}_\gamma/\mathrm{MS}_\epsilon$. Under the null hypothesis $H_0 : \gamma_{ij} = 0$ for all i and j, this ratio has an F-distribution with $(I-1)(J-1)$ and $n - IJ$ degrees of freedom. The null hypothesis is rejected for large values of this ratio. Interaction is indicated by nonparallelism of the treatment effects. In Figure 10.3, some possible patterns are indicated. The expected results of F-tests are given at the top of each graph. For example, pattern 1 shows NO–YES–NO, implying that the test for the main effect of variable 1 was not significant, the test for main effect of variable 2 was significant, and the test for interaction was not significant. It now becomes clear that if interaction is present, main effects are going to be difficult to interpret. For example, pattern 4 in Figure 10.3 indicates significant interaction but no significant main effects. But the significant interaction implies that at level 1 of variable 1 there is a significant difference in the main effect of variable 2. What is happening is that the effect of variable 2 is in the opposite direction at the second level of variable 1. This pattern is extreme. A more common pattern is that of pattern 6. How is this pattern to be interpreted? First, there is interaction; second, above the interaction there are significant main effects.

There are substantial practical problems associated with significant interaction patterns. For example, suppose that the two variables represent two drugs for pain relief administered simultaneously to a patient. With pattern 2, the inference would be that the two drugs together are more effective than either one acting singly. In pattern 4 (and pattern 3), the drugs are said to act *antagonistically*. In pattern 6, the drugs are said to act *synergistically*; the effect of both drugs combined is greater than the sum of each acting alone. (For some subtle problems associated with these patterns, see the discussion of transformations in Section 10.6.)

If interaction is not present, the main effects can be tested by means of the F-tests $\mathrm{MS}_\alpha/\mathrm{MS}_\epsilon$ and $\mathrm{MS}_\beta/\mathrm{MS}_\epsilon$ with $(I-1, n-IJ)$ and $(J-1, n-IJ)$ degrees of freedom, respectively. If a main effect is significant, the question arises: Which levels of the main effect differ significantly? At this point, a visual inspection of the levels may be sufficient to establish the pattern; in Chapter 12 we establish a more formal approach.

As usual, the test $\mathrm{MS}_\mu/\mathrm{MS}_\epsilon$ is of little interest, and this line is frequently omitted in an analysis of variance table.

Table 10.12 Layout for the Two-Way ANOVA[a]

Source of Variation	d.f.	SS[b]	MS	F-Ratio	d.f. of F-Ratio	E(MS)	Hypothesis Being Tested
Grand mean	1	$\text{SS}_\mu = n\overline{Y}^2_{\ldots}$	$\text{MS}_\mu = \text{SS}_\mu$	$\frac{\text{MS}_\mu}{\text{MS}_\epsilon}$	$(1, n - IJ)$	$\sigma^2 + n\mu^2$	$\mu = 0$
Row main effects	$I - 1$	$\text{SS}_\alpha = \sum n_{i\cdot}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\ldots})^2$	$\text{MS}_\alpha = \frac{\text{SS}_\alpha}{I-1}$	$\frac{\text{MS}_\alpha}{\text{MS}_\epsilon}$	$(I - 1, n - IJ)$	$\sigma^2 + \frac{\sum n_{i\cdot}\alpha_i^2}{I-1}$	$\alpha_i = 0$ for all i
Column main effects	$J - 1$	$\text{SS}_\beta = \sum n_{\cdot j}(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\ldots})^2$	$\text{MS}_\beta = \frac{\text{SS}_\beta}{J-1}$	$\frac{\text{MS}_\beta}{\text{MS}_\epsilon}$	$(J - 1, n - IJ)$	$\sigma^2 + \frac{\sum n_{\cdot j}\beta_j^2}{J-1}$	$\beta_j = 0$ for all j
Row × column interaction	$(I - 1)(J - 1)$	$\text{SS}_\gamma = \sum n_{ij}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot}.\overline{Y}_{\ldots})^2$	$\text{MS}_\gamma = \frac{\text{SS}_\gamma}{(I-1)(J-1)}$	$\frac{\text{MS}_\gamma}{\text{MS}_\epsilon}$	$((I - 1)(J - 1), n - IJ)$	$\sigma^2 + \frac{\sum n_{ij}\gamma_{ij}^2}{(I-1)(J-1)}$	$\gamma_{ij} = 0$ for all i and j, or $\mu_{ij} = u_i + v_j$
Residual	$n - IJ$	$\text{SS}_\epsilon = \sum(Y_{ijk} - \overline{Y}_{ij\cdot})^2$	$\text{MS}_\epsilon = \frac{\text{SS}_\epsilon}{n - IJ}$	—	—	σ^2	
Total	n	$\sum Y_{ijk}^2$	—				

[a] Model: $Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$ [where $\epsilon_{ijk} \sim$ iid $N(0, \sigma^2)$].
Data : $Y_{ijk} = \overline{Y}_{\ldots} + (\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\ldots}) + (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\ldots}) + (\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\ldots}) + (Y_{ijk} - \overline{Y}_{ij\cdot})$.
Equivalent model : $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$, where the Y_{ijk}'s are independent.

[b] Summation is over all subscripts displayed.

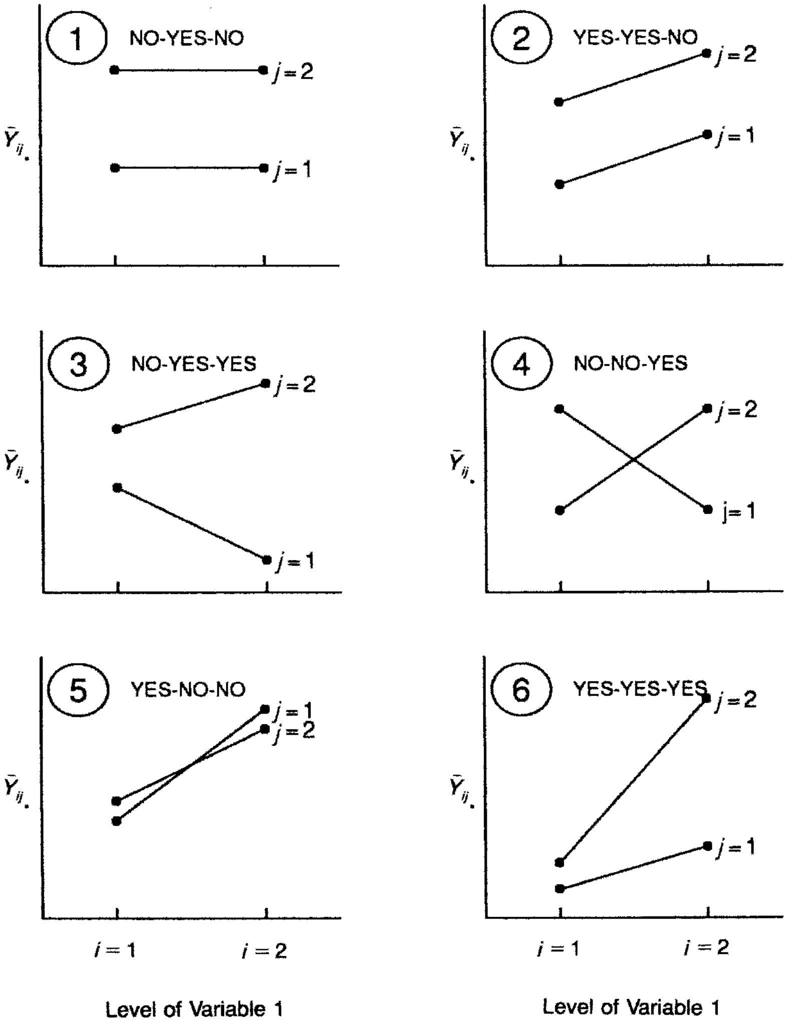

Figure 10.3 Some possible patterns for observed cell means in two-way ANOVA with two levels for each variable. Results of F-tests for main effects variable 1, variable 2, and interaction are indicated by YES or NO. See the text for a discussion.

Example 10.5. Nitrogen dioxide (NO_2) is an automobile emission pollutant, but less is known about its effects than those of other pollutants, such as particulate matter. Several animal models have been studied to gain an understanding of the effects of NO_2. Sherwin and Layfield [1976] studied protein leakage in the lungs of mice exposed to 0.5 part per million (ppm) NO_2 for 10, 12, and 14 days. Half of a total group of 44 animals was exposed to the NO_2; the other half served as controls. Control and experimental animals were matched on the basis of weight, but this aspect will be ignored in the analysis since the matching did not appear to influence the results. Thirty-eight animals were available for analysis; the raw data and some basic statistics are listed in Table 10.13.

The response is the percent of serum fluorescence. High serum fluorescence values indicate a greater protein leakage and some kind of insult to the lung tissue. The authors carried out t-tests and state that with regard to serum fluorescence, “no significant differences” were found.

Table 10.13 Serum Fluorescence Readings of Mice Exposed to Nitrogen Dioxide (NO_2) for 10, 12, and 14 Days Compared with Control Animals

	Serum Fluorescence		
	10 Days ($j = 1$)	12 Days ($j = 2$)	14 Days ($j = 3$)
Control ($i = 1$)	143	179	76
	169	160	40
	95	87	119
	111	115	72
	132	171	163
	150	146	78
	141	—	—
Exposed ($i = 2$)	152	141	119
	83	132	104
	91	201	125
	86	242	147
	150	209	200
	108	114	178
	75	—	—

n_{ij}

	j		
i	1	2	3
1	7	6	6
2	7	6	6

$Y_{ij\cdot}$

	j		
i	1	2	3
1	941	858	548
2	745	1039	873

$\overline{Y}_{ij\cdot}$

	j		
i	1	2	3
1	134.4	143.0	91.3
2	106.4	173.2	145.5

s_{ij}

	j		
i	1	2	3
1	24.7	35.5	43.2
2	32.1	51.0	37.1

The standard deviations are very similar, suggesting that the homogeneity of variance assumption is probably valid. It is a good idea again to graph the results to get some "feel" for the data, and this is done in Figure 10.4. We can see from this figure that there are no outlying observations that would invalidate the normality assumption of the two-way ANOVA model.

To obtain the entries for the two-way ANOVA table, we basically need six quantities:

$$n,\ Y_{\cdots},\quad \sum Y_{ijk}^2,\quad \sum \frac{Y_{i\cdot\cdot}^2}{n_{i\cdot}},\quad \sum \frac{Y_{\cdot j\cdot}^2}{n_{\cdot j}},\quad \sum \frac{Y_{ij\cdot}^2}{n_{ij}}$$

With these quantities, and using equations (20) and (21), the entire table can be computed. The values are as follows:

$$n = 38,\qquad Y_{\cdots} = 5004,\qquad \sum Y_{ijk}^2 = 730{,}828$$

$$\sum \frac{Y_{i\cdot\cdot}^2}{n_{i\cdot}} = 661{,}476.74,\qquad \sum \frac{Y_{\cdot j\cdot}^2}{n_{\cdot j}} = 671{,}196.74,\qquad \sum \frac{Y_{ij\cdot}^2}{n_{ij}} = 685{,}472.90$$

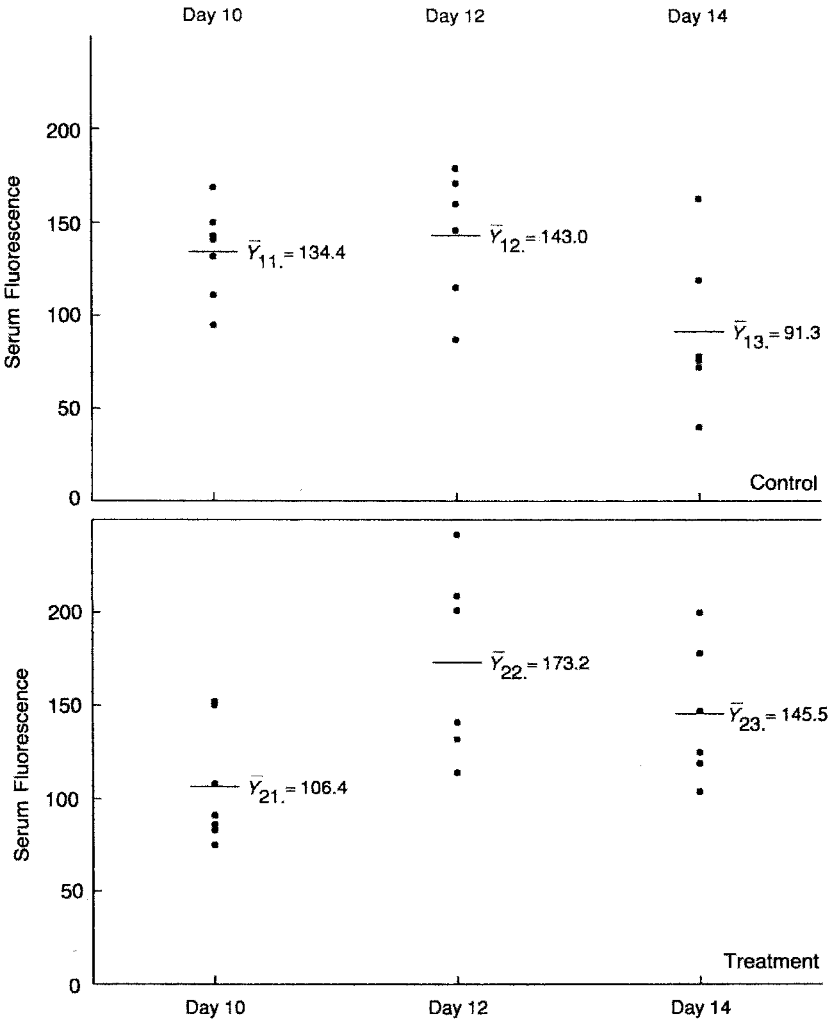

Figure 10.4 Serum fluorescence of mice exposed to nitrogen dioxide. (Data from Sherwin and Layfield [1976]; see Example 10.5.)

Sums of squares can now be calculated:

$$\text{SS}_\alpha = \text{SS}_{\text{TREATMENT}} = 661{,}476.74 - \frac{5004^2}{38} = 2528.95$$

$$\text{SS}_\beta = \text{SS}_{\text{DAYS}} = 671196.74 - \frac{5004^2}{38} = 12{,}248.95$$

$$\text{SS}_\gamma = \text{SS}_{\text{TREATMENT}\times\text{DAYS}} = 685{,}472.90 - \frac{5004^2}{38} - 2528.95 - 12{,}248.95 = 11{,}747.21$$

$$\text{SS}_\epsilon = \text{SS}_{\text{RESIDUAL}} = 730{,}828 - 685{,}472.90 = 45{,}355.10$$

(It can be shown that $\text{SS}_\epsilon = \sum(n_{ij} - 1)s_{ij}^2$. You can verify this for these data.) The ANOVA table is presented in Table 10.14.

Table 10.14 ANOVA **of Serum Fluorescence Levels of Mice Exposed to Nitrogen Dioxide (NO_2)**

Source of Variation	d.f.	SS	MS	F-Ratio	p-Value
Treatment	1	2,528.95	2528.95	1.78	> 0.10
Days	2	12,248.95	6124.48	4.32	< 0.05
Treatment × days	2	11,747.21	5873.60	4.14	< 0.05
Residual	32	45,355.10	1417.35	—	—
Total	37	71,880.21	—	—	—

Source: Data from Sherwin and Layfield [1976].

The MS for interaction is significant at the 0.05 level ($F_{2,32} = 4.14$, $p < 0.05$). How is this to be interpreted? The means $\overline{Y}_{ij\cdot}$ are graphed in Figure 10.5. There clearly is nonparallelism, and the model is not an additive one. But more should be said in order to interpret the results, particularly regarding the role of the control animals. Clearly, control animals were used to provide a measurement of background variation. The differences in mean fluorescence levels among the control animals indicate that the baseline response level changed from day 10 to day 14. If we consider the response of the animals exposed to nitrogen dioxide standardized by the control level, a different picture emerges. In Figure 10.5, the differences in means between exposed and unexposed animals is plotted as a dashed line with scale on the right-hand side of the graph. This line indicates that there is an increasing effect of exposure with time. The interpretation of the significant interaction effect then is, possibly, that exposure did induce increased protein leakage, with greater leakage attributable to longer exposure. This contradicts the authors' analysis of the data using t-tests. If the matching by weight was retained, it would

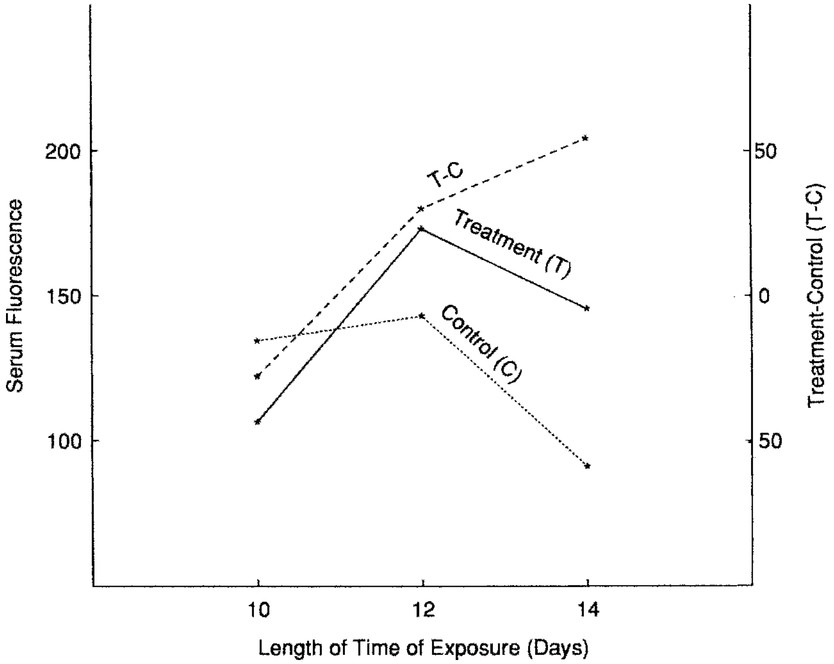

Figure 10.5 Mean serum fluorescence level of mice exposed to nitrogen dioxide, treatment vs. control. The difference (treatment – control) is given by the dashed line. (Data from Sherwin and Layfield [1976]; see Example 10.5.)

have been possible to consider the differences between exposed and control animals and carry out a one-way ANOVA on the differences. See Problem 10.5.

Two-Way ANOVA *from Means and Standard Deviations*

As in the one-way ANOVA, a two-way ANOVA can be reconstructed from means and standard deviations. Let $\overline{Y}_{ij\cdot}$ be the mean, s_{ij} the standard deviation, and n_{ij} the sample size associated with cell $ij (i = 1, \ldots , I, j = 1, \ldots , J)$, assuming a balanced design. Then

$$Y_{\cdots} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\overline{Y}_{ij\cdot}, \qquad Y_{i\cdot\cdot} = \sum_{j=1}^{J} n_{ij}\overline{Y}_{ij\cdot}, \qquad Y_{\cdot j\cdot} = \sum_{i=1}^{I} n_{ij}\overline{Y}_{ij\cdot}$$

Using equation (21), SS_α and SS_β can now be calculated. The term $\sum Y_{ij\cdot}^2/n_{ij}$ in SS_γ is equivalent to

$$\sum \frac{Y_{ij\cdot}^2}{n_{ij}} = \sum n_{ij}\overline{Y}_{ij\cdot}^2$$

Finally, SS_ϵ can be calculated from

$$\text{SS}_\epsilon = \sum (n_{ij} - 1)s_{ij}^2 \qquad (22)$$

Problems 10.22 and 10.23 deal with data presented in terms of means and standard deviations. There will be some round-off error in the two-way analysis constructed in this way, but it will not affect the conclusion.

It is easy to write a computer subroutine that produces such a table upon input of means, standard deviations, and sample sizes.

10.3.2 Randomized Block Design

In Chapter 2 we discussed the statistical concept of blocking. A block consists of a subset of homogeneous experimental units. The background variability among blocks is usually much greater than within blocks, and the experimental strategy is to assign all treatments randomly to the units of a block. A simple example of blocking is illustrated by the paired t-test. Suppose that two antiepileptic agents are to be compared. One possible (valid) design is to assign randomly half of a group of patients to one agent and half to the other. By this randomization procedure, the variability among patients is "turned" into error. Appropriate analyses are the two-sample t-test, the one-way analysis of variance, or a two-sample nonparametric test. However, if possible, a better design would be to test both drugs on the same patient; this would eliminate patient-to-patient variability, and comparisons are made within patients. The patients in this case act as *blocks*. A paired t-test or analogous nonparametric test is now appropriate. For this design to work, we would want to assign the drugs randomly within a patient. This would eliminate a possible additive sequence effect; hence, the term *randomized block design*. In addition, we would want to have a reasonably large time interval between drugs to eliminate possible carryover effects; that is, we cannot permit a treatment × period interaction. Other examples of naturally occurring blocks are animal litters, families, and classrooms. Constructed blocks could be made up of sets of subjects matched on age, race, and gender.

Blocking is done for two purposes:

1. To obtain smaller residual variability
2. To examine treatments under a wide range of conditions

A basic design principle is to partition a population of study units in such a way that background variability between blocks is maximized, and consequently, background variability within blocks is minimized.

Definition 10.8. In a *randomized block design*, each treatment is given once and only once in each block. Within a block, the treatments are assigned randomly to the experimental units.

Note that a randomized block design, by definition, is a balanced design: This is somewhat restrictive. For example, in animal experiments it would require litters to be of the same size.

The statistical model associated with the randomized block design is

$$Y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}, \qquad i = 1, \dots, I, \quad j = 1, \dots, J \tag{23}$$

and (**1**) $\sum \beta_i = \sum \tau_j = 0$ and (**2**) ϵ are iid $N(0, \sigma^2)$. In this model, β_i is the effect of block i and τ_j the effect of treatment j. In this model, as indicated, we assume no interaction between blocks and treatments (i.e., if there is a difference between treatments, the magnitude of this effect does not vary from block to block except for random variation). In Section 10.6 we discuss a partial check on the validity of the assumption of no interaction.

The analysis of variance table for this design is a simplified version of Table 10.12: The number of observations is the same in each block and for each treatment. In addition, there is no SS for interaction; another way of looking at this is that the SS for interaction is the error SS. The calculations are laid out in Table 10.15.

Tests of significance proceed in the usual way. The expected mean squares can be derived from Table 10.12, making use of the simpler design.

The computations for the randomized block design are very simple. You can verify that

$$\mathrm{SS}_\mu = \frac{Y_{..}^2}{n}, \qquad \mathrm{SS}_\beta = \frac{\sum Y_{i.}^2}{J} - \frac{Y_{..}^2}{n}, \qquad \mathrm{SS}_\tau = \frac{\sum Y_{.j}^2}{I} - \frac{Y_{..}^2}{n} \tag{24}$$

$$\mathrm{SS}_\epsilon = \sum Y_{ij}^2 - \frac{Y_{..}^2}{n} - \mathrm{SS}_\beta - \mathrm{SS}_\tau$$

Example 10.6. The pancreas, a large gland, secretes digestive enzymes into the intestine. Lack of this fluid results in bowel absorption problems (steatorrhea); this can be diagnosed by excess fat in feces. Commercial pancreatic enzyme supplements are available in three forms: capsule, tablets, and enteric-coated tablets. The enteric-coated tablets have a protective shell to prevent gastrointestinal reaction. Graham [1977] investigated the effectiveness of these three formulations in six patients with steatorrhea; the three randomly assigned treatments were preceded by a control period. For purposes of this example, we will consider the control period as a treatment, even though it was not randomized. The data are displayed in Table 10.16.

To use equation 4, we will need the quantities

$$Y_{..} = 618.6, \qquad \frac{\sum Y_{i.}^2}{4} = 21{,}532.80, \qquad \frac{\sum Y_{.j}^2}{6} = 17{,}953.02, \qquad \sum Y_{ij}^2 = 25{,}146.8$$

The analysis of variance table, omitting SS_μ, is displayed in Table 10.17.

The treatment effects are highly significant. A visual inspection of Table 10.16 suggests that capsules and tablets are the most effective, enteric-coated tablets less effective. The author points out that the "normal" amount of fecal fat is less than 6 g per day, suggesting that, at best, the treatments are palliative. The F-test for patients is also highly significant, indicating that the levels among patients varied considerably: Patient 4 had the lowest average level at 6.1 g in 24 hours; patient 5 had the highest level, with 47.1 g in 24 hours.

Table 10.15 Layout for the Randomized Block Design[a]

Source of Variation	d.f.	SS[b]	MS	F-Ratio	d.f. of F-Ratio	E(MS)	Hypothesis Being Tested
Grand mean	1	$\text{SS}_\mu = n\overline{Y}^2_{..}$	$\text{MS}_\mu = \text{SS}_\mu$	$\frac{\text{MS}_\mu}{\text{MS}_\epsilon}$	$(1, (I-1)(J-1))$	$\sigma^2 + ij\mu^2$	$\mu = 0$
Blocks	$I - 1$	$\text{SS}_\beta = J\sum(\overline{Y}_{i\cdot} - \overline{Y}_{..})^2$	$\text{MS}_\beta = \frac{\text{SS}_\beta}{I-1}$	$\frac{\text{MS}_\beta}{\text{MS}_\epsilon}$	$(I-1, (I-1)(J-1))$	$\sigma^2 + \frac{J\sum\beta_i^2}{I-1}$	$\beta_i = 0$ for all i
Treatments	$J - 1$	$\text{SS}_\tau = I\sum(\overline{Y}_{\cdot j} - \overline{Y}_{..})^2$	$\text{MS}_\tau = \frac{\text{SS}_\tau}{J-1}$	$\frac{\text{MS}_\tau}{\text{MS}_\epsilon}$	$(J-1, (I-1)(J-1))$	$\sigma^2 + \frac{I\sum\tau_j^2}{J-1}$	$\tau_j = 0$ for all j
Residual	$(I-1)(J-1)$	$\text{SS}_\epsilon = \sum(Y_{ij} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y}_{..})^2$	$\text{MS}_\epsilon = \frac{\text{SS}_\epsilon}{(I-1)(J-1)}$	—	—	σ^2	
Total	IJ	$\sum Y_{ij}^2$	—				

[a] Model: $Y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$ [where $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$].
Data : $Y_{ij} = \overline{Y}_{..} + (\overline{Y}_{i\cdot} - \overline{Y}_{..}) + (\overline{Y}_{\cdot j} - \overline{Y}_{..}) + (\overline{Y}_{ij} - Y_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y}_{..})$.
Equivalent model : $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$, where the Y_{ij}'s are independent.

[b] Summation is over all displayed subscripts.

Table 10.16 Effectiveness of Pancreatic Supplements on Fat Absorption in Patients with Steatorrhea (Grams/Day)

Case	None (Control)	Tablet	Capsule	Enteric-Coated Tablet	$Y_{i\cdot}$	$\overline{Y}_{i\cdot}$
1	44.5	7.3	3.4	12.4	67.6	16.9
2	33.0	21.0	23.1	25.4	102.5	25.6
3	19.1	5.0	11.8	22.0	57.9	14.5
4	9.4	4.6	4.6	5.8	24.4	6.1
5	71.3	23.3	25.6	68.2	188.4	47.1
6	51.2	38.0	36.0	52.6	177.8	44.4
$Y_{\cdot j}$	228.5	99.2	104.5	186.4	618.6	—
$\overline{Y}_{\cdot j}$	38.1	16.5	17.4	31.1	$\overline{Y}_{\cdot\cdot} = 25.8$	

Source: Data from Graham [1977].

Table 10.17 Randomized Block Analysis of Fecal Fat Excretion of Patients with Steatorrhea

Source of Variation	d.f.	SS	MS	F-Ratio	p-Value
Patients (blocks)	5	5588.38	1117.68	10.44	<0.001
Treatments	3	2008.60	669.53	6.26	<0.01
Residual	15	1605.40	107.03	—	—
Total	23	9202.38	—	—	—

Source: Data from Graham [1977].

10.3.3 Analyses of Randomized Block Designs Using Ranks

A nonparametric analysis of randomized block data using only the ranks was developed by Friedman [1937]. The model is that of equation (23), but the ϵ_{ij} are no longer required to be normally distributed. We assume that there are no ties in the data; for a small number of ties the average ranks may be used. The idea of the test is simple: If there are no treatment effects ($\tau_j = 0$ for all j), the ranks of the observations within a block are randomly distributed. For block i, let

$$R_{ij} = \text{rank of } Y_{ij} \text{ among } Y_{i1}, Y_{i2}, \dots, Y_{iJ}$$

The Friedman statistic for testing the null hypothesis $H_0 : \tau_j = 0$ (where $j = 1, \dots, J$) is

$$T_{\text{FR}} = 12I \sum_{j=1}^{J} \frac{(\overline{R}_{\cdot j} - \overline{R}_{\cdot\cdot})^2}{J(J+1)} \tag{25}$$

Computationally, the following formula is easier:

$$T_{\text{FR}} = \frac{12}{IJ(J+1)} \sum_{j=1}^{J} R_{\cdot j}^2 - 3(I)(J+1) \tag{26}$$

The null hypothesis is rejected for large values of T_{FR}. For small randomized block designs, the critical values of T_{FR} are tabulated; see, for example, Table 39 in Odeh et al. [1977], which goes up to $I = J = 6$. As the number of blocks becomes very large, the distribution of T_{FR}

approaches that of a χ^2-distribution with $(J - 1)$ degrees of freedom. See also Notes 10.1 and 10.2.

Example 10.6. (*continued*) Replacing the observations for each *individual* by their ranks produces Table 10.18. For individual 4, the two tied observations are replaced by the average of the two ranks. [As a check, the total $R..$ of ranks must be $R.. = IJ(J + 1)/2$. (Why?) For this example $I = 6$, $J = 4$, $IJ(J+1)/2 = (6 \cdot 4 \cdot 5)/2 = 60$, and $R.. = 22 + 8.5 + 9.5 + 20 = 60$.] The Friedman statistic, using equation (26), has the value

$$\begin{aligned} T_{\mathrm{FR}} &= \frac{12}{6 \times 4 \times 5}(22^2 + 8.5^2 + 9.5^2 + 20^2) - (3 \times 6 \times 5) \\ &= 104.65 - 90 = 14.65 \end{aligned}$$

This quantity is compared to a χ^2 distribution with 3 d.f. $(14.65/3 = 4.88)$; the p-value is $p = 0.0021$. From exact tables such as Odeh et al. [1977], the exact p-value is $p < 0.001$. The conclusion is the same as that of the analysis of variance in Section 10.3.2. Note also that the ranking of treatments in terms of the total ranks is the same as in Table 10.11. For an alternative rank analysis of these data, see Problem 10.20.

10.3.4 Types of ANOVA Models

In Section 10.2.2, two examples were mentioned of one-way analyses of variance. The first dealt with the age at which children begin to walk as a function of various training procedures; the second example dealt with patient hospitalization costs, based on an examination of some hospitals (treatments) selected randomly from all the hospitals in a large metropolitan area (from each hospital selected, a specified number of patient records are selected for cost analysis). The experimental design associated with the first example differs from the second: In a repetition of the first study, the same set of treatments could be used; in the second study, a new set of hospitals could presumably be selected; that is, the "treatment levels" are randomly selected from a larger set of treatment levels.

Definition 10.9. If the levels of a classification variable in an ANOVA situation are selected at random from a population, the variable is said to be a *random factor* or *random effect*. Factors with the levels fixed by those conducting the study or which are fixed classifications (e.g., gender) are called *fixed factors* or *fixed effects*.

Table 10.18 Rank Values for Supplement Use

	Treatment			
Case	Control	Tablet	Capsule	Enteric-Coated Tablet
1	4	2	1	3
2	4	1	2	3
3	3	1	2	4
4	4	1.5	1.5	3
5	4	1	2	3
6	3	2	1	4
$R._j$	22	8.5	9.5	20

Definition 10.10. ANOVA situations with all classification variables fixed are called *fixed effects models* (model I). If all the classification variables are random effects, the design is a *random effects model* (model II). If both random and fixed effects are present, the design is a *mixed effects model.*

Historically, no distinction was made between model I and II designs, in part due to identical analyses in simple situations and similar analyses in more complicated situations. Eisenhart [1947] was the first to describe systematically the differences between the two models. Some other examples of random effects models are:

1. A manufacturer of spectrophotometers randomly selects five instruments from its production line and obtains a series of replicated readings on each machine.
2. To estimate the maximal exercise performance in a healthy adult population, 20 subjects are selected randomly and 10 independent estimates of maximal exercise performance for each person are obtained.
3. To determine knowledge about the effect of drugs among sixth graders, a researcher randomly selects five sixth-grade classes from among the 100 sixth-grade classes in a large school district. Each child selected fills out a questionnaire.

How can we determine whether a design is model I or model II? The basic criterion deals with the population to which inferences are to be made. Another way of looking at this is to consider the number of times randomness is introduced (ideally). In Example 10.2 there are two sources of randomness: subjects and observations within subjects. If more than one "layer of randomness" has to be passed through in order to reach the population of interest, we have a random effects model.

An example of a mixed model is example 2 above with a further partitioning of subjects into male and female. The factor, gender, is fixed.

Sometimes a set of data can be modeled by either a fixed or random effects model. Consider example 1 again. Suppose that a cancer research center has bought the five instruments and is now running standardization experiments. For the purpose of the research center, the effects of machines are fixed effects.

To distinguish a random effects model from a fixed effects model, the components of the model are written as random variables. The two-way random effects ANOVA model with interaction is written as

$$Y_{ijk} = \mu + A_i + B_j + G_{ij} + e_{ijk}, \qquad i = 1, \dots, I, \quad j = 1, \dots, J, \quad k = 1, \dots, n_{ij} \quad (27)$$

The assumptions are:

1. e_{ijk} are iid $N(0, \sigma^2)$, as before.
2. A_i are iid $N(0, \sigma_\alpha^2)$.
3. B_j are iid $N(0, \sigma_\beta^2)$.
4. G_{ij} are iid $N(0, \sigma_\gamma^2)$.

The total variance can now be partitioned into several components (hence another term for these models: *components of variance models*). Assume that the experiment is balanced with $n_{ij} = m$ for all i and j. The difference between the fixed effect and random effect model is in the expected mean squares. Table 10.19 compares the EMS for both models, taking the EMS for the fixed effect model from Table 10.12.

The test for interaction is the same in both models. However, if interaction is present, to be valid the test for main effects in the random effects model must use MS_γ in the denominator rather than MS_ϵ.

Table 10.19 Comparison of Expected Mean Squares in the Two-Way ANOVA, Fixed Effect vs. Random Effect Models[a]

		EMS	
Source of Variation	d.f.	Fixed Effect	Random Effect
Row main effects	$I-1$	$\sigma^2 + \dfrac{Jm\sum\alpha_i^2}{I-1}$	$\sigma^2 + m\sigma_\gamma^2 + mJ\sigma_\alpha^2$
Column main effects	$J-1$	$\sigma^2 + \dfrac{Im\sum\beta_j^2}{J-1}$	$\sigma + m\sigma_\gamma^2 + mI\sigma_\beta^2$
Row × column interaction	$(I-1)(J-1)$	$\sigma^2 + \dfrac{IJm\sum\gamma_{ij}^2}{(I-1)(J-1)}$	$\sigma^2 + m\sigma_\gamma^2$
Residual	$n.. - IJ$	σ^2	σ^2

[a] There are ***m*** observations in each cell.

The null hypothesis

$$H_0 : \gamma_{ij} = 0 \qquad \text{all } i \text{ and } j$$

in the fixed effect model has as its counterpart,

$$H_0 : \sigma_\gamma^2 = 0$$

in the random effect model. In both cases the test is carried out using the ratio $\text{MS}_\gamma/\text{MS}_\epsilon$ with $(I-1)(J-1)$ and $n - IJ$ degrees of freedom. If interaction is not present, the tests for main effects are the same in both models. However, if H_0 is not rejected, the tests for main effects are different in the two models. In the random effects model the expected mean square for main effects now contains a term involving σ_γ^2. Hence the appropriate F-test involves MS_γ in the denominator rather than MS_ϵ; the degrees of freedom are changed accordingly.

Several comments can be made:

1. Frequently, the degrees of freedom associated with MS_γ are fewer than those of MS_ϵ, so that there is a loss of precision if MS_γ has to be used to test main effects.

2. From a design point of view, if m, I, and J can be chosen, it may pay to choose m small and I, J relatively large if a random effects model is appropriate. A minimum of two replicates per treatment combination is needed to obtain an estimate of σ^2. If possible, the rest of the observations should be allocated to the levels of the variables. This may not always be possible, due to costs or other considerations. If the total cost of the experiment is fixed, an algorithm can be developed for choosing the values of m, I, and J.

3. The difference between the fixed and random effects models for the two-way ANOVA designs is not as crucial as it seems. We have indicated caution in proceeding to the tests of main effects if interaction is present in the fixed model (see Figure 10.3 and associated discussion). In the random effects model, the same precaution holds. It is perhaps too strong to say that main effects should not be tested when interaction is present, but you should certainly be able to explain what information you hope to obtain from such tests after a full interpretation of the (significant) interaction.

4. Expected mean squares for an unbalanced random effects model are not derivable or are very complicated. A more useful approach is that of multiple regression, discussed in Chapter 11. See also Section 10.5.

5. For the randomized block design the MS_ϵ can be considered the mean square for interaction. Hence, in this case the F-tests are appropriate for both models. (Does this contradict the statement made in comment 3?) Note also that there is little interest in the test of block effects, except as a verification that the blocking was effective.

Good discussions about inference in the case of random effects models can be found in Snedecor and Cochran [1988] and Winer [1991].

10.4 REPEATED MEASURES DESIGNS AND OTHER DESIGNS

10.4.1 Repeated Measures Designs

Consider a situation in which blood pressures of two populations are to be compared. One person is selected at random from each population. The blood pressure of each of the two subjects is measured 100 times. How would you react to data analysis that used the two-sample t-test with two samples of size 100 and showed that the blood pressures differed in the two populations? The idea is ridiculous, but in one form or another appears frequently in the research literature. Where does the fallacy lie? There are two sources of variability: within individuals and among individuals. The variability within individuals is assumed incorrectly to represent the variability among individuals. Another way of saying this is that the 100 readings are not independent samples from the population of interest. They are repeated measurements on the same experimental unit. The repeated measures may be useful in this context in pinning down more accurately the blood pressure of the two people, but they do not make up for the small sample size. Another feature we want to consider is that the sequence of observations within the person cannot be randomized, for example, a sequence of measurements of growth. Thus, typically, we do not have a randomized block design.

Definition 10.11. In a *repeated measures design*, multiple (two or more) measurements are made sequentially on the same observational unit.

A repeated measures design usually is an example of a mixed model with the observational unit a random effect (e.g., persons or animals, and the treatments on the observational units fixed effects). Frequently, data from repeated measure designs are somewhat unbalanced and this makes the analysis more difficult. One approach is to summarize the repeated measures in some meaningful way by single measures and then analyze the single measures in the usual way. This is the way many computer programs analyze such data. We motivate this approach by an example. See Chapter 18 for further discussion.

Example 10.7. Hillel and Patten [1990] were interested in the effect of accessory nerve injury as result of neck surgery in cancer. The surgery frequently decreases the strength of the arm on the affected side. To assess the potential recovery, the unaffected arm was to be used as a control. But there is a question of the comparability of arms due to dominance, age, gender, and other factors. To assess this effect, 33 normal volunteers were examined by several measurements. The one discussed here is that of torque, or the ability to abduct (move or pull) the shoulder using a standard machine built for that purpose. The subjects were tested under three consecutive conditions (in order of increasing strenuousness): 90°, 60°, and 30° per second. The data presented in Table 10.20 are the best of three trials under each condition. For completeness, the age and height of each of the subjects are also presented. The researchers wanted answers to at least five questions, all dealing with differences between dominant and nondominant sides:

1. Is there a difference between the dominant and nondominant arms?
2. Does the difference vary between men and women?

Table 10.20 Peak Torque for 33 Subjects by Gender, Dominant Arm, and Age Group under Three Conditions

					90°		60°		30°	
Subject		Age	Height (in.)	Weight (lb)	DM[a]	ND[a]	DM	ND	DM	ND
Female	1	20	64	107	17	13	20	17	23	22
	2	23	68	140	25	25	28	29	31	31
	3	23	67	135	27	28	30	31	32	33
	4	23	67	155	23	28	27	29	27	32
	5	25	65	115	15	11	15	13	17	17
	6	26	68	147	27	17	25	21	32	27
	7	31	62	147	25	17	25	21	29	24
	8	31	66	137	19	15	17	17	21	19
	9	33	66	160	28	26	31	27	31	31
	10	36	66	118	23	23	26	27	27	25
	11	56	67	210	23	31	37	44	49	53
	12	59	67	130	15	17	17	19	20	20
	13	60	63	132	17	15	19	21	24	28
	14	60	64	180	15	15	17	19	19	21
	15	67	62	135	13	5	15	8	15	14
	16	73	62	124	11	9	13	13	19	17
Male	1	26	69	140	43	43	44	43	49	41
	2	28	71	175	45	43	48	45	53	52
	3	28	70	125	25	29	29	37	39	41
	4	28	70	175	39	41	49	47	55	44
	5	29	72	150	38	33	40	33	44	37
	6	30	68	145	53	41	51	40	59	44
	7	31	74	240	60	49	71	54	68	53
	8	32	67	168	32	31	37	31	39	30
	9	40	69	174	47	37	43	47	49	53
	10	41	72	190	33	25	29	25	39	27
	11	41	68	184	39	24	43	25	39	33
	12	56	70	200	21	11	23	12	33	24
	13	58	72	168	41	35	45	37	49	39
	14	59	73	170	31	32	31	31	35	38
	15	60	73	225	39	41	47	45	55	49
	16	68	67	140	31	23	33	27	37	33
	17	72	69	125	13	17	17	19	17	25

Source: Data from Hillel and Patten [1990].

[a] DM, dominant arm; ND, nondominant arm.

3. Does the difference depend on age, height, or weight?
4. Does the difference depend on treatment condition?
5. Is there interaction between any of the factors or variables mentioned in questions 1 to 4?

For purposes of this example, we only address questions 1, 2, 4, and 5, leaving question 3 for the discussion of analysis of covariance in Chapter 11.

The second to fourth columns in Table 10.21 contain the differences between the dominant and nondominant arms; the fifth to seventh columns are reexpressions of the three differences as follows. Let d90, d60, and d30 be the differences between the dominant and nondominant

Table 10.21 Differences in Torque under Three Conditions and Associated Orthogonal Contrasts[a]

		DM–ND			Orthogonal Contrasts		
		90°	60°	30°	Constant	Linear	Quadratic
Female	1	4	3	1	4.6	2.1	−0.4
	2	0	−1	0	−0.6	0.0	0.8
	3	−1	−1	−1	−1.7	0.0	0.0
	4	−5	−2	−5	−6.9	0.0	−2.4
	5	4	2	0	3.5	2.8	0.0
	6	10	4	5	11.0	3.5	2.9
	7	8	4	5	9.8	2.1	2.0
	8	4	0	2	3.5	1.4	2.4
	9	2	4	0	3.5	1.4	−2.4
	10	0	−1	2	0.6	−1.4	1.6
	11	−8	−7	−4	−11.0	−2.8	0.8
	12	−2	−2	0	−2.3	−1.4	0.8
	13	2	−2	−4	−2.3	4.2	0.8
	14	0	−2	−2	−2.3	1.4	0.8
	15	8	7	1	9.2	4.9	−2.0
	16	2	0	2	2.3	0.0	1.6
Male	1	0	1	8	5.2	−5.7	2.4
	2	2	3	1	3.5	0.7	−1.2
	3	−4	−8	−2	−8.1	−1.4	4.1
	4	−2	2	11	6.4	−9.2	2.0
	5	5	7	7	11.0	−1.4	−0.8
	6	12	11	15	21.9	−2.1	2.0
	7	11	17	15	24.8	−2.8	−3.3
	8	1	6	9	9.2	−5.7	−0.8
	9	10	−4	−4	1.2	9.9	5.7
	10	8	4	12	13.9	−2.8	4.9
	11	15	18	6	22.5	6.4	−6.1
	12	10	11	9	17.3	0.7	−1.2
	13	6	8	10	13.9	−2.8	0.0
	14	−1	0	−3	−2.3	1.4	−1.6
	15	−2	2	6	3.5	−5.7	0.0
	16	8	6	4	10.4	2.8	0.0
	17	−4	−2	−8	−8.1	2.8	−3.3

Source: Data from Hillel and Patten [1990].

[a] See Table 10.20 for notation.

arms under each of the three conditions. Then we define

$$\text{constant} = \frac{\text{d90} + \text{d60} + \text{d30}}{\sqrt{3}}$$

$$\text{linear} = \frac{\text{d90} - \text{d30}}{\sqrt{2}}$$

$$\text{quadratic} = \frac{\text{d90} - 2 \cdot \text{d60} + \text{d30}}{\sqrt{6}}$$

For example, for the first female subject, rounding off to one decimal place yields

$$\frac{4 + 3 + 1}{\sqrt{3}} = 4.6$$

$$\frac{4-1}{\sqrt{2}} = 9.9$$

$$\frac{4-2\times(3)+1}{\sqrt{6}} = -0.4$$

The first component clearly represents an average difference of dominance over the three conditions. The divisor is chosen to make the variance of this term equal to the variance of a single difference. The second term represents a slope within an individual. If the three conditions were considered as values of a predictor variable with values −1 (for 30°), 0 (for 60°), and 1 (for 90°), the slope would be expressed as in the second, or linear, term. The linear term assesses a possible trend in the differences over the three conditions within an individual. The last term, the quadratic term, fits a quadratic curve through the data assessing possible curvature or nonlinearity within an individual. This partitioning of the observations within an individual has the property that sums of squares are maintained. For example, for the first female subject,

$$4^2 + 3^2 + 1^2 = 26 = (4.6)^2 + (2.1)^2 + (-0.4)^2$$

except for rounding. (If you were to calculate these terms to more decimal places, you would find that the right side is identical to the left side.) In words, the variability in response within an individual has been partitioned into a constant component, a linear component, and a quadratic component. The questions posed can now be answered unambiguously since the three components have been constructed to be *orthogonal*, or uncorrelated. An analysis of variance is carried out on the three terms; unlike the usual analysis of variance, a term for the mean is included; results are summarized in Table 10.22. We start by discussing the analysis of the quadratic component. The analysis indicates that there are no significant differences between males and females in terms of the quadratic or nonlinear component. Nor is there an overall effect. Next, conclusions are similar for the linear effect. We conclude that there is no linear trend for abductions at $90°, 60°$, and 30°. This leaves the constant term, which indicates (1)

Table 10.22 ANOVA and Means of the Data in Table 10.21

Source of Variation		d.f.	SS	MS	F-Ratio
		Analysis of Variance			
Constant	Mean	1	900.7	900.7	13.3
	Gender	1	438.5	438.5	6.48
	Error 1	31	2099.2	67.72	
Linear	Mean	1	0.33	0.33	0.02
	Gender	1	33.43	33.43	2.43
	Error 2	31	426.0	13.74	
Quadratic	Mean	1	3.09	3.09	0.50
	Gender	1	0.70	0.70	0.11
	Error 3	31	191.2	6.17	

Means

		Constant	Linear	Quadratic
Female (n = 16)	Mean	1.306	1.138	0.456
	Standard deviation	5.920	2.121	1.609
Male (n = 17)	Mean	8.600	−0.876	0.165
	Standard deviation	9.917	4.734	3.085

that there is a significant gender effect of dominance ($F_{1,31} = 6.48$, $p < 0.05$) and an overall dominance effect. The average of the constant term for females is 1.31, for males is 8.6. One question that can be raised is whether the difference between female and male is a true gender difference or can be attributed to differences is size. An analysis of covariance can answer this question (see Problem 11.38).

Data from a repeated measures design often look like those of a randomized block design. The major difference is the way the data are generated. In the randomized block, the treatments are allocated randomly to a block. In the repeated measures design, this is not the case; not being possible, as in the case of observations over time, or because of experimental constraints, as in the example above. If the data are analyzed as a randomized block, care must be taken that the assumptions of the randomized block design are satisfied. The key assumption is that of *compound symmetry*: The sample correlations among treatments over subjects must all estimate the same population correlation. The randomization ensures this in the randomized block design. For example, for the data in Table 10.16, the correlations are as follows:

	Control	**Tablet**	**Capsule**
Tablet	0.658		
Capsule	0.599	0.960	
Coated tablet	0.852	0.784	0.833

These correlations are reasonably comparable. If the correlations are not assumed equal, a conservative F-test can be carried out by referring the observed value of F for treatments to an F-table with 1 and $(I - 1)$ [rather than $(J - 1)$ and $(I - 1)(J - 1)$] degrees of freedom). Alternatives to the foregoing two approaches include multivariate analyses. There is a huge literature on repeated measures analysis. The psychometric literature contains many papers on this topic. To explore this area, consult recent issues of journals such as *American Statistician*. One example is a paper by Looney and Stanley [1989]. See also Chapter 18.

10.4.2 Factorial Designs

An experimental layout that is very common in agricultural and nutritional studies is the balanced factorial design. It is less common in medical research, due to the ever-present risk of missing observations and ethical constraints.

Definition 10.12. In a *factorial design* each level of a factor occurs with every level of every other factor. Experimental units are assigned randomly to treatment combinations.

Suppose that there are three factors with levels $I = 3$, $J = 2$, and $K = 4$. Then there are $3 \times 2 \times 4 = 24$ treatment combinations. If there are three observations per combination, 72 experimental units are needed. Factorial designs, if feasible, are very economical and permit assessment of joint effects of treatments that are not possible with experiments dealing with one treatment at a time. The two-way analysis of variance can be thought of as dealing with a two-factor experiment. The generalization to three or more factors does not require new concepts or strategies, just increased computational complexity.

10.4.3 Hierarchical or Nested Designs

A hierarchical or nested design is illustrated by the following example. As part of a program to standardize measurement of the blood level of phenytoin, an antiepileptic drug, samples with known amounts of active ingredients are sent to four commercial laboratories for analysis. Each laboratory employs a number of technicians who make one or more determinations of the blood level. A possible layout is the following:

Laboratory	1	2	3	4
Technician	1 2	3 4 5	6 7	8 9
Assay	∧∧	∧∧∧	∧∧	∧∧

In this example, laboratory 2 employs three technicians who routinely do this assay; all other laboratories use two technicians. In laboratory 3, each technician runs three assays; in the other laboratories each technician runs two assays. There are three factors: laboratories, technicians, and assays; the arrangement is *not* factorial: there is no reason to match technician 1 with any technician from another laboratory.

Definition 10.13. In a *hierarchical or nested design* levels of one or more factors are subsampled within one or more other factors. In other words, the levels of one or more factors are not crossed with one or more other factors.

In the example above, the factors, "technicians" and "assay," are not "crossed" with the first factor but rather nested within that factor. For the factor "technician" to be "crossed," its levels would have to repeat within each level of "laboratory." That is why we deliberately labeled the levels of "technician" consecutively and introduced some imbalance. Determining whether a design is factorial or hierarchical is not always easy. If the first of the two technicians within a laboratory was the senior technician and the second (or second and third) a junior technician, then "technician" could perhaps be thought of as having two levels, "senior" and "junior," which could then be crossed with "laboratory." A second reason is that designs are sometimes mixed, having both factorial and hierarchical components. In the example above, if "technician" occurred at two levels, "technician" and "laboratory" could be crossed or factorial, but "assay" would continue to be nested within "technician."

10.4.4 Split-Plot Designs

A related experimental design is the split-plot design. We illustrate it with an example. We want to test the effect of physiotherapy in conjunction with drug therapy on the mobility of patients with arthritis. Patients are randomly assigned to physiotherapy, and each patient is given a standard drug and a placebo in random order. The experimental layout is as follows:

		Physiotherapy			
		***i* = 1 (Yes)**		***i* = 2 (No)**	
k	**Patient**	**1**	**2 · · · *J***	**1**	**2 · · · *J***
1	Drug	Y_{111}	— · · · —	Y_{211}	— · · · —
2	Placebo	Y_{112}	— · · · —	Y_{212}	— · · · —

The patients form the "whole plots" and the drug administration, the "split plot." These designs are characterized by almost separate analyses of specified effects. To illustrate in this example, let

$$D_{ij} = Y_{ij1} - Y_{ij2} \quad \text{and} \quad T_{ij} = Y_{ij1} + Y_{ij2}, \qquad i = 1, 2, \quad j = 1, \ldots, J$$

In words, D_{ij} is the difference between drug and placebo for patient j receiving physiotherapy level i; T_{ij} is the sum of readings for drug and placebo. Now carry out an analysis of variance (or two-sample t-test) on each of these variables; see Table 10.23.

Table 10.23 ANOVA Table for Split-Plot Design

		Interpretation of Split-Plot Analyses	
One-Way ANOVA	d.f.	Differences	Sums
Mean	1	Mean differences	Mean sums
Between groups	1	Differences × physiotherapy	Sums × physiotherapy
Within groups	$2(J-1)$	Differences within physiotherapy	Sums within physiotherapy
Total	$2J$	"Total"	"Total"

An analysis of variance of the sums is, in effect, an assessment of physiotherapy (averaged or summed over drug and placebo), that is, a comparison of $\overline{T}_{1\cdot}$ and $\overline{T}_{2\cdot\cdot}$.

The analysis of differences is very interesting. The assessment of the significance of "between groups" is a comparison of the average differences between drug and placebo with physiotherapy and without physiotherapy; that is, $\overline{D}_{1\cdot} - \overline{D}_{2\cdot}$ is a test for interaction. Additionally, the "mean differences" term can be used to test the hypothesis that $\overline{D}_{\cdot\cdot}$ comes from a population with mean zero, that is, it is a comparison of drug and placebo. This test makes sense only if the null hypothesis of no interaction is not rejected.

These remarks are intended to give you an appreciation for these designs. For more details, consult a text on design of experiments, such as Winer [1971].

10.5 UNBALANCED OR NONORTHOGONAL DESIGNS

In previous sections we have discussed balanced designs. The balanced design is necessary to obtain an additive partition of the sum of squares. If the design is not balanced, there are basically three strategies available; the first is to try to restore balance. If only one or two observations are "missing," this is a possible strategy, but if more than two or three are missing, a second or third alternative will have to be used. The second alternative is to use an unweighted means analysis. The third strategy is to use a multiple regression approach; this is discussed in detail in Section 11.10.

10.5.1 Causes of Imbalance

Perhaps the most important thing you can do in the case of unbalanced data is to reflect on the reason(s) for the imbalance. If the imbalance is due to some random mechanism unrelated to the factors under study, the procedures discussed below are appropriate. If the imbalance is due to a specific reason, perhaps related to the treatment, it will be profitable to think very carefully about the implications. Usually, such imbalance suggests a bias in the treatment effects. For example, if a drug has major side effects which cause patients to drop out of a study, the effect of the drug may be estimated inappropriately if only the remaining patients are used in the analysis; if one does the analysis only on patients for whom "all data are available," biased estimates may result.

10.5.2 Restoring Balance

Missing Data in the Randomized Block Design

Suppose that the ijth observation is missing in a randomized block design consisting of I blocks and J treatments. The usual procedure is to:

1. Estimate the missing data point by least squares using the formula

$$\widehat{Y}_{ij} = \frac{IY_{i\cdot} + JY_{\cdot j} - Y_{\cdot\cdot}}{(I-1)(J-1)} \tag{28}$$

where the row, column, and grand totals are those for the values present.

2. Carry out the usual analysis of variance on this augmented data set.
3. Reduce the degrees of freedom for MS_ϵ by 1.

If more than one observation is missing, say two or three, values are guessed for all but one, the latter is estimated by equation (28), a second missing value is deleted, and the process is repeated until convergence. The degrees of freedom for MS_ϵ are now reduced by the number of observations that are missing.

Example 10.6. (*continued*) We return to Table 10.11. Suppose that observation $Y_{31} = 19.1$ is missing and we want to estimate it. For this example, $I = 6$, $J = 4$, $Y_{3\cdot} = 38.8$, $Y_{\cdot 1} = 209.4$, and $Y_{\cdot\cdot} = 599.5$. We estimate Y_{31} by

$$\widehat{Y}_{31} = \frac{6(38.8) + 4(209.4) - 599.5}{(6-1)(4-1)} = 31.4$$

This value appears to be drastically different from 19.1; it is. It also indicates that there is no substitute for real data. The analysis of variance is not altered a great deal (see Table 10.24).

The F-ratios have not changed much from those in Table 10.12. So in this case, the conclusions are unchanged. Note that the degrees of freedom for residual are reduced by 1. This means that the critical values of the F-statistics are increased slightly. Therefore, this experiment has less power than the one without missing data.

Missing Data in Two-Way and Factorial Designs

If a cell in a two-way design has a missing observation, it is possible to replace the missing point by the mean for that cell, carry out the analysis as before, and subtract one degree of freedom for MS_ϵ. A second approach is to carry out an unweighted means analysis. We illustrate both procedures by means of an example.

Example 10.8. These data are part of data used in Wallace et al. [1977]. The observations are from a patient with prostatic carcinoma. The question of interest is whether the immune system of such a patient differs from that of noncarcinoma subjects. One way of assessing this is to stimulate in vitro the patient's lymphocytes with phytohemagglutinin (PHA). This causes blastic transformation. Of interest is the amount of blastogenic generation as measured by DNA incorporation of a radioactive compound. The data observed are the mean radioactive counts per minute both when stimulated with PHA and when not stimulated by PHA. As a control, the amount of PHA stimulation in a pooled sera of normal blood donors was used. To examine the response of a subject's lymphocytes, the quantity

$$\frac{\dfrac{\text{subject's mean count/minute stimulated with PHA}}{\text{subject's mean count/minute without PHA}}}{\dfrac{\text{normal sera mean count/minute stimulated with PHA}}{\text{normal sera mean count/minute without PHA}}} = \frac{X_{11}/X_{12}}{X_{21}/X_{22}} \tag{29}$$

Table 10.24 ANOVA **for Example 10.6**

Source of Variable	d.f.	SS	MS	F-Ratio
Patients (blocks)	5	5341.93	1068.39	9.90
Treatments	3	2330.30	776.77	7.20
Residual	14	1510.94	107.92	—
Total	22	9183.17	—	—

Table 10.25 DNA Incorporation of Sera of Patient with Prostatic Carcinoma Compared to Sera from Normal Blood Donors[a]

	Radioactivity (counts/min)	
Subject	With PHA	Without PHA
Patient sera	129,594 (11.772)	301 (5.707)
	143,687 (11.875)	333 (5.808)
	115,953 (11.661)	295 (5.687)
	103,098 (11.543)	285 (5.652)
	98,125 (11.494)	
Blood donor sera	43,125 (10.672)	247 (5.509)
	46,324 (10.743)	298 (5.697)
	42,117 (10.648)	387 (5.958)
	45,482 (10.725)	
	31,192 (10.348)	

[a] $\log_e$ of counts in parentheses.

was used. If the lymphocytes responded in the same way to the subject's sera and the pooled sera, the ratio should be approximately equal to 1. The data are displayed in Table 10.25.

There is a great deal of variability in the counts/minute values as related to level. In Section 10.6.3 we suggest that logarithms are appropriate for stabilization of the variability. There is a bonus involved in this case. Under the null hypothesis of no difference in patient and blood donor sera, the ratio in equation (28) is 1; that is,

$$H_0 : \frac{E(X_{11})/E(X_{12})}{E(X_{21})/E(X_{22})} = 1$$

This is equivalent to

$$H_0 : \log_e \frac{E(X_{11})/E(X_{12})}{E(X_{21})/E(X_{22})} = \log_e 1 = 0$$

or

$$\log_e E(X_{11}) - \log_e E(X_{12}) - \log_e E(X_{21}) + \log_e E(X_{22}) = 0 \tag{30}$$

Now define

$$Y_{ijk} = \log_e X_{ijk}, \qquad i = 1, 2, \quad j = 1, 2, \quad k = 1, \ldots , n_{ij}$$

It can be shown that equation (30) is zero only if the true interaction term is zero. Thus, the hypothesis that the patient's immune system does not differ from that of noncarcinoma subjects is translated into a null hypothesis about interaction involving the logarithms of the radioactive counts.

We finally get to the "missing data" problem. The data are not balanced: $n_{ij} \neq n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$ [we could delete one observation from the (1,2) cell, but considering the small numbers, we want to retain as much information as possible]. One strategy is to add an observation to cell (2,2) equal to the mean for that cell and adjust the degrees of freedom for interaction. The mean $\overline{Y}_{22\cdot}$ is 5.721. The analysis of variance becomes as shown in Table 10.26.

Note that the MS for error has 13 degrees of freedom, not 14. The MS for error will be the correct estimate using this procedure, but the MS for interaction (and main effects) will not be the same as the one obtained by techniques of Chapter 11. However, it should be close.

Table 10.26 ANOVA for the Missing Data Problem

Source	d.f.	SS	MS	F-Ratio	p-Value
Subject	1	1.4893	1.4893	—	—
PHA	1	131.0722	131.0722	—	—
PHA × subject	1	1.2247	1.2247	50.0	<0.001
Error	13	0.3184	0.02449	—	—
Total	16	—	—	—	—

10.5.3 Unweighted Means Analysis

The second approach is that of unweighted mean analysis. Again, assuming that the unequal cell frequencies are not due to treatment effects, the cell means are used and an average sample size calculated for each cell. The appropriate average sample size is given by the harmonic mean. In the context of our example, the harmonic mean is defined to be

$$\widetilde{n} = \frac{IJ}{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}$$

where n_{ij} is the number of observations in cell (i, j). The harmonic mean is used because the standard error of the mean of cell (i, j) is proportional to $1/n_{ij}$. All calculations for row and column effects are now based on cell means and the harmonic mean of the cell sample sizes. Write the cell means and marginal means as follows:

$$\begin{array}{|cc|c} \hline \overline{Y}_{11\cdot} & \overline{Y}_{12\cdot} & \widehat{M}_{1\cdot} \\ \overline{Y}_{21\cdot} & \overline{Y}_{22\cdot} & \widehat{M}_{2\cdot} \\ \hline \widehat{M}_{\cdot 1} & \widehat{M}_{\cdot 2} & \widehat{M}_{\cdot\cdot} \end{array}$$

The marginal and overall means are just the arithmetic average of the cell means, that is, the unweighted average (hence the name *unweighted mean analysis*). The row and column sums of squares are calculated as follows:

$$\begin{aligned} \mathrm{SS}_\alpha &= \widetilde{n}J \sum (\overline{M}_{i\cdot} - \overline{M}_{\cdot\cdot})^2 \\ \mathrm{SS}_\beta &= \widetilde{n}I \sum (\overline{M}_{\cdot j} - \overline{M}_{\cdot\cdot})^2 \\ \mathrm{SS}_\gamma &= \widetilde{n} \sum (\overline{Y}_{ij\cdot} - \overline{M}_{i\cdot} - \overline{M}_{\cdot j} + \overline{M}_{\cdot\cdot})^2 \end{aligned}$$

SS_ϵ is calculated in the usual way: $\mathrm{SS}_\epsilon = \sum (Y_{ijk} - \overline{Y}_{ij\cdot})^2$. For the example, the calculations are

$$\begin{array}{|cc|c} \multicolumn{3}{c}{\textit{Means}} \\ \hline 11.669000 & 5.713500 & 8.691250 \\ 10.627200 & 5.721333 & 8.174266 \\ \hline 11.148100 & 5.717416 & 8.432758 \end{array}$$

The harmonic mean $\widetilde{n}$ is

$$\widetilde{n} = \frac{(2)(2)}{1/5 + 1/4 + 1/5 + 1/3} = 4.067797$$

$$\mathrm{SS}_\mu = (4.067797)(2)\left[(8.691250 - 8.432758)^2 + (8.174266 - 8.432758)^2\right] = 1.0872$$

Table 10.27 ANOVA **Table for Unweighted Means**

Source	d.f.	SS	MS	F-Ratio	p-Value
Subject	1	1.0872	1.0872		
PHA	1	119.688	119.6888	1	
PHA × subject	1	1.1204	1.1204	45.7	<0.001
Error	13	0.3184	0.02449		
Total	16				

$$\text{SS}_\beta = (4.067797)(2)\left[(11.148100 - 8.432758)^2 + (5.717416 - 8.432758)^2\right] = 119.6888$$

$$\text{SS}_\gamma = (4.067797)\left[(4)(0.262408)^2\right] = 1.1204$$

making use of the fact that all the interaction deviations are equal in absolute value:

$$\overline{Y}_{11\cdot} - \overline{M}_{1\cdot} - \overline{M}_{\cdot 1} + \overline{M}_{\cdot\cdot} = 0.262408$$

$$\overline{Y}_{12\cdot} - \overline{M}_{1\cdot} - \overline{M}_{\cdot 2} + \overline{M}_{\cdot\cdot} = -0.262408, \ldots$$

The ANOVA table based on the unweighted means is shown in Table 10.27.

The conclusion remains unchanged. It turns out in this case that the test for interaction is identical to the multiple regression procedure of Chapter 11.

10.6 VALIDITY OF ANOVA MODELS

10.6.1 Assumptions in ANOVA Models

All the models considered in this chapter have assumed at least the following:

1. Homogeneity of variance
2. Normality of the residual error
3. Statistical independence of the residual errors
4. Linearity of the model

For example, consider again the model associated with the one-way analysis of variance (omitting the subscripts):

$$Y = \mu + \alpha + \epsilon$$

We assumed that (1) the error term ϵ had constant variance for all values of μ and α, and was normally distributed; (2) values of ϵ were randomly (independently) selected; and (3) the response Y was related linearly to μ, α, and ϵ.

In addition, the random effects and repeated measures models made assumptions about the covariances of the random factors and the residual error; other models assumed zero interaction (additivity).

If one or more of the assumptions does not hold, one of the following approaches is frequently used:

1. The data are analyzed by a method that makes fewer assumptions: for example, nonparametric analysis.

2. Part of the data is eliminated or not used, for example, extreme values (i.e., outliers) are deleted or replaced by less extreme values. Deletion usually induces bias.

3. The measurement variables are replaced by categorical variables and some kind of analysis of frequencies is carried out; for example, "age at first pregnancy" is replaced by "teenage mother: yes–no," and the number of observations in various categories is now the outcome variable.

4. A weighted analysis is done; for example, if the variance is not constant at all levels of response, the responses are weighted by the inverse of the variances. The log-linear models of Chapter 7 are an example of a weighting procedure.

5. The data are "transformed" to make the assumptions valid. Typical transformations are: logarithmic, square root, reciprocal, and arcsin $\sqrt{\quad}$. These transformations are nonlinear. Linear transformations do not alter the analysis of variance tests.

6. Finally, appeal is made to the "robustness" of the ANOVA and the analysis is carried out anyway. This is a little bit like riding a bicycle without holding onto the handle bars; it takes experience and courage. If you arrive safely, everyone is impressed, if not, they told you so.

The most common approach is to transform the data. There are advantages and disadvantages to transformations. A brief discussion is presented in the next section. In the other sections we present specific tests of the assumptions of the ANOVA model.

10.6.2 Transformations

Some statisticians recommend routine transformations of data before any analysis is carried out. We recommend the contrary approach; do not carry out transformations unless necessary, and then be very careful, particularly in estimation. We discuss this more fully below, but first we present some common transformations. Table 10.28 lists seven of the most commonly used transformations and one somewhat more specialized one. Each row in the table lists some of the characteristics of the transformation and its uses. A large number of these transformations are variance stabilizing. For example, if the variance of Y is $\lambda^2\mu_Y$, where λ is a constant and μ_Y is the expected value of Y, then $\sqrt{Y}$ tends to have a variance that is constant and equal to $\lambda^2/4$. Hence, this transformation is frequently associated with a Poisson random variable: in this case $\lambda = 1$, so that $\sqrt{Y}$ tends to have a variance of 1/4 regardless of the value of μ_Y. This result is approximate in that it holds for large values of μ_Y. However, the transformation works remarkably well even for small μ_Y, say, equal to 10. Freeman and Tukey [1950] have proposed a modification of the square root transformation which stabilizes the variance for even smaller values of μ_Y. Variance stabilizing transformations tend to be normalizing as well and can be derived explicitly as a function of the variance of the original variable.

The logarithmic transformation is used to stabilize the variance and/or change a multiplicative model into an linear model. When the standard deviation of Y is proportional to μ_Y the logarithmic transformation tends to stabilize the variance. The reciprocal transformation (one per observation) is used when the variance is proportional to μ_Y^4. These first three transformations deal with a progression in the dependence of the variance of Y on μ_Y: from μ_Y to μ_Y^4. The transformations consist of raising Y to an exponent from $Y^{1/2}$ to Y^{-1}. If we define the limit of Y^b to be $\log_e Y$ as b approaches 0, these transformations represent a gradation in exponents. A further logical step is to let the data determine the value of b. This transformation, Y^b, is an example of a power transformation. (*Power* here does not imply "powerful" but simply that Y is raised to the bth power.) See Note 10.4 for additional comments.

The next two transformations are used with proportions or rates. The first one of these is the ubiquitous logistic transformation, which is not variance stabilizing but does frequently induce linearity (cf. Section 7.5). The angle transformation is variance stabilizing but has a finite range; it is not used much anymore because computational power is now available to use the more complex but richer logistic transformation.

Table 10.28 Characteristics of Some Common Transformations of a Random Variable Y

	Range of	Variance^a						
$W = g(Y)$	Y	Y	W	Normalizing	Stabilizing	Linearity	Comments	Uses
$\sqrt{Y}$	$0 \leq Y \leq \infty$	$\lambda^2 \mu_Y$	$\lambda^2/4$	U	Y	—	for $\mu_Y < 10$ use : $W = 1/2(\sqrt{Y} + \sqrt{Y+1})$ (Freeman Tukey transformation)	Poisson
$\log_e Y$	$0 \leq Y \leq \infty$	$\lambda^2 \mu_Y^2$	λ^2	U	Y	C	Use $\log_e(Y+1)$ if zeroes occur	Wide range of Y, e.g., 1–1000
$\frac{1}{Y}$	$0 \leq Y \leq \infty$	$\lambda^2 \mu_Y^4$	λ^2	U	Y	C	Use $1/(Y+1)$ if zeroes occur	Survival time, response time
Y^b	$0 \leq Y \leq \infty$	—	1	Y	Y	C	Box–Cox transformation, Power transformation	Generalized transformation
$\log_e \frac{Y}{1-Y}$	$0 < Y < 1$	$\lambda^2 \mu_Y (1 - \mu_Y)$	$\frac{\lambda^2}{\mu_Y(1-\mu_Y)}$	U	N	C	Logit transformation	Logistic regression, binomial
$\arcsin \sqrt{Y}$	$0 \leq Y \leq 1$	$\lambda^2 \mu_Y (1 - \mu_Y)$	$\lambda^2/4$	U	Y	C	"Angle" transformation	Binomial
$1/2 \log_e \frac{1+Y}{1-Y}$	$-1 \leq Y \leq 1$	$\lambda^2 (1 - \mu_Y^2)^2$	λ^2	Y	Y	C	R. A. Fisher's Z-transformation	Normalize correlation coefficient
$\Phi^{-1}\left(\frac{\text{rank } Y}{n}\right)$	$-\infty \leq Y \leq \infty$	—	1	Y	Y	—	Normal scores transformation $(\text{Rank}(Y) - 1/2)/n$ is sometimes used	Nonparametric analysis

[a] C, could; N, no; U, usually; Y, yes.

The Fisher Z-transformation is used to transform responses whose range is between -1 and $+1$. It was developed specifically for the Pearson product-moment correlation coefficient and discussed in Chapter 9. Finally, we mention one transformation via ranks, the normal scores transformation. This transformation is used extensively in nonparametric analyses and discussed in Chapter 8.

There are benefits to the use of transformations. It is well to state them explicitly since we also have some critical comments. The benefits include the following:

1. Methods using the normal distribution can be used.
2. Tables, procedures, and computer programs are available.
3. A transformation derived for one purpose tends to achieve some other purposes as well—but not always.
4. Inferences (especially relating to hypothesis testing) can be made more easily.
5. Confidence intervals in the transformed scale can be "transformed back" (but estimates of standard errors cannot).

Transformations are more useful for testing purposes than for estimation. The following drawbacks of transformations should be kept in mind:

1. The order of statistics may not be preserved. Consider the following two sets of data: sample 1 : 1, 10; sample 2 : 5, 5. The arithmetic means are 5.5 and 5.0, respectively. The geometric means (i.e., the antilogarithms of the arithmetic mean of the logarithms of the observations) are 3.2 and 5.0, respectively. Hence, the ordering of the *means* is not preserved by the transformation (the ordering of the *raw* data is preserved).
2. Contrary to some, we think that there may be a "natural scale" of measurement. Some examples of variables with a natural scale of measurement are "life expectancy" measured in years, days, or months; cost of medical care in dollars; number of accidents attributable to alcoholism. Administrators or legislators may not be impressed with, or willing to think about, the cost of medical care in terms of "square root of millions of dollars expended."
3. Closely related is the problem of bias. An obvious approach to the criticism in our discussion of drawback 2 is to do the analysis in the transformed units and then transform back to the original scale. Unfortunately, this introduces bias as mentioned in our discussion of drawback 1. Formally, if Y is the variable of interest and $W = g(Y)$ its transform, then it is usually the case that

$$E(W) \neq g(E(Y))$$

   There are ways of assessing this bias and eliminating it but such methods are cumbersome and require an additional layer of computations, something the transformation was often designed to reduce!
4. Finally, many of the virtues of transformations are asymptotic virtues; they are approached as the sample size becomes very large. This should be kept in mind when analyzing relatively small data sets.

10.6.3 Testing of Homogeneity of Variance

It is often the case that the variance or standard deviation is proportional to the mean level of response. There are two common situations where this occurs. First, where the range of response varies over two or more orders of magnitude; second, in situations where the range of response is bounded, on the left, the right or both. Examples of the former are Poisson random variables; examples of the latter, responses such as proportions, rates, or random variables that cannot be negative.

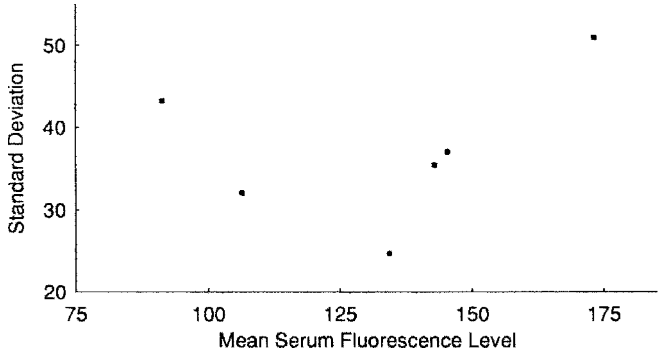

Figure 10.6 Mean serum fluorescence level and standard deviation. (Data from Sherwin and Layfield [1976]; see Example 10.5.)

The simplest verification of homogeneity of variance is provided by a graph, plotting the variance or standard deviation vs. the level of response.

Example 10.5. (*continued*) In Table 10.8, the means and standard deviations of serum fluorescence readings of mice exposed to nitrogen dioxide are given. In Figure 10.6 the standard deviations are plotted against the means of the various treatment combinations. This example does not demonstrate any pattern between the standard deviation and the cell means. It would not be expected because the range of the cell means is fairly small.

Example 10.9. A more interesting example is the data of Quesenberry et al. [1976] discussed in Problem 3.14. Samples of peanut kernels were analyzed for aflatoxin levels. Each sample was divided into 15 or 16 subsamples. There was considerable variability in mean levels and corresponding standard deviations.

A plot of means vs. standard deviations displays an increasing pattern, suggesting a logarithmic transformation to stabilize the variance. This transformation as well as two other transformations ($\sqrt{Y}, Y^{1/4}$) are summarized in Table 10.29. Means vs. standard deviations are

Table 10.29 Aflatoxin Levels in Peanut Kernels: Means and Standard Deviations for 11 Samples Using Transformations

		Mean and Standard Deviation of Aflatoxin Level							
		Y		$W = Y^{1/4}$		$W = \sqrt{Y}$		$W = \log Y$	
Sample	n	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	16	110	25.6	3.2	0.192	10.4	1.24	4.7	0.240
2	16	79	20.6	3.0	0.204	8.8	1.19	4.3	0.281
3	16	21	3.9	2.1	0.109	4.5	0.45	3.0	0.213
4	16	33	12.2	2.4	0.192	5.7	0.96	3.4	0.311
5	15	32	10.6	2.4	0.194	5.6	0.92	3.4	0.328
6	16	15	2.7	2.0	0.089	3.8	0.35	2.7	0.183
7	15	33	6.2	2.4	0.111	5.8	0.54	3.5	0.183
8	16	31	2.8	2.4	0.054	5.6	0.26	3.4	0.092
9	16	17	4.2	2.0	0.129	4.1	0.51	2.8	0.261
10	16	8	3.1	1.7	0.143	2.9	0.49	2.1	0.339
11	15	84	17.7	3.0	0.164	9.1	0.98	4.4	0.221

Source: Data from Quesenberry et al. [1976].

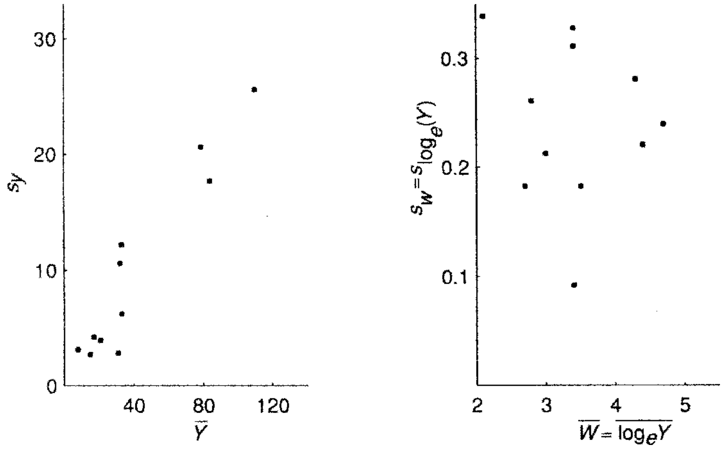

Figure 10.7 Means vs. standard deviation, arithmetic and logarithmic scales. (Data from Wallace et al. [1977]; see Example 10.8.)

plotted in Figure 10.7. The first pattern clearly indicates a linear trend; the plot for the data expressed as logarithms suggests very little pattern. This does not prove that the lognormal model is appropriate. Quesenberry et al. [1976], in fact, considered two classes of models: the 11 samples are from normal distributions with means and variances $\mu_i, \sigma_i^2, i = 1, \dots, 11$; the second class of models assumes that the logarithms of the aflatoxin levels for the 11 samples come from normal distributions with means and variances $\gamma_i, \theta^2, i = 1, \dots, 11$.

On the basis of their analysis, they conclude that the normal models are more appropriate. The cost is, of course, that 10 more parameters have to be estimated. Graphs of means vs. standard deviation for the $\sqrt{Y}$ and $Y^{1/4}$ scale still suggest a relationship.

The tests of homogeneity of variance developed here are graphical. There are more formal tests. All of the tests assume normality and are sensitive to departure from normality. In view of the robustness of the analysis of variance to heterogeneity of variance, Box [1953] remarked that "...to make the preliminary tests on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port." There are four common tests of homogeneity of variance, associated with the names of Hartley, Cochran, Bartlett, and Scheffé. Only the first two are described here, they will be adequate for most purposes. For a description of the other tests see, for example, Winer [1971]. Suppose that there are k samples with sample size n_i and sample variance $s_i^2, i = 1, \dots, k$. For the moment, assume that all n_i are equal to n. Hartley's test calculates

$$F_{\text{MAX}} = \frac{s^2_{\text{maximum}}}{s^2_{\text{minimum}}}$$

Cochran's test calculates

$$C = \frac{s^2_{\text{maximum}}}{\sum s_i^2}$$

In the absence of software for computing critical values, both statistics can be referred to appropriate tables in the Web appendix. If the sample sizes are not equal, the tables can be entered with the minimum sample size to give a conservative test and with the maximum

Table 10.30 Calculations for Example 10.9

Scale	$F_{\max}$	C
Y	$\left(\dfrac{25.6}{2.7}\right)^2 = 89.9$	$\dfrac{(25.7)^2}{1758.1} = 0.38$
$\sqrt{Y}$	$\left(\dfrac{1.24}{0.26}\right)^2 = 22.7$	$\dfrac{(1.24)^2}{9.787} = 0.16$
$Y^{1/4}$	$\left(\dfrac{0.204}{0.054}\right)^2 = 14.1$	$\dfrac{(0.204)^2}{0.252} = 0.16$
$\log_e Y$	$\left(\dfrac{0.339}{0.092}\right)^2 = 13.6$	$\dfrac{(0.339)^2}{0.694} = 0.17$
Critical value at 0.05 level	5.8	0.15

sample size to give a "liberal" test (i.e., the null hypothesis is rejected more frequently than the nominal significance level).

Example 10.9. (*continued*) For the transformations considered, the F_{MAX} test and C test statistics are as shown in Table 10.30.

The critical values have been obtained by interpolation. The F_{MAX} test indicates that none of the transformations achieve satisfactory homogeneity of variance, validating one of Quesenberry et al.'s conclusions. The Cochran test suggests that there is little to choose between the three transformations.

A question remains: How valid is the analysis of variance under heterogeneity of variance? Box [1953] indicates that for three treatments a ratio of 3 in the maximum-to-minimum *population* variance does not alter the significance level of the test appreciably (one-way ANOVA model with $n_{i\cdot} = 5, I = 3$). The analysis of variance is therefore reasonably robust with respect to deviation from homogeneity of variance.

10.6.4 Testing of Normality in ANOVA

Tests of normality are not as common or well developed as tests of homogeneity of variance. There are at least two reasons: first, they are not as crucial because even if the underlying distribution of the data is not normal, appeal can be made to the central limit theorem. Second, it turns out that fairly large sample sizes are needed (say, $n > 50$) to discriminate between distributions. Again, most tests are graphical.

Consider for simplicity the one-way analysis of variance model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, n_i$$

By assumption the ϵ_{ij} are iid $N(0, \sigma^2)$. The ϵ_{ij} are estimated by

$$\epsilon_{ij} = Y_{ij} - \overline{Y}_{i\cdot}$$

The e_{ij} are normally distributed with population mean zero; $\sum e_{ij}^2/(n - I)$ is an unbiased estimate of σ^2 but the e_{ij} are not statistically independent. They can be made statistically independent, but it is not worthwhile for testing the normality. Some kind of normal probability plot is usually made and a decision made based on a visual inspection. Frequently, such a plot is used to identify outliers. Before giving an example, we give a simple procedure which is based on the use of order statistics.

Definition 10.14. Given a sample of n observations, $Y_1, Y_2, \ldots, Y_n$, the *order statistics* $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ are the values ranked from lowest to highest.

Now suppose that we generate samples of size n from an $N(0, 1)$ distribution and average the values of the order statistics.

Definition 10.15. *Rankits* are the expected values of the order statistics of a sample of size n from an $N(0, 1)$ distribution. That is, let $Z_{(1)}, \ldots, Z_{(n)}$ be the order statistics from an $N(0, 1)$ population; then the rankits are $E(Z_{(1)}), E(Z_{(2)}), \ldots, E(Z_{(n)})$.

Rankits have been tabulated in Table A.13. A plot of the order statistics of the residuals against the rankits is equivalent to a normal probability plot. A reasonable approximation for the ith rankit is given by the formula

$$E(Z_{(i)}) \doteq 4.91[p^{0.14} - (1 - p)^{0.14}] \tag{31}$$

where

$$p = \frac{i - 3/8}{n + 1/4}$$

For a discussion, see Joiner and Rosenblatt [1971]. To illustrate its use we return to Example 10.1. A one-way analysis of variance was constructed for these data and we now want to test the normality assumption.

Example 10.1. (*continued*) The distribution of ages at which infants first walked [discussed in Section 10.2.1 (see Table 10.1)] is now analyzed for normality. The residuals $Y_{ij} - \overline{Y}_{i\cdot}$ for the 23 observations are:

−1.125	−0.375	−0.208	0.900
−0.625	−1.375	0.292	−0.850
−0.375	−1.375	−2.708	−0.350
−0.125	0.375	−0.208	1.150
2.875	−0.875	1.542	−0.850
−0.625	3.625	1.292	

Note that the last observation has been omitted again so that we are working with the 23 observations given in the paper. These observations are now ranked from smallest to largest to be plotted on probability paper. To illustrate the use of rankits, we will calculate the expected values of the 23 normal (0,1) order statistics using equation (31). The 23 order statistics for e_{ij}, $e_{(ij)}$, and the corresponding rankits are presented in Table 10.31.

For example, the largest deviation is −2.708; the expected value of $Z_{(1)}$ associated with this deviation is calculated as follows:

$$p = \frac{1 - 3/8}{23 + 1/4} = 0.02688$$

$$\begin{aligned} E(Z_{(1)}) &= 4.91[(0.02688)^{0.14} - (1 - 0.02688)^{0.14}] \\ &= -1.93 \end{aligned}$$

The rankits and the ordered residuals are plotted in Figure 10.8. What do we do with this graph? Is there evidence of nonnormality?

Table 10.31 Order Statistics for Example 10.1

$e_{(ij)}$	$E(Z_{(ij)})$	$e_{(ij)}$	$E(Z_{(ij)})$	$e_{(ij)}$	$E(Z_{(ij)})$
−2.708	−1.93	−0.625	−0.33	0.375	0.57
−1.375	−1.48	−0.375	−0.22	0.900	0.70
−1.375	−1.21	−0.375	−0.11	1.150	0.84
−1.125	−1.01	−0.350	0.0	1.292	1.01
−0.875	−0.84	−0.208	0.11	1.542	1.21
−0.850	−0.70	−0.208	0.22	2.875	1.48
−0.850	−0.57	−0.125	0.33	3.625	1.93
−0.625	−0.44	−0.292	0.44		

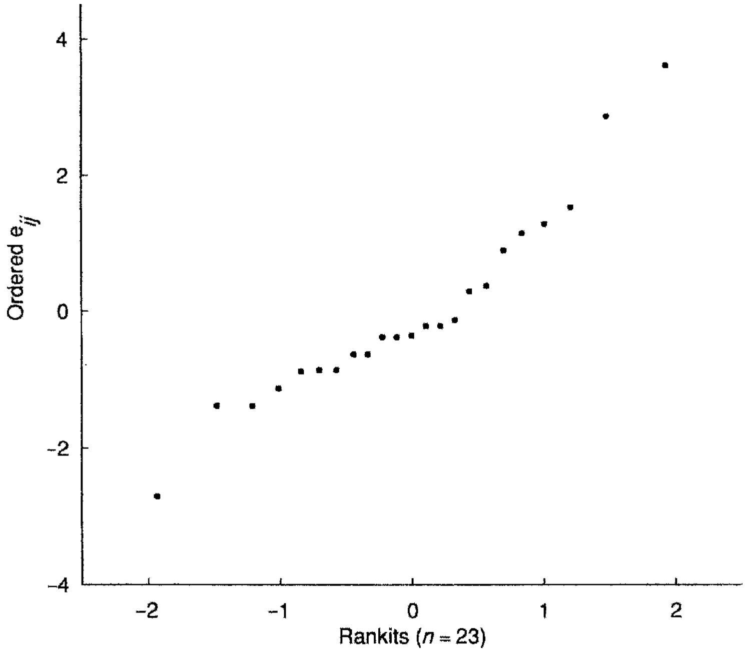

Figure 10.8 Normal probability plot of residuals from linear model. (Data from Zelazo et al. [1972]; see Example 10.1.)

There does seem to be some excessive deviation in the tails. The question is: How important is it? One way to judge this would be to generate many plots for normal and nonnormal data and compare the plots to develop a visual "feel" for the data. This has been done by Daniel and Wood [1999] and Daniel [1976]. Comparison of this plot with the plots in Daniel and Wood suggests that these data deviate moderately from normality. For a further discussion, see Section 11.8.1.

More formal tests of normality can be carried out using the Kolmogorov–Smirnov test of Chapter 8. A good test is based on the Pearson product-moment correlation of the order statistics and corresponding rankits. If the residuals are normally distributed, there should be a very high correlation between the order statistics and the rankits. The (null) hypothesis of normality is rejected when the correlation is *not large enough*. Weisberg and Bingham [1975] show that this is a very effective procedure. The critical values for the correlation have been tabulated; see, for example, Ryan et al. [1980]. For $n \geq 15$, the critical value is on the order of 0.95 or more. This is a simple number to remember. For Example 10.1, the correlation between the order statistics of the residuals, $e_{(ij)}$ and the rankits $E(Z_{(ij)})$ is $r = 0.9128$ for $n = 23$. This is somewhat lower than the critical value of 0.95 again, suggesting that the residuals are "not quite" normally distributed.

10.6.5 Independence

One of the most difficult assumptions to verify is that of statistical independence of the residuals. There are two problems. First, tests of independence for continuous variables are difficult to implement. Frequently, such tests are, in fact, tests of no correlation among the residuals, so that if the errors are normally distributed and uncorrelated, they are independent. Second, the observed residuals in the analysis of variance have a built-in dependence due to the constraints on the linear model. For example, in the one-way analysis of variance with I treatments and, say, $n_{i\cdot} = m$ observations per treatment, there are mI residuals but only $(m - 1)I$ degrees of freedom; this induces some correlation among the residuals. This is not an important dependence and can be taken care of.

Tests for dependence usually are tests for serial correlation (i.e., correlation among adjacent values). This assumes that the observations can be ordered in space or time. The most common test statistic for serial correlation is the Durbin–Watson statistic. See, for example, Draper and Smith [1998]. Computer packages frequently will print this statistic assuming that the observations are entered in the same sequence in which they were obtained. This, of course, is rarely the case and the statistic and its value should not be used. Such "free information" is sometimes hard to ignore; the motto for computer output is *caveat lector* (let the reader beware).

10.6.6 Linearity in ANOVA

Like independence, linearity is difficult to verify. In Example 10.7 we illustrated a multiplicative model. The model was transformed to a linear (nonadditive) model by considering the logarithm of the original observations. Other types of nonlinear models are discussed in Chapters 11 to 15. Evidence for a nonlinear model may consist of heterogeneity of variance or interaction. However, this need not always be the case. Scheffé [1959] gives the following example. Suppose that there are $I + J + 1$ independent Poisson variables defined as follows: $U_1, U_2, \ldots, U_I$ have means $\alpha_1, \alpha_2, \ldots, \alpha_I$; $V_1, V_2, \ldots, V_J$ have means $\beta_1, \beta_2, \ldots, \beta_J$; and W has mean γ. Let $Y_{ij} = W + U_i + V_j$; then $E(Y_{ij}) = \gamma + \alpha_i + \beta_j$; that is, we have an additive, linear model. But $\text{var}(Y_{ij}) = \gamma + \alpha_i + \beta_j$, so that there is heterogeneity of variance (unless all the α_i are equal and all the β_j are equal). The square root transformation destroys the linearity and the additivity. Scheffé [1959] states: "It is not obvious whether Y or $\sqrt{Y}$ is more nearly normal ... but in the present context it hardly matters." A linear model is frequently assumed to be appropriate for a set of data without any theoretical basis. It may be a reasonable first-order approximation to the "state of nature" but should be recognized as such.

Sometimes a nonlinear model can be derived from theoretical principles. The form of the model may then suggest a transformation to linearity. But as the example above illustrates, the transformation need not induce other required properties of ANOVA models, or may even destroy them.

Another strategy for testing linearity is to add some nonlinear terms to the model and then test their significance. In Sections 11.7 and 11.8 we elaborate on this strategy.

10.6.7 Additivity

The term *additivity* is used somewhat ambiguously in the statistical literature. It is sometimes used to describe the transformation of a multiplicative model to a linear model. The effects of the treatment variables become "additive" rather than multiplicative. We have called such a transformation a *linearizing transformation*. It is not always possible to find such a transformation (see Section 11.10.5). We have reserved the term *additivity* for the additive model illustrated by the two-way analysis of variance model (see Definition 10.4). A test for additivity then becomes a test for "no interaction." Scheffé [1959] proves that transformations to additivity exists for a very broad class of models.

The problem is that the existence of interaction may be of key concern. Consider Example 10.8. The existence of interaction in this example is taken as evidence that the immune system of a patient with prostatic carcinoma differed from that of normal blood donors. This finding has important implications for a theory of carcinogenesis. These data are an example of the importance of expressing observations in an appropriate scale. Of course, what evidence is there that the logarithms of the radioactive count is the appropriate scale? There is some arbitrariness, but the original model was stated in terms of percentage changes, and this implies constant changes on a logarithmic scale.

So the problem has been pushed back one step: Why state the original problem in terms of percentage changes? The answer must be found in the experimental situation and the nature of the data. Ultimately, the researcher will have to provide justification for the initial model used.

This discussion has been rather philosophical. One other situation will be considered: the randomized block design. There is no test for interaction because there is only one observation per cell. Tukey [1949] suggested a procedure that is an example of a general class of procedures. The validity of a model is evaluated by considering an enlarged model and testing the significance of the terms in the enlarged model. To be specific, consider the randomized block design model of equation (23):

$$Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J$$

Tukey [1949] embedded this model in the "richer" model

$$Y_{ij} = \mu + \beta_i + \tau_j + \lambda\beta_i\tau_j + \varepsilon_{ij}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J \tag{32}$$

He then proposed to test the null hypothesis,

$$H_0 : \lambda = 0$$

as a test for nonadditivity. Why this form? It is the simplest nonlinear effect involving both blocks and treatments. The term λ is estimated and tested as follows. Let the model without interaction be estimated by

$$Y_{ij} = \overline{Y}.. + b_i + t_j + e_{ij}$$

where

$$b_i = \overline{Y}_i. - \overline{Y}.., t_j = \overline{Y}._j - \overline{Y}.. \quad \text{and} \quad e_{ij} = Y_{ij} - \overline{Y}.. - b_i - t_j$$

We have the usual constraints,

$$\sum b_i = \sum t_j = 0$$

and

$$\sum_i e_{ij} = \sum_j e_{ij} = 0 \qquad \text{for all } i \text{ and } j$$

Now define

$$X_{ij} = b_i t_j, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J \tag{33}$$

It can be shown that the least squares estimate, $\widehat{\lambda}$, of λ is

$$\widehat{\lambda} = \frac{\sum X_{ij} Y_{ij}}{\sum X_{ij}^2} \tag{34}$$

Since $\overline{X} = 0$ (why?), the quantity $\widehat{\lambda}$ is precisely the regression of Y_{ij} on X_{ij}. The sum of squares for regression is the sum of squares for nonadditivity in the ANOVA table:

$$\text{SS}_\lambda = \text{SS}_{\text{nonadditivity}} = \frac{\left(\sum X_{ij} Y_{ij}\right)^2}{\sum X_{ij}^2} \tag{35}$$

The ANOVA table for the randomized block design including the test for nonadditivity is displayed in Table 10.32. As expected, the SS_λ has one degree of freedom since we are estimating a slope. But who "pays" for the one degree of freedom? A little reflection indicates that it must come out of the error term; the number of constraints on the block and treatment effects remain the same. A graph of Y_{ij} vs. X_{ij} (or equivalently, e_{ij} vs. X_{ij}) will indicate whether there is any pattern.

The idea of testing models within larger models as a way of testing the validity of the model is discussed further in Section 11.8.2.

Example 10.6. (*continued*) We now apply the Tukey test for additivity to the experiment assessing the effect of pancreatic supplements on fat absorption in patients with steatorrhea, discussed in Section 10.3.2. We need to calculate SS_λ from equation (35) and this involves the regression of Y_{ij} on X_{ij}, where X_{ij} is defined by equation (33). To save space we calculate only a few of the X_{ij}. For example,

$$\begin{aligned} X_{11} &= \left(\overline{Y}_{1\cdot} - \overline{Y}_{\cdot\cdot}\right)\left(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot\cdot}\right) \\ &= (16.9 - 25.775)(38.083 - 25.775) \\ &= -109.2 \end{aligned}$$

and

$$\begin{aligned} X_{23} &= \left(\overline{Y}_{2\cdot} - \overline{Y}_{\cdot\cdot}\right)\left(\overline{Y}_{\cdot 3} - \overline{Y}_{\cdot\cdot}\right) \\ &= (25.625 - 25.775)(17.417 - 25.775) \\ &= 1.3 \end{aligned}$$

(Note that a few more decimal places for the means are used here as compared to Table 10.15.) A graph of Y_{ij} vs. X_{ij} is presented in Figure 10.9. The estimate of the slope is

$$\begin{aligned} \widehat{\lambda} &= \frac{\sum X_{ij} Y_{ij}}{\sum X_{ij}^2} \\ &= \frac{(-109.2)(44.5) + (82.0)(7.3) + \cdots + (98.8)(52.6)}{(-109.2)^2 + (82.0)^2 + \cdots + (98.8)^2} \\ &= \frac{13{,}807}{467{,}702} \\ &= 0.029521 \end{aligned}$$

SS_λ is

$$\text{SS}_\lambda = \frac{(13{,}807)^2}{467{,}702} = 407.60$$

The analysis of variance is tabulated in Table 10.33.

Table 10.32 ANOVA of Randomized Block Design Incorporating Tukey Test for Additivity[a]

Source of Variation	d.f.	SS[b]	MS	F-Ratio	d.f.	E(MS)	Hypothesis Tested
Grand mean	1	$\text{SS}_\mu = n\overline{Y}_{\cdot\cdot}^2$	$\text{MS}_\mu = \text{SS}_\mu$	$\dfrac{\text{MS}_\mu}{\text{MS}_\epsilon}$	$(1, IJ - I - J)$	$n\mu^2 + \sigma^2$	$\mu = 0$
Blocks	$I - 1$	$\text{SS}_\beta = J\sum(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$	$\text{MS}_\beta = \dfrac{\text{SS}_\beta}{I-1}$	$\dfrac{\text{MS}_\beta}{\text{MS}_\epsilon}$	$(I - 1, IJ - I - J)$	$\dfrac{J\sum\beta_i^2}{I-1} + \sigma^2$	$\beta_i = 0$ all i
Treatments	$J - 1$	$\text{SS}_\tau = I\sum(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot})^2$	$\text{MS}_\tau = \dfrac{\text{SS}_\tau}{J-1}$	$\dfrac{\text{MS}_\tau}{\text{MS}_\epsilon}$	$(J - 1, IJ - I - J)$	$\dfrac{I\sum\tau_j^2}{J-1} + \sigma^2$	$\tau_i = 0$ all j
Nonadditivity	1	$\text{SS}_\lambda^c = \dfrac{\left(\sum X_{ij}Y_{ij}\right)^2}{\sum X_{ij}^2}$	$\text{MS}_\lambda = \text{SS}_\lambda$	$\dfrac{\text{MS}_\lambda}{\text{MS}_\epsilon}$	$(1, IJ - I - J)$	$\lambda^2 C^d + \sigma^2$	$\lambda = 0$
Residual	$IJ - I - J$	$\text{SS}_\epsilon =$ by subtraction	$\text{MS}_\epsilon = \dfrac{\text{SS}_\epsilon}{IJ - I - J}$				
Total	IJ	$\sum Y_{ij}^2$					

[a] Model: $Y_{ij} = \mu + \beta_i + \tau_j + \lambda\beta_i\tau_j + \epsilon_{ij}$ [$\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$]. Data: $Y_{ij} = \overline{Y}_{\cdot\cdot} + (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) + (Y_{\cdot j} - Y_{\cdot\cdot}) + \widehat{\lambda}X_{ij} + \widetilde{e}_{ij}$ (residual obtained by subtraction).

[b] Summation is over all subscripts displayed.

[c] $X_{ij} = (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot})$.

[d] C is a constant that depends on the cell means; it is zero if the additive model holds.

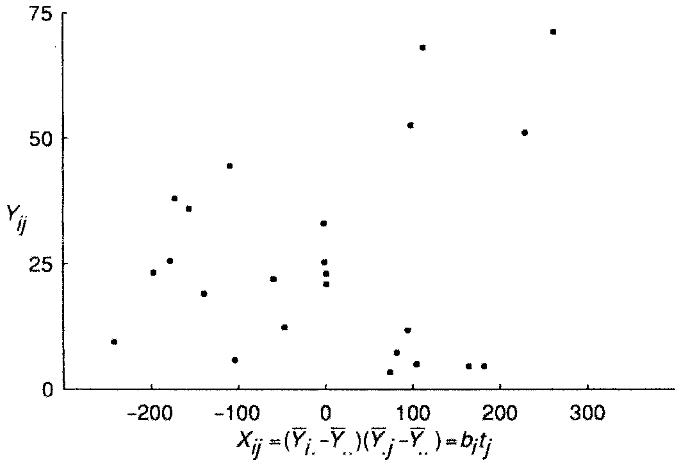

Figure 10.9 Plot of the Tukey test for additivity. See the text for an explanation.

Table 10.33 Randomized Block Analysis with Tukey Test for Additivity of Fecal Fat Excretion of Patients with Steatorrhea

Source of Variation	d.f.	SS	MS	F-Ratio	p-Value
Patients	5	5588.38	1117.68	13.1	<0.001
Treatments	3	2008.60	669.53	7.83	<0.01
Additivity	1	407.60	407.60	4.76	$0.025 < p < 0.05$
Residual	14	1197.80	85.557		
Total	23	9202.38			

Source: Data from Graham [1977].

The test for additivity indicates significance at the 0.05 level ($p = 0.047$); thus there is some evidence that the data cannot be represented by an additive model. Tukey [1949] related the constant a in Y^a (power transformation) to the degree of nonadditivity by the following formula:

$$\widehat{a} = 1 - \widehat{\lambda Y}\ldots$$

The quantity $\widehat{a}$ is a statistic and hence a random variable. For a particular set of data, the confidence interval on $\widehat{a}$ will tend to be fairly wide; hence, a "nice" value of "a" is usually chosen. For the example, $\widehat{a} = 1 - (0.029521)(25.775) = 0.239$. A "nice" value for " a" is thus 0.25, or even 0.20.

10.6.8 Strategy for Analysis of Variance

It is useful to have a checklist in carrying out an ANOVA. Not every item on the list needs to be considered, nor necessarily in the order given, but you will find it useful to be reminded of these items:

1. Describe how the data were generated: from what population? To what population will inferences be made? State explicitly at what steps in the data generation randomness entered.

2. Specify the ANOVA null hypotheses, alternative hypotheses; whether the model is fixed, random, or mixed.
3. Graph the data to get some idea of treatment effects, variability, and possible outliers.
4. If necessary, test the homogeneity of variance and the normality.
5. If ANOVA is inappropriate on the data as currently expressed, consider alternatives. If transformations are used, repeat steps 2 and 4.
6. Carry out the ANOVA. Calculate F-ratios. Watch out for F-ratios much less than 1; they usually indicate an inappropriate model.
7. State conclusions and limitations.
8. If null hypotheses are not rejected, consider the power of the study and of the analysis.
9. For more detailed analyses and estimation procedures, see Chapter 12.

NOTES

10.1 Ties in Nonparametric Analysis of Variance (One-Way and Randomized Block)

As indicated, both the Kruskal–Wallis and the Friedman tests are conservative in the presence of ties. The adjustment procedure is similar to those used in Chapter 8, equation (4). For the Kruskal–Wallis situation, let

$$C_{\text{KW}} = \frac{\sum_{l=1}^{L} \left(t_l^3 - t_l\right)}{n^3 - n}$$

where L is the number of groups of tied ranks and t_l is the number of ties in group l, $l = 1, \dots, L$. Then the statistic T_{KW} [equation (13)] is adjusted to $T_{\text{ADJ}} = T_{\text{KW}}/(1 - C_{\text{KW}})$. Since $0 \leq C_{\text{KW}} \leq 1$, $T_{\text{ADJ}} \geq T_{\text{KW}}$. Hence, if the null hypothesis is rejected with T_{KW}, it will certainly be rejected with T_{ADJ} since the degrees of freedom remain unchanged. Usually, C_{KW} will be fairly small: Suppose that there are 10 tied observations in an ANOVA of 20 observations; in this case $C_{\text{KW}}(10^3 - 10)/(20^3 - 20) = 0.1241$, so that $T_{\text{ADJ}} = T_{\text{KW}}/(1 - 0.1241) = 1.14T_{\text{KW}}$. The adjusted value is only 14% larger than the value of T_{KW} even in this extreme situation. (If the 10 ties are made up of five groups of two ties each, the adjustment is less than 0.5%).

A similar adjustment is made for the Friedman statistic, given by equations (25) and (26). In this case,

$$C_{\text{FR}} = \frac{\sum_{i=1}^{I} \sum_{l=1}^{L_i} \left(t_{il}^3 - t_{il}\right)}{I(J^3 - J)}$$

where t_{il} is the number of ties in group l within block i and untied values within a block are counted as a group of size 1. (Hence $\sum_{l=1}^{L_i} t_{il} = J$ for every i.) The adjusted Friedman statistic, T_{ADJ}, is $T_{\text{ADJ}} = T_{\text{FR}}/(1 - C_{\text{FR}})$. Again, unless there are very many ties, the adjustment factor, C_{FR} will be relatively small.

10.2 Nonparametric Analyses with Ordered Alternatives

All the tests considered in this chapter have been "omnibus tests"; that is, the alternative hypotheses have been general. In the one-way ANOVA, the null hypothesis is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$, the alternative hypothesis $H_1 : \mu_i \neq \mu_i'$ for at least one i and i'. Since the power of a test is determined by the alternative hypothesis, we should be able to "do better" using more specific alternative hypotheses. One such hypothesis involves ordered alternatives. For the one-way ANOVA (see Section 10.2), let $H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_I$ with at least one strict inequality. A regression-type parametric analysis could be carried out by coding the categories $X = 1$, $X = 2, \ldots , X = I$.

A nonparametric test of H_0 against an ordered alternative H_1 was developed by Terpstra and Jonckheere (see, e.g., Hollander and Wolfe [1999]). The test is based on the Mann–Whitney statistic (see Section 8.6). The Terpstra–Jonckheere statistic is

$$T_{\text{TJ}} = \sum_{i=1}^{I-1} \sum_{k=i+1}^{I} M_{ik} = \sum_{i<k} M_{ik}$$

where M_{ik} is the number of pairs with the observation in group i less than that of group $k(i < k)$ among the $n_i n_k$ pairs.

Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$, the statistic T_{TJ} has a distribution that approaches a normal distribution as n becomes large, with mean and variance given by

$$E\,[T_{\text{TJ}}] = \frac{n^2 - \sum n_i^2}{4}$$

and

$$\text{var}[T_{\text{TJ}}] = \frac{[n^2(2n+3) - \sum n_i^2(2n_i+3)]}{72}$$

where $n = n_1 + n_2 + \cdots + n_I$. See Problems 10.3 and 10.11 for an application.

In Section 10.3.3, a nonparametric analysis of randomized block design was presented to test the null hypothesis $H_0 : \tau_1 = \tau_2 = \cdots = \tau_J = 0$. Again, we consider an ordered alternative, $H_1 : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_J$ with at least one strict inequality. Using the notation of Section 10.3.3, let $R_{\cdot j}$ = sum of ranks for treatment j. Page [1963] developed a nonparametric test of H_0 against H_1. The statistic $T_{\text{PAGE}} = \sum_{j=1}^{J} jR_{\cdot j}$ under the null hypothesis approaches a normal distribution (as I become large) with mean and variance

$$E\,[T_{\text{PAGE}}] = \frac{IJ^2(J+1)}{4}$$

and

$$\text{var}\,[T_{\text{PAGE}}] = \frac{I(J^3 - J)^2}{144(J-1)}$$

10.3 Alternative Rank Analyses

Conover and Iman [1981] in a series of papers have advocated a very simple rank analysis: Replace observations by their ranks and then carry out the usual parametric analysis. These procedures must be viewed with caution when models are nonadditive [Akritas, 1990] and discussion in Chapter 8. Hettmansperger and McKean [1978] provide an illustration of another class of rank-based analytical procedures that can be developed. There are three steps in this type of approach:

1. Define a robust or nonparametric estimate of dispersion.
2. State an appropriate statistical model for the data.
3. Given a set of data, estimate the values of the parameters of the model to minimize the robust estimate of dispersion.

A drawback of such procedures is that estimates cannot be written explicitly, and more important, the estimation procedure is nonlinear, requiring a computer to carry it out. However, with the increasing availability of microcomputers, it will only be a matter of time until software will be developed, making such procedures widely accessible.

It is possible to run a parametric analysis of the raw data routinely and compare it with some alternative rank analysis. If the two analyses do not agree, the data should be examined more carefully to determine the cause of the discrepant results; usually, it will be due to the *nonnormality* of the data. The researcher then has two choices: if the nonnormality is thought to be a characteristic of the biological system from which the data came, the rank analysis would be preferred. On the other hand, if the nonnormality is due to outliers (see Chapter 8), there are other options available, all the way from redoing the experiment (more carefully this time), to removing the outliers, to sticking with the analysis of the ranks. Clearly, there are financial, ethical, and professional costs and risks. What should *not* be done in the case of disagreement is to pick the analysis that conforms, in some sense, to the researcher's preconceptions or desires.

10.4 Power Transformation

Let Y^δ be a transformation of Y. The assumption is that Y^δ is normally distributed with mean μ (which will depend on the experimental model) and variance σ^2. The SS_ε will now be a function of δ. It can be shown that the appropriate quantity to be minimized is

$$L(\delta) = \frac{n}{2}\text{SS}_\varepsilon - \sum \ln(\delta y^\delta)$$

and defined to be

$$= \frac{n}{2}\text{SS}_\epsilon - \sum \ln y$$

for $\delta = 0$ (corresponding to the logarithmic transformation). Typically, this equation is solved by trial and error. With a computer this can be done quickly. Usually, there will be a range of values of δ over which the values of $L(\delta)$ will be close to the minimum; it is customary then to pick a value of δ that is simple. For example, if the minimum of $L(\delta)$ occurs at $\delta = 0.49$, the value chosen will be $\delta = 0.50$ to correspond to the square root transformation. For an example, see Weisberg [1985]. Empirical evidence suggests that the value of δ derived from the data is frequently close to some "natural" rescaling of the data. (This may just be a case of perfect 20/20 hindsight.)

PROBLEMS

For Problems 10.1 to 10.23, carry out one or more of the following tasks. Additional tasks are indicated at each problem.

(a) State an appropriate ANOVA model, null hypotheses, and alternative hypotheses. State whether the model is fixed, random, or mixed. Define the population to which inferences are to be made.

(b) Test the assumption of homogeneity of variance.

(c) Test the assumption of normality using a probability plot.

(d) Test the assumption of normality correlating residuals and rankits.

(e) Graph the data. Locate the cell means on the graph.

(f) Transform the data. Give a rationale for transformation.

(g) Carry out the analysis of variance. State conclusions and reservations. Compare with the conclusions of the author(s) of the paper. If possible, estimate the power if the results are not significant.

(h) Carry out a nonparametric analysis of variance. Compare with conclusions of parametric analysis.

(i) Partition each observation into its component parts [see, e.g., equations (4) and (19)] and verify that the sum of squares of each component is equal to one of the sums of squares in the ANOVA table.

(j) Construct the ANOVA table from means and standard deviations (or standard errors). Do relevant parts of (g).

10.1 Olsen et al. [1975] studied "morphine and phenytoin binding to plasma proteins in renal and hepatic failure." Twenty-eight subjects with uremia were classified into four groups. The percentage of morphine that was bound is the endpoint.

$$\begin{aligned}
\text{Chronic } (n_1 = 18):&\ 31.5, 35.1, 32.1, 34.2, 26.7, 31.9, 30.8,\\
&\ 27.3, 27.3, 29.0, 30.0, 36.4, 39.8, 32.0, 35.9, 29.9, 32.2, 31.8\\
\text{Acute } (n_2 = 2):&\ 31.6, 28.5\\
\text{Dialysis } (n_3 = 3):&\ 29.3, 32.1, 26.9\\
\text{Anephric } (n_4 = 5):&\ 26.5, 22.7, 27.5, 24.9, 23.4
\end{aligned}$$

(a) Do tasks (a) to (e) and (g) to (i).

(b) In view of the nature of the response variable (percent of morphine bound), explain why, strictly speaking, the assumption of homogeneity of variance cannot hold.

10.2 Graham [1977] assayed 16 commercially available pancreatic extracts for six types of enzyme activity. See also Example 10.6. Data for one of these enzymes, proteolytic activity, are presented here. The 16 products were classified by packaging form: capsule, tablet, and enteric-coated tablets. The following data were obtained:

	Proteolytic Activity (U/unit)						
Tablet ($n = 5$)	6640	4440	240	990	410		
Capsule ($n = 4$)	6090	5840	110	195			
Coated tablet ($n = 7$)	1800	1420	980	1088	2200	870	690

(a) Do tasks (a) to (e) and (g) to (i).

(b) Is there a transformation that would make the variance more homogeneous? Why is this unlikely to be the case? What is peculiar about the values for the coated tablets?

10.3 The following data from Rifkind et al. [1976] consist of antipyrine clearance of males suffering from β-thalassemia, a chronic type of anemia. In this disease, abnormally thin red blood cells are produced. The treatment of the disease has undesirable side effects, including liver damage. Antipyrine is a drug used to assess liver function with a high clearance rate, indicating satisfactory liver function. These data deal with the antipyrine clearance rate of 10 male subjects classified according to pubertal stage.

The question is whether there is any significant difference in clearance rate among the pubertal stages (I = infant; V = adult).

Pubertal Stage	Clearance Rate (Half-Life in Hours)				
I	7.4	5.6	3.7	6.6	6.0
IV	10.9	12.2			
V	11.3	10.0	13.3		

(a) Do tasks (a) to (e) and (g) to (i).

***(b)** Assuming that the antipyrine clearance rate increases with age, carry out a nonparametric test for trend (see Note 10.2). What is the alternative hypothesis in this case?

10.4 It is known that organisms react to stress. A more recent discovery is that the immune system's response is altered as a function of stress. In a paper by Keller et al. [1981], the immune response of rats as measured by the number of circulating lymphocytes (cells per milliliter $\times 10^{-6}$) was related to the level of stress. The following data are taken from this paper:

Group	Number of Rats	Mean Number of Lymphocytes	SE
Home-cage control	12	6.64	0.80
Apparatus control	12	4.84	0.70
Low shock	12	3.98	1.13
High shock	12	2.92	0.42

(a) Do tasks (a), (b), (e), and (j).

(b) The authors state: "a significant lymphocytopenia [$F(3, 44) = 3.86$, $p < 0.02$] was induced by the stressful conditions." Does your F-ratio agree with theirs?

(c) Sharpen the analysis by considering a trend in the response levels as a function of increasing stress level.

10.5 This problem deals with the data in Table 10.8. The authors of the paper state that the animals were matched on the basis of weight but that there was no correlation with weight. Assume that the data are presented in the order in which the animals were matched, that is, $Y_{111} = 143$ is matched with $Y_{211} = 152$; in general, Y_{1jk} is matched with Y_{2jk}.

(a) Construct a table of differences $D_{jk} = Y_{2jk} - Y_{1jk}$.

(b) Carry out a one-way ANOVA on the differences; include SS_μ in your table.

(c) Interpret SS_μ for these data.

(d) State your conclusions and compare them with the conclusions of Example 10.5.

(e) Relate the MS(between groups) in the one-way ANOVA to one of the MS terms in Table 10.14. Can you identify the connection and the reason for it?

***(f)** We want to correlate the Y_{1jk} observations with the Y_{2jk} observations, but the problem is that the response level changes from day to day, which would induce a correlation. So we will use the following "trick." Calculate $Y^*_{ijk} = Y_{ijk} - \overline{Y}_{ij\cdot}$;

and correlate Y^*_{1jk} with Y^*_{2jk}. Test this correlation using a t-test with $16 - 1 = 15$ degrees of freedom. Why $16 - 1$? There are $7 - 1 = 6$ independent pairs for day 10, 5 each for day 12, and day 14, for a total of 16. Since the observations sum to zero already, we subtract one more degree of freedom for estimating the correlation. If matching was not effective, this correlation should be zero.

10.6 Ross and Bras [1975] investigated the relationship between diet and length of life in 121 randomly bred rats. After 21 days of age, each rat was given a choice of several diets *ad libitum* for the rest of its life. The daily food intake (g/day) was categorized into one of six intervals, so that an equal number of rats (except for the last interval) appeared in each interval. The response variable was life span in days. The following data were obtained:

Mean food intake (g/day)	18.3	19.8	20.7	21.6	22.4	24.1
Food intake category	1	2	3	4	5	6
Number of rats	20	20	20	20	20	21
Mean life span (days)	733	653	630	612	600	556
Standard error	117	126	111	115	113	106

(a) Carry out tasks (a), (b), (e), and (j).

(b) Can this be thought of as a regression problem? How would the residual MS from regression be related to the MS error of the analysis of variance?

***(c)** Can you relate in detail the ANOVA procedure and the regression analysis; particularly an assessment of a nonlinear trend?

10.7 The following data from Florey et al. [1977] are the fasting serum insulin levels for adult Jamaican females after an overnight fast:

	Fasting Serum Insulin Level (μU/mL)			
Age	25–34	35–44	45–54	55–64
Number	73	97	74	53
Mean	22.9	26.2	22.0	23.8
SD	10.3	13.0	7.4	10.0

(a) Do tasks (a), (b), (e), and (j).

(b) Why did the authors partition the ages of the subjects into intervals? Are there other ways of summarizing and analyzing the data? What advantages or disadvantages are there to your alternatives?

10.8 The assay of insulin was one of the earliest topics in bioassay. A variety of methods have been developed over the years. In the mid-1960s an assay was developed based on the fact that insulin stimulates glycogen synthesis in mouse diaphragm tissue, in vitro. A paper by Wardlaw and van Belle [1964] describes the statistical aspects of this assay. The data in this problem deal with a qualitative test for insulin activity. A pool of 36 hemidiaphragms was collected for each day's work and the tissues incubated in tubes containing medium with or without insulin. Each tube contained three randomly selected diaphragms. For parts of this problem we ignore tube effects and assume that each treatment was based on six hemidiaphragms. Four unknown samples were

Table 10.34 Glycogen Content Data

Medium Only		Standard Insulin (0.5 mU/mL)		Test Preparation							
				A		B		C		D	
280	290	460	465	470	480	430	300	510	505	310	290
240	275	400	460	440	390	385	505	610	570	350	330
225	350	470	470	425	445	380	485	520	570	250	300

Source: Data adapted from Wardlaw and van Belle [1964].

assayed. Since the diaphragms synthesize glycogen in medium, a control preparation of medium only was added as well as a standard insulin preparation. The glycogen content (optical density in anthrone TEST × 1000) data are given in Table 10.34.

(a) Carry out tasks (a) to (e) and (g) to (i). (To simplify the arithmetic if you are using a calculator, divide the observations by 100.)

(b) Each column in the data layout represents one tube in which the three hemidiaphragms were incubated so that the design of the experiment is actually hierarchical. To assess the effect of tubes, we partition the SS_ε (with 30 d.f.) into two parts: SS(between tubes within preparations) = $SS_{BT(WP)}$ with six degrees of freedom (why?) and SS(within tubes) = SS_{WT} with 24 degrees of freedom (why?). The latter SS can be calculated by considering each tube as a treatment. The former can then be calculated as $SS_{BT(WP)} = SS_\varepsilon - SS_{WT}$. Carry out this analysis and test the null hypothesis that the variability between tubes within preparations is the same as the within-tube variability.

10.9 Schizophrenia is one of the psychiatric illnesses that is thought to have a definite physiological basis. Lake et al. [1980] assayed the concentration of norepinephrine in the cerebrospinal fluid of patients (NE in CSF) with one of three types of schizophrenia and controls. They reported the following means and *standard errors*:

NE in CSF (pg/mL)	Control Group	Schizophrenic Group		
		Paranoid	Undifferentiated	Schizoaffective
N	29	14	10	11
Mean	91	144	101	122
Standard error	6	20	11	21

Carry out tasks (a), (b), (e), and (j).

10.10 Corvilain et al. [1971] studied the role of the kidney in the catabolism (conversion) of insulin by comparing the metabolic clearance rate in seven control subjects, eight patients with chronic renal failure, and seven anephric (without kidneys) patients. The data for this problem consist of the plasma insulin concentrations (ng/mL) at 45 and 90 min after the start of continuous infusion of labeled insulin. A low plasma concentration is associated with a high metabolic clearance rate, as shown in Table 10.35.

(a) Consider the plasma insulin concentration at 45 minutes. Carry out tasks (a) to (e) and (g) to (i).

Table 10.35 Plasma Concentration Data (ng/mL)

Control			Renal Failure			Anephric		
Patient	45	90	Patient	45	90	Patient	45	90
1	3.7	3.8	1	3.0	4.2	1	6.7	9.6
2	3.4	4.2	2	3.1	3.9	2	2.6	3.4
3	2.4	3.1	3	4.4	6.1	3	3.4	—[a]
4	3.3	4.4	4	5.1	7.0	4	4.0	5.1
5	2.4	2.9	5	1.9	3.5	5	3.1	4.2
6	4.8	5.4	6	3.4	5.7	6	2.7	3.8
7	3.2	4.1	7	2.9	4.3	7	5.3	6.6
			8	3.8	4.8			

[a] Missing observation.

(b) Consider the plasma insulin concentration at 90 minutes. Carry out tasks (a) to (e) and (g) to (i).

(c) Calculate the difference in concentrations between 90 and 45 minutes for each patient. Carry out tasks (a) to (e) and (g) to (i). Omit Patient 3 in the anephric group.

(d) Graph the means for the three groups at 45 and 90 minutes on the same graph. What is the overall conclusion that you draw from the three analyses? Were all three analyses necessary? Would two of three have sufficed? Why or why not?

10.11 We return to the data of Zelazo et al. [1972] one more time. Carry out the Terpstra–Jonckheere test for ordered alternatives as discussed in Note 10.2. Justify the use of an ordered alternative hypothesis. Discuss in terms of power the reason that this analysis does indicate a treatment effect, in contrast to previous analyses.

10.12 One of the problems in the study of SIDS is the lack of a good animal model. Baak and Huber [1974] studied the guinea pig as a possible model observing the effect of lethal histamine shock on the guinea pig thymus. The purpose was to determine if changes in the thymus of the guinea pig correspond to pathological changes observed in SIDS victims. In the experiment 40 animals (20 male, 20 female) were randomly assigned either to "control" or "histamine shock." On the basis of a Wilcoxon two-sample test—which ignored possible gender differences—the authors concluded that the variable medullary blood vessel surface (mm^2/mm^3) did not differ significantly between "control" and "histamine shock." The data below have been arranged to keep track of gender differences.

	Control					**Histamine Shock**				
Female	6.4	6.2	6.9	6.9	5.4	8.4	10.2	6.2	5.4	5.5
	7.5	6.1	7.3	5.9	6.8	7.3	5.2	5.1	5.7	9.8
Male	4.3	7.5	5.2	4.9	5.7	7.5	6.7	5.7	4.9	6.8
	4.3	6.4	6.2	5.0	5.0	6.6	6.9	11.8	6.7	9.0

(a) Do tasks (a) to (e), (g), and (i).

(b) Replace the observations by their ranks and repeat the analysis of variance. Compare your conclusions with those of part (a).

10.13 In tumor metastasis, tumor cells spread from the original site to other organs. Usually, a particular tumor will spread preferentially to specific organs. There are two possibilities as to how this may occur: The tumor cells gradually adapt to the organ to which they have spread, or tumor cells that grow well at this organ are selected preferentially. Nicolson and Custead [1982] studied this problem by comparing the metastatic potential of melanoma tumor cells mechanically lodged in the lungs of mice or injected intravenously and allowed to metastasize to the lung. Each of these cell lines was then harvested and injected subcutaneously. The numbers of pulmonary tumor colonies were recorded for each of three treatments: original line (control), mechanical placement (adaptation), and selection. The data in Table 10.36 were obtained in three experiments involving 84 mice.

Table 10.36 Experimental Data for Three Treatments

	Number of Pulmonary Tumor Colonies										
Experiment	Control				Adaption				Selection		
1	0	4	20	32	0	3	20	7	92	141	
		0	9	22		0	6	24	64	96	149
		1	11	31		2	14	29	79	100	151
2	0	8	31	41	0	10	13	0	101	132	
		3	8	32		0	11	14	52	109	136
		6	22	39		5	12	14	89	110	140
3	0	4	36	49	0	11	21	30	79	111	
		0	18	39		0	13	27	46	89	114
		2	29	42		3	13	28	51	100	114

(a) Carry out tasks (a) to (g). You may want to try several transformations: for example, $\sqrt{\ }$, $Y^{1/4}$. An appropriate transformation is logarithmic. To avoid problems with zero values, use $\log(Y + 1)$.

(b) How would you interpret a significant "experiment × treatment" interaction?

10.14 A paper by Gruber [1976] evaluated interactions between two analgesic agents: fenoprofen and propoxyphene. The design of the study was factorial with respect to drug combinations. Propoxyphene (P) was administered in doses of 0, 5, 100, and 150 mg.; fenoprofen (F) in doses of 0, 200, 400, and 600 mg. Each combination of the two factors was studied. In addition, postepisiotomy postpartum patients were categorized into one of four pain classes: "little," "some," "lot," and "terrible" pain; for each of the 16 medication combinations, 8, 10, 10, and 2 patients in the four pain classes were used. The layout of the number of patients could be constructed as shown in Table 10.37.

(a) One response variable was "analgesic score" for a medication combination. Table 10.38 is a partial ANOVA table for this variable. Fill in the lines in the table, completing the table.

(b) The total analgesic score for the 16 sets of 30 patients classified by the two drug levels is given in Table 10.39. Carry out a "randomized block analysis" on these total scores dividing the sums of squares by 30 to return the analysis to a single reading status. Link this analysis with the table in part (a). You have, in effect, partitioned the SS for medications in that table into three parts. Test the significance of the *three* mean squares.

(c) Graph the mean analgesia score (per patient) by plotting the dose on the x-axis for fenoprofen, indicating the levels of the propoxyphene dose in the graph. State your conclusions.

Table 10.37 Design of Medication Combinations

	Medication Combination						
Pain Level	$(0P, 0F)$	$(0P, 200F)$	$\cdots$	$(0P, 600F)$	$(50P, 0F)$	$\cdots$	$(150P, 600F)$
"Little"	8	8	$\cdots$	8	8	$\cdots$	8
"Some"	10	10	$\cdots$	10	10	$\cdots$	10
"Lot"	10	10	$\cdots$	10	10	$\cdots$	10
"Terrible"	2	2	$\cdots$	2	2	$\cdots$	2

Table 10.38 ANOVA Table for Analgesic Score

Source	d.f.	SS	MS	F-Ratio	P-Value
Pain class	—	3,704	—	—	—
Medications	—	9,076	—	—	—
Interaction	—	3,408	—	—	—
Residual	—	—	—		
Total	479	41,910			

Table 10.39 Total Analgesia Score

	Fenoprofen Calcium Dose (mg)			
Propoxyphene Dose (mg)	0	200	400	600
0	409	673	634	756
50	383	605	654	785
100	496	773	760	755
150	496	723	773	755

10.15 Although the prescription, "Take two aspirins, drink lots of fluids, and go to bed," is usually good advice, it is known that aspirin induces "microbleeding" in the gastrointestinal system, as evidenced by minute amounts of blood in the stool. Hence, there is constant research to develop other anti-inflammatory and antipyretic (fever-combating) agents. Arsenault et al. [1976] reported on a new agent, R-803, studying its effect in a Latin square design, comparing it to placebo and aspirin (900 mg, q.i.d). For purposes of this exercise the data are extracted in the form of a randomized block design. Each subject received each of three treatments for a week. We will assume that the order was random. The variable measured is the amount of blood lost in mL/day as measured over a week.

	Mean Blood Loss (ml/day)								
Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Placebo	0.45	0.54	0.69	0.53	3.03	0.78	0.14	0.82	0.96
R-803	0.82	0.39	0.67	1.19	1.18	1.07	0.49	0.14	0.80
Aspirin	18.00	6.46	6.19	6.52	7.18	9.39	6.93	1.57	4.03

(a) Do tasks (a) to (e) and (g) to (i).

(b) Carry out the Tukey test for additivity. What are your conclusions?

Table 10.40 COHb Data for Problem 10.16

	No. Hours Since Beginning of Exposure				
Subject	0	2	4	6	8
1	4.4	4.9	5.2	5.7	5.7
2	3.3	5.3	6.9	7.0	8.8
3	5.0	6.4	7.2	7.7	9.3
4	5.3	5.3	7.4	7.0	8.3
5	4.1	6.8	9.6	11.5	12.0
6	5.0	6.0	6.8	8.3	8.1
7	4.6	5.2	6.6	7.4	7.1

10.16 Occupational exposures to toxic substances are being investigated more and more carefully. Ratney et al. [1974] studied the effect of daily exposure of 180 to 200 ppm of methylene chloride on carboxyhemoglobin (COHb) measured during the workday. The COHb data (% COHb) for seven subjects measured five times during the day is given in Table 10.40.

- **(a)** Carry out tasks (a), (c) to (e), and (g) to (i).
- **(b)** Suppose that the observation for subject 3 at time 6 ($Y_{34} = 7.7$) is missing. Estimate its value and redo the ANOVA.
- **(c)** Carry out the Tukey test for additivity.
- **(d)** Carry out the Page test for trend (see Note 10.2).
- **(e)** Why do the data not form a randomized block design?
- **(f)** Could this problem be treated by regression methods, where X = hours since exposure and Y = % COHb? Why or why not?
- **(g)** Calculate all 10 pairwise correlations between the treatment combinations. Do they look "reasonably close"?

10.17 Wang et al. [1976] studied the effects on sleep of four hypnotic agents and a placebo. The preparations were: lorazepam 2 and 4 mg, and flurazepam 15 and 30 mg. Each of 15 subjects received all five treatments in a random order in five-night units. The analysis of variance of length of sleep is presented here.

Source	**d.f.**	**SS**	**MS**	***F*-Ratio**	***p*-Value**
Treatments	—	—	12.0	—	—
Patients	—	—	14.8	—	—
Residual	—	—	2.2		
Total	74	—			

- **(a)** Do task (a).
- **(b)** Fill in the missing values in the ANOVA table.
- **(c)** State your conclusions.
- **(d)** The article does not present any raw data or means. How satisfactory is this in terms of clinical significance?

10.18 High blood pressure is a known risk factor for cardiovascular disease, and many drugs are now on the market that provide symptomatic as well as therapeutic relief. One of

Table 10.41 Blood Pressure Data (mmHg) for Problem 10.18

	Recumbent		Upright	
Patient	Placebo	Propranolol	Placebo	Propranolol
N.F.	96	71	73	87
A.C.	96	85	104	76
P.D.	92	89	83	90
J.L.	97	110	101	85
G.P.	104	85	112	94
A.H.	100	73	101	93
C.L.	93	81	88	85

these drugs is propranolol. Hamet et al. [1973] investigated the effect of propranolol in labile hypertension. Among the variables studied was mean blood pressure measured in mmHg (diastolic $+1/3$ pulse pressure). A placebo treatment was included in a double-blind fashion. The blood pressure was measured in the recumbent and upright positions. The blood pressure data is given in Table 10.41.

(a) Assuming that the treatments are just four treatments, carry out tasks (a) to (e) and (g) to (i) (i.e., assume a randomized block design).

(b) The sum of squares for treatments (3 d.f.) can be additively partitioned into three parts: SS_{DRUG}, $\text{SS}_{\text{POSITION}}$, and $\text{SS}_{\text{DRUG}\times\text{POSITION}}$, each with one degree of freedom. To do this, construct an "interaction table" of treatment totals.

$$\text{SS}_{\text{DRUGS}} = \frac{1340^2}{14} + \frac{1204^2}{14} - \frac{2544^2}{28} = 660.57$$

$$\text{SS}_{\text{POSITION}} = \frac{1272^2}{14} + \frac{1272^2}{14} - \frac{2544^2}{28} = 0[sic]$$

$$\begin{aligned}\text{SS}_{\text{DRUGS}\times\text{POSITION}} &= \frac{678^2}{7} + \frac{594^2}{4} + \frac{662^2}{7} + \frac{610^2}{7} - \frac{2544^2}{28} \\ &\quad - \text{SS}_{\text{DRUGS}} - \text{SS}_{\text{POSITION}} = 36.57\end{aligned}$$

Expand the ANOVA table to include these terms. (The $\text{SS}_{\text{POSITION}} = 0$ is most unusual; the raw data are as reported in the table.)

(c) This analysis could have been carried out as a series of three paired t-tests as follows: for each subject, calculate the following three quantities "$++--$," "$+-+-$," and "$+--+$." For example, for subject N.F. "$++--$" $= 96 + 71 - 73 - 87 = 7$, "$+-+-$" $= 96 - 71 + 73 - 87 = 11$, and "$+--+$" $= 96 - 71 - 73 + 87 = 39$. These quantities represent effects of position, drug treatment, and interaction, respectively, and are called *contrasts* (see Chapter 12 for more details). Each contrast can be tested against zero by means of a one-sample t-test. Carry out these t-tests. Compare the variances for each contrast; one assumption in the analysis of variance is that these contrast variances all estimate the same variance. How is the sum of the contrast variances related to the SS_ε in the ANOVA?

(d) Let d_1 be the sum of the observations associated with the pattern $++--$, d_2 the sum of the observations associated with the pattern $+-+-$, and d_3 the sum of the observations associated with the pattern $+ - - +$. How is $(d_1^2 + d_2^2 + d_3^2)$ related to $\text{SS}_{\text{TREATMENT}}$?

10.19 Consider the data in Example 10.5. Rank all 38 observations from lowest to highest and carry out the usual analysis of variance on these ranks. Compare your p-values with the p-values of Table 10.14. In view of Note 10.3, does this analysis give you some concern?

10.20 Consider the data of Table 10.16 dealing with the effectiveness of pancreatic supplements on fat absorption. Rank all of the observations from 1 to 24 (i.e., ignoring both treatment and block categories).

(a) Carry out an analysis of variance on the ranks obtained above.

(b) Compare your analysis with the analysis using the Friedman statistic. What is a potential drawback in the analysis of part (a)?

(c) Return to the Friedman ranks in Section 10.3.3 and carry out an analysis of variance on them. How is the Friedman statistic related to SS_τ of the ANOVA of the Friedman ranks?

10.21 These data are from the same source as those in Problem 10.3. We add data for females to generate the two-way layout shown in Table 10.42.

Table 10.42 Two-Way Layout for Problem 10.21

	Antipyrine Clearance (Half-Life in Hours)					
	Stage I		Stage IV		Stage V	
Males	7.4	5.6	3.7	10.9	11.3	13.3
	6.6	6.0		12.2	10.0	
Females	9.1	6.3	7.1	11.0	8.3	
	11.3	9.4	7.9		4.3	

(a) Do tasks (a) to (d).

(b) Graph the data. Is there any suggestion of interaction? Of main effects?

(c) Carry out a weighted means analysis.

(d) Partition each observation into its component parts and verify that the sums of squares are *not* additive.

10.22 Fuertes-de la Haba et al. [1976] measured intelligence in offspring of oral and nonoral contraceptive users in Puerto Rico. In the early 1960s, subjects were randomly assigned to oral conceptive use or other methods of birth control. Subsequently, mothers with voluntary pregnancies were identified and offspring between ages 5 and 7 were administered a Spanish–Puerto Rican version of the Wechsler Intelligence Scale for Children (WISC). Table 10.43 lists the data for boys only, taken from the article.

(a) Carry out tasks (a), (b), and (e).

(b) Do an unweighted means analysis. Interpret your findings.

(c) The age categories have obviously been "collapsed." What effect could such a collapsing have on the analysis? (Why introduce age as a variable since IQ is standardized for age?)

(d) Suppose that we carried out a contingency table analysis on the cell frequencies. What could such an analysis show?

Table 10.43 Data for Problem 10.22

	Age Groups (Years)		
	5	6	7–8
Oral contraceptive WISC score			
n	9	18	14
Mean	81.44	88.50	76.00
SD	9.42	11.63	9.29
Other birth control WISC score			
n	11	28	21
Mean	82.91	87.75	83.24
SD	10.11	10.85	9.60

Table 10.44 Data for Problem 10.23

	Gender	
	Boys	Girls
Oral contraceptive WISC score		
n	41	55
Mean	82.68	86.87
SD	11.78	14.66
Other birth control WISC score		
n	60	54
Mean	85.28	85.83
SD	10.55	12.22

10.23 The data in Table 10.44 are also from the article by Fuertes-de la Haba [1976] but have been "collapsed over age" and are presented by treatment (type of contraceptive) by gender. The response variable is, again, Wechsler IQ score.

(a) Carry out tasks (a), (b), and (e).

(b) Do an unweighted means analysis.

(c) Compare your conclusions with those of Problem 10.22.

10.24 This problem considers some implications of the additive model for the two-way ANOVA as defined by equation (18) and illustrated in Example 10.4.

(a) Graph the means of Example 10.4 by using the level of the second variable for the abscissa. Interpret the difference in the patterns.

(b) How many degrees of freedom are left for the means assuming that the model defined by equation (18) holds?

(c) We now want to define a nonadditive model retaining the values of the α's, β's, and μ, equivalently, retaining the same marginal and overall means. You are free to vary any of the cell means subject to the constraints above. Verify that you can manipulate only four cell means. After changing the cell means, calculate for each cell ij the quantity $Y_{ij} = \mu - \alpha_i - \beta_j$. What are some characteristics of these quantities?

(d) Graph the means derived in part (c) and compare the pattern obtained with that of Figure 10.2.

***10.25** This problem is designed to give you some experience with algebraic manipulation. It is not designed to teach you algebra but to provide additional insight into the mathematical structure of analysis of variance models. You will want to take this medicine in small doses.

(a) Show that equation (5) follows from the model defined by equation (4).

(b) Prove equations (6) and (7).

(c) Prove equations (10) to (12) starting with the components of equation (5).

(d) Consider equation (17). Let $\mu_i = \sum n_{ij}\mu_{ij}/n_{i\cdot}$, and so on. Relate α_i and β_j to $\mu_{i\cdot}$ and $\mu_{\cdot j}$.

(e) For the two-way ANOVA model as defined by equation (21), show that $\text{SS}_\varepsilon = \text{SS}_{\text{ERROR}} = \sum(n_{ij} - 1)s_{ij}^2$, where s_{ij}^2 is the variance of the observations in cell (i, j).

(f) Derive the expected mean squares for MS_α and MS_γ in the fixed and random effects models, as given in Table 10.19.

REFERENCES

Akritas, M. G. [1990]. The rank transform in some two-factor designs. *Journal of the American Statistical Association*, **85**: 73–78.

Arsenault, A., Le Bel, E., and Lussier, E. [1976]. Gastrointestinal microbleeding in normal subjects receiving acetylsalicylic acid, placebo and R-803, a new antiinflammatory agent, in a design balanced for residual effects. *Journal of Clinical Pharmacology*, **16**: 473–480. Used with permission from J.B. Lippincott Company.

Baak, J. P. A., and Huber, J. [1974]. Effects of lethal histamine shock in the guinea pig thymus. In *SIDS 1974, Proceedings of the Francis E. Camps International Symposium of Sudden and Unexpected Deaths in Infancy*, R. R. Robertson (ed.). Canadian Foundation for the Study of Infant Death, Toronto, Ontario, Canada.

Barboriak, J. J., Rimm, A., Tristani, F. E., Walker, J. R., and Lepley, D., Jr. [1972]. Risk factors in patients undergoing aorta-coronary bypass surgery. *Journal of Thoracic and Cardiovascular Surgery*, **64**: 92–97.

Beyer, W. H. (ed.) [1968]. *CRC Handbook of Tables for Probability and Statistics*. CRC Press, Cleveland, OH.

Box, G. E. P. [1953]. Non-normality and tests on variances. *Biometrika*, **40**: 318–335. Used with permission of the Biometrika Trustees.

Chikos, P. M., Figley, M. M., and Fisher, L. D. [1977]. Visual assessment of total heart volume and chamber size from standard chest radiographs. *American Journal of Roentgenology*, **128**: 375–380. Copyright © 1977 by the American Roentgenology Society.

Conover, W. J., and Iman, R. L. [1981]. Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician*, **35**: 124–133.

Corvilain, J., Brauman, H., Delcroix, C., Toussaint, C., Vereerstraeten, P., and Franckson, J. R. M. [1971]. Labeled insulin catabolism in chronic renal failure and the anephric state. *Diabetes*, **20**: 467–475.

Daniel, C. [1976]. *Applications of Statistics to Industrial Experiments*. Wiley, New York.

Daniel, C., and Wood, F. [1999]. *Fitting Equations to Data*, 2nd ed. Wiley, New York.

Draper, N. R., and Smith, H. [1998]. *Applied Regression Analysis*, 3rd ed. Wiley, New York.

Eisenhart, C. [1947]. The assumptions underlying the analysis of variance. *Biometrics*, **3**: 1–21.

Fisher, R. A. [1950]. *Statistical Methods for Research Workers*, 11th ed. Oliver & Boyd, London.

Florey, C. du V., Milner, R. D. G., and Miall, W. I. [1977]. Serum insulin and blood sugar levels in a rural population of Jamaican adults. *Journal of Chronic Diseases*, **30**: 49–60. Used with permission of Pergamon Press, Inc.

Freeman, M. F., and Tukey, J. W. [1950]. Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, **21**: 607–611.

Friedman, M. [1937]. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, **32**: 675–701.

Fuertes-de la Haba, A., Santiago, G., and Bangdiwala, I. S. [1976]. Measured intelligence in offspring of oral and non-oral contraceptive users. *American Journal of Obstetrics and Gynecology*, **7**: 980–982.

Graham, D. Y. [1977]. Enzyme replacement therapy of exocrine pancreatic insufficiency in man. *New England Journal of Medicine*, **296**: 1314–1317.

Gruber, C. M., Jr. [1976]. Evaluating interactions between fenoprofen and propoxyphene: analgesic and adverse reports by postepisiotomy patients. *Journal of Clinical Pharmacology*, **16**: 407–417. Used with permission from J.B. Lippincott Company.

Hamet, P., Kuchel, O., Cuche, J. L., Boucher, R., and Genest, J. [1973]. Effect of propranolol on cyclic AMP excretion and plasma renin activity in labile essential hypertension. *Canadian Medical Association Journal*, **1**: 1099–1103.

Hettmansperger, T. P., and McKean, J. W. [1978]. Statistical inference based on ranks. *Psychometrika*, **43**: 69–79.

Hillel, A., and Patten, C. [1990]. Effects of age and gender on dominance for lateral abduction of the shoulder. Unpublished data; used by permission.

Hollander, M., and Wolfe, D. A. [1999]. *Nonparametrical Statistical Methods*, 2nd ed. Wiley, New York.

Joiner, B. L., and Rosenblatt, J. R. [1971]. Some properties of the range in samples from Tukey's symmetric lambda distributions. *Journal of the American Statistical Association*, **66**: 394–399.

Keller, S. E., Weiss, J. W., Schleifer, S. J., Miller, N. E., and Stein, M. [1981]. Suppression of immunity by stress. *Science*, **213**: 1397–1400. Copyright © 1981 by the AAAS.

Kruskal, W. H., and Wallis, W. A. [1952]. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47**: 583–621.

Lake, C. R., Sternberg, D. E., van Kammen, D. P., Ballenger, J. C., Ziegler, M. G., Post, R. M., Kopin, I. J., and Bunney, W. E. [1980]. Schizophrenia: elevated cerebrospinal fluid norepinephrine. *Science*, **207**: 331–333. Copyright © 1980 by the AAAS.

Looney, S. W., and Stanley, W. B. [1989]. Exploratory repeated measures analysis for two or more groups. *American Statistician*, **43**: 220–225.

Nicolson, G. L., and Custead, S. E. [1982]. Tumor metastasis is not due to adaptation of cells to a new organ environment. *Science*, **215**: 176–178. Copyright © 1982 by the AAAS.

Odeh, R. E., Owen, D. B., Birnbaum, Z. W., and Fisher, L. D. [1977]. *Pocket Book of Statistical Tables*. Marcel Dekker, New York.

Olsen, G. D., Bennett, W. M., and Porter, G. A. [1975]. Morphine and phenytoin binding to plasma proteins in renal and hepatic failure. *Clinical Pharmaceuticals and Therapeutics*, **17**: 677–681.

Page, E. B. [1963]. Ordered hypotheses for multiple treatments: a significance test for linear ranks. *Journal of the American Statistical Association*, **58**: 216–230.

Quesenberry, P. D., Whitaker, T. B., and Dickens, J. W. [1976]. On testing normality using several samples: an analysis of peanut aflatoxin data. *Biometrics*, **32**: 753–759.

Ratney, R. S., Wegman, D. H., and Elkins, H. B. [1974]. In vivo conversion of methylene chloride to carbon monoxide. *Archives of Environmental Health*, **28**: 223–226. Reprinted with permission of the Helen Dwight Reid Educational Foundation. Published by Heldref Publications, 4000 Albemarle Street, N.W., Washington DC 20016. Copyright © 1974.

Rifkind, A. B., Canale, V., and New, M. I. [1976]. Antipyrine clearance in homozygous beta-thalassemia. *Clinical Pharmaceuticals and Therapeutics*, **20**: 476–483.

Ross, M. H., and Bras, G. [1975]. Food preference and length of life. *Science*, **190**: 165–167. Copyright © 1975 by the AAAS.

Ryan, T. A., Jr., Joiner, B. L., and Ryan, B. F. [1980]. *Minitab Reference Manual*, Release 1/10/80. Statistics Department, Pennsylvania State University, University Park, PA.

Scheffé, H. [1959]. *The Analysis of Variance*. Wiley, New York.

Sherwin, R. P., and Layfield, L. J. [1976]. Protein leakage in lungs of mice exposed to 0.5 ppm nitrogen dioxide: a fluorescence assay for protein. *Archives of Environmental Health*, **31**: 116–118.

Snedecor, G. W., and Cochran, W. G. [1988]. *Statistical Methods*, 8th ed. Iowa State University Press, Ames, IA.

Tukey, J. W. [1949]. One degree of freedom for additivity. *Biometrics*, **5**: 232–242.

Wallace S. S, Fisher, L. D., and Tremann, J. A. [1977]. Unpublished manuscript.

Wang, R. I. H., Stockdale, S. L., and Hieb, E. [1976]. Hypnotic efficacy of lorazepam and flurazepam. *Clinical Pharmaceuticals and Therapeutics*, **19**: 191–195.

Wardlaw, A. C., and van Belle, G. [1964]. Statistical aspects of the mouse diaphragm test for insulin. *Diabetes*, **13**: 622–634.

Weisberg, S. [1985]. *Applied Linear Regression*, 2nd ed. Wiley, New York.

Weisberg, S., and Bingham, C. [1975]. Approximate analysis of variance test for non-normality suitable for machine calculation. *Technometrics*, **17**: 133–134.

Winer, B. J. [1991]. *Statistical Principles in Experimental Design*, 3rd ed. McGraw-Hill, New York.

Zelazo, P. R., Zelazo, N. A., and Kalb, S. [1972]. “Walking” in the newborn. *Science*, **176**: 314–315.

CHAPTER 11

Association and Prediction: Multiple Regression Analysis and Linear Models with Multiple Predictor Variables

11.1 INTRODUCTION

We looked at the linear relationship between two variables, say X and Y, in Chapter 9. We learned to estimate the regression line of Y on X and to test the significance of the relationship. Summarized by the correlation coefficient, the square of the correlation coefficient is the percent of the variability explained.

Often, we want to predict or explain the behavior of one variable in terms of more than one variable, say k variables $X_1, \ldots, X_k$. In this chapter we look at situations where Y may be explained by a linear relationship with the explanatory or predictor variables $X_1, \ldots, X_k$. This chapter is a generalization of Chapter 9, where only one explanatory variable was considered. Some additional considerations will arise. With more than one potential predictor variable, it will often be desirable to find a simple model that explains the relationship. Thus we consider how to select a subset of predictor variables from a large number of potential predictor variables to find a reasonable predictive equation. Multiple regression analyses, as the methods of this chapter are called, are one of the most widely used tools in statistics. If the appropriate limitations are kept in mind, they can be useful in understanding complex relationships. Because of the difficulty of calculating the estimates involved, most computations of multiple regression analyses are performed by computer. For this reason, this chapter includes examples of output from multiple regression computer runs.

11.2 MULTIPLE REGRESSION MODEL

In this section we present the multiple regression mathematical model. We discuss the methods of estimation and the assumptions that are needed for statistical inference. The procedures are illustrated with two examples.

11.2.1 Linear Model

Definition 11.1. A *linear equation* for the variable Y in terms of $X_1, \ldots, X_k$, is an equation of the form

$$Y = a + b_1 X_1 + \cdots + b_k X_k \tag{1}$$

The values of $a, b_1, \ldots, b_k$, are fixed constant values. These values are called *coefficients*.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

Suppose that we observe Y and want to model its behavior in terms of independent, predictor, explanatory, or covariate variables, $X_1, \ldots, X_k$. For a particular set of values of the covariates, the Y value will not be known with certainty. As before, we model the expected value of Y for given or known values of the X_j. Throughout this chapter, we consider the behavior of Y for fixed, known, or observed values for the X_j. We have a multiple linear regression model if the expected value of Y for the known $X_1, \ldots, X_k$ is linear. Stated more precisely:

Definition 11.2. Y has a linear regression on $X_1, \ldots, X_k$ if the expected value of Y for the known X_j values is linear in the X_j values. That is,

$$E(Y|X_1, \ldots, X_k) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k \tag{2}$$

Another way of stating this is the following. Y is equal to a linear function of the X_j, plus an error term whose expectation is zero:

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \tag{3}$$

where

$$E(\varepsilon) = 0$$

We use the Greek letters α and β_j for the population parameter values and Latin letters a and b_j for the estimates to be described below. Analogous to definitions in Chapter 9, the number α is called the *intercept* of the equation and is equal to the expected value of Y when all the X_j values are zero. The β_j coefficients are the regression coefficients.

11.2.2 Least Squares Fit

In Chapter 9 we fitted the regression line by choosing the estimates a and b to minimize the sum of squares of the differences between the Y values observed and those predicted or modeled. These differences were called *residuals*; another way of explaining the estimates is to say that the coefficients were chosen to minimize the sum of squares of the residual values. We use this same approach, for the same reasons, to estimate the regression coefficients in the multiple regression problem. Because we have more than one predictor or covariate variable and multiple observations, the notation becomes slightly more complex. Suppose that there are n observations; we denote the observed values of Y for the ith observation by Y_i and the observed value of the jth variable X_j by X_{ij}. For example, for two predictor variables we can lay out the data in the array shown in Table 11.1.

Table 11.1 Data Layout for Two Predictor Variables

Case	Y	X_1	X_2
1	Y_1	X_{11}	X_{12}
2	Y_2	X_{21}	X_{22}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	Y_i	X_{i1}	X_{i2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	Y_n	X_{n1}	X_{n2}

The following definition extends the definition of least squares estimation to the multiple regression situation.

Definition 11.3. Given data $(Y_i, X_{i1}, \ldots, X_{ik})$, $i = 1, \ldots, n$, the *least squares fit* of the regression equation chooses $a, b_1, \ldots, b_k$ to minimize

$$\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$$

where $\widehat{Y_i} = a + b_1X_{i1} + \cdots + b_kX_{ik}$. The b_j are the *(sample) regression coefficients*, a is the *sample intercept*. The difference $Y_i - \widehat{Y_i}$ is the ith *residual*.

The actual fitting is usually done by computer, since the solution by hand can be quite tedious. Some details of the solution are presented in Note 11.1.

Example 11.1. We consider a paper by Cullen and van Belle [1975] dealing with the effect of the amount of anesthetic agent administered during an operation. The work also examines the degree of trauma on the immune system, as measured by the decreasing ability of lymphocytes to transform in the presence of mitogen (a substance that enhances cell division). The variables measured (among others) were X_1, the duration of anesthesia (in hours); X_2, the trauma factor (see Table 11.2 for classification); and Y, the percentage depression of lymphocyte transformation following anesthesia. It is assumed that the amount of anesthetic agent administered is directly proportional to the duration of anesthesia. The question of the influence of each of the two predictor variables is the crucial one, which will not be answered in this section. Here we consider the combined effect. The set of 35 patients considered for this example consisted of those receiving general anesthesia. The basic data are reproduced in Table 11.3. The predicted values and deviations are calculated from the least squares regression equation, which was $Y = -2.55 + 1.10X_1 + 10.38X_2$.

11.2.3 Assumptions for Statistical Inference

Recall that in the simple linear regression models of Chapter 9, we needed assumptions about the distribution of the error terms before we proceeded to statistical inference, that is, before we tested hypotheses about the regression coefficient using the F-test from the analysis of variance table. More specifically, we assumed:

Simple Linear Regression Model Observe (X_i, Y_i), $i = 1, \ldots, n$. The model is

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{4}$$

Table 11.2 Classification of Surgical Trauma

0	Diagnostic or therapeutic regional anesthesia; examination under general anesthesia
1	Joint manipulation; minor orthopedic procedures; cystoscopy; dilatation and curettage
2	Extremity, genitourinary, rectal, and eye procedures; hernia repair; laparoscopy
3	Laparotomy; craniotomy; laminectomy; peripheral vascular surgery
4	Pelvic extenteration; jejunal interposition; total cystectomy

Table 11.3 Effect of Duration of Anesthesia (X_1) and Degree of Trauma (X_2) on Percentage Depression of Lymphocyte Transformation following Anesthesia (Y)

Patient	X_1: Duration	X_2: Trauma	Y: Percent Depression	Predicted Value of Y	$Y - \widehat{Y}$ Residual
1	4.0	3	36.7	33.0	3.7
2	6.0	3	51.3	35.2	16.1
3	1.5	2	40.8	19.9	20.9
4	4.0	2	58.3	22.6	35.7
5	2.5	2	42.2	21.0	21.2
6	3.0	2	34.6	21.5	13.1
7	3.0	2	77.8	21.5	56.3
8	2.5	2	17.2	21.0	−3.8
9	3.0	3	−38.4	31.9	−70.3
10	3.0	3	1.0	31.9	−30.9
11	2.0	3	53.7	20.8	22.9
12	8.0	3	14.3	37.4	−23.1
13	5.0	4	65.0	44.5	20.5
14	2.0	2	5.6	20.4	−14.8
15	2.5	2	4.4	21.0	−16.6
16	2.0	2	1.6	20.4	−18.8
17	1.5	2	6.2	19.9	−13.7
18	1.0	1	12.2	8.9	3.3
19	3.0	3	29.9	31.9	−2.0
20	4.0	3	76.1	33.0	43.1
21	3.0	3	11.5	32.0	−20.5
22	3.0	3	19.8	31.9	−12.1
23	7.0	4	64.9	46.7	18.2
24	6.0	4	47.8	45.6	2.2
25	2.0	2	35.0	20.4	14.6
26	4.0	2	1.7	22.6	−20.9
27	2.0	2	51.5	20.4	31.1
28	1.0	1	20.2	8.9	11.3
29	1.0	1	−9.3	8.9	−18.2
30	2.0	1	13.9	10.0	3.9
31	1.0	1	−19.0	8.9	−27.9
32	3.0	1	−2.3	11.1	−13.4
33	4.0	3	41.6	33.0	8.6
34	8.0	4	18.4	47.8	−29.4
35	2.0	2	9.9	20.4	−10.5
Total	112.5	83	896.1	896.3	−0.2[a]
Mean	3.21	2.37	25.60	25.60	−0.006

[a] Zero except for round-off error.

or

$$Y_i = E(Y_i|X_i) + \varepsilon_i$$

where the "error" terms ε_i are statistically independent of each other and all have the same normal distribution with mean zero and variance σ^2; that is, $\varepsilon_i \sim N(0, \sigma^2)$.

Using this model, it is possible to set up the analysis of variance table associated with the regression line. The ANOVA table has the following form:

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	*F*-Ratio
Regression	1	$\text{SS}_{\text{REG}} = \sum_i (\widehat{Y}_i - \overline{Y})^2$	$\text{MS}_{\text{REG}} = \text{SS}_{\text{REG}}$	$\dfrac{\text{MS}_{\text{REG}}}{\text{MS}_{\text{RESID}}}$
Residual	$n - 2$	$\text{SS}_{\text{RESID}} = \sum_i (Y_i - \widehat{Y}_i)^2$	$\text{MS}_{\text{RESID}} = \dfrac{\text{SS}_{\text{RESID}}}{n-2}$	
Total	$n - 1$	$\sum_i (Y_i - \overline{Y}_i)^2$		

The mean square for residual is an estimate of the variance σ^2 about the regression line. (In this chapter we change notation slightly from that used in Chapter 9. The quantity σ^2 used here is the variance about the regression line. This was σ_1^2 in Chapter 9.)

The F-ratio is an F-statistic having numerator and denominator degrees of freedom of 1 and $n - 2$, respectively. We may test the hypothesis that the variable X has linear predictive power for Y, that is, $\beta \neq 0$, by using tables of critical values for the F-statistic with 1 and $n - 2$ degrees of freedom. Further, using the estimate of the variance about the regression line MS_{RESID}, it was possible to set up confidence intervals for the regression coefficient β.

For multiple regression equations of the current chapter, the same assumptions needed in the simple linear regression analyses carry over in a very direct fashion. More specifically, our assumptions for the multiple regression model are the following.

Multiple Regression Model Observe $(Y_i, X_{i1}, \ldots, X_{ik})$, $i = 1, 2, \ldots, n$ (n observations). The distribution of Y_i for fixed or known values of $X_{i1}, \ldots, X_{ik}$ is

$$Y_i = E(Y_i | X_{i1}, \ldots, X_{ik}) + \varepsilon_i \tag{5}$$

where $E(Y_i|X_{i1}, \ldots, X_{ik}) = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$ or $Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$. The ε_i are statistically independent and all have the same normal distribution with mean zero and variance σ^2; that is, $\varepsilon_i \sim N(0, \sigma^2)$.

With these assumptions, we use a computer program to find the least squares estimate of the regression coefficients. From these estimates we have the predicted value for Y_i given the values of $X_{i1}, \ldots, X_{ik}$. That is,

$$\widehat{Y}_i = a + b_1 X_{i1} + \cdots + b_k X_{ik} \tag{6}$$

Using these values, the ANOVA table for the one-dimensional case generalizes. The ANOVA table in the multidimensional case is now the following:

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	*F*-Ratio
Regression	k	$\text{SS}_{\text{REG}} = \sum_i (\widehat{Y}_i - \overline{Y})^2$	$\text{MS}_{\text{REG}} = \dfrac{\text{SS}_{\text{REG}}}{k}$	$\dfrac{\text{MS}_{\text{REG}}}{\text{MS}_{\text{RESID}}}$
Residual	$n - k - 1$	$\text{SS}_{\text{RESID}} = \sum_i (Y_i - \widehat{Y}_i)^2$	$\text{MS}_{\text{RESID}} = \dfrac{\text{SS}_{\text{RESID}}}{n-k-1}$	
Total	$n - 1$	$\sum_i (Y_i - \overline{Y}_i)^2$		

For the ANOVA table and multiple regression model, note the following:

1. If $k = 1$, there is one X variable; the equations and ANOVA table reduce to that of the simple linear regression case.

2. The F-statistic tests the hypothesis that the regression line has no predictive power. That is, it tests the hypothesis

$$H_0{:}\beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{7}$$

This hypothesis says that all of the beta coefficients are zero; that is, the X variables do not help to predict Y. The alternative hypothesis is that one or more of the regression coefficients $\beta_1, \ldots, \beta_k$ are nonzero. Under the null hypothesis, H_0, the F-statistic, has an F-distribution with k and $n-k-1$ degrees of freedom. Under the alternative hypotheses that one or more of the β_j are nonzero, the F-statistic tends to be too large. Thus the hypothesis that the regression line has predictive power is tested by using tables of the F-distribution and rejection when F is too large.

3. The residual sum of squares is an estimate of the variability about the regression line; that is, it is an estimate of σ^2. Introducing notation similar to that of Chapter 9, we write

$$\hat{\sigma}^2 = S^2_{Y\cdot X_1,\ldots,X_k} = \text{MS}_{\text{RESID}} = \frac{\sum_i (Y_i - \widehat{Y}_i)^2}{n-k-1} \tag{8}$$

4. Using the estimated value of σ^2, it is possible to find estimated standard errors for the b_j, the estimates of the regression coefficients β_j. The estimated standard error is associated with the t distribution with $n-k-1$ degrees of freedom. The test of $\beta_j = 0$ and an appropriate $100(1-\alpha)\%$ confidence interval are given by the following equations. To test H_j: $\beta_j = 0$ at significance level α, use two-sided critical values for the t-distribution with $n-k-1$ degrees of freedom and the test statistic

$$t = \frac{b_j}{\text{SE}(b_j)} \tag{9}$$

where b_j and $\text{SE}(b_j)$ are taken from computer output. Reject H_j if

$$|t| \geq t_{n-k-1,1-\alpha/2}$$

A $100(1-\alpha)\%$ confidence interval for β_j is given by

$$b_j \pm \text{SE}(b_j) t_{n-k-1,1-\alpha/2} \tag{10}$$

These two facts follow from the pivotal variable

$$t = \frac{b_j - \beta_j}{\text{SE}(b_j)}$$

which has a t-distribution with $n-k-1$ degrees of freedom.

5. Interpretations of the estimated coefficients in a multiple regression equation must be done cautiously. Recall (from the simple linear regression chapter) that we used the example of height and weight; we noted that if we managed to get the subjects to eat and/or diet to change their weight, this would not have any substantial effect on a person's height despite a relationship between height and weight in the population. Similarly, when we look at the estimated multiple regression equation, we can say that for the observed X values, the regression coefficients β_j have the following interpretation. If all of the X variables except for one, say X_j, are kept fixed, and if X_j changes by one unit, the expected value of Y changes by β_j. Let us consider this statement again for emphasis. *If all the X variables except for one X variable, X_j, are held constant, and the observation has X_j changed by an amount 1, the expected value of Y_i changes by the amount β_j.* This is seen by looking at the difference in the expected values:

$$\alpha + \beta_1 X_1 + \cdots + \beta_j (X_j + 1) + \cdots + \beta_k X_k - (\alpha + \cdots + \beta_j X_j + \cdots + \beta_k X_k) = \beta_j$$

This does not mean that when the regression equation is estimated, by changing X by a certain amount we can therefore change the expected value of Y. Consider a medical example where X_j might be systolic blood pressure and other X variables are other measures of physiological performance. Any maneuvers taken to change X_j might also result in changing some or all of the other X's in the population. The change in Y of β_j holds for the distribution of X's in the population sampled. By changing the values of X_j we might change the overall relationship between the Y_i's and the X_j's, so that the estimated regression equation no longer holds. (Recall again the height and weight example for simple linear regression.) For these reasons, interpretations of multiple regression equations must be made tentatively, especially when the data result from observational studies rather than controlled experiments.

6. If two variables, say X_1 and X_2, are closely related, it is difficult to estimate their regression coefficients because they tend to get confused. Take the extreme case where the variables X_1 and X_2 are actually the same value. Then if we look at $\beta_1 X_1 + \beta_2 X_2$ we can factor out the X_1 variable that is equal to X_2. That is, if $X_1 = X_2$, then $\beta_1 X_1 + \beta_2 X_2 = (\beta_1 + \beta_2) X_1$. We see that β_1 and β_2 are not determined uniquely in this case, but any values for β_1 and β_2 whose sum is the same will give the "same" regression equation. More generally, if X_1 and X_2 are very closely associated in a linear fashion (i.e., if their correlation is large), it is very difficult to estimate the betas. This difficulty is referred to as *collinearity*. We return to this fact in more depth below.

7. In Chapter 9 we saw that the assumptions of the simple linear regression model held if the two variables X and Y have a bivariate normal distribution. This fact may be extended to the considerations of this chapter. If the variables $Y, X_1, \ldots, X_k$ have a multivariate normal distribution, then conditionally upon knowing the values of $X_1, \ldots, X_k$, the assumptions of the multiple regression model hold. Note 11.2 has more detail on the multivariate normal distribution. We shall not go into this in detail but merely mention that if the variables have a multivariate normal distribution, any one of the variables has a normal distribution, any two of the variables have a bivariate normal distribution, and any linear combination of the variables also has a normal distribution.

These generalizations of the findings for simple linear regression are illustrated in the next section, which presents several examples of multiple regression.

11.2.4 Examples of Multiple Regression

Example 11.1. (*continued*) We modeled the percent depression of lymphocyte transformation following anesthesia by using the duration of the anesthesia in hours and trauma factor. The least squares estimates of the regression coefficients, the estimated standard errors and the ANOVA table are given below.

Constant or Variable j	b_j	SE(b_j)
Duration of anesthesia	1.105	3.620
Trauma factor	10.376	7.460
Constant	−2.555	12.395

Source	**d.f.**	**SS**	**MS**	**F-Ratio**
Regression	2	4,192.94	2,096.47	3.18
Residual	32	21,070.09	658.44	
Total	34	25,263.03		

From tables of the F-distribution, we see that at the 5% significance level the critical value for 2 and 30 degrees of freedom is 3.32, while for 2 and 40 degrees of freedom it is 3.23. Thus,

$F_{2,32,0.95}$ is between 3.23 and 3.32. Since the observed F-ratio is 3.18, which is smaller at the 5% significance level, we would not reject a null hypothesis that the regression equation has no contribution to the prediction. (Why is the double negative appropriate here?) This being the case, it would not pay to proceed further to examine the significance of the individual regression coefficients. (You will note that a standard error for the constant term in the regression is also given. This is also a feature of the computer output for most multiple regression packages.)

Example 11.2. This is a continuation of Example 9.1 regarding malignant melanoma of the skin in white males. We saw that mortality was related to latitude by a simple linear regression equation and also to contiguity to an ocean. We now consider the modeling of the mortality result using a multiple regression equation with both the "latitude" variable and the "contiguity to an ocean" variable. When this is done, the following estimates result:

Constant or Variable	b_j	SE(b_j)
Latitude in degrees	−5.449	0.551
Contiguity to ocean (1 = contiguous to ocean, 0 = does not border ocean)	18.681	5.079
Constant	360.28	22.572

Source	**d.f.**	**SS**	**MS**	***F*-Ratio**
Regression	2	40,366.82	20,183.41	69.96
Residual	46	13,270.45	288.49	
Total	48	53,637.27		

The F critical values at the 0.05 level with 2 and 40 and 2 and 60 degrees of freedom are 3.23 and 3.15, respectively. Thus the F-statistic for the regression is very highly statistically significant. This being the case, we might then wonder whether or not the significance came from one variable or whether both of the variables contributed to the statistical significance. We first test the significance of the latitude variable at the 5% significance level and also construct a 95% confidence interval. $t = -5.449/0.551 = -9.89$, $|t| > t_{48,0.975} \doteq 2.01$; reject $\beta_1 = 0$ at the 5% significance level. The 95% confidence interval is given by $-5.449 \pm 2.01 \times 0.551$ or $(-6.56, -4.34)$.

Consider a test of the significance of β_2 at the 1% significance level and a 99% confidence interval for β_2. $t = 18.681/5.079 = 3.68$, $|t| > t_{48,0.995} \doteq 2.68$; reject $\beta_2 = 0$ at the 1% significance level. The 99% confidence interval is given by $18.681 \pm 2.68 \times 5.079$ or $(5.07, 32.29)$.

In this example, from the t statistic we conclude that both latitude in degrees and contiguity to the ocean contribute to the statistically significant relationship between the melanoma of the skin mortality rates and the multiple regression equation.

Example 11.3. The data for this problem come from Problems 9.5 to 9.8. These data consider maximal exercise treadmill tests for 43 active women. We consider two possible multiple regression equations from these data. Suppose that we want to predict or explain the variability in $VO_{2\ \text{MAX}}$ by using three variables: X_1, the duration of the treadmill test; X_2, the maximum heart rate attained during the test; and X_3, the height of the subject in centimeters. Data resulting from the least squares fit are:

Covariate or Constant	b_j	**SE**(b_j)	t($t_{39,0.975} \doteq$ **2.02**)
Duration (seconds)	0.0534	0.00762	7.01
Maximum heart rate (beats/min)	−0.0482	0.05046	−0.95
Height (cm)	0.0199	0.08359	0.24
Constant	6.954	13.810	

				F**-Ratio**
Source	**d.f.**	**SS**	**MS**	($F_{3,39,0.95} \doteq$ **2.85**)
Regression	3	644.61	214.87	21.82
Residual	39	384.06	9.85	
Total	42	1028.67		

Note that the overall F-test is highly significant, 21.82, compared to a 5% critical value for the F-distribution with 3 and 39 degrees of freedom of approximately 2.85. When we look at the t statistic for the three individual terms, we see that the t value for duration, 7.01, is much larger than the corresponding 0.05 critical value of 2.02. The other two variables have values for the t statistic with absolute value much less than 2.02. This raises the possibility that duration is the only variable of the three that contributes to the predictive equation. Perhaps we should consider a model where we predict the maximum oxygen consumption in terms of duration rather than using all three variables. In sections to follow, we consider the question of selecting a "best" predictive equation using a subset of a given set of potential explanatory or predictor variables.

Example 11.3. (*continued*) We use the same data but consider the dependent variable to be age. We shall try to model this from three explanatory, or independent, or predictor variables. Let X_1 be the duration of the treadmill test in seconds; let X_2 be $\text{VO}_{2\ \text{MAX}}$, the maximal oxygen consumption; and let X_3 be the maximum heart rate during the treadmill test. Analysis of these data lead to the following:

			t**-Statistic**
Covariate or Constant	b_j	**SE**(b_j)	($t_{39,0.975} \doteq$ **2.02**)
Duration	−0.0524	0.0268	−1.96
$\text{VO}_{2\ \text{MAX}}$	−0.633	0.378	−1.67
Maximum heart rate	−0.0884	0.119	−0.74
Constant	106.51	18.63	

				F**-Ratio**
Source	**d.f.**	**SS**	**MS**	($F_{3,39,0.95} \doteq$ **2.85**)
Regression	3	2256.97	752.32	13.70
Residual	39	2142.19	54.93	
Total	42	4399.16		

The overall F value of 13.7 is very highly statistically significant, indicating that if one has the results of the treadmill test, including duration, $\text{VO}_{2\ \text{MAX}}$, and maximum heart rate, one can gain a considerable amount of knowledge about the subject's age. Note, however, that when we look at the p-values for the individual variables, not one of them is statistically significant!

How can it be that the overall regression equation is very highly statistically significant but none of the variables individually can be shown to have contributed at the 5% significance level? This paradox results because the predictive variables are highly correlated among themselves; they are *collinear*, as mentioned above. For example, we already know from Chapter 9 that the duration and $\text{VO}_{2\ \text{MAX}}$ are highly correlated variables; there is much overlap in their predictive information. We have trouble showing that the prediction comes from one or the other of the two variables.

11.3 LINEAR ASSOCIATION: MULTIPLE AND PARTIAL CORRELATION

The simple linear regression equation was very closely associated with the correlation coefficient between the two variables; the square of the correlation coefficient was the proportion of the variability in one variable that could be explained by the other variable using a linear predictive equation. In this section we consider a generalization of the correlation coefficient.

11.3.1 Multiple Correlation Coefficient

In considering simple linear regression, we saw that r^2 was the proportion of the variability of the Y_i about the mean that could be explained from the regression equation. We generalize this to the case of multiple regression.

Definition 11.4. The *squared multiple correlation coefficient*, denoted by R^2, is the proportion of the variability in the dependent variable Y that may be accounted for by the multiple regression equation. Algebraically,

$$R^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Since

$$\sum_i (Y_i - \overline{Y})^2 = \sum_i (Y_i - \widehat{Y}_i)^2 + \sum_i (\widehat{Y}_i - \overline{Y}_i)^2$$

$$R^2 = \frac{\text{SS}_{\text{REG}}}{\text{SS}_{\text{TOTAL}}} = \frac{\sum_i (\widehat{Y}_i - \overline{Y})^2}{\sum_i (Y_i - \overline{Y})^2} \tag{11}$$

Definition 11.5. The positive square root of R^2 is denoted by R, the *multiple correlation coefficient*.

The multiple correlation coefficient may also be computed as the correlation between the Y_i and the estimated best linear predictor, $\widehat{Y}_i$. If the data come from a multivariate sample rather than having the X's fixed by experimental design, the quantity R is an estimate of the correlation between Y and the best linear predictor for Y in terms of $X_1, \ldots, X_k$, that is, the correlation between Y and $\mathrm{a} + \mathrm{b}_1 X_1 + \cdots + \mathrm{b}_k X_k$. The population correlation will be zero if and only if all the regression coefficients $\beta_1, \ldots, \beta_k$ are equal to zero. Again, the value of R^2 is an estimate (for a multivariate sample) of the square of the correlation between Y and the best linear predictor for Y in the overall population. Since the population value for R^2 will be zero if and only if the multiple regression coefficients are equal to zero, a test of the statistical significance of R^2 is the F-test for the regression equation. R^2 and F are related (as given by the definition of R^2 and the F test in the analysis of variance table). It is easy to show that

$$R^2 = \frac{kF}{kF + n - k - 1}, \qquad F = \frac{(n-k-1)R^2}{k(1-R^2)} \tag{12}$$

The multiple correlation coefficient thus has associated with it the same degrees of freedom as the F distribution: k and $n - k - 1$. Statistical significance testing for R^2 is based on the statistical significance test of the F-statistic of regression.

At significance level α, reject the null hypothesis of the no linear association between Y and $X_1, \ldots, X_k$ if

$$R^2 \geq \frac{kF_{k,n-k-1,1-\alpha}}{kF_{k,n-k-1,1-\alpha} + n - k - 1}$$

where $F_{k,n-k-1,1-\alpha}$ is the $1 - \alpha$ percentile for the F-distribution with k and $n - k - 1$ degrees of freedom.

For any of the examples considered above, it is easy to compute R^2. Consider the last part of Example 11.3, the active female exercise test data, where duration, VO_2 $_{\text{MAX}}$, and the maximal heart rate were used to "explain" the subject's age. The value for R^2 is given by $2256.97/4399.16 = 0.51$; that is, 51% of the variability in Y (age) is explained by the three explanatory or predictor variables. The multiple regression coefficient, or positive square root, is 0.72.

The multiple regression coefficient has the same limitations as the simple correlation coefficient. In particular, if the explanatory variables take values picked by an experimenter and the variability about the regression line is constant, the value of R^2 may be increased by taking a large spread among the explanatory variables $X_1, \ldots, X_k$. The value for R^2, or R, may be presented when the data do *not* come from a multivariate sample; in this case it is an indicator of the amount of the variability in the dependent variable explained by the covariates. *It is then necessary to remember that the values do not reflect something inherent in the relationship between the dependent and independent variables, but rather, reflect a quantity that is subject to change according to the value selection for the independent or explanatory variables.*

Example 11.4. Gardner [1973] considered using environmental factors to explain and predict mortality. He studied the relationship between a number of socioenvironmental factors and mortality in county boroughs of England and Wales. Rates for all sizable causes of death in the age bracket 45 to 74 were considered separately. Four social and environmental factors were used as independent variables in a multiple regression analysis of each death rate. The variables included social factor score, "domestic" air pollution, latitude, and the level of water calcium. He then examined the residuals from this regression model and considered relating the residual variability to other environmental factors. The only factors showing sizable and consistent correlation were the long-period average rainfall and latitude, with rainfall being the more significant variable for all causes of death. When rainfall was included as a fifth regressor variable, no new factors were seen to be important. Tables 11.4 and 11.5 give the regression coefficients, not for the raw variables but for standardized variables.

These data were developed for 61 English county boroughs and then used to predict the values for 12 other boroughs. In addition to taking the square of the multiple correlation coefficient for the data used for the prediction, the correlation between observed and predicted values for *the other 12 boroughs* were calculated. Table 11.5 gives the results of these data.

This example has several striking features. Note that Gardner tried to fit a variety of models. This is often done in multiple regression analysis, and we discuss it in more detail in Section 11.8. Also note the dramatic drop (!) in the amount of variability in the death rate that can be explained between the data used to fit the model and the data used to predict values for other boroughs. This may be due to several sources. First, the value of R^2 is always nonnegative and can only be zero if variability in Y can be perfectly predicted. In general, R^2 tends to be too large. There is a value called *adjusted* R^2, which we denote by R_a^2, which takes this effect into account.

Table 11.4 Multiple Regression[a] of Local Death Rates on Five Socioenvironmental Indices in the County Boroughs[b]

Gender/Age Group	Period	Social Factor Score	"Domestic" Air Pollution	Latitude	Water Calcium	Long Period Average Rainfall
Males/45–64	1948–1954	0.16	0.48***	0.10	−0.23	0.27***
	1958–1964	0.19*	0.36***	0.21**	−0.24**	0.30***
Males/65–74	1950–1954	0.24*	0.28*	0.02	−0.43***	0.17
	1958–1964	0.39**	0.17	0.13	−0.30**	0.21
Females/45–64	1948–1954	0.16	0.20	0.32**	−0.15	0.40***
	1958–1964	0.29*	0.12	0.19	−0.22*	0.39***
Females/65–74	1950–1954	0.39***	0.02	0.36***	−0.12	0.40***
	1958–1964	0.40***	−0.05	0.29***	−0.27**	0.29**

[a] A standardized partial regression coefficients given; that is, the variables are reduced to the same mean (0) and variance (1) to allow values for the five socioenvironmental indices in each cause of death to be compared. The higher of two coefficients is not necessarily the more significant statistically.

[b] $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table 11.5 Results of Using Estimated Multiple Regression Equations from 61 County Boroughs to Predict Death Rates in 12 Other County Boroughs

Gender/Age Group	Period	$\widehat{R}^2$	r_2^a
Males/45–64	1948–1954	0.80	0.12
	1958–1964	0.84	0.26
Males/65–74	1950–1954	0.73	0.09
	1958–1964	0.76	0.25
Females/45–64	1948–1954	0.73	0.46
	1958–1964	0.72	0.48
Females/65–74	1950–1954	0.80	0.53
	1958–1964	0.73	0.41

[a] r is the correlation coefficient in the second sample between the value predicted for the dependent variable and its observed value.

This estimate of the population, R^2, is given by

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \tag{13}$$

For the Gardner data on males from 45 to 64 during the time period 1948–1954, the adjusted R^2 value is given by

$$R_a^2 = 1 - (1 - 0.80)\left(\frac{61-1}{61-5}\right) = 0.786$$

We see that this does not account for much of the drop. Another possible effect may be related to the fact that Gardner tried a variety of models; in considering multiple models, one may get a very good fit just by chance because of the many possibilities tried. The most likely explanation, however, is that a model fitted in one environment and then used in another setting may lose much predictive power because *variables important to one setting may not be as important in another setting*. As another possibility, there could be an important variable that is not even known by the person analyzing the data. If this variable varies between the original data set and the new data set, where one desires to predict, extreme drops in predictive power may occur. As a general rule of thumb, *the more complex the model, the less transportable the model is in time and/or space*. This example illustrates that whenever possible, when fitting a multivariate model including multiple linear regression models, if the model is to be used for prediction it is useful to try the model on an independent sample. Great degradation in predictive power is not an unusual occurrence.

In one example above, we had the peculiar situation that the relationship between the dependent variable age and the independent variables duration, $\mathrm{VO}_{2\ \mathrm{MAX}}$, and maximal heart rate was such that there was a very highly statistically significant relationship between the regression equation and the dependent variable, but at the 5% significance level we were not able to demonstrate the statistical significance of the regression coefficients of any of the three independent variables. That is, we could not demonstrate that any of the three predictor variables actually added statistically significant information to the prediction. We mentioned that this may occur because of high correlations between variables. This implies that they contain much of the same predictive information. In this case, estimation of their individual contribution is very difficult. This idea may be expressed quantitatively by examining the variance of the estimate for a regression coefficient, say β_j. This variance can be shown to be

$$\mathrm{var}(b_j) = \frac{\sigma^2}{[x_j^2](1 - R_j^2)} \tag{14}$$

In this formula σ^2 is the variance about the regression line and $[x_j^2]$ is the sum of the squares of the difference between the values observed for the jth predictor variable and its mean (this bracket notation was used in Chapter 9). R_j^2 is the square of the multiple correlation coefficient between X_j as dependent variable and the other predictor variables as independent variables. Note that if there is only one predictor, R_j^2 is zero; in this case the formula reduces to the formula of Chapter 9 for simple linear regression. On the other hand, if X_j is very highly correlated with other predictor variables, we see that the variance of the estimate of b_j increases dramatically. This again illustrates the phenomenon of *collinearity*. A good discussion of the problem may be found in Mason [1975] as well as in Hocking [1976].

In certain circumstances, more than one multiple regression coefficient may be considered at one time. It is then necessary to have notation that explicitly gives the variables used.

Definition 11.6. The multiple correlation coefficient of Y with the set of variables $X_1, \ldots ,$ X_k is denoted by

$$R_{Y(X_1,\ldots,X_k)}$$

when it is necessary to explicitly show the variables used in the computation of the multiple correlation coefficient.

11.3.2 Partial Correlation Coefficient

When two variables are related linearly, we have used the correlation coefficient as a measure of the amount of association between the two variables. However, we might suspect that a relationship between two variables occurred because they are both related to another variable. For example, there may be a positive correlation between the density of hospital beds in a geographical area and an index of air pollution. We probably would not conjecture that the number of hospital beds increased the air pollution, although the opposite could conceivably be true. More likely, both are more immediately related to population density in the area; thus we might like to examine the relationship between the density of hospital beds and air pollution after controlling or adjusting for the population density. We have previously seen examples where we controlled or adjusted for a variable. As one example this was done in the combining of 2 × 2 tables, using the various strata as an adjustment. A partial correlation coefficient is designed to measure the amount of linear relationship between two variables after adjusting for or controlling for the effect of some set of variables. The method is appropriate when there are linear relationships between the variables and certain model assumptions such as normality hold.

Definition 11.7. The *partial correlation coefficient* of X and Y adjusting for the variables $X_1, \dots , X_k$ is denoted by $\rho_{X,Y.X_1,\dots,X_k}$. The sample partial correlation coefficient of X and Y adjusting for $X_1, \dots , X_k$ is denoted by $r_{X,Y.X_1,\dots,X_k}$. The partial correlation coefficient is the correlation of Y minus its best linear predictor in terms of the X_j variables with X minus its best linear predictor in terms of the X_j variables. That is, letting $\widehat{Y}$ be a predicted value of Y from multiple linear regression of Y on $X_1, \dots , X_k$ and letting $\widehat{X}$ be the predicted value of X from the multiple linear regression of X on $X_1, \dots , X_k$, the partial correlation coefficient is the correlation of $X - \widehat{X}$ and $Y - \widehat{Y}$.

If all of the variables concerned have a multivariate normal distribution, the partial correlation coefficient of X and Y adjusting for $X_1, \dots , X_k$ is the correlation of X and Y conditionally upon knowing the values of $X_1, \dots , X_k$. The conditional correlation of X and Y in this multivariate normal case is the same for each fixed set of the values for $X_1, \dots , X_k$ and is equal to the partial correlation coefficient.

The statistical significance of the partial correlation coefficient is equivalent to testing the statistical significance of the regression coefficient for X if a multiple regression is performed with Y as a dependent variable with $X, X_1, \dots , X_k$ as the independent or explanatory variables. In the next section on nested hypotheses, we consider such significance testing in more detail.

Partial regression coefficients are usually estimated by computer, but there is a simple formula for the case of three variables. Let us consider the partial correlation coefficient of X and Y adjusting for a variable Z. In terms of the correlation coefficients for the pairs of variables, the partial correlation coefficient in the population and its estimate from the sample are given by

$$\rho_{X,Y\cdot Z} = \frac{\rho_{X,Y} - \rho_{X,Z}\rho_{Y,Z}}{\sqrt{(1-\rho^2_{X,Z})(1-\rho^2_{Y,Z})}}$$

$$r_{X,Y.Z} = \frac{r_{X,Y} - r_{X,Z}r_{Y,Z}}{\sqrt{(1-r^2_{X,Z})(1-r^2_{Y,Z})}} \qquad (15)$$

We illustrate the effect of the partial correlation coefficient by the exercise data for active females discussed above. We know that age and duration are correlated. For the data above, the correlation coefficient is −0.68913. Let us consider how much of the linear relationship between age and duration is left if we adjust out the effect of the oxygen consumption, $\text{VO}_{2\ \text{MAX}}$, for the same data set. The correlation coefficients for the sample are as follows:

$$r_{\text{AGE, DURATION}} = -0.68913$$

$$r_{\text{AGE, VO}_2\ \text{MAX}} = -0.65099$$

$$r_{\text{DURATION, VO}_2\ \text{MAX}} = 0.78601$$

The partial correlation coefficient of age and duration adjusting $\text{VO}_{2\ \text{MAX}}$ using the equation above is estimated by

$$r_{\text{AGE,DURATION}\cdot\text{VO}_2\ \text{MAX}} = \frac{-0.68913 - [(-0.65099)(-0.78601)]}{\sqrt{[1-(-0.65099)^2][1-(0.78601)^2]}} = -0.37812$$

If we consider the corresponding multiple regression problem with a dependent variable of age and independent variables duration and $\text{VO}_2\ _{\text{MAX}}$, the t-statistic for duration is -2.58. The two-sided 0.05 critical value is 2.02, while the critical value at significance level 0.01 is 2.70. Thus, we see that the p-value for statistical significance of this partial correlation coefficient is between 0.05 and 0.01.

11.3.3 Partial Multiple Correlation Coefficient

Occasionally, one wants to examine the linear relationship, that is, the correlation between one variable, say Y, and a second group of variables, say $X_1, \dots, X_k$, while adjusting or controlling for a third set of variables, $Z_1, \dots, Z_p$. If it were not for the Z_j variables, we would simply use the multiple correlation coefficient to summarize the relationship between Y and the X variables. The approach taken is the same as for the partial correlation coefficient. First subtract out for each variable its best linear predictor in terms of the Z_j's. From the remaining residual values compute the multiple correlation between the Y residuals and the X residuals. More formally, we have the following definition.

Definition 11.8. For each variable let $\widehat{Y}$ or $\widehat{X}_j$ denote the least squares linear predictor for the variable in terms of the quantities $Z_1, \dots, Z_p$. The best linear predictor for a sample results from the multiple regression of the variable on the independent variables $Z_1, \dots, Z_p$. The *partial multiple correlation coefficient* between the variable Y and the variables $X_1, \dots, X_k$ adjusting for $Z_1, \dots, Z_p$ is the multiple correlation between the variable $Y - \widehat{Y}$ and the variables $X_1 - \widehat{X}_1, \dots, X_k - \widehat{X}_k$. The partial multiple correlation coefficient of Y and $X_1, \dots, X_k$ adjusting for $Z_1, \dots, Z_p$ is denoted by

$$R_{Y(X_1,\dots,X_k).Z_1,\dots,Z_p}$$

A significance test for the partial multiple correlation coefficient is discussed in Section 11.4. The coefficient is also called the *multiple partial correlation coefficient.*

11.4 NESTED HYPOTHESES

In the second part of Example 11.3, we saw a multiple regression equation where we could not show the statistical significance of individual regression coefficients. This raised the possibility of reducing the complexity of the regression equation by eliminating one or more variables from the predictive equation. When we consider such possibilities, we are considering what is called a *nested hypothesis*. In this section we discuss nested hypotheses in the multiple regression setting. First we define nested hypotheses; we then introduce notation for nested hypotheses in multiple regression. In addition to notation for the hypotheses, we need notation for the various sums of squares involved. This leads to appropriate F-statistics for testing nested hypotheses. After we understand nested hypotheses, we shall see how to construct F-tests for the partial correlation coefficient and the partial multiple correlation coefficient. Furthermore, the ideas of nested hypotheses are used below in stepwise regression.

Definition 11.9. One hypothesis, say hypothesis H_1, is *nested* within a second hypothesis, say hypothesis H_2, if whenever hypothesis H_1 is true, hypothesis H_2 is also true. That is to say, hypothesis H_1 is a special case of hypothesis H_2.

In our multiple regression situation most nested hypotheses will consist of specifying that some subset of the regression coefficients β_j have the value zero. For example, the larger first hypothesis might be H_2, as follows:

$$H_2\colon Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

The smaller (nested) hypothesis H_1 might specify that some subset of the β's, for example, the last $k-j$ betas corresponding to variables $X_{j+1}, \ldots, X_k$, are all zero. We denote this hypothesis by H_1.

$$H_1\colon Y = \alpha + \beta_1 X_1 + \cdots + \beta_j X_j + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

In other words, H_2 holds *and*

$$\beta_{j+1} = \beta_{j+2} = \cdots = \beta_k = 0$$

A more abbreviated method of stating the hypothesis is the following:

$$H_1\colon \beta_{j+1} = \beta_{j+2} = \cdots = \beta_k = 0 | \beta_1, \ldots, \beta_j$$

To test such nested hypotheses, it will be useful to have a notation for the regression sum of squares for any subset of independent variables in the regression equation. If variables $X_1, \ldots, X_j$ are used as explanatory or independent variables in a multiple regression equation for Y, we denote the regression sum of squares by

$$\text{SS}_{\text{REG}}(X_1, \ldots, X_j)$$

We denote the residual sum of squares (i.e., the total sum of squares of the dependent variable Y about its mean minus the regression sum of squares) by

$$\text{SS}_{\text{RESID}}(X_1, \ldots, X_j)$$

If we use more variables in a multiple regression equation, the sum of squares explained by the regression can only increase, since one potential predictive equation would set all the regression coefficients for the new variables equal to zero. This will almost never occur in practice if for no other reason than the random variability of the error term allows the fitting of extra regression coefficients to explain a little more of the variability. The increase in the regression sum of squares, however, may be due to chance. The F-test used to test nested hypotheses looks at the increase in the regression sum of squares and examines whether it is plausible that the increase could occur by chance. Thus we need a notation for the increase in the regression sum of squares. This notation follows:

$$\text{SS}_{\text{REG}}(X_{j+1}, \ldots, X_k | X_1, \ldots, X_j) = \text{SS}_{\text{REG}}(X_1, \ldots, X_k) - \text{SS}_{\text{REG}}(X_1, \ldots, X_j)$$

This is the sum of squares attributable to $X_{j+1}, \ldots, X_k$ after fitting the variables $X_1, \ldots, X_j$. With this notation we may proceed to the F-test of the hypothesis that adding the last $k-j$ variables does not increase the sum of squares a statistically significant amount beyond the regression sum of squares attributable to $X_1, \ldots, X_k$.

Assume a regression model with k predictor variables, $X_1, \ldots, X_k$. The F-statistic for testing the hypothesis

$$H_1\colon \beta_{j+1} = \cdots = \beta_k = 0 | \beta_1, \ldots, \beta_j$$

is

$$F = \frac{\mathrm{SS}_{\mathrm{REG}}(X_{j+1}, \ldots , X_k | X_1, \ldots , X_j)/(k - j)}{\mathrm{SS}_{\mathrm{RESID}}(X_1, \ldots , X_k)/(n - k - 1)}$$

Under H_1, F has an F-distribution with $k - j$ and $n - k - 1$ degrees of freedom. Reject H_1 if $F > F_{k-j,n-k-1,1-\alpha}$, the $1 - \alpha$ percentile of the F-distribution.

The partial correlation coefficient is related to the sums of squares as follows. Let X be a predictor variable in addition to $X_1, \ldots , X_k$.

$$r^2_{X,Y \cdot X_1,\ldots,X_k} = \frac{\mathrm{SS}_{\mathrm{REG}}(X | X_1, \ldots , X_k)}{\mathrm{SS}_{\mathrm{RESID}}(X_1, \ldots , X_k)} \tag{16}$$

The sign of $r_{X,Y \cdot X_1,\ldots,X_k}$ is the same as the sign of the X regression coefficient when Y is regressed on X, $Y \cdot X_1, \ldots , X_k$. The F-test for statistical significance of $r_{X,Y \cdot X_1,\ldots,X_k}$ uses

$$F = \frac{\mathrm{SS}_{\mathrm{REG}}(X | X_1, \ldots , X_k)}{\mathrm{SS}_{\mathrm{RESID}}(X, X_1, \ldots , X_k)/(n - k - 2)} \tag{17}$$

Under the null hypothesis that the partial correlation is zero (or equivalently, that $\beta_X = 0 | \beta_1, \ldots , \beta_k$), F has an F-distribution with 1 and $n - k - 2$ degrees of freedom. F is sometimes called the *partial F-statistic*. The t-statistic for the statistical significance of β_X is related to F by

$$t^2 = \frac{\beta_X^2}{\mathrm{SE}(\beta_X)^2} = F$$

Similar results hold for the partial multiple correlation coefficient. The correlation is always positive and its square is related to the sums of squares by

$$R^2_{Y(X_1,\ldots,X_k)\cdot Z_1,\ldots,Z_p} = \frac{\mathrm{SS}_{\mathrm{REG}}(X_1, \ldots , X_k | Z_1, \ldots , Z_p)}{\mathrm{SS}_{\mathrm{RESID}}(Z_1, \ldots , Z_p)} \tag{18}$$

The F-test for statistical significance uses the test statistic

$$F = \frac{\mathrm{SS}_{\mathrm{REG}}(X_1, \ldots , X_k | Z_1, \ldots , Z_p)/k}{\mathrm{SS}_{\mathrm{RESID}}(X_1, \ldots , X_k, Z_1, \ldots , Z_p)/(n - k - p - 1)} \tag{19}$$

Under the null hypothesis that the population partial multiple correlation coefficient is zero, F has an F-distribution with k and $n - k - p - 1$ degrees of freedom. This test is equivalent to testing the nested multiple regression hypothesis:

$$H\colon \beta_{X_1} = \cdots = \beta_{X_k} = 0 | \beta_{Z_1}, \ldots , \beta_{Z_p}$$

Note that in each case above, the contribution to R^2 after adjusting for additional variables is the increase in the regression sum of squares divided by the residual sum of squares after taking the regression on the adjusting variables. The corresponding F-statistic has a numerator degrees of freedom equal to the number of predictive variables added, or equivalently, the number of additional parameters being estimated. The denominator degrees of freedom are equal to the number of observations minus the total number of parameters estimated. The reason for the -1 in the denominator degrees of freedom in equation (19) is the estimate of the constant in the regression equation.

Example 11.3. (*continued*) We illustrate some of these ideas by returning to the 43 active females who were exercise-tested. Let us compute the following quantities:

$$r_{\text{VO}_2\ \text{MAX,DURATION}\,\cdot\,\text{AGE}}$$

$$R^2_{\text{AGE(VO}_2\ \text{MAX, HEART RATE)}\,\cdot\,\text{DURATION}}$$

To examine the relationship between VO2 $_{\text{MAX}}$ and duration adjusting for age, let duration be the dependent or response variable. Suppose that we then run two multiple regressions: one predicting duration using only age as the predictive variable and a second regression using both age and VO2 $_{\text{MAX}}$ as the predictive variable. These runs give the following data: for Y = duration and X_1 = age:

Covariate or Constant	b_j	SE(b_j)	***t*-statistic** ($t_{41,0.975} \doteq 2.02$)
Age	−5.208	0.855	−6.09
Constant	749.975	39.564	

Source	**d.f.**	**SS**	**MS**	***F*-Ratio** ($F_{1,41,0.95} \doteq 4.08$)
Regression of duration on age	1	119,324.47	119,324.47	37.08
Residual	41	131,935.95	3,217.95	
Total	42	251,260.42		

and for Y = duration, X_1 = age, and X_2 = VO2 $_{\text{MAX}}$:

Covariate or Constant	b_j	SE(b_j)	***t*-statistic** ($t_{40,0.975} \doteq 2.09$)
Age	−2.327	0.901	−2.583
VO2 $_{\text{MAX}}$	9.151	1.863	4.912
Constant	354.072	86.589	

Source	**d.f.**	**SS**	**MS**	***F*-Ratio** ($F_{2,40,0.95} \doteq 3.23$)
Regression of duration on age and VO2 $_{\text{MAX}}$	2	168,961.48	84,480.74	41.06
Residual	40	82,298.94	2,057.47	
Total	42	251,260.42		

Using equation (16), we find the square of the partial correlation coefficient:

$$\begin{aligned} r^2_{\text{VO}_2\ \text{MAX, DURATION}\cdot\text{AGE}} &= \frac{168{,}961.48 - 119{,}324.47}{131{,}935.95} \\ &= \frac{49{,}637.01}{131{,}935.95} \\ &= 0.376 \end{aligned}$$

Since the regression coefficient for VO2 MAX is positive (when regressed with age) having a value of 9.151, the positive square root gives r:

$$r_{\mathrm{VO_2\ MAX,\ DURATION\ \cdot\ AGE}} = +\sqrt{0.376} = 0.613$$

To test the statistical significance of the partial correlation coefficient, equation (17) gives

$$F = \frac{168{,}961.48 - 119{,}324.467}{82{,}298.94/(43 - 1 - 1 - 1)} = 24.125$$

Note that $t^2_{\mathrm{VO_2MAX}} = 24.127 = F$ within round-off error. As $F_{1,40,0.999} = 12.61$, this is highly significant ($p < 0.001$). In other words, the duration of the treadmill test and the maximum oxygen consumption are significantly related even after adjustment for the subject's age.

Now we turn to the computation and testing of the partial multiple correlation coefficient. To use equations (18) and (19), we need to regress age on duration, and also regress age on duration, VO2 MAX, and the maximum heart rate. The ANOVA tables follow. For age regressed upon duration:

Source	d.f.	SS	MS	F-Ratio ($F_{1,41,0.95} \doteq 4.08$)
Regression	1	2089.18	2089.18	37.08
Residual	41	2309.98	56.34	
Total	42	4399.16		

and for age regressed upon duration, VO2 MAX, and maximum heart rate:

Source	d.f.	SS	MS	F-Ratio ($F_{3,39,0.95} \doteq 2.85$)
Regression	3	2256.97	752.32	13.70
Residual	39	2142.19	54.93	
Total	42	4399.16		

From equation (18),

$$R^2_{\mathrm{AGE(VO_2\ MAX,\ HEART\ RATE)\ \cdot\ DURATION}} = \frac{2256.97 - 2089.18}{2309.98} = 0.0726$$

and $R = \sqrt{R^2} = 0.270$.

The F-test, by equation (19), is

$$F = \frac{(2256.97 - 2089.18)/2}{2142.19/(43 - 2 - 1 - 1)} = 1.53$$

As $F_{2,39,0.90} \doteq 2.44$, we have not shown statistical significance even at the 10% significance level. In words: VO2 MAX and maximum heart rate have no more additional linear relationship with age, after controlling for the duration, than would be expected by chance variability.

11.5 REGRESSION ADJUSTMENT

A common use of regression is to make inference regarding a specific predictor of inference from observational data. The primary explanatory variable can be a treatment, an environmental exposure, or any other type of measured covariate. In this section we focus on the common biomedical situation where the predictor of interest is a treatment or exposure, but the ideas naturally generalize to any other type of explanatory factor.

In observational studies there can be many uncontrolled and unmeasured factors that are associated with seeking or receiving treatment. A naive analysis that compares the mean response among treated individuals to the mean response among nontreated subjects may be distorted by an unequal distribution of additional key variables across the groups being compared. For example, subjects that are treated surgically may have poorer function or worse pain prior to their being identified as candidates for surgery. To evaluate the long-term effectiveness of surgery, each patient's functional disability one year after treatment can be measured. Simply comparing the mean function among surgical patients to the mean function among patients treated nonsurgically does not account for the fact that the surgical patients probably started at a more severe level of disability than the nonsurgical subjects. When important characteristics systematically differ between treated and untreated groups, crude comparisons tend to distort the isolated effect of treatment. For example, the average functional disability may be higher among surgically treated subjects compared to nonsurgically treated subjects, even though surgery has a beneficial effect for each person treated since only the most severe cases may be selected for surgery. Therefore, without adjusting for important predictors of the outcome that are also associated with being given the treatment, unfair or invalid treatment comparisons may result.

11.5.1 Causal Inference Concepts

Regression models are often used to obtain comparisons that "adjust" for the effects of other variables. In some cases the adjustment variables are used purely to improve the precision of estimates. This is the case when the adjustment covariates are not associated with the exposure of interest but are good predictors of the outcome. Perhaps more commonly, regression adjustment is used to alleviate bias due to confounding. In this section we review causal inference concepts that allow characterization of a well-defined estimate of treatment effect, and then discuss how regression can provide an adjusted estimate that more closely approximates the desired causal effect.

To discuss causal inference concepts, many authors have used the *potential outcomes framework* [Neyman, 1923; Rubin, 1974; Robins, 1986]. With any medical decision we can imagine the outcome that would result if each possible future path were taken. However, in any single study we can observe only one realization of an outcome per person at any given time. That is, we can only measure a person's response to a single observed and chosen history of treatments and exposures. We can still envision the hypothetical, or "potential" outcome that would have been observed had a different set of conditions occurred. An outcome that we believe could have happened but was not actually observed is called a *counterfactual outcome*. For simplicity we assume two possible exposure or treatment conditions. We define the *potential outcomes* as:

$Y_i(0)$: reponse for subject i at a specific measurement time after treatment $X = 0$ is experienced

$Y_i(1)$: reponse for subject i at a specific measurement time after treatment $X = 1$ is experienced

Given these potential outcomes, we can define the *causal effect* for subject i as

$$\text{causal effect for subject } i : \Delta_i = Y_i(1) - Y_i(0)$$

The causal effect Δ_i measures the difference in the outcome for subject i if they were given treatment $X = 1$ vs. the outcome if they were given treatment $X = 0$. For a given population of N subjects, we can define the *average causal effect* as

$$\overline{\Delta} = \frac{1}{N} \sum_{i=1}^{N} \Delta_i$$

The average causal effect is a useful overall summary of the treatment under study. Individual causal effects would be useful for selecting the best intervention for a given person. In general, we can only reliably estimate average causal effects for specific populations of subjects. Using covariates, we may try to narrow the population such that it closely approximates the particular persons identified for possible treatment.

There are a number of important implications associated with the potential outcomes framework:

1. In any given study we can only observe either $Y_i(0)$ or $Y_i(1)$ and not both. We are assuming that $Y_i(0)$ and $Y_i(1)$ represent outcomes under different treatment schemes, and in nature we can only realize one treatment and one subsequent outcome per subject.
2. Each subject is assumed to have an individual causal effect of treatment, Δ_i. Thus, there is no assumption of a single effect of treatment that is shared for all subjects.
3. Since we cannot observe $Y_i(0)$ and $Y_i(1)$, we cannot measure the individual treatment effect Δ_i.

Example 11.4. Table 11.6 gives a hypothetical example of potential outcomes. This example is constructed to approximate the evaluation of surgical and nonsurgical interventions for treatment of a herniated lumbar disk (see Keller et al. [1996] for an example). The outcome represents a measure of functional disability on a scale of 1 to 10, where the intervention has a beneficial effect by reducing functional disability. Here $Y_i(0)$ represents the postintervention outcome if subject i is given a conservative nonsurgical treatment and $Y_i(1)$ represents the postintervention outcome if subject i is treated surgically. Since only one course of treatment

Table 11.6 Hypothetical Example of Potential Outcomes and Individual Causal Effects

	Potential Outcome		Causal		Potential Outcome		Causal
Subject i	$Y_i(0)$	$Y_i(1)$	Effect Δ_i	Subject i	$Y_i(0)$	$Y_i(1)$	Effect Δ_i
1	4.5	2.7	−1.8	11	7.5	5.1	−2.3
2	3.1	1.0	−2.1	12	6.7	5.2	−1.5
3	3.9	2.0	−1.9	13	6.0	4.4	−1.6
4	4.3	2.2	−2.1	14	5.6	3.2	−2.4
5	3.3	1.5	−1.9	15	6.5	4.0	−2.4
6	3.3	0.8	−2.5	16	7.7	6.0	−1.8
7	4.0	1.5	−2.5	17	7.1	5.1	−2.1
8	4.9	3.2	−1.7	18	8.3	6.0	−2.3
9	3.8	2.0	−1.9	19	7.0	4.6	−2.4
10	3.6	2.0	−1.6	20	6.9	5.3	−1.5
				Mean	5.40	3.39	−2.01

is actually administered, these outcomes are conceptual and only one can actually be measured. The data are constructed such that the effect of surgical treatment is a reduction in the outcome. For example, the individual causal effects range from a −1.5- to a −2.5-point difference between the outcome if treated and the outcome if untreated. The average causal effect for this group is −2.01. To be interpreted properly, the population over which we are averaging needs to be detailed. For example, if these subjects represent veterans over 50 years of age, then −2.01 represents the average causal effect for this specific subpopulation. The value −2.01 may not generalize to represent the average causal effect for other populations (i.e., nonveterans, younger subjects).

Although we cannot measure individual causal effects, we can estimate average causal effects if the mechanism that assigns treatment status is essentially an unbiased random mechanism. For example, if $P[X_i = 1 \mid Y_i(0), Y_i(1)] = P(X_i = 1)$, the mean of a subset of observations, $Y_i(1)$, observed for those subjects with $X_i = 1$ will be an unbiased estimate of the mean for the entire population if all subjects are treated. Formally, the means observed for the treatment, $X = 1$, and control, $X = 0$, groups can be written as

$$\overline{Y}_1 = \frac{1}{n_1}\sum_{j=1}^{N} Y_j(1) \cdot 1(X_j = 1)$$

$$\overline{Y}_0 = \frac{1}{n_0}\sum_{j=1}^{N} Y_j(0) \cdot 1(X_j = 0)$$

where $n_1 = \sum_j 1(X_j = 1)$, $n_0 = \sum_j 1(X_j = 0)$, and $1(X_j = 0)$, $1(X_j = 1)$ are indicator functions denoting assignment to control and treatment, respectively. For example, if we assume that $P(X_i = 1) = 1/2$ and that $n_1 = n_0 = N/2$, then with random allocation to treatment,

$$\begin{aligned} E(\overline{Y}_1) &= \frac{1}{N/2}\sum_{j=1}^{N} Y_j(1) \cdot E[1(X_j = 1)] \\ &= \frac{1}{N/2}\sum_{j=1}^{N} Y_j(1) \cdot 1/2 \\ &= \frac{1}{N}\sum_{j} Y_j(1) \\ &= \mu_1 \end{aligned}$$

where we define μ_1 as the mean for the population if all subjects receive treatment. A similar argument shows that $E(\overline{Y}_0) = \mu_0$, the mean for the population if all subjects were not treated. Essentially, we are assuming the existence of parallel and identical populations, one of which is treated and one of which is untreated, and sample means from each population under simple random sampling are obtained.

Under random allocation of treatment and control status, the observed means $\overline{Y}_1$ and $\overline{Y}_0$ are unbiased estimates of population means. This implies that the sample means can be used to estimate the average causal effect of treatment:

$$\begin{aligned} E(\overline{Y}_1 - \overline{Y}_0) &= E(\overline{Y}_1) - E(\overline{Y}_1) \\ &= \mu_1 - \mu_0 \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_i Y_i(1) - \frac{1}{N}\sum_i Y_i(0) \\
&= \frac{1}{N}\sum_i [Y_i(1) - Y_i(0)] \\
&= \frac{1}{N}\sum_i \Delta_i \\
&= \overline{\Delta}
\end{aligned}
$$

Example 11.5. An example of the data observed from a hypothetical randomized study that compares surgical ($X = 1$) to nonsurgical ($X = 0$) interventions is presented in Table 11.7. Notice that for each subject, only one of $Y_i(0)$ or $Y_i(1)$ is observed, and therefore a treatment vs. control comparison can only be calculated using the group averages rather than using individual potential outcomes. Since the study was randomized, the difference in the averages observed is a valid (unbiased) estimate of the average causal effect of surgery. The mean difference observed in this experimental realization is -1.94, which approximates the unobservable target value of $\overline{\Delta} = -2.01$ shown in Table 11.6. In this example the key random variable is the treatment assignment, and because the study was randomized, the distribution for the treatment assignment indicator, $X_i = 0/1$, is completely known and independent of the potential outcomes.

Often, inference regarding the benefit of treatment is based on observational data where the assignment to $X = 0$ or $X = 1$ is not controlled by the investigator. Consequently, the factors

Table 11.7 Example of Data that would Be Observed in a Randomized Treatment Trial

Subject		Outcome Observed		
i	Assignment	$Y_i(0)$	$Y_i(1)$	Difference
1	0	4.5		
2	1		1.0	
3	1		2.0	
4	1		2.2	
5	0	3.3		
6	1		0.8	
7	1		1.5	
8	0	4.9		
9	0	3.8		
10	0	3.6		
11	1		5.1	
12	0	6.7		
13	0	6.0		
14	0	5.6		
15	0	6.5		
16	1		6.0	
17	1		5.1	
18	0	8.3		
19	1		4.6	
20	1		5.3	
Mean		5.48	3.42	−1.94

that drive treatment assignment need to be considered if causal inference is to be attempted. If sufficient covariate information is collected, regression methods can be used to control for confounding.

Definition 11.10. *Confounding* refers to the presence of an additional factor, Z, which when not accounted for leads to an association between treatment, X, and outcome, Y, that does not reflect a causal effect. Confounding is ultimately a "confusion" of the effects of X and Z. For a variable Z to be a confounder, it must be associated with X in the population, be a predictor of Y in the control ($X = 0$) group, and not be a consequence of either X or Y.

This definition indicates that confounding is a form of selection bias leading to biased estimates of the effect of treatment or exposure (see Rothman and Greenland [1998, Chap. 8] for a thorough discussion of confounding and for specific criteria for the identification of a confounding factor). Using the potential outcomes framework allows identification of the research goal: estimating the average causal effect, $\overline{\Delta}$. When confounding is present, the expected difference between $\overline{Y}_1$ and $\overline{Y}_0$ is no longer equal to the desired average causal effect, and additional analytical approaches are required to obtain approximate causal effects.

Example 11.6. Table 11.8 gives an example of observational data where subjects in stratum 2 are more likely to be treated surgically than subjects in stratum 1. The strata represent a baseline assessment of the severity of functional disability. In many settings those subjects with more severe disease or symptoms are treated with more aggressive interventions, such as surgery. Notice that both potential outcomes, $Y_i(0)$ and $Y_i(1)$, tend to be lower for subjects in stratum 1 than for subjects in stratum 2. Despite the fact that subjects in stratum 1 are much less likely to actually receive surgical intervention, treatment with surgery remains a beneficial intervention for both strata 1 and 2 subjects. The benefit of treatment for all subjects is apparent in the negative individual causal effects shown in Table 11.6. The imbalanced allocation of more severe cases to surgical treatment leads to crude summaries of $\overline{Y}_1 = 4.46$ and $\overline{Y}_0 = 4.32$. Thus the subjects who receive surgery have a slightly higher posttreatment mean functional score than those subjects who do not receive surgery. Does this comparison indicate the absence of a causal effect of surgery? The overall comparison is based on a treated group that has 80% of subjects drawn from stratum 2, the more severe group, while the control group has only 20% of subjects from stratum 2. The crude comparison of $\overline{Y}_1$ to $\overline{Y}_0$ is roughly a comparison of the posttreatment functional scores among severe subjects (80% of the $X = 1$ group) to the posttreatment functional scores among less severe subjects (80% of the $X = 0$ group). It is "unfair" to attribute the crude difference between treatment groups solely to the effect of surgery since the groups are clearly not comparable. A mixing of the effect of surgery with the effect of baseline severity is an illustration of bias due to confounding. The observed difference $\overline{Y}_1 - \overline{Y}_0 = 0.14$ is a distorted estimate of the average causal effect, $\overline{\Delta} = -2.01$.

11.5.2 Adjustment for Measured Confounders

There are several statistical methods that can be used to adjust for measured confounders. The goal of adjustment is to obtain an estimate of the treatment effect that more closely approximates the average causal effect. Commonly used methods include:

1. *Stratified methods*. In stratified methods the sample is broken into *strata*, $k = 1, 2, \ldots, K$, based on the value of a covariate, Z. Within each stratum, k, a treatment comparison can be calculated. Let $\delta^{(k)} = \overline{Y}_1^{(k)} - \overline{Y}_0^{(k)}$, where $\overline{Y}_1^{(k)}$ is the mean among treated subjects in strata k, and $\overline{Y}_0^{(k)}$ is the mean among control subjects in strata k. An overall summary of the stratum-specific treatment contrasts can be computed using a simple or weighted average of the stratum-specific comparisons, $\overline{\delta} = \sum_{k=1}^{K} w_k \cdot \delta^{(w_k)}$, where w_k is a weight. In the example presented in Table 11.8

Table 11.8 Example of an Observational Study Where Factors That Are Associated with the Potential Outcomes Are Predictive of the Treatment Assignment

Subject		Outcome Observed			
i	Assignment	$Y_i(0)$	$Y_i(1)$	Stratum	Difference
1	1		2.7	1	
2	0	3.1		1	
3	0	3.9		1	
4	1		2.2	1	
5	0	3.3		1	
6	0	3.3		1	
7	0	4.0		1	
8	0	4.9		1	
9	0	3.8		1	
10	0	3.6		1	
Mean		3.74	2.45		−1.29
11	1		5.1	2	
12	1		5.2	2	
13	1		4.4	2	
14	0	5.6		2	
15	1		4.0	2	
16	0	7.7		2	
17	1		5.1	2	
18	1		6.0	2	
19	1		4.6	2	
20	1		5.3	2	
Mean		6.65	4.96		−1.69
Overall mean		4.32	4.46		0.14

the subjects are separated into two strata, and mean differences of $\delta^{(1)} = -1.29$ and $\delta^{(2)} = -1.69$ are obtained comparing treatment and controls within strata 1 and strata 2, respectively. These estimates are much closer to the true average causal effect of $\overline{\Delta} = -2.01$ in Table 11.6 than the comparison of crude means, $\overline{Y}_1 - \overline{Y}_0 = 0.14$.

2. *Regression analysis*. Regression methods extend the concept of stratification to allow use with continuously measured adjustment variables and with multiple predictor variables. A regression model

$$E(Y \mid X, Z) = \alpha + \beta_1 X + \beta_2 Z$$

can be used to obtain an estimate of treatment, X, that adjusts for the covariate Z. Using the regression model, we have

$$\beta_1 = E(Y \mid X = 1, Z = z) - E(Y \mid X = 0, Z = z)$$

indicating that the parameter β_1 represents the average or common treatment comparison formed within groups determined by the value of the covariate, $Z = z$.

3. *Propensity score methods*. Propensity score methods are discussed by Rosenbaum and Rubin [1983]. In this approach the *propensity score*, $P(X = 1 \mid Z)$, is estimated using logistic regression or discriminant analysis, and then used either as a stratifying factor, a covariate in regression, or a matching factor (see Little and Rubin [2000] and the references therein for further detail on use of the propensity score for adjustment).

The key assumption that is required for causal inference is the "no unmeasured confounding" assumption. This states that for fixed values of a covariate, Z_i (this may be multiple covariates), the assignment to treatment, $X_i = 1$, or control, $X_i = 0$, is unrelated to the potential outcomes. This assumption can be stated as

$$P[X_i = 1 \mid Y_i(0), Y_i(1), Z_i] = P[X_i = 1 \mid Z_i]$$

One difficult aspect of this concept is the fact that we view potential outcomes as being measured after the treatment is given, so how can the potential outcomes predict treatment assignment? An association can be induced by another variable, such as Z_i. For example, in the surgical example presented in Table 11.8, an association between potential outcomes and treatment assignment is induced by the baseline severity. The probability that a subject is assigned $X_i = 1$ is predicted by baseline disease severity, and the potential outcomes are associated with the baseline status. Thus, if we ignore baseline severity, treatment assignment X_i is associated with both $Y_i(0)$ and $Y_i(1)$. The goal of collecting covariates Z_i is to measure sufficient predictors of treatment such that within the strata defined by Z_i, the treatment assignment is approximately randomized. A causal interpretation for effects formed using observational data requires the assumption that there is no unmeasured confounding within any strata. This assumption cannot be verified empirically.

Example 11.1. (*continued*) We return to the data from Cullen and van Belle [1975]. We use the response variable DMPA, the disintegrations per minute of lymphocytes measured after surgery. We focus on the effect of anesthesia used for the surgery: $X = 0$ for general anesthesia and $X = 1$ for local anesthesia. The following crude analysis uses a regression of DMPA on anesthesia (X), which is equivalent to the two-sample t-test:

	Coefficient	**SE**	***t***	***p*-Value**
Intercept	109.03	11.44	9.53	<0.001
Anesthesia	38.00	15.48	2.45	0.016

The analysis suggests that local anesthesia leads to a mean DMPA that is 38.00 units greater than the mean DMPA when general anesthesia is used. This difference is statistically significant with p-value 0.016.

Recall that these data are comprised of patients undergoing a variety of surgical procedures that are broadly classified using the variable TRAUMA, whose values 0 to 4 were introduced in Table 11.2. The type of anesthesia that is used varies by procedure type and therefore TRAUMA, as shown in Table 11.9. From this table we see that use of local anesthesia occurs more frequently for TRAUMA 0, 1, or 2, and that general anesthesia is used more frequently for TRAUMA 3 or 4. In addition, in earlier analyses we have found TRAUMA to be associated with the outcome. Thus, the crude analysis of anesthesia that estimates a 38.00 unit (S.E. = 15.48) effect of local anesthesia is confounded by TRAUMA and does not reflect an average causal effect. To adjust for TRAUMA, we use regression with the indicator variables, TRAUMA(j) = 1 if TRAUMA = j and 0 otherwise, for $j = 1, 2, 3, 4$. We use a model that includes an intercept and therefore do not also include an indicator for TRAUMA 0. The regression results are shown in Table 11.10.

After controlling for TRAUMA, the estimated comparison of local to general anesthesia within TRAUMA groups is 23.47 (S.E. = 18.24), and this difference is no longer statistically significant. This example shows that for causal analysis of observational data, any factors that are associated with treatment and associated with the outcome need to be considered in the analysis. In order to use 23.47 as the average causal effect of anesthesia, we would need to justify the required

Table 11.9 Anesthesia Use by Type of TRAUMA

	Anesthesia		
TRAUMA	0 = General	1 = Local	Total
0	0	11	11
1	6	12	18
2	14	16	30
3	11	3	14
4	4	0	4
Total	35	42	77

Table 11.10 Regression Results with Anesthesia and Trauma Predictors

	Coefficient	SE	t	p-Value
Intercept	129.53	27.40	4.73	<0.001
Anesthesia	23.47	18.24	1.29	0.202
TRAUMA 1	3.66	26.66	0.14	0.891
TRAUMA 2	−13.68	25.38	−0.54	0.592
TRAUMA 3	−25.34	30.86	−0.82	0.414
TRAUMA 4	−67.28	43.60	−1.54	0.127

assumption of no additional measured or unmeasured confounding factors. The assumption of no unmeasured confounding can only be supported by substantive considerations specific to the study design and the scientific process under investigation. Finally, since there are no empirical contrasts comparing local to general anesthesia within the TRAUMA 0 and TRAUMA 4 strata, we would need to either consider the average causal effect as only pertaining to the TRAUMA 1, 2, and 3 groups, or be willing to extrapolate to the TRAUMA 0 and 4 groups.

11.5.3 Model Selection Issues

One of the most difficult and controversial issues regarding the use of regression models is the procedure for specifying which variables are to be used to control for confounding. The epidemiological and biostatistical literature has introduced and evaluated several schemes for choosing adjustment variables. In the next section we discuss methods that can be used to identify a parsimonious explanatory or predictive model. However, the motivation for selecting covariates to control for confounding is different from the goal of identifying a good predictive model. To control for confounding, we identify adjustment variables in order to remove bias in the regression estimate for a predictor of primary interest, typically a treatment or exposure variable.

Pocock et al. [2002] discuss covariate choice issues in the analysis of data from clinical trials. The authors note that post hoc choice of covariates may not be done objectively and thus leads to estimates that reflect the investigators bias (e.g., choose to control for a variable if it makes the effect estimate larger!). In addition, simulation studies have shown that popular automatic variable-selection schemes can lead to biased estimates and distorted significance levels [Mickey and Greenland, 1989; Maldonado and Greenland, 1993; Sun et al., 1996; Hurvich and Tsai, 1990].

Kleinbaum [1994] discusses the a priori specification of the covariates to be used for regression analysis. The main message is that substantive considerations should drive the specification of the regression model when confirmatory estimation and inference are desired. This position is also supported by Raab et al. [2000].

11.5.4 Further Reading

Little and Rubin [2000] provide a comprehensive review of causal inference concepts. These authors also discuss the importance of the *stable unit treatment assumption* that is required for causal inference.

An overview of causal inference and discussion of the use of graphs for representing causal relationships are given in the text by Pearl [2000].

11.6 SELECTING A "BEST" SUBSET OF EXPLANATORY VARIABLES

11.6.1 The Problem

Given a large number of potential explanatory variables, one can sometimes select a smaller subset that explains the variability in the dependent variable. We have seen examples above where it appears that one or more of the variables in a multiple regression do not contribute, beyond an amount consistent with chance, to the explanation of the variability in the dependent variable. Thus, consider a response variable Y with a large number of potential predictor variables X_j. How should we choose a "best" subset of variables to explain the Y variability? This topic is addressed in this section. If we knew the number of predictor variables we wanted, we could use some criterion for the best subset. One natural criterion from the concepts already presented would be to choose the subset that gives the largest value for R^2. Even then, selection of the subset can be a formidable task. For example, suppose that there are 30 predictor variables and a subset of 10 variables is wanted; there are

$$\binom{30}{10} = 30{,}045{,}015$$

possible regression equations that have 10 predictor variables. This is not a routinely manageable number even with modern high-speed computers. Furthermore, in many instances we will not know how many possible variables we should place into our prediction equation. If we consider all possible subsets of 30 variables, there are over 1 billion possible combinations for the prediction. Thus once again, one cannot examine all subsets. There has been much theoretical work on selecting the best subset according to some criteria; the algorithms allow one to find the best subset without looking explicitly at all of the possible subsets. Still, for large numbers of variables, we need another procedure to select the predictive subset.

A further complication arises when we have a very large number of observations; then we may be able to show statistically that all of the potential predictor variables contribute additional information to explain the variability in the dependent variable Y. However, the large majority of the predictor variables may add so little to the explanation that we would prefer a much smaller subset that explains almost as much of the variability and gives a much simpler model. In general, simple models are desirable because they may be used more readily, and often when applied in a different setting, turn out to be more accurate than a model with a large number of variables.

In summary, the task before us in this section is to consider a means of choosing a subset of predictor variables from a pool of potential predictor variables.

11.6.2 Approaches to the Problem That Consider All Possible Subsets of Explanatory Variables

We discuss two approaches and then apply both approaches to an example. The first approach is based on the following idea: If we have the appropriate predictive variables in a multiple regression equation, plus possibly some other variables that have no predictive power, then the residual mean square for the model will estimate σ^2 the variability about the true regression line.

On the other hand, if we do not contain enough predictive variables, the residual mean square will contain additional variability due to the poor multiple regression fit and will tend to be too large. We want to use this fact to allow us to get some idea of the number of variables needed in the model. We do this in the following way. Suppose that we consider all possible predictions for some fixed number, say p, of the total possible number of predictor variables. Suppose that the correct predictive equation has a much smaller number of variables than p. Then when we look at all of the different subsets of p predictor variables, most of them will contain the *correct* variables for the predictive equation plus other variables that are not needed. In this case, the mean square residual will be an estimate of σ^2. If we average all of the mean square residuals for the equations with p variables, since most of them will contain the correct predictive variables, we should get an estimate fairly close to σ^2. We examine the mean square residuals by plotting the average mean square residuals for all the regression equations using p variables vs. p. As p becomes large, this average value should tend to level off at the true residual variability. By drawing a horizontal line at approximately the value where things average out, we can get some idea of the residual variability. We would then search for a simple model that has approximately this asymptotic estimate of σ^2. That is, we expect a picture such as Figure 11.1.

The second approach, due to C. L. Mallows, is called *Mallow's C_p statistic*. In this case, let p equal the number of predictive variables in the model, *plus one*. This is a change from the preceding paragraph, where p was the number of predictive variables. The switch to this notation is made because in the literature for Mallow's C_p, this is the value used. The statistic is as follows:

$$C_p(\text{model with } p-1 \text{ explanatory variables})$$
$$= \frac{\text{SS}_{\text{RESID}}(\text{model})}{\text{MS}_{\text{RESID}}(\text{using all possible predictors})} - (N - 2p)$$

where MS_{RESID} (using all possible predictors) is the residual mean square when the dependent variable Y is regressed on all possible independent predictors; SS_{RESID} (model) is the residual sum of squares for the possible model being considered (this model uses $p-1$ explanatory variables), N is the total number of observations, and p is the number of explanatory variables in the model plus one.

To use Mallow's C_p, we compute the value of C_p for each possible subset of explanatory variables. The points (C_p, p) are then plotted for each possible model. The following facts about the C_p statistics are true:

1. If the model fits, the expected value for each C_p is approximately p.
2. If C_p is larger than p, the difference, $C_p - p$, gives approximately the amount of bias in the sum of squares involved in the estimation. The bias occurs because the estimating

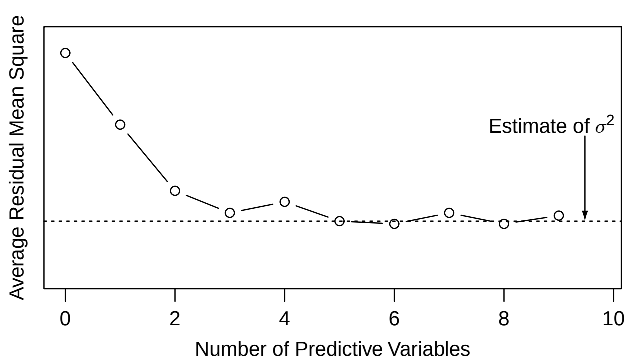

Figure 11.1 Average residual mean square as a function of the number of predictive variables.

predictive equation is not the true equation and thus estimates something other than the correct Y value.

3. The value of C_p itself gives an overall estimate of the sum of the squares of the average difference between correct Y values and the Y values predicted from the model. This difference is composed of two parts, one part due to bias because the estimating equation is not correct (and cannot be correct if the wrong variables are included), and a second part because of variability in the estimate. If the expected value of Y may be modeled by a few variables, there is a cost to adding more variables to the estimation procedure. In this case, statistical noise enters into the estimation of the additional variables, so that by using the more complex estimated predictive equation, future predictions would be off by more.
4. Thus what we would like to look for in our plot is a value C_p that is close to the 45° line, $C_p = p$. Such a value would have a low bias. Further, we would like the value of C_p itself to be small, so that the total error sum of squares is not large. The nicest possible case occurs when we can more or less satisfy both demands at the same time.
5. If we have to choose between a C_p value, which is close to p, or one that is smaller but above p, we are choosing between an equation that has a small bias (when $C_p = p$) but in further prediction is likely to have a larger predictive error, and a second equation (the smaller value for C_p) which in the future prediction is more likely to be close to the true value but where we think that the estimated predictive equation is probably biased. Depending on the use of the model, the trade-off between these two ills may or may not be clearcut.

Example 11.1. (*continued*) In this example we return to the data of Cullen and van Belle [1975]. We shall consider the response variable, DPMA, which is the disintegrations per minute of lymphocytes after the surgery. The viability of the lymphocytes was measured in terms of the uptake of nutrients that were labeled radioactively. A large number of disintegrations per minute suggests a high cell division rate, and thus active lymphocytes. The potential predictive variables for explaining the variability in DPMA are trauma factor (as discussed previously), duration (as discussed previously), the disintegrations per minute before the surgery, labeled DPMB, and the lymphocyte count in thousands per cubic millimeter before the surgery, LYMPHB, as well as the lymphocyte count in thousands per cubic millimeter after the surgery, LYMPHA. Let these variables have the following labels: $Y =$ DPMA; $X_1 =$ DURATION; $X_2 =$ TRAUMA; $X_3 =$ DPMB; $X_4 =$ LYMPHB; $X_5 =$ LYMPHA.

Table 11.11 presents the results for the 32 possible regression runs using subsets of the five predictor variables. For each run the value of p, C_p, the residual mean square, the average residual mean square for runs with the same number of variables, the multiple R^2, and the adjusted R^2, R_a^2, are presented. For a given number of variables, the entries are ordered in terms of increasing values of C_p. Note several things in Table 11.11. For a fixed number, $p - 1$, of predictor variables, if we look at the values for C_p, the residual mean square, R^2, and R_a^2, we see that as C_p increases, the residual mean square increases while R^2 and R_a^2 decrease. This relationship is a mathematical fact. Thus, if we know how many predictor variables, p, we want in our equation, any of the following six criteria for the best subset of predictor variables are equivalent:

1. Pick the predictive equation with a minimum value of C_p.
2. Pick the predictive equation with the minimum value of the residual mean square.
3. Pick the predictive equation with the maximum value of the multiple correlation coefficient, R^2.
4. Pick the predictive equation with the maximum value of the adjusted multiple correlation coefficient, R_a^2.
5. Pick the predictive equation with a maximum sum of squares due to regression.
6. Pick the predictive equation with the minimum sum of squares for the residual variability.

Table 11.11 Results from the 32 Regression Runs on the Anesthesia Data of Cullen and van Belle [1975]

Numbers of Explanatory Variables in Predictive Equation	p	C_p	Residual Mean Square	Residual Average Mean Square	R^2	R_a^2
None	1	60.75	4047	4047	0	0
3	2	5.98	1645		0.606	0.594
1		49.45	3578		0.142	0.116
2		57.12	3919	3476	0.060	0.032
4		60.48	4069		0.024	−0.005
5		62.70	4168		0.000+	−0.030
2,3	3	2.48	1444		0.664	0.643
1,3		2.82	1459		0.661	0.639
3,5		6.26	1617		0.624	0.600
3,4		6.91	1647		0.617	0.593
1,4		48.37	3549	2922	0.175	0.123
1,2		51.06	3672		0.146	0.093
1,5		51.43	3689		0.142	0.088
2,4		56.32	3914		0.090	0.033
2,5		59.10	4041		0.060	0.001
4,5		62.39	4192		0.024	−0.036
2,3,4	4	3.03	1422		0.680	0.648
1,3,4		3.32	1435		0.677	0.645
1,3,5		3.36	1438		0.676	0.645
2,3,5		3.52	1445		0.674	0.643
1,2,3		3.96	1466	2396	0.670	0.639
3,4,5		7.88	1651		0.628	0.592
1,2,4		50.03	3647		0.178	0.099
1,4,5		50.15	3653		0.177	0.097
1,2,5		52.98	3787		0.146	0.064
2,4,5		57.75	4013		0.096	0.008
1,2,3,4	5	4.44	1440		0.686	0.644
1,3,4,5		4.64	1450		0.684	0.642
2,3,4,5		4.69	1453	1913	0.683	0.641
1,2,3,5		4.83	1460		0.682	0.640
1,2,4,5		51.91	3763		0.180	0.070
1,2,3,4,5	6	6	1468	1468	0.691	0.637

The C_p data are more easily assimilated if we plot them. Figure 11.2 is a C_p plot for these data. The line $C_p = p$ is drawn for reference. Recall that points near this line have little bias in terms of the fit of the model; for points above this line we have biased estimates of the regression equation. We see that there are a number of models that have little bias. All things being equal, we prefer as small a C_p value as possible, since this is an estimate of the amount of variability between the true values and predicted values, which takes into account two components, the bias in the estimate of the regression line as well as the residual variability due to estimation. For this plot we are in the fortunate position of the lowest C_p value showing no bias. In addition, a minimal number of variables are involved. This point is circled, and going back to Table 11.11, corresponds to a model with $p = 3$, that is, two predictor variables. They are variables 2 and 3, the TRAUMA variable, and DPMB, the lymphocyte count in thousands per cubic millimeters before the surgery. This is the model we would select using Mallow's C_p approach.

We now turn to the average residual mean square plot to see if that would help us to decide how many variables to use. Figure 11.3 gives this plot. We can see that this plot does not level

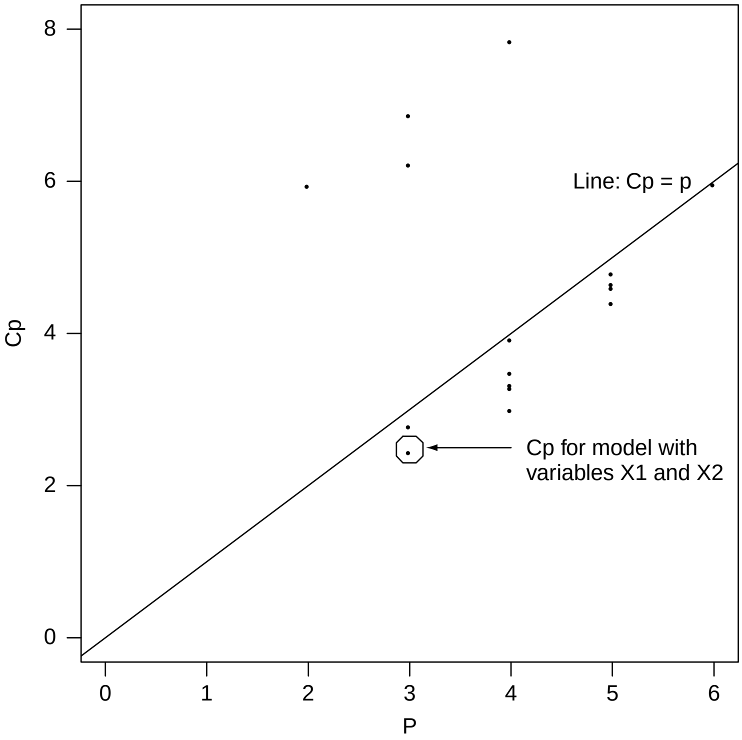

Figure 11.2 Mallow’s C_p plot for the data of Cullen and van Belle [1975]. Only points with $C_p < 8$ are plotted.

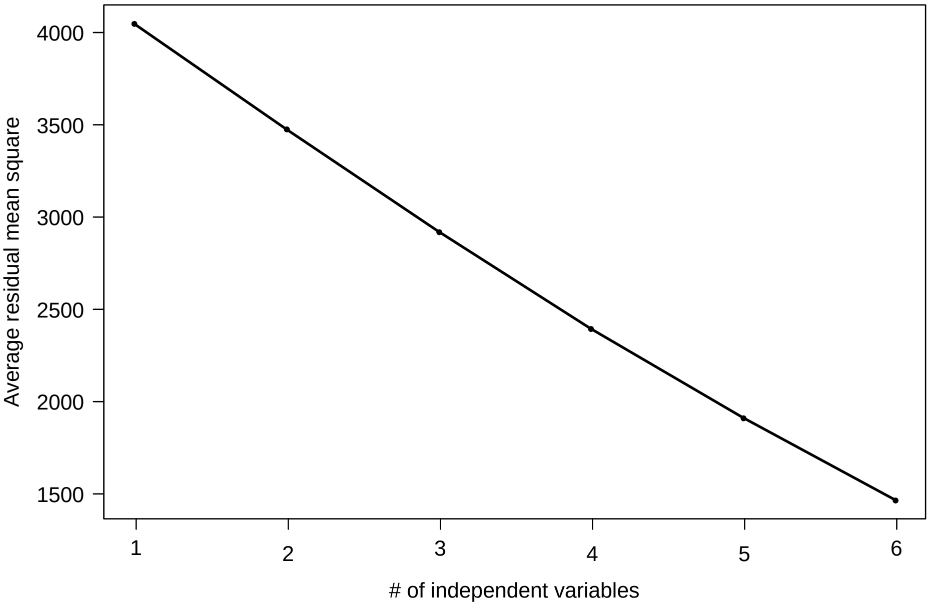

Figure 11.3 Average mean square plot for the Cullen and van Belle data [1975].

out but decreases until we have five variables. Thus this plot does not help us to decide on the number of variables we might consider in the final equation. If we look at Table 11.11, we can see why this happens. Since the final model has two predictive variables, even with three variables, many of the subsets, namely four, do not include the most predictive variable, variable 3, and thus have very large mean squares. We have not considered enough variables in the model above and beyond the final model for the curve to level out. With a relatively small number of potential predictor variables, five in this model, the average residual mean square plot is usually not useful.

Suppose that we have too many predictor variables to consider all combinations; or suppose that we are worried about the problem of looking at the huge number of possible combinations because we feel that the multiple comparisons may allow random variability to have too much effect. In this case, how might we proceed? In the next section we discuss one approach to this problem.

11.6.3 Stepwise Procedures

In this section we consider building a multiple regression model variable by variable.

Step 1

Suppose that we have a dependent variable Y and a set of potential predictor variables, X_i, and that we try to explain the variability in Y by choosing only one of the predictor variables. Which would we want? It is natural to choose the variable that has the largest squared correlation with the dependent variable Y. Because of the relationships among the sums of squares, this is equivalent to the following step.

Step 2

1. Choose i to maximize r^2_{Y,X_i}.
2. Choose i to maximize $\text{SS}_{\text{REG}}(X_i)$.
3. Choose i to minimize $\text{SS}_{\text{RESID}}(X_i)$.

By renumbering our variables if necessary, we can assume that the variable we picked was X_1. Now suppose that we want to add one more variable, say X_i, to X_1, to give us as much predictive power as possible. Which variable shall we add? Again we would like to maximize the correlation between Y and the predicted value of Y, $\widehat{Y}$; equivalently, we would like to maximize the multiple correlation coefficient squared. Because of the relationships among the sums of squares, this is equivalent to any of the following at this next step.

Step 3

X_1 is in the model; we now find $X_i (i \neq 1)$.

1. Choose i to maximize $R^2_{Y(X_1,X_i)}$.
2. Choose i to maximize $r^2_{Y,X_i.X_1}$.
3. Choose i to maximize $\text{SS}_{\text{REG}}(X_1, X_i)$.
4. Choose i to maximize $\text{SS}_{\text{REG}}(X_i|X_1)$.
5. Choose i to minimize $\text{SS}_{\text{RESID}}(X_1, X_i)$.

Our stepwise regression proceeds in this manner. Suppose that j variables have entered. By renumbering our variables if necessary, we can assume without loss of generality that the variables that have entered the predictive equation are $X_1, \ldots , X_j$. If we are to add one more variable to the predictive equation, which variable might we add? As before, we would like to add the variable that makes the correlation between Y and the predictor variables as large as possible. Again, because of the relationships between the sums of squares, this is equivalent to any of the following:

Step $j + 1$

$X_1, \ldots, X_j$ are in the model; we want $X_i (i \neq 1, \ldots, j)$.

1. Choose i to maximize $R^2_{Y(X_1,\ldots,X_j,X_i)}$.
2. Choose i to maximize $r^2_{Y,X_i \cdot X_1,\ldots,X_j}$.
3. Choose i to maximize $\text{SS}_{\text{REG}}(X_1, \ldots, X_j, X_i)$.
4. Choose i to maximize $\text{SS}_{\text{REG}}(X_i | X_1, \ldots, X_j)$.
5. Choose i to minimize $\text{SS}_{\text{RESID}}(X_1, \ldots, X_j, X_i)$.

If we continue in this manner, eventually we will use all of the potential predictor variables. Recall that our motivation was to select a simple model. Thus we would like a small model; this means that we would like to stop at some step before we have included all of our potential predictor variables. How long shall we go on including predictor variables in this model? There are several mechanisms for stopping. We present the most widely used stopping rule. We would not like to add a new variable if we cannot show statistically that it adds to the predictive power. That is, if in the presence of the other variables already in the model, there is no statistically significant relationship between the response variable and the next variable to be added, we will stop adding new predictor variables. Thus, the most common method of stopping is to test the significance of the partial correlation of the next variable and the response variable Y after adjusting for the variables entered previously. We use the partial F-test as discussed above. Commonly, the procedure is stopped when the p-value for the F level is greater than some fixed level; often, the fixed level is taken to be 0.05. This is equivalent to testing the statistical significance of the partial correlation coefficient. The partial F-statistic in the context of regression analysis is also often called the *F to enter*, since the value of F, or equivalently its p-value, is used as a criteria for entering the equation.

Since the F-statistic always has numerator degrees of freedom 1 and denominator degrees of freedom $n - j - 2$, and n is usually much larger than j, the appropriate critical value is effectively the F critical value with 1 and ∞ degrees of freedom. For this reason, rather than using a p-value, often the entry criterion is to enter variables as long as the F-statistic itself is greater than some fixed amount.

Summarizing, we stop when:

1. The p-value for $r^2_{Y,X_i.X_1,\ldots,X_j}$ is greater than a fixed level.
2. The partial F-statistic

$$\frac{\text{SS}_{\text{REG}}(X_i | X_1, \ldots, X_j)}{\text{SS}_{\text{RESID}}(X_1, \ldots, X_j, X_i)/(n - j - 2)}$$

is less than some specified value, or its p-value is greater than some fixed level.

All of this is summarized in Table 11.12; we illustrate by an example.

Example 11.3. (*continued*) Consider the active female exercise data used above. We shall perform a stepwise regression with VO_2 MAX as the dependent variable and DURATION, MAXIMUM HEART RATE, AGE, HEIGHT, and WEIGHT as potential independent variables. Table 11.13 contains a portion of the BMDP computer output for this run.

Table 11.12 Stepwise Regression Procedure (Forward) Selection for p Variable Case

Step	Variable Entered[a]	Intercept and Slopes Calculated[b]	Total SS Attributable to Regression	Contribution of Entered Variable to Regression	F-Ratio to Test Significance of Entered Variable
1	X_1	$a^{(1)}, b_1^{(1)}$	$\text{SS}_{\text{REG}}(X_1)$	$\text{SS}_{\text{REG}}(X_1)$	$\dfrac{\text{SS}(X_1)(n-2)}{\text{SS}_{\text{RESID}}(X_1)} = F_{1,n-2}$
2	X_2	$a^{(2)}, b_1^{(2)}, b_2^{(2)}$	$\text{SS}_{\text{REG}}(X_1, X_2)$	$\text{SS}_{\text{REG}}(X_2 \| X_1)$	$\dfrac{\text{SS}(X_2 \| X_1)(n-3)}{\text{SS}_{\text{RESID}}(X_1, X_2)} = F_{1,n-3}$
3	X_3	$a^{(3)}, b_1^{(3)}, b_2^{(3)}, b_3^{(3)}$	$\text{SS}_{\text{REG}}(X_1, X_2, X_3)$	$\text{SS}_{\text{REG}}(X_3 \| X_1, X_2)$	$\dfrac{\text{SS}(X_3 \| X_1, X_2)(n-4)}{\text{SS}_{\text{RESID}}(X_1, X_2, X_3)} = F_{1,n-4}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	X_j	$a^{(j)}, b_1^{(j)}, b_2^{(j)}, \ldots, b_j^{(j)}$	$\text{SS}_{\text{REG}}(X_1, X_2, \ldots, X_j)$	$\text{SS}_{\text{REG}}(X_j \| X_1, \ldots, X_{j-1})$	$\dfrac{\text{SS}(X_j \| X_1, \ldots, X_{j-1})(n-j-1)}{\text{SS}_{\text{RESID}}(X_1, \ldots, X_j)} = F_{1,n-j-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p	X_p	$a^{(p)}, b_1^{(p)}, b_2^{(p)}, \ldots, b_p^{(p)}$	$\text{SS}_{\text{REG}}(X_1, X_2, \ldots, X_p)$	$\text{SS}_{\text{REG}}(X_p \| X_1, \ldots, X_{p-1})$	$\dfrac{\text{SS}(X_p \| X_1, \ldots, X_p)(n-p-1)}{\text{SS}_{\text{RESID}}(X_1, \ldots, X_p)} = F_{1,n-p-1}$

[a] To simplify notation, variables are labeled by the step at which they entered the equation.

[b] The superscript notation indicates that the estimate of α changes from step to step, as well as the estimates of $\beta_1, \beta_2, \ldots, \beta_{p-1}$.

Table 11.13 Stepwise Multiple Linear Regression for the Data of Example 11.3

```
STEP NO.      0
---------------
STD. ERROR OF EST.     4.9489

ANALYSIS OF VARIANCE

                                   SUM OF SQUARES    DF    MEAN SQUARE
                       RESIDUAL    1028.6670         42      24.49208

                       VARIABLES IN EQUATION FOR VO2MAX

                                       STD. ERROR   STD REG                F
VARIABLE        COEFFICIENT   OF COEFF     COEFF      TOLERANCE TO REMOVE LEVEL
(Y-INTERCEPT       29.05349)

                          VARIABLES NOT IN EQUATION
                                      PARTIAL                  F
         VARIABLE   CORR.        TOLERANCE         TO ENTER        LEVEL
DUR          1         0.78601      1.00000               66.28            1
HR           3         0.33729      1.00000                5.26            1
AGE          4       −0.65099       1.00000               30.15            1
HT           5       −0.29942       1.00000                4.04            1
WT           6       −0.12618       1.00000                0.66            1

STEP NO.     1
--------------
VARIABLE ENTERED      1 DUR
MULTIPLE R                  0.7860
MULTIPLE R-SQUARE           0.6178
ADJUSTED R-SQUARE           0.6085
STD. ERROR OF EST.          3.0966


ANALYSIS OF VARIANCE
                               SUM OF SQUARES       DF       MEAN SQUARE        F RATIO
REGRESSION                        635.51730          1      635.5173              66.28
RESIDUAL                          393.15010         41        9.589027

                   VARIABLES IN EQUATION FOR VO2MAX

                                     STD. ERROR  STD REG                          F
VARIABLE         COEFFICIENT    OF COEFF     COEFF    TOLERANCE  TO REMOVE  LEVEL
(Y-INTERCEPT        3.15880)
DUR  1               0.05029          0.0062        0.786        1.00000          66.28           1
                 VARIABLES NOT IN EQUATION
                 PARTIAL                             F
VARIABLE       CORR.          TOLERANCE          TO ENTER   LEVEL
HR      3      −0.14731           0.72170               0.89          1
AGE    4      −0.24403           0.52510               2.53          1
HT      5        0.01597           0.86364               0.01          1
WT      6      −0.32457           0.99123               4.71          1
```

(*continued overleaf*)

Table 11.13 (*continued*)

```
STEP NO.    2
--------------

VARIABLE ENTERED      6 WT
MULTIPLE R              0.8112
MULTIPLE R-SQUARE       0.6581
ADJUSTED R-SQUARE       0.6410
STD. ERROR OF EST.      2.9654

ANALYSIS OF VARIANCE

                  SUM OF SQUARES     DF      MEAN SQUARE  F RATIO
REGRESSION           676.93490        2         338.4675    38.49
RESIDUAL             351.73250       40           8.793311

              VARIABLES IN EQUATION FOR VO2MAX

                                  STD. ERROR  STD REG                   F
VARIABLE         COEFFICIENT   OF COEFF    COEFF     TOLERANCE  TO REMOVE  LEVEL
(Y-INTERCEPT   10.30026)
DUR  1              0.05150     0.0059        0.805   0.99123       75.12      1
WT    6            -0.12659     0.0583       -0.202   0.99123        4.71      1

                     VARIABLES NOT IN EQUATION

                      PARTIAL                      F
VARIABLE     CORR.   TOLERANCE     TO ENTER    LEVEL
HR           3       -0.08377      0.68819      0.28     1
AGE          4       -0.24750      0.52459      2.54     1
HT           5        0.20922      0.66111      1.79     1
```

The 0.05 F critical value with degrees of freedom 1 and 42 is approximately 4.07. Thus at step 0, duration, maximum heart rate, and age are all statistically significantly related to the dependent variable $\text{VO}_{2\ \text{MAX}}$.

We see this by examining the F-to-enter column in the output from step 0. This is the F-statistic for the square of the correlation between the individual variable and the dependent variable. In step 0 up on the left, we see the analysis of variance table with only the constant coefficient. Under partial correlation we have the correlation between each variable and the dependent variable. At the first step, the computer program scans the possible predictor variables to see which has the highest absolute value of the correlation with the dependent variable. This is equivalent to choosing the largest F-to-enter. We see that this variable is DURATION. In step 1, DURATION has entered the predictive equation. Up on the left, we see the multiple R, which in this case is simply the correlation between the $\text{VO}_{2\ \text{MAX}}$ and DURATION variables, the value for R^2, and the standard error of the estimate; this is the estimated standard deviation about the regression line. This value squared is the mean square for the residual, or the estimate for σ^2 if this is the correct model. Below this is the analysis of variance table, and below this, the value of the regression coefficient, 0.050, for the DURATION variable. The standard error of the regression coefficient is then given. The standardized regression coefficient is the value of the regression coefficient if we had replaced DURATION by its standardized value. The value F-to-remove in a stepwise regression is the statistical significance of the partial correlation between the variable in the model and the dependent variable when adjusting for other variables in the model. The left-hand side lists the variables not already in the equation. Again we have the partial correlations between the potential predictor variables and the dependent variable after adjusting for the variables in the model, in this case one variable, DURATION. Let us focus on the variable AGE at step 0 and at step 1. In step 0 there was a very highly statistically significant relationship between VO2 $_{\text{MAX}}$ and AGE, the F-value being 30.15. After DURATION enters the predictive equation, in step 1 we see that the statistical significance has disappeared, with the F-to-enter decreasing to 2.53. This occurs because AGE is very closely related to DURATION and is also highly related to VO2 $_{\text{MAX}}$. The explanatory power of AGE may, equivalently, be explained by the explanatory power of DURATION. We see that *when a variable does not enter a predictive model, this does not mean that the variable is not related to the dependent variable but possibly that other variables in the model can account for its predictive power*. An equivalent way of viewing this is that the partial correlation has dropped from -0.65 to -0.24. There is another column labeled "tolerance". The tolerance is 1 minus the square of the multiple correlation between the particular variable being considered and all of the variables already in the stepwise equation. Recall that if this correlation is large, it is very difficult to estimate the regression coefficient [see equation (14)]. The tolerance is the term $(1 - R_j^2)$ in equation (14). If the tolerance becomes too small, the numerical accuracy of the model is in doubt.

In step 1, scanning the F-to-enter column, we see the variable WEIGHT, which is statistically significantly related to VO2 $_{\text{MAX}}$ at the 5% level. This variable enters at step 2. After this variable has entered, there are no statistically significant relationships left between the variables not in the equation and the dependent variable after adjusting for the variables in the model. The stepwise regression would stop at this point unless directed to do otherwise.

It is possible to modify the stepwise procedure so that rather than starting with 0 variables and building up, we start with all potential predictive variables in the equation and work down. In this case, at the first step we discard from the model the variable whose regression coefficient has the largest p-value, or equivalently, the variable whose correlation with the dependent variable after adjusting for the other variables in the model is as small as possible. At each step, this process continues removing a variable as long as there are variables to remove from the model that are not statistically significantly related to the response variable at some particular level. The procedure of adding in variables that we have discussed in this chapter is called a *step-up stepwise procedure*, while the opposite procedure of removing variables is called a *step-down stepwise procedure*. Further, as the model keeps building, it may be that a variable entered earlier in the stepwise procedure no longer is statistically significantly related to the dependent variable in the presence of the other variables. For this reason, when performing a step-up regression, most regression programs have the ability at each step to remove variables that are no longer statistically significant. All of this aims at a simple model (in terms of the number of variables) which explains as much of the variability as possible. The step-up and step-down procedures do not look at as many alternatives as the C_p plot procedure, and thus may not be as prone to overfitting the data because of the many models considered. If we perform a step-up or step-down fit for the anesthesia data discussed above, the resulting model is the same as the model picked by the C_p plot.

11.7 POLYNOMIAL REGRESSION

We motivate this section by an example. Consider the data of Bruce et al. [1973] for 44 active males with a maximal exercise treadmill test. The oxygen consumption VO2 $_{\text{MAX}}$ was regressed on, or explained by, the age of the participants. Figure 11.4 shows the residual plot.

Examination of the residual plot shows that the majority of the points on the left are positive with a downward trend. The points on the right have generally higher values with an upward trend. This suggests that possibly the simple linear regression model does not fit the data well.

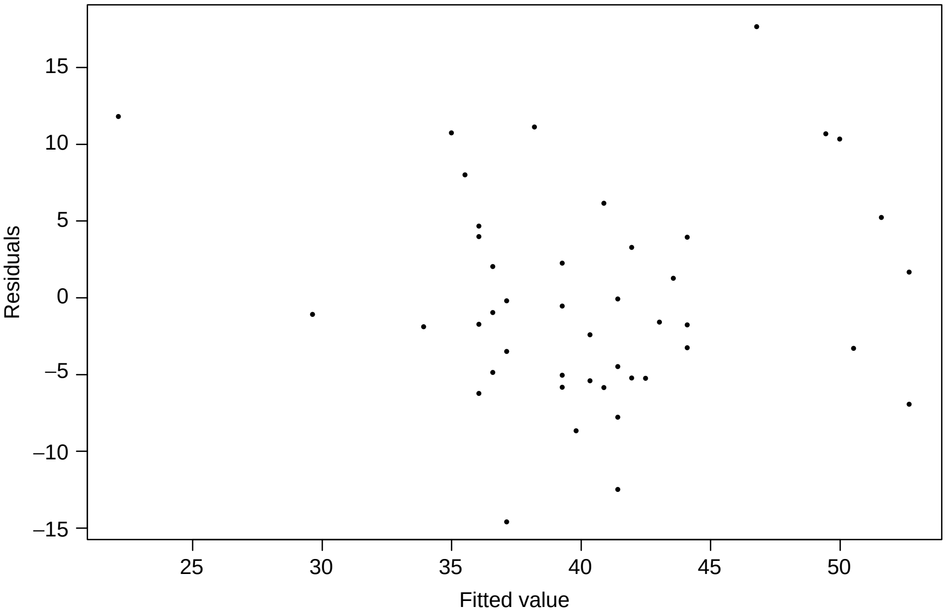

Figure 11.4 Residual plot of the regression of VO_2 MAX on age, active males.

The fact that the residuals come down and then go up suggests that possibly rather than being linear, the regression curve should be a second-order curve, such as

$$Y = a + b_1 X + b_2 X^2 + e$$

Note that this equation looks like a multiple linear regression equation. We could write this equation as a multiple regression equation,

$$Y = a + b_1 X_1 + b_2 X_2 + e$$

with $X_1 = X$ and $X_2 = X^2$. This simple observation allows us to fit polynomial equations to data by using multiple linear regression techniques. Observe what we are doing with multiple linear regression: The equation must be linear in the unknown parameters, but we may insert *known* functions of an explanatory variable. If we create the new variables $X_1 = X$ and $X_2 = X^2$ and run a multiple regression program, we find the following results:

Variable or Constant	b_j	**SE**(b_j)	t**-statistic** ($t_{41,0.975} \doteq 2.02$)
Age	−1.573	0.452	−3.484
Age^2	0.011	0.005	2.344
Constant	89.797	11.023	

We note that both terms age and age^2 are statistically significant. Recall that the t-test for the age^2 term is equivalent to the partial correlation of the age squared, with VO_2 MAX adjusting for the effect of age. This is equivalent to considering the hypothesis of linear regression *nested* within the hypothesis of quadratic regression. Thus, we reject the hypothesis of linear regression and could use this quadratic regression formula. A plot of the residuals using the quadratic regression shows no particular trend and is not presented here. One might wonder, now that we have a second-order term, whether perhaps a third-order term might help the situation. If we run a multiple regression with three variables ($X_3 = X^3$), the following results obtain:

Variable or Constant	b_j	SE(b_j)	t-statistic ($t_{40,0.975} \doteq 2.02$)
Age	−0.0629	2.3971	−0.0264
Age^2	−0.0203	0.0486	−0.4175
Age^3	0.0002	0.0003	0.6417
Constant	1384.49	783.15	

Since the age^3 term, which tests the nested hypothesis of the quadratic equation within the cubic equation, is nonsignificant, we may accept the quadratic equation as appropriate.

Figure 11.5 is a scatter diagram of the data as well as the linear and quadratic curves. Note that the quadratic curve is higher at the younger ages and levels off more around 50 to 60. Within the high range of the data, the quadratic or second-order curve increases. This may be an artifact of the curve fitting because all physiological knowledge tells us that the capacity for conditioning does not increase with age, although some subjects may improve their exercise performance with extra training. Thus, the second-order curve would seem to indicate that in a population of healthy active males, the decrease in VO_2 MAX consumption is not as rapid at the higher ages as at the lower ages. This is contrary to the impression that one would get from a linear fit. One would not, however, want to use the quadratic curve to extrapolate beyond or even to the far end of the data in this particular example.

We see that the real restrictions of multiple regression is not that the equation be linear in the variables observed, but rather that it be linear in the unknown coefficients. The coefficients

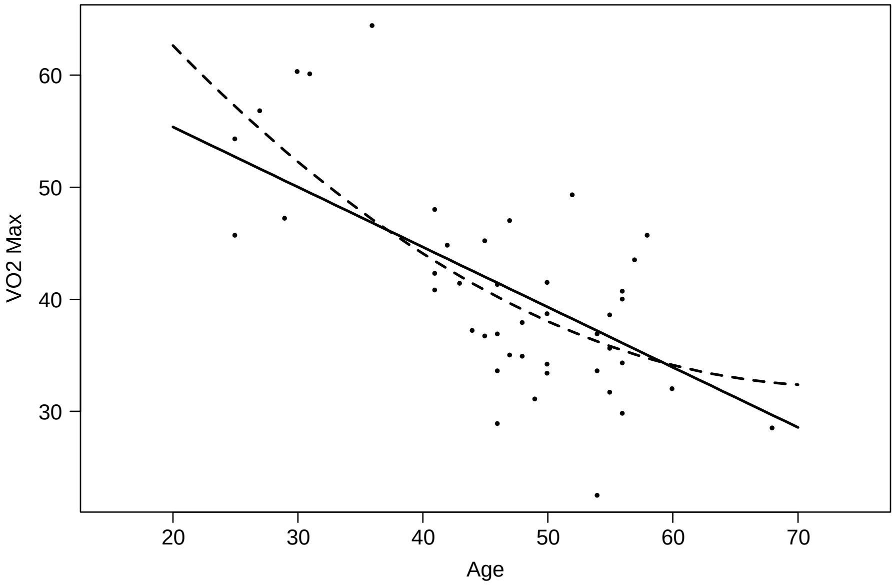

Figure 11.5 Active males with treadmill test: linear (solid line) and quadratic (dashed line) fits. (From Bruce et al. [1973].)

may be multiplied by known functions of the observed variables; this makes a variety of models possible. For example, with *two variables* we could also consider as an alternative to a linear fit (as given below) a second-order equation or polynomial in two variables:

$$Y = a + b_1X_1 + b_2X_2 + e$$

(linear in X_1 and X_2), and

$$Y = a + b_1X_1 + b_2X_2 + b_3X_1^2 + b_4X_1X_2 + b_5X_2^2 + e$$

(a second-order polynomial in X_1 and X_2).

Other functions of variables may be used. For example, if we observe a response that we believe is a periodic function of the variable X with a period of length L, we might try an equation of the form

$$Y = a + b_1 \sin\frac{\pi X}{L} + b_2 \cos\frac{\pi X}{L} + b_3 \sin\frac{2\pi X}{L} + b_4 \cos\frac{2\pi X}{L} + e$$

The important point to remember is that not only can polynomials in variables be fit, but any model may be fit where the response is a linear function of known functions of the variables involved.

11.8 GOODNESS-OF-FIT CONSIDERATIONS

As in the one-dimensional case, we need to check the fit of the regression model. We need to see that the form of the model roughly fits the data observed; if we are engaged in statistical inference, we need to see that the error distribution looks approximately normal. As in simple linear regression, one or two outliers can greatly skew the results; also, an inappropriate functional form can give misleading conclusions. In doing multiple regression it is harder than in simple linear regression to check the assumptions because there are more variables involved. We do not have nice two-dimensional plots that display our data completely. In this section we discuss some of the ways in which multiple regression models may be examined.

11.8.1 Residual Plots and Normal Probability Plots

In the multiple regression situation, a variety of plots may be useful. We discussed in Chapter 9 the residual plots of the predicted value for Y vs. the residual. Also useful is a normal probability plot of the residuals. This is useful for detecting outliers and for examining the normality assumption. Plots of the residual as a function of the independent or explanatory variables may point out a need for quadratic terms or for some other functional form. It is useful to have such plots even for potential predictor variables not entered into the predictive equation; they might be omitted because they are related to the response variable in a nonlinear fashion. This might be revealed by such residual plots.

Example 11.3. (*continued*) We return to the healthy normal active females. Recall that the VO2_{MAX} in a stepwise regression was predicted by DURATION and WEIGHT. Other variables considered were MAXIMUM HEART RATE, AGE, and HEIGHT. We now examine some of the residual plots as well as normal probability plots. The left panel of Figure 11.6 is a plot of residuals vs. fitted values. The residuals look fairly good except for the point circled on the right-hand margin, which lies farther from the value of zero than the rest of the points. The right-hand panel gives the square of the residuals. These values will have approximately a chi-square distribution with

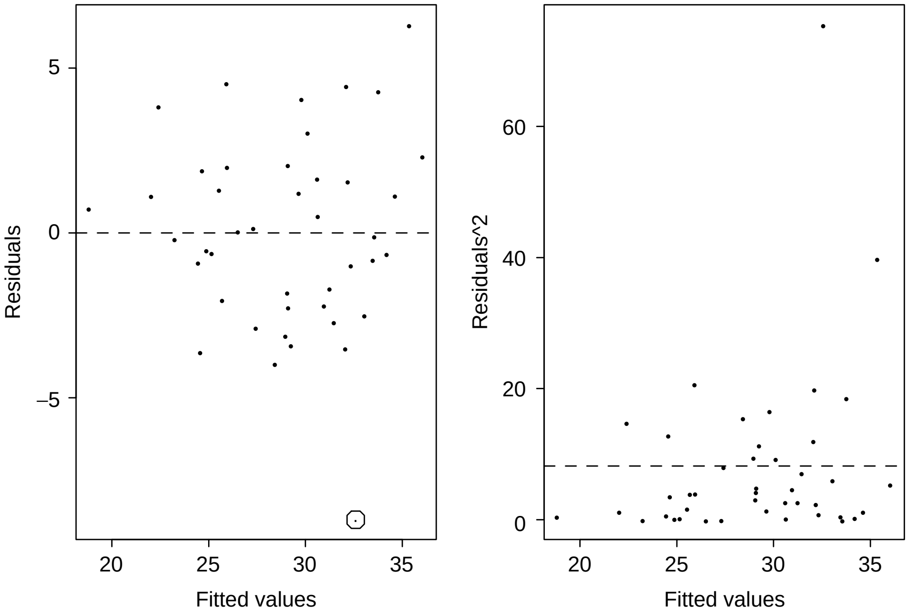

Figure 11.6 Residual plots.

one degree of freedom if normality holds. If the model is correct, there will not be a change in the variance with increasing predicted values. There is no systematic change here. However, once again the one value has a large deviation.

Figure 11.7 gives the normal probability plot for the residuals. In this output, the values predicted are on the horizontal axis rather than on the vertical axis, as plotted previously. Again, the residuals look quite nice except for the point on the far left; this point corresponds to the circled value in Figure 11.6. This raises the possibility of rerunning the analysis omitting the one outlier to see what effect it had on the analysis. We discuss this below after reviewing more graphical data.

Figures 11.8 to 11.12 deal with the residual values as a function of the five potential predictor variables. In each figure the left-hand panel presents the observed and predicted values for the data points and the right-hand panel for the observed values of those data present the residual values. In Figure 11.7, for DURATION, note that the values predicted are almost linear. This is because most of the predictive power comes from the DURATION variable, so that the value predicted is not far removed from a linear function of DURATION. The residual plot looks nice, with the possible exception of the outlier. In Figure 11.8, with respect to WEIGHT, we have the same sort of behavior as we do in the last three figures for AGE, MAXIMAL HEART RATE, and HEIGHT. In no case does there appear to be systematic unexplained variability than might be explained by adding a quadratic term or other terms to the equation.

If we rerun these data removing the potential outlier, the results change as given below.

	All Data		Removing the Outlier Point	
Variable or Constant	b_j	t	b_j	t
DURATION	0.0515	8.67	0.0544	10.17
WEIGHT	−0.127	−2.17	−0.105	−2.02
Constant	10.300		7.704	

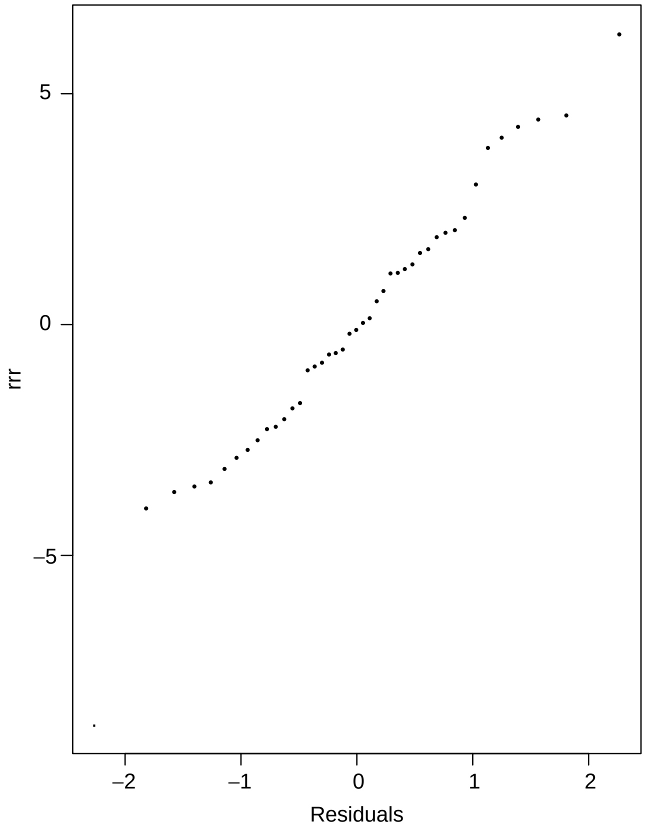

Figure 11.7 Normal residual plot.

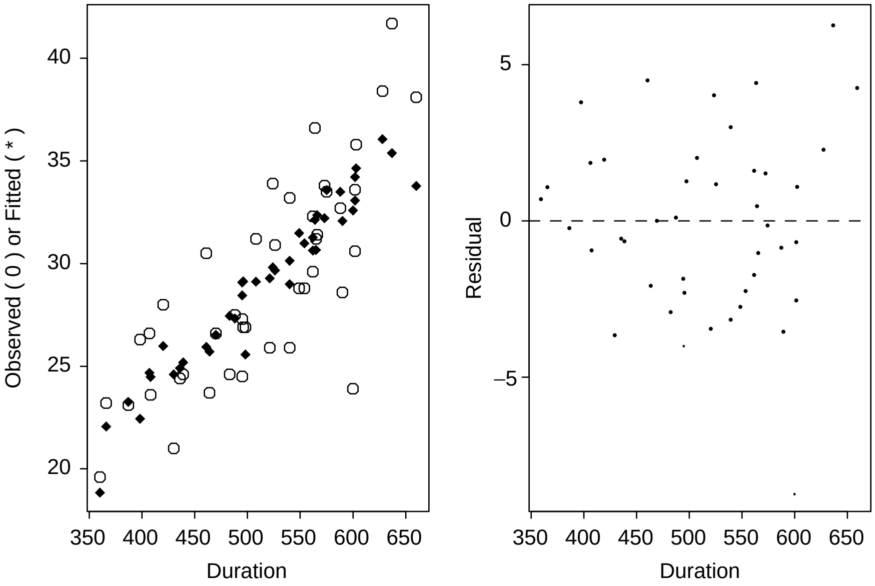

Figure 11.8 Duration vs. residual plots.

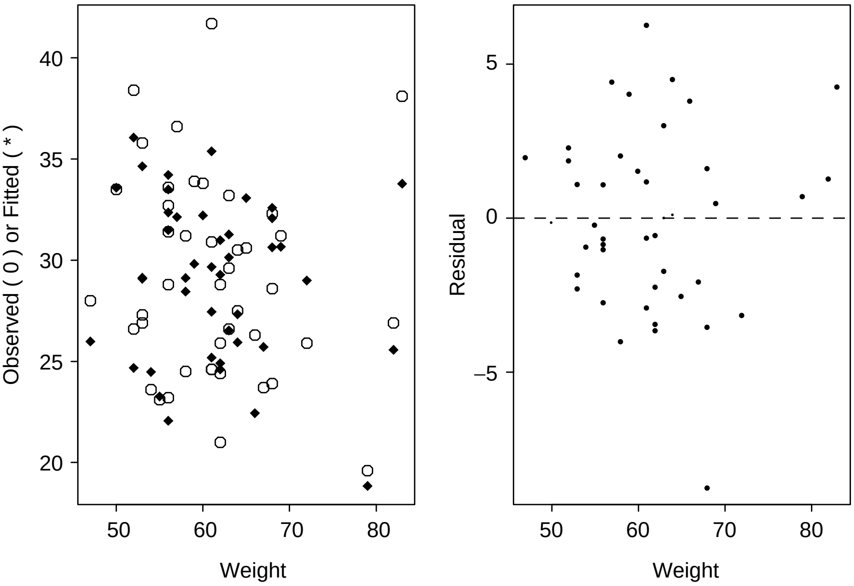

Figure 11.9 Weight vs. residual plots.

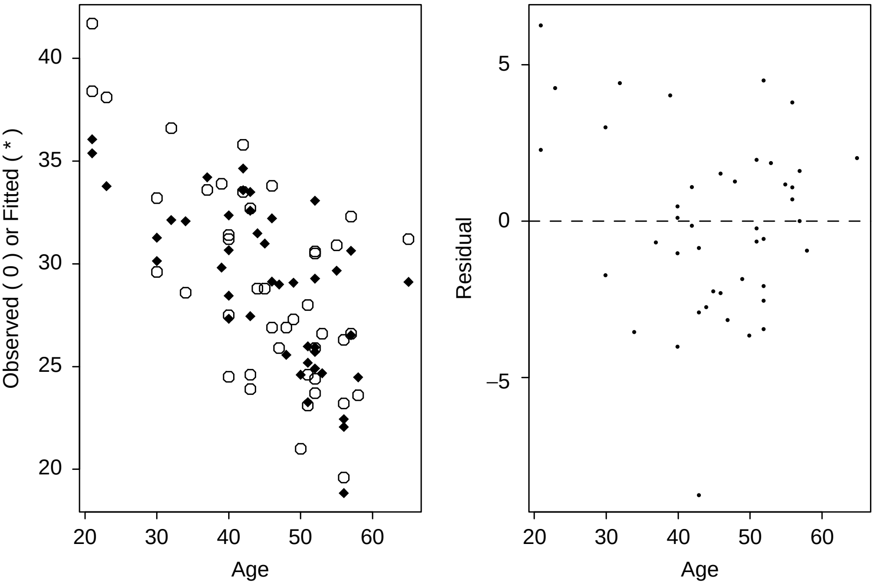

Figure 11.10 Age vs. residual plots.

We see a moderate change in the coefficient for WEIGHT; the change increases the importance of DURATION. The t statistic for WEIGHT is now right on the precise edge of statistical significance of the 0.05 level. Thus, although the original model did not mislead us, part of the contribution from WEIGHT came from the data point that was removed. This brings up the issue of how such data might be presented in a scientific paper or talk. One possibility would be to present both results and discuss the issue. The removal of outlying values may allow one to get a

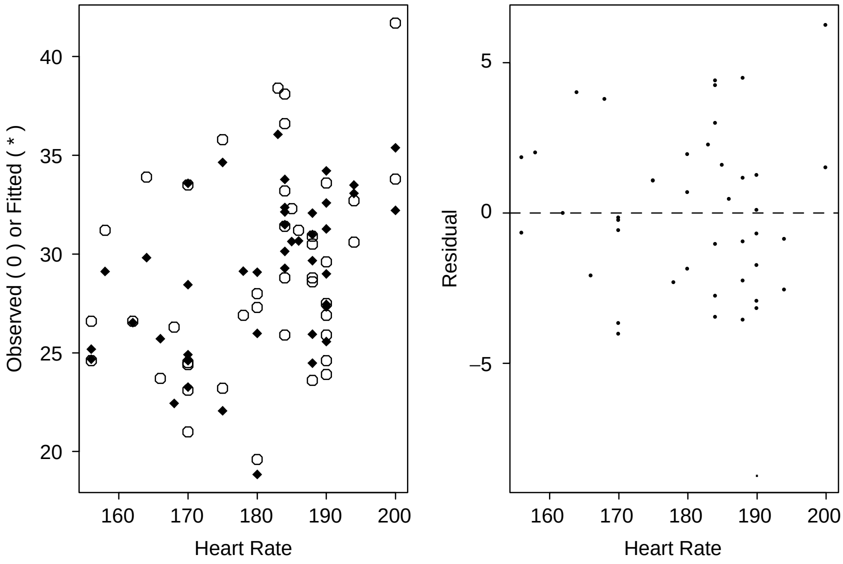

Figure 11.11 Maximum heart rate vs. residual plots.

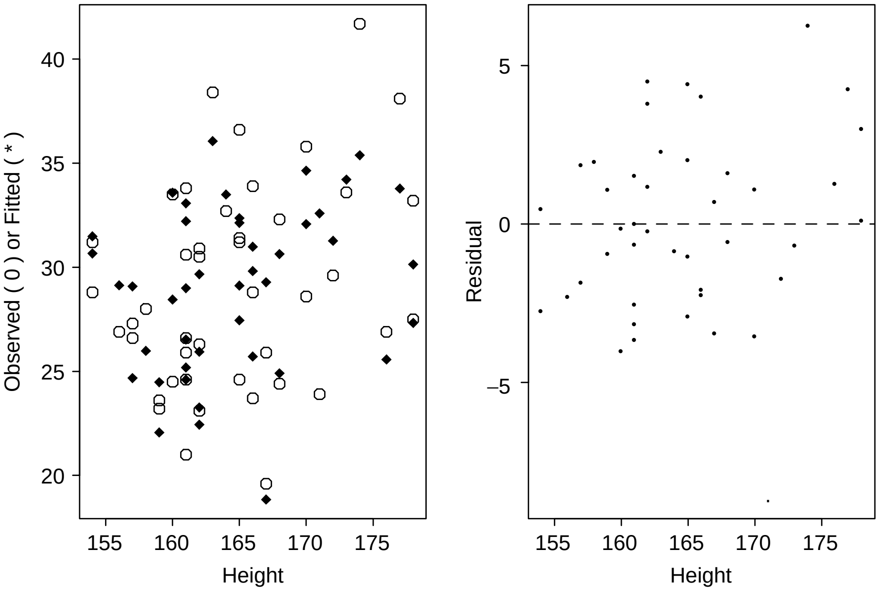

Figure 11.12 Height vs. residual plots.

closer fit to the data, and in this case the residual variability decreased from an estimated σ^2 of 2.97 to 2.64. Still, if the outlier is not considered to be due to bad data, but rather is due to an exceptional individual, in applying such relationships, other exceptional individuals may be expected to appear. In such cases, interpretation necessarily becomes complex. This shows, again, that although there is a nice precision to significance levels, in practice, interpretation of the statistical analysis is an art as well as a science.

11.8.2 Nesting in More Global Hypothesis

Since it is difficult to inspect multidimensional data visually, one possibility for testing the model fit is to embed the model in a more global hypothesis; that is, nest the model used within a more general model. One example of this would be adding quadratic terms and cross-product terms as discussed in Section 11.7. The number of such possible terms goes up greatly as the number of variables increases; this luxury is available only when there is a considerable amount of data.

11.8.3 Splitting the Samples; Jackknife Procedures

An estimated equation will fit data better than the true population equation because the estimate is designed to fit the data at hand. One way to get an estimate of the precision in a multiple regression model is to split the sample size into halves at random. One can estimate the parameters from one-half of the data and then predict the values for the remaining unused half of the data. The evaluation of the fit can be performed using the other half of the data. This gives an unbiased estimate of the appropriateness of the fit and the precision. There is, however, the problem that one-half of the data is "wasted" by not being used for the estimation of the parameters. This may be overcome by estimating the precision in this split-sampling manner but then presenting final estimates based on the entire data set.

Another approach, which allows more precision in the estimate, is to delete subsets of the data and to estimate the model on the remaining data; one then tests the fit on the smaller subsets removed. If this is done systematically, for example by removing one data point at a time, estimating the model using the remaining data and then examining the fit to the data point omitted, the procedure is called a *jackknife procedure* (see Efron [1982]). Resampling from the observed data, the *bootstrap* method may also be used [Efron and Tibshirani, 1986]. We will not go further into such issues here.

11.9 ANALYSIS OF COVARIANCE

11.9.1 Need for the Analysis of Covariance

In Chapter 10 we considered the analysis of variance. Associated with categorical classification variables, we had a continuous response. Let us consider the simplest case, where we have a one-way analysis of variance consisting of two groups. Suppose that there is a continuous variable X in the background: a covariate. For example, the distribution of the variable X may differ between the groups, or the response may be very closely related to the value for the variable X. Suppose further that the variable X may be considered a more fundamental cause of the response pattern than the grouping variable. We illustrate some of the potential complications by two figures.

On the left-hand side of Figure 11.13, suppose that we have data as shown. The solid circles show the response values for group 1 and the crosses the response values for group 2. There is clearly a difference in response between the two groups. Suppose that we think that it is not the grouping variable that is responsible but the covariate X. On the right-hand side we see a possible pattern that could lead to the response pattern given. In this case we see that the observations from both groups 1 and 2 have the same response pattern *when the value of X is taken into account*; that is, they both fall around one fixed regression line. In this case, the difference observed between the groups may alternatively be explained because they differ in the covariate value X. Thus in certain situations, in the analysis of variance one would like to adjust for potential differing values of a covariate. Another way of stating the same thing is: *In certain analysis of variance situations there is a need to remove potential bias, due to the fact that categories differ in their values of a covariate X.* (See also Section 11.5.)

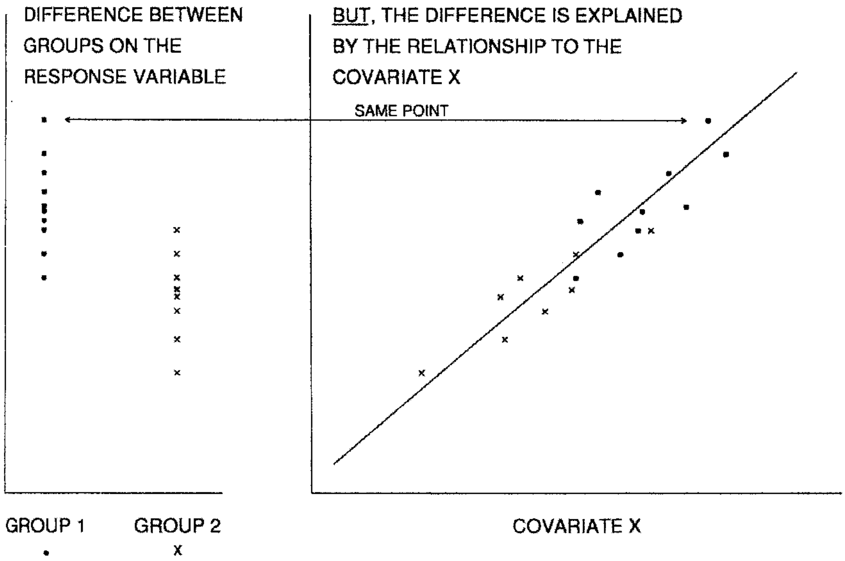

Figure 11.13 One-way analysis of variance with two categories: group difference because of bias due to different distribution on the covariate X.

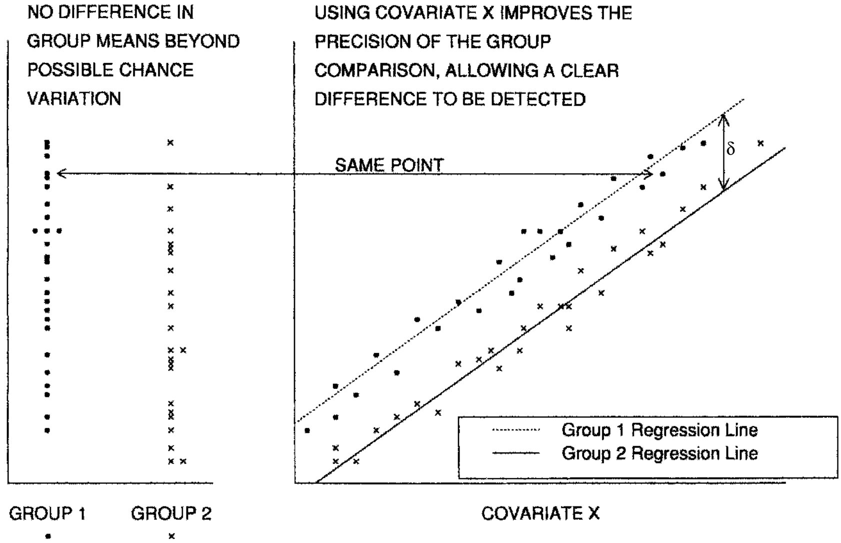

Figure 11.14 Two groups with close distribution on the covariate X. By using the relationship of the response to X separately in each group, a group difference obscured by the variation in X is revealed.

Figure 11.14 shows a pattern of observations on the left for groups 1 and 2. There is no difference between the response in the groups given the variability of the observations. Consider the same points, however, where we consider the relationship to a covariate X as plotted on the right. The right-hand figure shows that the two groups have parallel regression lines that differ by an amount delta. Thus for a fixed value of the covariate X, on the average, the observations from the two groups differ. In this plot, there is clearly a statistically significant difference between the two groups because their regression lines will clearly have different intercepts. Although the two groups have approximately the same distribution of the covariate values, if we consider the covariate we are able to improve the precision of the comparison between the two groups. On the left, most of the variability is not due to intrinsic variability within the groups, but rather is due to the variability in the covariate X. On the right, when the covariate X is taken into account, we can see that there is a difference. Thus a second reason for considering covariates in the analysis of variance is: *Consideration of a covariate may improve the precision of the comparison of the categories in the analysis of variance.*

In this section we consider methods that allow one or more covariates to be taken into account when performing an analysis of variance. Because we take into account those variables that vary with the variables of interest, the models and the technique are called the *analysis of covariance.*

11.9.2 Analysis of Covariance Model

In this section we consider the one-way analysis of covariance. This is a sufficient introduction to the subject so that more general analysis of variance models with covariates can then be approached.

In the one-way analysis of covariance, we observe a continuous response for each of a fixed number of categories. Suppose that the analysis of variance model is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where $i = 1, \ldots, I$ indexes the I categories; α_i, the category effect, satisfies $\sum_i \alpha_i = 0$; and $j = 1, \ldots, n_i$ indexes the observations in the ith category. The ε_{ij} are independent $N(0, \sigma^2)$ random variables.

Suppose now that we wish to take into account the effect of the continuous covariate X. As in Figures 11.13 and 11.14, we suppose that the response is linearly related to X, where the slope of the regression line, γ, is the same for each of the categories (see Figure 11.15). That is, our analysis of covariance model is

$$Y_{ij} = \mu + \alpha_i + \gamma X_{ij} + \varepsilon_{ij} \tag{20}$$

with the assumptions as before.

Although we do not pursue the matter, the analogous analysis of covariance model for the two-way analysis of variance without interaction may be given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma X_{ijk} + \varepsilon_{ijk}$$

Analysis of covariance models easily generalize to include more than one covariate. For example, if there are p covariates to adjust for, the appropriate equation is

$$Y_{ij} = \mu + \alpha_i + \gamma_1 X_{ij}(1) + \gamma_2 X_{ij}(2) + \cdots + \gamma_p X_{ij}(p) + \varepsilon_{ij}$$

where $X_{ij}(k)$ is the value for the kth covariate when the observation comes from the ith category and the jth observation in that category. Further, if the response is not linear, one may model a different form of the response. For example, the following equation models a quadratic response to the covariate X_{ij}:

$$Y_{ij} = \mu + \alpha_i + \gamma_1 X_{ij} + \gamma_2 X_{ij}^2 + \epsilon_{ij}$$

In each case in the analysis of covariance, *the assumption is that the response to the covariates is the same within each of the strata or cells for the analysis of covariance.*

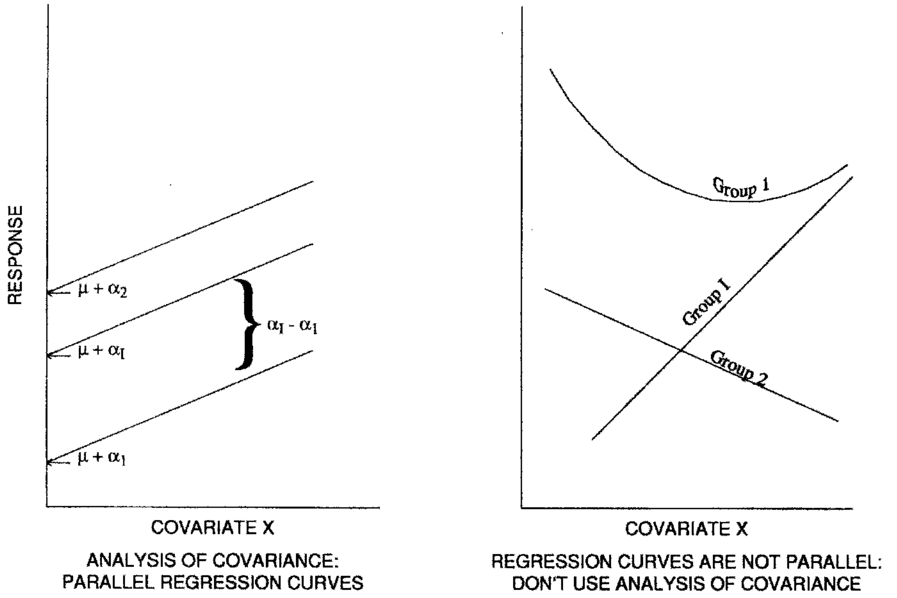

Figure 11.15 Parallel regression curves are assumed in the analysis of covariance.

It is possible to perform both the analysis of variance and the analysis of covariance by using the methods of multiple linear regression analysis, as given earlier in this chapter. The trick to thinking of an analysis of variance problem as a multiple regression problem is to use *dummy* or *indicator variables*, which allow us to consider the unknown parameters in the analysis of variance to be parameters in a multiple regression model.

Definition 11.11. A *dummy*, or *indicator variable* for a category or condition is a variable taking the value 1 if the observation comes from the category or satisfies the condition; otherwise, taking the value zero.

We illustrate this definition with two examples. A dummy variable for the male gender is

$$X = \begin{cases} 1, & \text{if the subject is male} \\ 0, & \text{otherwise} \end{cases}$$

A series of dummy variables for blood types (A, B, AB, O) are

$$X_1 = \begin{cases} 1, & \text{if the blood type is A} \\ 0, & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1, & \text{if the blood type is B} \\ 0, & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1, & \text{if the blood type is AB} \\ 0, & \text{otherwise} \end{cases}$$

$$X_4 = \begin{cases} 1, & \text{if the blood type is O} \\ 0, & \text{otherwise} \end{cases}$$

By using dummy variables, analysis of variance models may be turned into multiple regression models. We illustrate this by an example.

Consider a one-way analysis of variance with three groups. Suppose that we have two observations in each of the first two groups and three observations in the third group. Our model is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{21}$$

where i denotes the group and j the observation within the group. Our data are Y_{11}, Y_{12}, Y_{21}, Y_{22}, Y_{31}, Y_{32}, and Y_{33}. Let X_1, X_2, and X_3 be indicator variables for the three categories.

$$X_1 = \begin{cases} 1, & \text{if the observation is in group 1} \\ 0, & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1, & \text{if the observation is in group 2} \\ 0, & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1, & \text{if the observation is in group 3} \\ 0, & \text{otherwise} \end{cases}$$

Then equation (21) becomes (omitting subscript on Y and e)

$$Y = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon \tag{22}$$

Note that X_1, X_2, and X_3 are related. If $X_1 = 0$ and $X_2 = 0$, then X_3 must be 1. Hence there are only two independent dummy variables. In general, for k groups there are $k-1$ independent dummy variables. This is another illustration of the fact that the k treatment effects in the one-way analysis of variance have $k-1$ degrees of freedom. Our data, renumbering the Y_{ij} to be Y_k, $k = 1, \ldots, 7$, are given in Table 11.14. For technical reasons, we do not estimate equation (22). Since

$$\sum_i X_i = 1, \qquad R^2_{X_1(X_2,X_3)} = 1$$

Recall that we cannot estimate regression coefficients well if the multiple correlation is near 1. Instead, an equivalent model

$$Y = \delta + \gamma_1 X_1 + \gamma_2 X_2 + \epsilon$$

is used. Here $\delta = \mu + \alpha_3$, $\gamma_1 = \alpha_1 - \alpha_3$, and $\gamma_2 = \alpha_2 - \alpha_3$. That is, all effects are compared relative to group 3. We may now use a multiple regression program to perform the one-way analysis of variance.

To move to an analysis of covariance, we use $Y = \delta + \gamma_1 X_1 + \gamma_2 X_2 + \beta X + \epsilon$, where X is the covariate. If there is no group effect, we have the same expected value (for fixed X) regardless of the group; that is, $\gamma_1 = \gamma_2 = 0$.

Table 11.14 Data Using Dummy Variables

Y_k	Y_{ij}	X_1	X_2	X_3
Y_1	Y_{11}	1	0	0
Y_2	Y_{12}	1	0	0
Y_3	Y_{21}	0	1	0
Y_4	Y_{22}	0	1	0
Y_5	Y_{31}	0	0	1
Y_6	Y_{32}	0	0	1
Y_7	Y_{33}	0	0	1

More generally, for I groups the model is

$$Y = \delta + \gamma_1 X_1 + \cdots + \gamma_{I-1} X_{I-1} + \beta X + \epsilon$$

The null hypothesis is $H_0{:}\gamma_1 = \gamma_2 = \cdots = \gamma_{I-1} = 0$. This is tested using nested hypotheses. Let $\text{SS}_{\text{REG}}(X)$ be the regression sum of squares for the model $Y = \delta + \beta X + e$. Let

$$\text{SS}_{\text{REG}}(\gamma|X) = \text{SS}_{\text{REG}}(X_1, \ldots, X_{I-1}, X) - \text{SS}_{\text{REG}}(X)$$

and

$$\text{SS}_{\text{RESID}}(\gamma, X) = \text{SS}_{\text{TOTAL}} - \text{SS}_{\text{REG}}(X_1, \ldots, X_{I-1}, X)$$

The analysis of covariance table is:

Source	**d.f.**	**SS**	**MS**	***F*-Ratio**
Regression on X	1	$\text{SS}_{\text{REG}}(X)$	$\text{MS}_{\text{REG}}(X)$	$\dfrac{\text{MS}_{\text{REG}}(X)}{\text{MS}_{\text{RESID}}}$
Groups adjusted for X	$I-1$	$\text{SS}_{\text{REG}}(\gamma\|X)$	$\text{MS}_{\text{REG}}(\gamma\|X)$	$\dfrac{\text{MS}_{\text{REG}}(\gamma\|X)}{\text{MS}_{\text{RESID}}}$
Residual	$n-I-1$	$\text{SS}_{\text{RESID}}(\gamma\|X)$	MS_{RESID}	
Total	$n-1$	SS_{TOTAL}		

The F-test for the equality of group means has $I - 1$ and $n - I - 1$ degrees of freedom. If there is a statistically significant group effect, there is an interest in the separation of the parallel regression lines. The regression lines are:

Group	**Line**
1	$\widehat{\delta} + \widehat{\gamma_1} + \widehat{\beta}X$
2	$\widehat{\delta} + \widehat{\gamma_2} + \widehat{\beta}X$
$\vdots$	$\vdots$
$I-1$	$\widehat{\delta} + \widehat{\gamma_{I-1}} + \widehat{\beta}X$
I	$\widehat{\delta} + \widehat{\beta}X$

where the "hat" denotes the usual least squares multiple regression estimate. Customarily, these values are calculated for X equal to the average X value over all the observations. These values are called *adjusted means* for the group. This is in contrast to the mean observed for the observations in each group. Note again that group I is the reference group. It may sometimes be useful to rearrange the groups to have a specific group be the reference group. For example, suppose that there are three treatment groups and one reference group. Then the effects γ_1, γ_2, and γ_3 are, naturally, the treatment effects relative to the reference group.

We illustrate these ideas with two examples. In each example there are two groups ($I = 2$) and one covariate for adjustment.

Example 11.1. (*continued*) The data of Cullen and van Belle [1975] are considered again. In this case a larger set of data is used. One group received general anesthesia ($n_1 = 35$) and another group regional anesthesia ($n_2 = 42$). The dependent variable, Y, is the percent

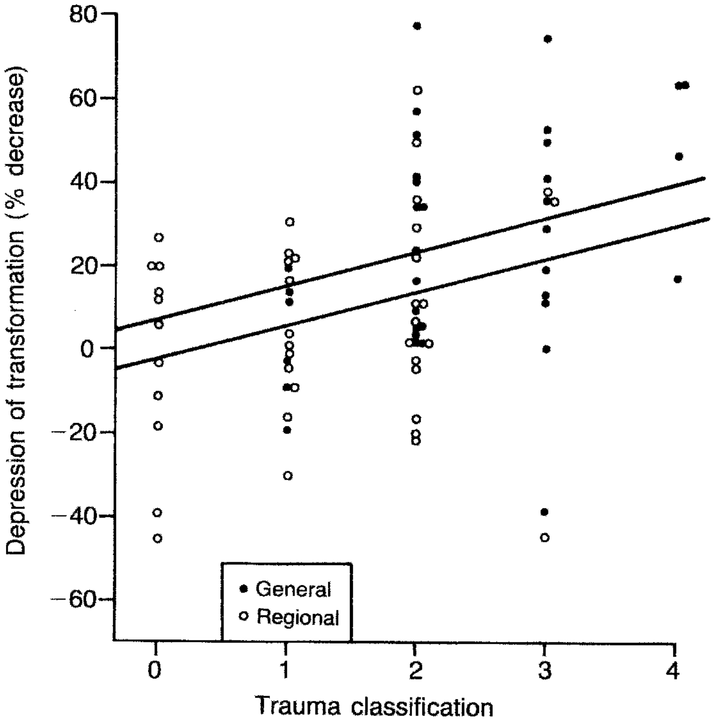

Figure 11.16 Relationship of postoperative depression of lymphocyte transformation to the level of trauma. Each point represents the response of one patient.

depression of lymphocyte transformation following surgery. The covariate, X, is the degree of trauma of the surgical procedure.

Figure 11.16 shows the data with the estimated analysis of covariance regression lines. The top line is the regression line for the general anesthesia group (which had a higher average trauma, 2.4 vs. 1.4). The analysis of covariance table is:

Source	d.f.	SS	MS	F-Ratio
Regression on trauma	1	4,621.52	4,621.52	7.65
General vs. regional anesthesia adjusted for trauma	1	1,249.78	1,249.78	2.06
Residual	74	44,788.09	605.24	
Total	76	56,201.52		

Note that trauma is significantly related to the percent depression of lymphocyte transformation, $F = 7.65 > F_{1,74,0.95}$. In testing the adjusted group difference,

$$F = 2.06 < 3.97 = F_{1,74,0.95}$$

so there is not a statistically significant difference between regional and general anesthesia after adjusting for trauma.

The two regression lines are

$$Y_1 = 25.6000 + 8.4784(X - 2.3714)$$

$$Y_2 = 6.7381 + 8.4784(X - 1.2619)$$

At the average value of $\overline{X} = 1.7552$, the predicted or adjusted means are

$$\widehat{Y}_1 = 25.6000 + (-5.1311) = 20.47$$

$$\widehat{Y}_2 = 6.7381 + (4.2757) = 11.01$$

The original difference is $\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot} = 25.6000 - 6.7381 = 18.86$. The adjusted (nonsignificant) difference is $\widehat{Y}_1 - \widehat{Y}_2 = 20.47 - 11.01 = 9.46$, a considerable drop. In fact the unadjusted one-way analysis of variance, or equivalently unpaired t-test, is significant: $p < 0.01$. The observed difference may be due to bias in the differing amount of surgical trauma in the two groups.

Example 11.8. Do men and women use the same level of oxygen when their maximal exercise limit is the same? The Bruce et al. [1973] maximal exercise data are used. The limit of exercise is expressed by the duration on the treadmill. Thus we wish to know if there is a $VO_{2\ MAX}$ difference between genders when adjusting for the duration of exercise. The analysis of covariance table is:

Source	**d.f.**	**SS**	**MS**	***F*-Ratio**
Duration	1	6049.51	6049.51	504.97
Gender, adjusting for duration	1	229.83	229.83	19.18
Residual	84	1006.05	11.98	
Total	86	7285.39		

The gender difference is highly statistically significant after adjusting for the treadmill duration. The estimated regression lines are:

$$\text{Females:} \qquad VO_{2\ MAX} = -1.59 + 0.0595 \times \text{duration}$$

$$\text{Males:} \qquad VO_{2\ XMAX} = 2.27 + 0.0595 \times \text{duration}$$

The overall duration mean is 581.89. The means are:

	$VO_{2\ MAX}$ Means	
	Observed	**Adjusted**
Female	29.05	33.03
Male	40.80	36.89

The fact that at maximum exercise normal males use more oxygen per unit of body weight is not accounted for entirely by their average longer duration on the treadmill (647 s vs. 515 s). Even when adjusting for duration, more oxygen per kilogram per minute is used.

Model assumptions may be tested by residual plots and normal probability plots as above. One assumption was that the regression lines were parallel. This may be tested by using the model (in the one-way ANOVA)

$$Y = \delta + \gamma_1 X_1 + \cdots + \gamma_{I-1} X_{I-1} + \beta X + \beta_1 X \cdot X_1 + \cdots + \beta_I X \cdot X_I + \epsilon$$

If an observation is in group $i (i = 1, \ldots, I - 1)$, this reduces to

$$Y = \delta + \gamma_i + \beta_i X + \epsilon$$

Nested within this model is the special case $\beta_1 = \beta_2 = \cdots = \beta_I$.

Source	d.f.	SS	MS	*F*-Ratio
Model with $\gamma_1, \ldots, \gamma_{I-1}, \beta$	I	SS_{REG} $(\gamma_1, \ldots, \gamma_{I-1})$	$\text{MS}_{\text{REG}}(\gamma' \text{s})$	
Model with $\gamma_1, \ldots, \gamma_{I-1}, \beta, \beta_1, \ldots, \beta_I$; extra ss	$I - 1$	SS_{REG} $(\beta_1, \ldots, \beta_I \vert \gamma_1, \ldots, \gamma_{I-1}, \beta)$	$\text{MS}_{\text{REG}}(\beta_i' \text{s} \vert \gamma_i' \text{s}, \beta)$	$\dfrac{\text{MS}_{\text{REG}}(\beta_i' \text{s} \vert \gamma_i' \text{s}, \beta)}{\text{MS}_{\text{RESID}}(\gamma_i' \text{s}, \beta_i' \text{s})}$
Residual	$n - 2I$	$\text{SS}_{\text{RESID}}(\gamma_1, \ldots, \gamma_{I-1}, \beta_1, \ldots, \beta_I)$	$\text{MS}_{\text{RESID}}(\gamma_i' \text{s}, \beta_i' \text{s})$	
Total	$n - 1$	SS_{TOTAL}		

For the exercise test example, we have:

Source	d.f.	SS	MS	*F*-Ratio
Model with group, equal slopes, and duration	2	6279.34	3139.67	
Model with unequal slopes (minus SS for nested equal-slope model)	1	29.40	29.40	2.50
Residual	83	976.65	11.77	
Total	86	7285.39		

As $F = 2.50 < F_{1,83,0.95}$, the hypothesis of equal slopes (parallelism) is reasonable and the analysis of covariance was appropriate. This use of a nested hypothesis is an example of the method of Section 11.8.2 for testing the goodness of fit of a model.

11.10 ADDITIONAL REFERENCES AND DIRECTIONS FOR FURTHER STUDY

11.10.1 There Are Now Many References on Multiple Regression Methods

Draper and Smith [1981] present extensive coverage of the topics of this chapter, plus much more material and a large number of examples with solutions. The text is on a more advanced mathematical level, making use of matrix algebra. Kleinbaum and Kupper [1998] present material on a level close to that of this chapter; taking more pages for the topics of this chapter, they have a more leisurely presentation. The text is an excellent supplementary reference to the material of this chapter. Another useful text is Daniel and Wood [1999].

11.10.2 Time-Series Data

It would appear that the multiple regression methods of this chapter would apply when one of the explanatory variables is time. This may be true in certain limited cases, but it is not usually true. Analyzing data with time as an independent variable is called *time-series analysis*. Often, in time, the errors are dependent at different time points. Box, Jenkins, and Reinsel [1994] are one source for time-series methods.

11.10.3 Causal Models: Structural Models and Path Analysis

In many studies, especially observational studies of human populations, one might conjecture that certain variables contribute in a causal fashion to the value of another variable. For example, age and gender might be hypothesized to contribute to hospital bed use, but not vice versa. In a statistical analysis, bed use would be modeled as a linear function of age and gender plus other unexplained variability. If only these three variables were considered, we would have a multiple regression situation. Bed use with other variables might be considered an explanatory variable for number of nursing days used. *Structural models* consist of a series of multiple regression equations; the equations are selected to model conjectured causal pathways. The models do not prove causality but can examine whether the data are consistent with certain causal pathways.

Three books addressing structural models (from most elementary to more complex) are Li [1975], Kaplan [2000], and Goldberger and Duncan [1973]. Issues of causality are addressed in Blalock [1985], Cook et al. [2001], and Pearl [2000].

11.10.4 Multivariate Multiple Regression Models

In this chapter we have analyzed the response of one dependent variable as explained by a linear relationship with multiple independent or predictor variables. In many circumstances there are multiple (more than one) dependent variables whose behavior we want to explain in terms of the independent variables. When the models are linear, the topic is called *multivariate multiple regression*. The mathematical complexity increases, but in essence each dependent variable is modeled by a separate linear equation. Morrison [1976] and Timm [1975] present such models.

11.10.5 Nonlinear Regression Models

In certain fields it is not possible to express the response of the dependent variable as a linear function of the independent variables. For example, in pharmacokinetics and compartmental analysis, equations such as

$$Y = \beta_1 e^{\beta_2 x} + \beta_3 e^{\beta_4 x} + e$$

and

$$Y = \frac{\beta_1}{x - \beta_2} + e$$

may arise where the β_i's are unknown coefficients and the e is an error (unexplained variability) term. See van Belle et al. [1989] for an example of the latter equation. Further examples of *nonlinear* regression equations are given in Chapters 13 and 16.

There are computer programs for estimating the unknown parameters.

1. The estimation proceeds by trying to get better and better approximations to the "best" (maximum likelihood) estimates. Sometimes the programs do not come up with an estimate; that is, they do not converge.
2. Estimation is much more expensive (in computer time) than it is in the linear models program.
3. The interpretation of the models may be more difficult.
4. It is more difficult to check the fit of many of the models visually.

NOTES

11.1 Least Squares Fit of the Multiple Regression Model

We use the sum of squares notation of Chapter 9. The regression coefficients b_j are solutions to the k equations

$$\begin{aligned}
[x_1^2]b_1 + [x_1x_2]b_2 + \cdots + [x_1x_k]b_k &= [x_1y] \\
[x_1x_2]b_1 + [x_2^2]b_2 + \cdots + [x_2x_k]b_k &= [x_2y] \\
&\vdots \\
[x_1x_k]b_1 + [x_2x_k]b_2 + \cdots + [x_k^2]b_k &= [x_ky]
\end{aligned}$$

For readers familiar with matrix notation, we give a Y vector and covariate matrix.

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1k} \\ X_{21} & \cdots & X_{2k} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nk} \end{pmatrix}$$

The b_j are given by

$$\begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

where the prime denotes the matrix transpose and -1 denotes the matrix inverse. Once the b_j's are known, a is given by

$$a = \overline{Y} - (b_1\overline{X}_1 + \cdots + b_k\overline{X}_k)$$

11.2 Multivariate Normal Distribution

The density function for *multivariate normal distribution* is given for those who know matrix algebra. Consider jointly distributed variables

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_p \end{pmatrix}$$

written as a vector. Let the *mean vector* and *covariance matrix* be given by

$$\mu = \begin{pmatrix} E(Z_1) \\ \vdots \\ E(Z_p) \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \operatorname{var}(Z_1) & \operatorname{cov}(Z_1, Z_2) & \cdots & \operatorname{cov}(Z_1, Z_p) \\ \vdots & & & \vdots \\ \operatorname{cov}(Z_p, Z_1) & \cdots & \cdots & \operatorname{var}(Z_p) \end{pmatrix}$$

The density is

$$f(z_1, \ldots, z_p) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp[-(Z-\mu)' \sum^{-1}(Z-\mu)/2]$$

where $|\Sigma|$ is the determinant of Σ and -1 denotes the matrix inverse. See Graybill [2000] for much more information about the multivariate normal distribution.

Table 11.15 ANOVA Table Incorporating Pure Error

Source	d.f.	SS	MS	***F*-Ratio**
Regression	p	SS_{REG}	MS_{REG}	$\dfrac{\text{MS}_{\text{REG}}}{\text{MS}_{\text{RESID}}}$
Residual	$n-p-1$	SS_{RESID}	MS_{RESID}	
*Model	*$\text{d.f.}_{\text{MODEL}}$	SS_{MODEL}	MS_{MODEL}	$\dfrac{\text{MS}_{\text{MODEL}}}{\text{MS}_{\text{PURE ERROR}}}$
*Pure error	*$\text{d.f.}_{\text{PURE ERROR}}$	$\text{SS}_{\text{PURE ERROR}}$	$\text{MS}_{\text{PURE ERROR}}$	
Total	$n-1$	SS_{TOTAL}		

11.3 Pure Error

We have seen that it is difficult to test goodness of fit without knowing at least one large model that fits the data. This allows estimation of the residual variability. There is a situation where one can get an accurate estimate of the residual variability without any knowledge of an appropriate model. Suppose that for some fixed value of the X_i's, there are *repeated* measurements of Y. These Y variables will be multiple independent observations with the same mean and variance. By subtracting the sample mean for the point in question, we can estimate the variance. More generally, if more than one X_i combination has multiple observations, we can pool the sum of squares (as in one-way ANOVA) to estimate the residual variability.

We now show how to partition the sum of squares. Suppose that there are K combinations of the covariates X_i for which we observe two or more Y values. Let Y_{ik} denote the ith observation ($i = 1, 2, \ldots, n_k$) at the kth covariate values. Let $\overline{Y}_k$ be the mean of the Y_{ik}:

$$\overline{Y}_k = \sum_{i=1}^{n_k} \frac{Y_{ik}}{n_k}$$

We define the pure error sum of squares and model of squares as follows:

$$\text{SS}_{\text{PURE ERROR}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (Y_{ik} - \overline{Y}_k)^2$$

$$\text{SS}_{\text{MODEL FIT}} = \text{SS}_{\text{RESID}} - \text{SS}_{\text{PURE ERROR}}$$

Also,

$$\text{MS}_{\text{PURE ERROR}} = \frac{\text{SS}_{\text{PURE ERROR}}}{\text{d.f.}_{\text{PURE ERROR}}}$$

$$\text{MS}_{\text{MODEL FIT}} = \frac{\text{SS}_{\text{MODEL}}}{\text{d.f.}_{\text{MODEL}}}$$

where

$$\text{d.f.}_{\text{PURE ERROR}} = \sum_{k=1}^{K} n_k - K$$

$$\text{d.f.}_{\text{MODEL}} = n + K - \sum_{k=1}^{K} n_k - p - 1$$

n is the total number of observations, and p is the number of covariates in the multiple regression model. The analysis of variance table becomes that shown in Table 11.15. The terms with an asterisk further partition the residual sum of squares. The F-statistic $\text{MS}_{\text{MODEL}}/\text{MS}_{\text{PURE ERROR}}$ with $\text{d.f.}_{\text{MODEL}}$ and $\text{d.f.}_{\text{PURE ERROR}}$ degrees of freedom tests the model fit. If the model is not rejected as unsuitable, the usual F-statistic tests whether or not the model has predictive power (i.e., whether all the $\beta_i = 0$).

PROBLEMS

Problems 11.1 to 11.7 deal with the fitting of one multiple regression equation. Perform each of the following tasks as indicated. Note that various parts are from different sections of the chapter. For example, tasks (e) and (f) are discussed in Section 11.8.

- **(a)** Find the t-value for testing the statistical significance of each of the regression coefficients. Do we reject $\beta_j = 0$ at the 5% significance level? At the 1% significance level?
- **(b)** **i.** Construct a 95% confidence interval for each β_j.
  ii. Construct a 99% confidence interval for each β_j.
- **(c)** Fill in the missing values in the analysis of variance table. Is the regression significant at the 5% significance level? At the 1% significance level?
- **(d)** Fill in the missing values in the partial table of observed, predicted, and residual values.
- **(e)** Plot the residual plot of Y vs. $Y - \widehat{Y}$. Interpret your plot.
- **(f)** Plot the normal probability plot of the residual values. Do the residuals seem reasonably normal?

11.1 The 94 sedentary males with treadmill tests of Problems 9.9 to 9.12 are considered here. The dependent and independent variables were $Y = \text{VO}_2\ _{\text{MAX}}$, X_1 = duration, X_2 = maximum heart rate, X_3 = height, X_4 = weight.

Constant or Covariate	b_j	SE(b_j)
X_1	0.0510	0.00416
X_2	0.0191	0.0258
X_3	−0.0320	0.0444
X_4	0.0089	0.0520
Constant	2.89	11.17

Source	d.f.	SS	MS	F-Ratio
Regression	?	4314.69	?	?
Residual	?	?	?	
Total	?	5245.31		

Do tasks (a), (b-i), and (c). What is R^2?

11.2 The data of Mehta et al. [1981] used in Problems 9.13 to 9.22 are used here. The aorta platelet aggregation percent under dipyridamole, using epinephrine, was regressed on the control values in the aorta and coronary sinus. The results were:

Constant or Covariate	b_j	SE(b_j)
Aorta control	−0.0306	0.301
Coronary sinus control	0.768	0.195
Constant	15.90	

Source	d.f.	SS	MS	F-Ratio
Regression	?	?	?	?
Residual	?	231.21	?	
Total	?	1787.88		

Y	$\widehat{Y}$	Residual	Y	$\widehat{Y}$	Residual
89	81.58	7.42	69	?	?
45	?	?	83	88.15	−5.15
96	86.68	?	84	88.03	−4.03
70	?	2.34	85	88.92	−3.92

Do tasks (a), (b-ii), (c), (d), (e), and (f) [with small numbers of points, the interpretation in (e) and (f) is problematic].

11.3 This problem uses the 20 aortic valve surgery cases of Chapter 9; see the introduction to Problems 9.30 to 9.33. The response variable is the end diastolic volume adjusted for body size, EDVI. The two predictive variables are the EDVI before surgery and the systolic volume index, SVI, before surgery; Y = EDVI postoperatively, X_1 = EDVI preoperatively, and X_2 = SVI preoperatively. See the following tables and Table 11.16. Do tasks (a), (b-i), (c), (d), (f). Find R^2.

Constant or Covariate	b_j	SE(b_j)
X_1	0.889	0.155
X_2	−1.266	0.337
Constant	65.087	

Source	d.f.	SS	MS	F-Ratio
Regression	?	21,631.66	?	?
Residual	?	?	?	
Total	?	32,513.75		

Problems 11.4 to 11.7 refer to data of Hossack et al. [1980, 1981]. Ten normal men and 11 normal women were studied during a maximal exercise treadmill test. While being exercised they had a catheter (tube) inserted into the pulmonary (lung) artery and a short tube into the left radial or brachial artery. This allowed sampling and observation of

Table 11.16 Data for Problem 11.3

Y	$\widehat{Y}$	Residual	Y	$\widehat{Y}$	Residual
111	112.8	0.92	70	84.75	−14.75
56	?	?	149	165.13	−16.13
93	?	−39.99	55	?	?
160	148.78	11.22	91	88.89	2.11
111	?	5.76	118	103.56	−11.56
83	86.00	?	63	?	?
59	?	4.64	100	86.14	13.86
68	93.87	?	198	154.74	43.26
119	62.27	56.73	176	166.39	9.61
71	86.72	?			

arterial pressures and the oxygen content of the blood. From this, several parameters as described below were measured or calculated. The data for the 11 women are given in Table 11.17; the data for the 10 normal men are displayed in Table 11.18. Descriptions of the variables follow.

- *Activity:* a subject who routinely exercises three or more times per week until perspiring was active (Act); otherwise, the subject was sedentary (Sed).
- Wt: weight in kilograms.
- Ht: height in centimeters.
- $\mathrm{VO}_{2\mathrm{MAX}}$: oxygen (in millimeters per kilogram of body weight) used in 1 min at maximum exercise.
- FAI: functional aerobic impairment. For a patient's age and activity level (active or sedentary) the expected treadmill duration (ED) is estimated from a regression equation. The excess of observed duration (OD) to expected duration (ED) as a percentage of ED is the FAI. FAI = 100 × (OD − ED)/ED.
- $\dot{Q}_{MAX}$: output of the heart in liters of blood per minute at maximum.
- HR_{MAX}: heart rate in beats per minute at maximum exercise.
- SV_{MAX}: volume of blood pumped out of the heart in milliliters during each stroke (at maximum cardial output).
- CaO_2: oxygen content of the arterial system in milliliters of oxygen per liter of blood.
- $C\overline{v}O_2$: oxygen content of the venous (vein) system in milliliters of oxygen per liter of blood.
- $a\overline{v}O_2D_{MAX}$: difference in the oxygen content (in milliliters of oxygen per liter of blood) between the arterial system and the venous system (at maximum exercise); thus, $\mathrm{a\overline{v}O_2D_{MAX}} = \mathrm{CaO_2} - \mathrm{C\overline{v}O_2}$.
- $\overline{P}_{\mathrm{SA,\ MAX}}$: average pressure in the arterial system at the end of exercise in milliliters of mercury (mmHg).
- $\overline{P}_{\mathrm{PA,\ MAX}}$: average pressure in the pulmonary artery at the end of exercise in mmHg.

Table 11.17 Physical and Hemodynamic Variables in 11 Normal Women

Case	Activity	Age (yr)	Wt	Ht	$VO_{2\ MAX}$	FAI	Q_{MAX}	HR_{MAX}	SV_{MAX}	CaO_2	$C\bar{v}O_2$	$a\bar{v}O_2D_{MAX}$	$\overline{P}_{SA,MAX}$	$\overline{P}_{PA,MAX}$
1	Sed	45	63.2	163	28.81	−12	12.43	194	64	193	46	147	109	27
2	Sed	52	56.6	166	24.04	−3	12.19	158	87	181	73	108	137	16
3	Sed	43	65.0	155	26.66	−1	11.52	194	59	212	61	151	?	30
4	Sed	51	58.2	161	24.34	−3	10.78	188	63	173	41	132	154	15
5	Sed	61	74.1	167	21.42	−6	11.71	178	66	198	62	136	140	29
6	Sed	52	69.0	161	26.72	−15	12.89	188	72	193	50	143	125	30
7	Sed	60	50.9	166	23.74	−15	10.94	164	68	160	42	118	95	26
8	Sed	56	66.0	158	28.72	−31	13.93	184	81	168	52	136	148	21
9	Sed	56	66.0	165	20.77	6	10.25	166	62	171	53	118	102	27
10	Sed	51	64.3	168	24.77	−4	11.98	176	68	187	54	133	152	38
11	Act	28	55.5	160	47.72	−37	14.36	200	76	202	31	171	132	25
Mean		50.5	62.6	163	27.07	−11	12.09	181	70	187	51	136	129	26
SD		9.3	6.7	4.1	7.34	13	1.27	14	9	15	10	18	21	7

Table 11.18 Physical and Hemodynamic Variables in 10 Normal Men

Case	Age (yr)	Wt	Ht	VO_2 MAX	FAI	$\dot{Q}_{MAX}$	HR_{MAX}	SV_{MAX}	$\overline{P}_{SA,MAX}$	$\overline{P}_{PA,MAX}$
1	64	73.6	170	30.3	−4	13.4	156	85	114	24
2	61	90.9	191	27.1	12	17.8	156	115	104	30
3	38	76.8	180	44.4	5	19.4	190	102	100	24
4	62	92.7	185	24.6	18	15.8	173	91	78	33
5	59	92.0	183	41.2	−18	21.1	167	127	133	36
6	47	83.2	185	48.9	−20	22.4	173	132	160	22
7	24	69.8	178	62.1	−2	24.9	188	133	127	25
8	26	78.6	191	50.9	5	20.1	169	119	115	15
9	54	95.9	183	33.2	9	19.2	154	125	108	31
10	20	83.0	176	32.5	34	15.0	196	77	120	18
Mean	46	83.7	182	39.2	4	18.9	169	114	117	26
SD	17	8.9	7	12.0	16	3.5	21	25	22	7

11.4 For the 10 men, let $Y = VO_{2\ MAX}$, $X_1 =$ weight, $X_2 = HR_{MAX}$, and $X_3 = SV_{MAX}$. (In practice, one would not use three regression variables with only 10 data points. This is done here so that the small data set may be presented in its entirety.)

Constant or Covariate	b_j	SE(b_j)
Weight	−0.699	0.128
HR_{MAX}	0.289	0.078
SV_{MAX}	0.448	0.0511
Constant	−1.454	

Source	d.f.	SS	MS	F-Ratio
Regression	?	?	?	?
Residual	?	55.97	?	
Total	?	1305.08		

Y	$\widehat{Y}$	Residual	Y	$\widehat{Y}$	Residual
30.3	30.38	−0.08	48.9	?	−0.75
27.1	?	−4.64	62.1	63.80	−1.70
44.4	45.60	−1.20	50.9	45.88	?
24.6	24.65	?	33.2	32.15	1.05
41.2	39.53	1.67	32.5	?	?

Do tasks (a), (c), (d), (e), and (f).

11.5 After examining the normal probability plot of residuals, the regression of Problem 11.4 was rerun omitting cases 2 and 8. In this case we find:

Constant or Covariate	b_j	SE(b_j)
Weight	−0.615	0.039
HR_{MAX}	0.274	0.024
SV_{MAX}	0.436	0.015
Constant	−4.486	

Source	d.f.	SS	MS	F-Ratio
Regression	?	1017.98	?	?
Residual	?	?	?	
Total	?	1021.18		

Y	$\widehat{Y}$	Residual	Y	$\widehat{Y}$	Residual
30.3	?	?	48.9	49.35	?
44.4	?	−0.45	62.1	?	?
24.6	25.62	?	33.2	33.28	−0.08
41.2	?	1.09	32.5	31.77	0.73

Do tasks (a), (b-i), (c), (d), and (f). *Comment*: The very small residual (high R^2) indicates that the data are very likely highly "over fit." Compute R^2.

11.6 Selection of the regression variables of Problems 11.4 and 11.5 was based on Mallow's C_p plot. With so few cases, the multiple comparison problem looms large. As an independent verification, we try the result on the data of the 11 normal women. We find:

Constant or Covariate	b_j	SE(b_j)
Weight	−0.417	0.201
HR_{MAX}	0.441	0.098
SV_{MAX}	0.363	0.160
Constant	−51.96	

Source	d.f.	SS	MS	F-Ratio
Regression	?	419.96	?	?
Residual	?	117.13	?	
Total	?	?		

Y	$\widehat{Y}$	Residual	Y	$\widehat{Y}$	Residual
28.81	?	−1.75	23.72	23.89	−0.15
24.04	?	−1.72	28.72	31.14	−2.42
26.66	27.99	?	20.77	16.30	4.46
24.34	29.63	?	24.77	23.60	1.17
21.42	?	?	47.72	40.77	6.95
26.72	?	?			

Do tasks (a), (b-i), (c), (d), (e), and (f). Do (e) or (f) look suspicious? Why?

11.7 Do another run with the data of Problem 11.6 omitting the last point.

Constant or Covariate	b_j	SE(b_j)
Weight	−0.149	0.074
HR_{MAX}	0.233	0.042
SV_{MAX}	0.193	0.056
Constant	−20.52	

Source	**d.f.**	**SS**	**MS**	***F*-Ratio**
Regression	?	?	?	?
Residual	?	?	?	
Total	?	?		

Note the large change in the b_j's when omitting the outlier.

Y	$\widehat{Y}$	**Residual**	Y	$\widehat{Y}$	**Residual**
28.81	27.54	1.27	26.72	?	−0.11
24.04	24.59	−0.55	23.72	?	0.57
26.66	?	?	28.72	28.08	?
24.34	26.70	−2.36	20.77	20.23	?
21.42	?	?	24.77	23.96	0.81

Do tasks (a), (c), and (d). Find R^2. Do you think the female findings roughly support the results for the males?

11.8 Consider the regression of Y on $X_1, X_2, \ldots, X_6$. Which of the following five hypotheses are *nested* within other hypotheses?

$$H_1\colon \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

$$H_2\colon \beta_1 = \beta_5 = 0$$

$$H_3\colon \beta_1 = \beta_5$$

$$H_4\colon \beta_2 = \beta_5 = \beta_6 = 0$$

$$H_5\colon \beta_5 = 0$$

11.9 Consider a hypothesis H_1 nested within H_2. Let R_1^2 be the multiple correlation coefficient for H_1 and R_2^2 the multiple correlation coefficient for H_2. Suppose that there are n observations and H_2 regresses on Y and $X_1, \ldots, X_k$, while H_1 regresses Y only on the first j X_i's ($j < k$). Show that the F statistic for testing $\beta_{j+1} = \cdots = \beta_k = 0$ may be written as

$$F = \frac{(R_2^2 - R_1^2)/(k-j)}{(1 - R_2^2)/(n-k-1)}$$

Table 11.19 Simple Correlation Coefficients between Nine Variables for Black Men, United States, 1960–1962[a]

Variable	1	2	3	4	5	6	7	8	9
1. Height	—								
2. Weight	0.34	—							
3. Right triceps skinfold	−0.04	0.61	—						
4. Infrascapular skinfold	−0.05	0.72	0.72	—					
5. Arm girth	0.10	0.89	0.60	0.70	—				
6. Glucose	−0.20	−0.05	0.09	0.10	−0.03	—			
7. Cholesterol	−0.08	0.15	0.17	0.20	0.17	0.12	—		
8. Age	−0.23	−0.09	−0.05	0.02	−0.10	0.37	0.34	—	
9. Systolic blood pressure	−0.18	0.11	0.07	0.12	0.12	0.29	0.20	0.47	—
10. Diastolic blood pressure	−0.09	0.17	0.08	0.16	0.18	0.20	0.17	0.33	0.79

[a] Number of observations for samples: $N = 358$ and $\underline{N = 349}$. Figures underlined were derived from persons in the sample for whom glucose and cholesterol measurements were available.

Florey and Acheson [1969] studied blood pressure as it relates to physique, blood glucose, and serum cholesterol separately for males and females, blacks and whites. Table 11.19 presents sample correlation coefficients for black males on the following variables:

- *Height:* in inches
- *Weight:* in pounds
- *Right triceps skinfold:* in thickness in centimeters of skin folds on the back of the right arm, measured with standard calipers
- *Infrascapular skinfold:* skinfold thickness on the back below the tip of the right scapula
- *Arm girth:* circumference of the loose biceps
- *Glucose:* taken 1 hour after a challenge of 50 g of glucose in 250 cm^3 of water
- *Total serum cholesterol concentration*
- *Age:* in years
- *Systolic blood pressure* (mmHg)
- *Diastolic blood pressure* (mmHg)

An additional variable considered was the *ponderal index*, defined to be the height divided by the cube root of the weight. Note that the samples sizes varied because of a few uncollected blood specimens. For Problem 11.10, use $N = 349$.

11.10 Using the Florey and Acheson [1969] data above, the correlation squared of systolic blood pressure, variable 9, with the age and physical variables (variables 1, 2, 3, 4, 5, and 8) is 0.266. If we add variables 6 and 7, the blood glucose and cholesterol variables, R^2 increases to 0.281. Using the result of Problem 11.9, is this a statistically significant difference?

11.11 Suppose that the following description of a series of multiple regression runs was presented. Find any incorrect or inconsistent statements (if they occur). Forty-five people were given a battery of psychological tests. The dependent variable of self-image was analyzed by multiple regression analysis with five predictor variables: 1, tension index; 2, perception of success in life; 3, IQ; 4, aggression index; and 5, a hypochondriacal index. The multiple correlation with variables 1, 4, and 5 was -0.329, $p < 0.001$. When variables 2 and 3 were added to the predictive equation, $R^2 = 0.18$, $p > 0.05$. The relationship of self-image to the variables was complex; the correlation with variables 2 and 3 was low (0.03 and -0.09, respectively), but the multiple correlation of self-image with variables 2 and 3 was higher than expected, $R^2 = 0.22$, $p < 0.01$.

11.12 Using the definition of R^2 (Definition 11.4) and the multiple regression F test in Section 11.2.3, show that

$$R^2 = \frac{kF}{kF + n - k - 1}$$

and

$$F = \frac{(n - k - 1)R^2}{k(1 - R^2)}$$

Haynes et al. [1978] consider the relationship of psychological factors and coronary heart disease. As part of a long ongoing study of coronary heart disease, the Framingham study, from 1965 to 1967, questionnaires were given to 1822 individuals. Of particular interest was type A behavior. Roughly speaking, type A individuals feel considerable time pressure, are very driving and aggressive, and feel a need for perfection. Such behavior has been linked with coronary artery disease. The questions used in this study follow. The scales (indicated by the superscript numbers) are explained following the questions.

Psychosocial Scale and Items Used in the Framingham Study

Note: The superscript numbers in this list refer to the response sets that follow item 17.

1. Framingham type A behavior pattern:
   Traits and qualities which describe you:^{1}

   Being hard-driving and competitive
   Usually pressed for time
   Being bossy and dominating
   Having a strong need to excel in most things
   Eating too quickly

   Feeling at the end of an average day of work:

   Often felt very pressed for time
   Work stayed with you so you were thinking about it after hours
   Work often stretched you to the very limits of your energy and capacity
   Often felt uncertain, uncomfortable, or dissatisfied with how you were doing

   Do you get upset when you have to wait for anything?
2. Emotional lability:
   Traits and qualities which describe you:^{1}

   Having feelings easily hurt
   Getting angry very easily

Getting easily excited
Getting easily sad or depressed
Worrying about things more than necessary

Do you cry easily?
Are you easily embarrassed?
Are your feeling easily hurt?
Are you generally a high-strung person?
Are you usually self-conscious?
Are you easily upset?
Do you feel sometimes that you are about to go to pieces?
Are you generally calm and not easily upset?

3. Ambitiousness:
Traits and qualities which describe you:[1]

Being very socially ambitious
Being financially ambitious
Having a strong need to excel in most things

4. Noneasygoing:
Traits and qualities which describe you:[1]

Having a sense of humor
Being easygoing
Having ability to enjoy life

5. Nonsupport from boss:
Boss (the person directly above you):[2]

Is a person you can trust completely
Is cooperative
Is a person you can rely upon to carry his or her load
Is a person who appreciates you
Is a person who interferes with you or makes it difficult for you to get your work done
Is a person who generally lets you know how you stand
Is a person who takes a personal interest in you

6. Marital dissatisfaction:

Everything considered, how happy would you say that your marriage has been?[3]
Everything considered, how happy would you say that your spouse has found your marriage to be?[3]
About marriage, are you more satisfied, as satisfied, or less satisfied than most of your close friends are with their marriages?[4]

7. Marital disagreement:
How often do you and your spouse disagree about:[5]

Handling family finances or money matters
How to spend leisure time
Religious matters
Amount of time that should be spent together

Gambling
Sexual relations
Dealings with in-laws
On bringing up children
Where to live
Way of making a living
Household chores
Drinking

8. Work overload:
Regular line of work fairly often involves:[2]

Working overtime
Meeting deadlines or rigid time schedules

9. Aging worries:
Worry about:[6]

Growing old
Retirement
Sickness
Death
Loneliness

10. Personal worries:
Worry about:[6]

Sexual problems
Change of life
Money matters
Family problems
Not being a success

11. Tensions:
Often troubled by feelings of tenseness, tightness, restlessness, or inability to relax?[5]

Often bothered by nervousness or shaking?
Often have trouble sleeping or falling asleep?
Feel under a great deal of tension?
Have trouble relaxing?
Often have periods of restlessness so that you cannot sit for long?
Often felt difficulties were piling up too much for you to handle?

12. Reader's daily stress:
At the end of the day I am completely exhausted mentally and physically[1]

There is a great amount of nervous strain connected with my daily activities
My daily activities are extremely trying and stressful
In general I am usually tense and nervous

13. Anxiety symptoms:
Often become tired easily or feel continuously fatigued?[2]

Often have giddiness or dizziness or a feeling of unsteadiness?

Often have palpitations, or a pounding or racing heart?

Often bothered by breathlessness, sighing respiration or difficulty in getting a deep breath?

Often have poor concentration or vagueness in thinking?

14. Anger symptoms:
When really angry or annoyed:[7]

Get tense or worried
Get a headache
Feel weak
Feel depressed
Get nervous or shaky

15. Anger-in:
When really angry or annoyed:[7]

Try to act as though nothing much happened
Keep it to yourself
Apologize even though you are right

16. Anger-out:
When really angry or annoyed:[7]

Take it out on others
Blame someone else

17. Anger-discuss:
When really angry or annoyed:[7]

Get it off your chest
Talk to a friend or relative

Response Sets

1. Very well, fairly well, somewhat, not at all
2. Yes, no
3. Very happy, happy, average, unhappy, very unhappy
4. More satisfied, as satisfied, less satisfied
5. Often, once in a while, never
6. A great deal, somewhat, a little, not at all
7. Very likely, somewhat likely, not too likely

The correlations between the indices are reported in Table 11.20.

11.13 We use the Haynes et al. [1978] data of Table 11.20. The multiple correlation squared of the Framingham type A variable with all 16 of the other variables is 0.424. Note the high correlations for variables 2, 3, 14, 15, and 17.

$$R^2_{1(2,3,14,15,17)} = 0.352$$

Table 11.20 Correlations among 17 Framingham Psychosocial Scales with Continuous Distributions

Psychosocial Scales	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Framingham type A		0.43	0.31	0.09	0.12	0.23	0.29	0.06	0.27	0.32	−0.04	0.19	0.11	0.47	0.42	0.24	0.34
2. Emotional lability			0.12	0.26	0.08	0.05	0.21	0.12	0.37	0.31	0.10	0.23	0.11	0.43	0.61	0.42	0.60
3. Ambitiousness				−0.23	0.01	0.01	−0.05	−0.04	0.04	0.06	0.08	0.03	0.09	0.12	0.06	−0.01	0.07
4. Noneasygoing					0.05	0.03	0.15	0.22	0.18	0.17	−0.12	0.16	0.00	0.19	0.22	0.17	0.18
5. Nonsupport from boss						0.11	0.11	−0.01	0.09	0.10	−0.06	−0.01	−0.02	0.12	0.10	0.06	0.06
6. Work overload							0.11	−0.07	0.04	0.06	−0.03	−0.07	0.04	0.15	0.11	0.02	0.06
7. Marital disagreement								0.44	0.33	0.47	−0.08	0.15	−0.01	0.21	0.22	0.18	0.19
8. Marital dissatisfaction									0.12	0.25	0.00	0.02	−0.02	0.11	0.12	0.13	0.13
9. Aging worries										0.53	0.01	0.16	0.04	0.27	0.33	0.29	0.31
10. Personal worries											−0.05	0.19	0.03	0.31	0.33	0.21	0.31
11. Anger-in												−0.18	−0.07	0.06	0.11	0.12	0.18
12. Anger-out													0.11	0.11	0.13	0.09	0.19
13. Anger-discuss														0.08	0.10	0.06	0.12
14. Daily stress															0.51	0.34	0.41
15. Tension																0.49	0.61
16. Anxiety symptoms																	0.45
17. Anger symptoms																	

Source: Data from Haynes et al. [1978].

(a) Is there a statistically significant ($p < 0.05$) gain in R^2 by adding the remainder of the variables?

(b) Find the partial correlation of variables 1 and 2 after adjusting for variable 15. That is, what is the correlation of the Framingham type A index and emotional lability if adjustment is made for the amount of tension?

Stoudt et al. [1970] report on the relationship between certain body size measurements and anthropometric indices. As one would expect, there is considerable correlation among such measurements. The details of the measurements are reported in the reference above. The correlation for women are given in Table 11.21.

11.14 This problem deals with partial correlations.

(a) For the Stoudt et al. [1970] data, the multiple correlation of seat breadth with height and weight is 0.64826. Find

$$r_{\text{seat breadth, height.weight}} \quad \text{and} \quad r_{\text{seat breadth, weight.height}}$$

(b) The Florey and Acheson [1969] data show that the partial multiple correlation between systolic blood pressure and the two predictor variables glucose and cholesterol adjusting for the weight and measurement variables is

$$R^2_{9(6,7).1,2,3,4,5,8} = 0.207, \qquad R = 0.144$$

What are the numerator and denominator degrees of freedom for testing statistical significance? What is (approximately) the 0.05 (0.01) critical value? Find F in terms of R^2. Do we reject the null hypothesis of no correlation at the 5% (1%) level?

11.15 Suppose that you want to regress Y on $X_1, X_2, \ldots, X_8$. There are 73 observations. Suppose that you are given the following sums of squares:

$$\text{SS}_{\text{TOTAL}}, \quad \text{SS}_{\text{REG}}(X_1), \quad \text{SS}_{\text{REG}}(X_4), \quad \text{SS}_{\text{REG}}(X_1, X_5),$$
$$\text{SS}_{\text{REG}}(X_3, X_6), \quad \text{SS}_{\text{REG}}(X_7, X_8), \quad \text{SS}_{\text{REG}}(X_1, X_5, X_6),$$
$$\text{SS}_{\text{REG}}(X_1, X_3, X_6), \quad \text{SS}_{\text{REG}}(X_4, X_7, X_8), \quad \text{SS}_{\text{REG}}(X_3, X_5, X_6, X_8),$$
$$\text{SS}_{\text{REG}}(X_3, X_4, X_7, X_8), \quad \text{SS}_{\text{REG}}(X_3, X_5, X_6, X_7, X_8)$$

For each of the following: (1) state that the quantity cannot be estimated, or (2) show (a) how to compute the quantity in terms of the sums of squares, and (b) give the F-statistic in terms of the sums of squares, and give the degrees of freedom.

(a) r^2_{Y,X_3}

(b) $R^2_{Y(X_1,X_5,X_6)}$

(c) $R^2_{Y(X_1,X_5,X_6).X_3}$

(d) $R^2_{Y(X_3,X_4,X_7,X_8)}$

(e) $r^2_{Y,X_6.X_1,X_5}$

(f) $R^2_{Y(X_5,X_6).X_3,X_4}$

(g) $R^2_{Y(X_3,X_4).X_7,X_8}$

(h) $R^2_{Y(X_3,X_5,X_6).X_7,X_8}$

Table 11.21 Correlations for Women Regarding Body Size

Body Measurement	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Sitting height, erect	0.907	0.440	0.364	0.585	0.209	0.347	0.231	−0.032	0.204	0.350	0.059	−0.076	0.052	0.057	−0.063	0.772	0.197	−0.339
2. Sitting height, normal		0.420	0.352	0.533	0.199	0.327	0.230	−0.029	0.197	0.317	0.045	−0.091	0.034	0.064	−0.063	0.729	0.165	−0.300
3. Knee height			0.747	0.023	0.196	0.689	0.585	0.106	0.254	0.406	0.180	−0.121	0.128	0.100	0.041	0.782	0.322	−0.128
4. Popliteal height				−0.095	−0.141	0.429	0.387	−0.200	−0.101	0.255	−0.126	0.166	−0.219	−0.193	−0.248	0.723	−0.035	−0.196
5. Elbow rest height					0.293	0.051	−0.045	0.143	0.275	0.094	0.179	0.111	0.222	0.191	0.150	0.258	0.253	−0.177
6. Thigh clearance height						0.465	0.352	0.597	0.609	0.370	0.594	0.523	0.641	0.539	0.541	0.137	0.693	−0.026
7. Buttock–knee length							0.786	0.413	0.552	0.426	0.441	0.410	0.450	0.343	0.296	0.609	0.620	−0.036
8. Buttock–popliteal length								0.326	0.390	0.341	0.371	0.333	0.355	0.269	0.243	0.514	0.490	−0.005
9. Elbow to elbow breadth									0.696	0.331	0.878	0.870	0.835	0.619	0.751	−0.070	0.844	0.393
10. Seat breadth										0.327	0.680	0.666	0.746	0.614	0.596	0.137	0.805	0.187
11. Biacromial diameter											0.433	0.301	0.331	0.209	0.243	0.407	0.443	−0.116
12. Chest girth												0.862	0.843	0.615	0.762	0.016	0.882	0.317
13. Waist girth													0.803	0.589	0.747	−0.090	0.844	0.432
14. Right arm girth														0.740	0.774	−0.026	0.888	0.272
15. Right arm skinfold															0.755	−0.022	−0.641	0.203
16. Infrascapular skinfold																−0.136	0.729	0.278
17. Height																	0.189	−0.289
18. Weight																		0.204
19. Age																		

Source: Data from Stoudt et al. [1970].

11.16 Suppose that in the Framingham study [Haynes et al., 1978] we want to examine the relationship between type A behavior and anger (as given by the four anger variables). We would like to be sure that the relationship does not occur because of joint relationships with the other variables; that is, we want to adjust for all the variables other than type A (variable 1) and the anger variables 11, 12, 13, and 17.

(a) What quantity would you use to look at this?

(b) If the value (squared) is 0.019, what is the value of the F-statistic to test for significance? The degrees of freedom?

11.17 Suppose that using the Framingham data, we decide to examine emotional lability. We want to see how it is related to four areas characterized by variables as follows:

Work :	variables 5 and 6
Worry and anxiety :	variables 9, 10, and 16
Anger :	variables 11, 12, 13, and 17
Stress and tension :	variables 14 and 15

(a) To get a rough idea of how much relationship one might expect, we calculate

$$R^2_{2(5,6,9,10,16,11,12,13,17,14,15)} = 0.49$$

(b) To see which group or groups of variables may be contributing the most to this relationship, we find

$$\begin{aligned} R^2_{2(5,6)} &= 0.01 \quad \text{work} \\ R^2_{2(9,10,16)} &= 0.26 \quad \text{worry/anxiety} \\ R^2_{2(11,12,13,17)} &= 0.38 \quad \text{anger} \\ R^2_{2(14,15)} &= 0.39 \quad \text{stress/tension} \end{aligned}$$

(c) As the two most promising set of variables were the anger and the stress/tension, we compute

$$R^2_{2(11,12,13,14,15,17)} = 0.48$$

(i) Might we find a better relationship (larger R^2) by working with indices such as the average score on variables 11, 12, 13, and 17 for the anger index? Why or why not?

(ii) After using the anger and stress/tension variables, is there statistical significance left in the relationship of lability and work and work/anxiety? What quantity would estimate this relationship? (In Chapter 14 we show some other ways to analyze these data.)

11.18 The Jensen et al. [1980] data of 19 subjects were used in Problems 9.23 to 9.29. Here we consider the data before training. The exercise $\text{VO}_{2,\ \text{MAX}}$ is to be regressed upon three variables.

$$Y = \text{VO}_{2,\ \text{MAX}}$$

$$X_1 = \text{maximal ejection fraction}$$

$$X_2 = \text{maximal heart rate}$$

$$X_3 = \text{maximal systolic blood pressure}$$

The residual mean square with all three variables in the model is 73.40. The residual sums of squares are:

$$\begin{aligned} SS_{\text{RESID}}(X_1, X_2) &= 1101.58 \\ SS_{\text{RESID}}(X_1, X_3) &= 1839.80 \\ SS_{\text{RESID}}(X_2, X_3) &= 1124.78 \\ SS_{\text{RESID}}(X_1) &= 1966.32 \\ SS_{\text{RESID}}(X_2) &= 1125.98 \\ SS_{\text{RESID}}(X_3) &= 1885.98 \end{aligned}$$

(a) For each model, compute C_p.

(b) Plot C_p vs. p and select the best model.

(c) Compute and plot the average mean square residual vs. p.

11.19 The 20 aortic valve cases of Problem 11.3 give the data about the values of C_p and the residual mean square as shown in Table 11.22.

Table 11.22 Mallow's C_p for Subset of Data from Example 11.3

Numbers of the Explanatory Variables	p	C_p	Residual Mean Square	Numbers of the Explanatory Variables	p	C_p	Residual Mean Square
None	1	14.28	886.99	2,4,5	4	2.29	468.36
				1,4,5		2.41	472.20
4	2	3.87	578.92	3,4,5		2.69	481.50
5		11.60	804.16	1,3,4		6.91	619.81
3		13.63	863.16	1,2,4		6.91	619.90
2		14.14	877.97	2,3,4		7.80	648.81
1		16.00	932.21	2,3,5		14.14	856.68
				1,3,5		14.40	866.45
4,5	3	0.72	454.10	1,2,5		14.45	866.75
1,4		4.94	584.23	1,2,3		15.21	891.72
2,4		5.82	611.35				
3,4		5.87	612.75	1,2,4,5	5	4.05	491.14
1,5		12.76	825.45	2,3,4,5		4.16	494.92
3,5		12.96	831.53	1,3,4,5		4.41	503.66
2,5		13.17	838.17	1,2,3,4		8.90	660.65
2,3		13.23	839.87	1,2,3,5		15.83	903.14
1,3		15.60	912.88				
1,2		15.96	924.03	1,2,3,4,5	6	6	524.37

(a) Plot Mallow's C_p plot and select the "best" model.

(b) Plot the average residual mean square vs. p. Is it useful in this context? Why or why not?

11.20 The blood pressure, physique, glucose, and serum cholesterol work of Florey and Acheson [1969] was mentioned above. The authors first tried using a variety of regression analyses. It was known that the relationship between age and blood pressure is often curvilinear, so an age^2 term was used as a potential predictor variable. After exploratory analyses, stepwise regression of blood pressure (systolic or diastolic) upon five variables (age, age^2, ponderal index, glucose, and cholesterol) was run. The four regressions (black and white, female and male) for systolic blood pressure are given in Tables 11.23 to 11.26. The "standard error of the estimate" is the estimate of σ^2 at each stage.

(a) For the black men, give the values of the partial F-statistics and the degrees of freedom as each variable entered the equation.

(b) Are the F values in part (a) significant at the 5% significance level?

(c) For a fixed ponderal index of 32 and a glucose level of 125 mg%, plot the regression curve for systolic blood pressure for white women aged 20 to 70.

(d) Can you determine the partial correlation of systolic blood pressure and glucose adjusting for age in black women from these data? If so, give the value.

***(e)** Consider all the multiple regression R^2 values of systolic blood pressure with subsets of the five variables used. For white males and these data, give all possible

Table 11.23 Selected Regression Statistics for Systolic Blood Pressure and Selected Independent Variables of White Men, United States, 1960–1962[a]

		Multiple				
Step	Variables Entered	R	R^2	Increase in R^2	Regression Coefficient	Standard Error of Estimate
1	Age squared	0.439	0.193	0.193	0.0104	17.9551
2	Ponderal index	0.488	0.238	0.045	−6.1775	17.4471
3	Glucose	0.499	0.249	0.011	0.0500	17.3221
4	Cholesterol	0.503	0.253	0.004	0.0351	17.2859
5	Age	0.507	0.257	0.004	−0.5136	17.2386

[a] Dependent variable, systolic blood pressure. Constant term = 194.997; $N = 2599$.

Table 11.24 Selected Regression Statistics for Systolic Blood Pressure and Selected Independent Variables of Black Men, United States, 1960–1962[a]

		Multiple				
Step	Variables Entered	R	R^2	Increase in R^2	Regression Coefficient	Standard Error of Estimate
1	Age squared	0.474	0.225	0.225	0.6685	21.9399
2	Ponderal index	0.509	0.259	0.034	−6.4515	21.4769
3	Glucose	0.523	0.273	0.014	0.0734	21.3048

[a] Dependent variable = systolic blood pressure. Constant term = 180.252; $N = 349$.

Table 11.25 Selected Regression Statistics for Systolic Blood Pressure and Selected Independent Variables of White Women, United States, 1960–1962[a]

		Multiple				
Step	Variables Entered	R	R^2	Increase in R^2	Regression Coefficient	Standard Error of Estimate
1	Age squared	0.623	0.388	0.388	0.00821	18.9317
2	Ponderal index	0.667	0.445	0.057	−7.3925	18.0352
3	Glucose	0.676	0.457	0.012	0.0650	17.8445

[a] Dependent variable = systolic blood pressure. Constant term = 193.260; $N = 2931$.

Table 11.26 Selected Regression Statistics for Systolic Blood Pressure and Selected Independent Variables of Black Women, United States, 1960–1962[a]

		Multiple				
Step	Variables Entered	R	R^2	Increase in R^2	Regression Coefficient	Standard Error of Estimate
1	Age squared	0.590	0.348	0.348	0.9318	24.9930
2	Ponderal index	0.634	0.401	0.053	0.1388	23.9851
3	Glucose	0.656	0.430	0.029	−6.0723	23.4223

[a] Dependent variable = systolic blood pressure. Constant term = 153.149; $N = 443$.

inequalities that are *not* of the obvious form

$$R^2_{Y(X_{i_1},\ldots,X_{i_m})} \le R^2_{Y(X_{j_1},\ldots,X_{j_n})}$$

where $X_{i_1}, \ldots, X_{i_m}$ is a subset of $X_{j_1}, \ldots, X_{j_n}$.

11.21 From a correlation matrix it is possible to compute the order in which variables enter a stepwise multiple regression. The partial correlations, F statistics, and regression coefficients for the standardized variables (except for the constant) may be computed. The first 18 women's body dimension variables (as given in Stoudt et al. [1970] and mentioned above) were used. The dependent variable was weight, which we are trying to predict in terms of the 17 measured dimension variables. Because of the large sample size, it is "easy" to find statistical significance. In such cases the procedure is sometimes terminated while statistically significant predictor variables remain. In this case, the addition of predictor variables was stopped when R^2 would increase by less than 0.01 for the next variable. The variable numbers, the partial correlation with the dependent variable (conditioning upon variables in the predictive equation) for the variables not in the model, and the corresponding F-value for step 0 are given in Table 11.27, those for step 1 in Table 11.28, those for step 5 in Table 11.29, and those for the final step in Table 11.30.

- **(a)** Fill in the question marks in Tables 11.27 and 11.28.
- **(b)** Fill in the question marks in Table 11.29.
- **(c)** Fill in the question marks in Table 11.30.
- **(d)** Which variables entered the predictive equation?
- ***(e)** What can you say about the proportion of the variability in weight explained by the measurements?

Table 11.27 Values for Step 0[a]

var	PCORR	F-Ratio[a]	var	PCORR	F-Ratio[a]
1	0.1970	144.506	10	0.8050	6589.336
2	?	100.165	11	0.4430	873.872
3	0.3230	?	12	0.8820	12537.104
4	−0.0350	4.390	13	0.8440	8862.599
5	0.2530	244.755	14	0.8880	13346.507
6	0.6930	3306.990	15	0.6410	2496.173
7	0.6200	?	16	0.7290	4059.312
8	0.4900	1130.830	17	0.1890	132.581
9	?	8862.599			

[a] The F-statistics have 1 and 3579 d.f.

Table 11.28 Values for Step 1[a]

var	PCORR	F-Ratio[a]	var	PCORR	F-Ratio[a]
1	0.3284	432.622	9	0.4052	?
2	0.2933	?	10	0.4655	989.824
3	0.4568	943.565	11	0.3435	478.797
4	0.3554	517.351	12	0.5394	1467.962
5	0.1246	56.419	13	0.4778	1058.297
6	?	501.893	15	−0.0521	9.746
7	0.5367	1447.655	16	?	74.882
8	0.4065	708.359	17	0.4614	967.603

[a] The F-statistics have 1 and 3578 d.f.

Table 11.29 Values for Step 5[a]

var	PCORR	F-Ratio[a]	var	PCORR	F-Ratio[a]
1	?	323.056	8	0.0051	0.093
2	0.2285	196.834	9	0.0083	0.252
3	0.1623	96.676	11	0.1253	?
4	0.1157	48.503	15	−0.1298	61.260
5	?	183.520	16	−0.0149	?
6	0.2382	214.989	17	0.3131	388.536

[a] The F-statistics have 1 and ? d.f.

Table 11.30 Values for the Final Step[a]

var	PCORR	F-Ratio[a]	var	PCORR	F-Ratio[a]
1	?	5.600	8	−0.0178	1.143
2	−0.0289	2.994	9	0.0217	1.685
3	−0.0085	0.263	11	0.0043	0.067
4	−0.0172	1.062	15	−0.1607	94.635
5	0.0559	?	16	−0.0034	0.042

[a] The F-statistics have 1 and 3572 d.f.

(f) What can you say about the p-value of the next variable that would have entered the stepwise equation? (Note that this small p has less than 0.01 gain in R^2 if entered into the predictive equation.)

11.22 Data from Hossack et al. [1980, 1981] for men and women (Problems 11.4 to 11.7) were combined. The maximal cardiac output, Q_{DOT}, was regressed on the maximal oxygen uptake, $\mathrm{VO}_{2\ \mathrm{MAX}}$. From other work, the possibility of a curvilinear relationship was entertained. Polynomials of the zeroth, first, second, and third degree (or highest power of X) were considered. Portions of the BMDP output are presented below, with appropriate questions (see Figures 11.17 to 11.19).

(a) *Goodness-of-fit test*: For the polynomial of each degree, a test is made for additional information in the orthogonal polynomials of higher degree, with data as shown in Table 11.31. The numerator sum of squares for each of these tests is the sum of squares attributed to all orthogonal polynomials of higher degree,

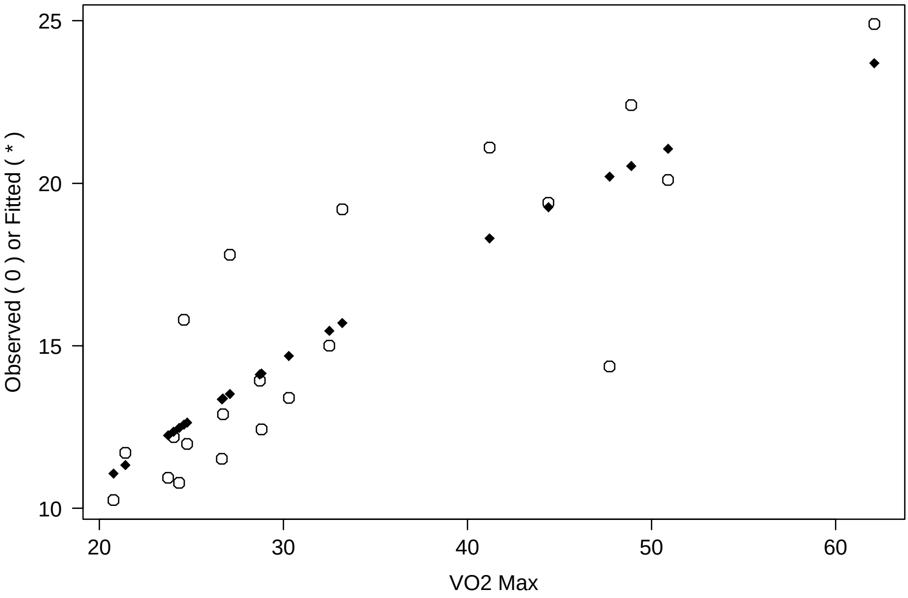

Figure 11.17 Polynomial regression of QDOT on VO2 $_{\text{MAX}}$. Figure for Problem 11.22.

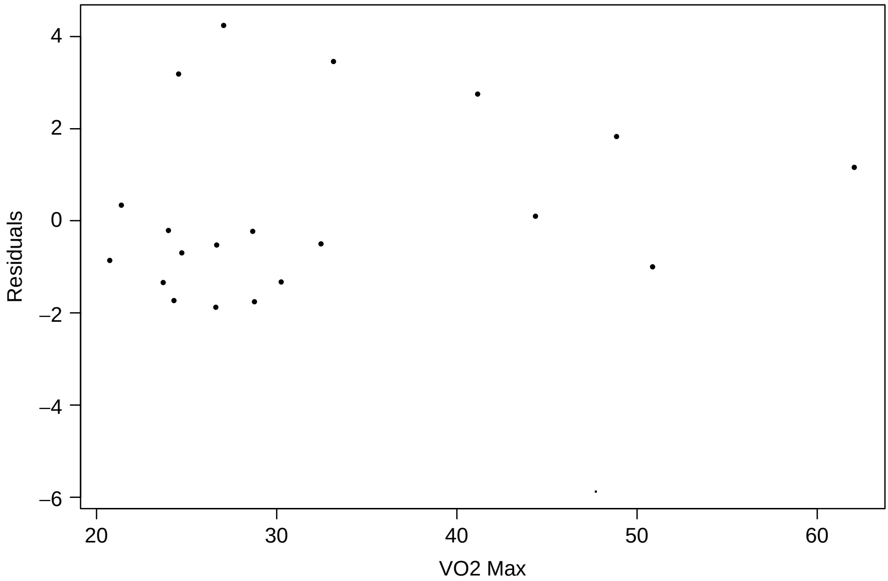

Figure 11.18 Figure for Problem 11.22.

and the denominator sum of squares is the residual sum of squares from the fit to the highest-degree polynomial (fit to all orthogonal polynomials). A significant F-statistic thus indicates that a higher-degree polynomial should be considered. What degree polynomial appears most appropriate? Why do the degrees of freedom in Table 11.31 add up to more than the total number of observations (21)?

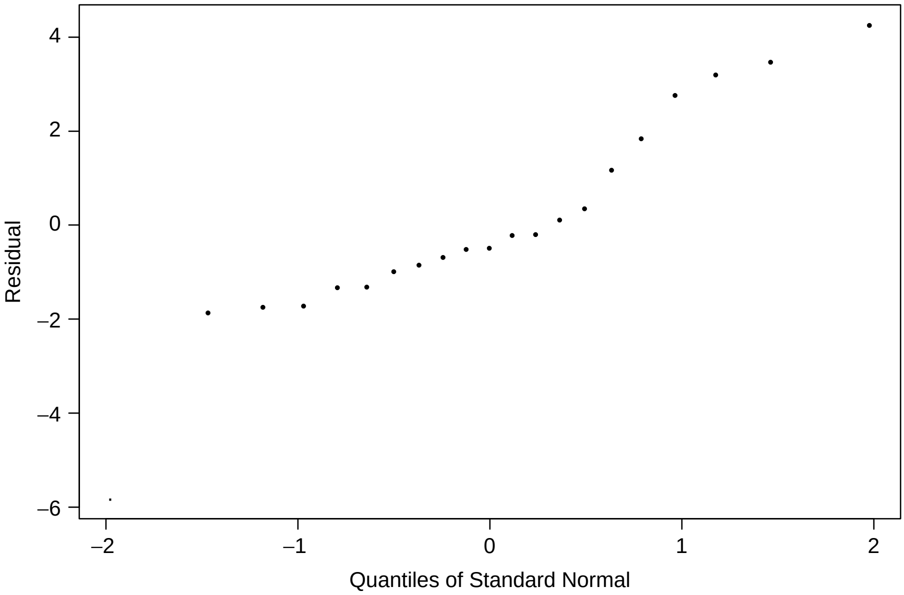

Figure 11.19 Figure for Problem 11.22.

Table 11.31 Goodness of Fit for Figure 11.22

Degree	SS	d.f.	MS	F-Ratio	Tail Probability
0	278.50622	4	69.62656	12.04	0.00
1	12.23208	3	4.07736	0.70	0.56
2	10.58430	2	5.29215	0.91	0.42
3	5.22112	1	5.22112	0.90	0.36
Residual	92.55383	16	5.78461		

(b) For a linear equation, the coefficients, observed and predicted values, residual plot, and normal residual are:

Degree	Regression Coefficient	Standard Error	t-Value
0	4.88737	1.58881	3.08
1	0.31670	0.04558	6.95

What would you conclude from the normal probability plot? Is the most outlying point a male or female? Which subject number in its table?

(c) For those with access to a polynomial regression program: Rerun the problem, removing the outlying point.

11.23 As in Problem 11.22, this problem deals with a potential polynomial regression equation. Weight and height were collected from a sample of the U.S. population in surveys done in

Table 11.32 Weight by Height Distribution for Men 25–34 Years of Age, Health Examination Survey, 1960–1962[a]

	Number of Examinees at Weight (lb)										
Height (in.)	Total	Under 130	130–139	140–149	150–159	160–169	170–179	180–189	190–199	200–209	210+
Total	675	39	50	78	93	92	87	74	56	48	58
<63	11	3	2	2	4	—	—	—	—	—	—
63	11	2	2	1	4	1	1	—	—	—	—
64	34	10	4	5	5	4	3	1	—	1	1
65	28	6	3	—	7	2	6	1	—	2	1
66	67	6	7	8	11	14	9	2	5	2	3
67	70	4	6	17	9	11	5	5	5	5	3
68	120	5	14	18	25	11	13	13	12	5	4
69	80	1	5	9	10	11	14	11	8	8	3
70	103	2	4	9	9	17	16	14	9	8	15
71	48	—	1	5	4	7	7	7	4	5	8
72	57	—	2	2	4	8	8	8	9	5	11
≥73	46	—	—	2	1	6	5	12	4	7	9

[a] Height without shoes; weight partially clothed; clothing weight estimated as averaging 2 (lb).

Table 11.33 Number of Men Aged 25–34 Years by Weight for Height; United States, 1971–1974[a]

		Number of Examinees at Weight (lb)									
Height (in.)	Total	Under 130	130–139	140–149	150–159	160–169	170–179	180–189	190–199	200–209	210+
Total	804	33	54	86	129	102	103	84	72	42	99
<63	6	1	3	1	—	—	1	—	—	—	—
63	17	4	3	5	3	—	—	—	1	—	1
64	23	3	5	8	2	1	1	1	1	1	—
65	41	5	6	7	11	3	3	1	2	2	1
66	70	5	10	11	11	10	9	5	6	2	1
67	86	3	10	6	19	15	11	9	5	4	4
68	92	5	4	15	12	15	14	13	7	2	5
69	120	3	5	10	26	17	22	8	10	4	15
70	112	2	5	12	15	14	11	18	13	10	12
71	73	2	1	8	14	10	8	7	13	1	9
72	69	—	2	1	10	9	8	9	5	6	19
≥73	95	—	—	2	2	8	15	13	9	10	32

[a] Height without shoes; weight partially clothed; clothing weight estimated as averaging 2 (lb).

Table 11.34 Coefficients and t-values for Problem 11.23

Degree	Regression Coefficient	Standard Error	t-Value
0	61.04225	0.60868	100.29
1	0.04408	0.00355	12.40
0	50.89825	3.85106	13.22
1	0.16548	0.04565	3.62
2	−0.00036	0.00013	−2.67
0	34.30283	25.84667	1.33
1	0.46766	0.46760	1.00
2	−0.00216	0.00278	−0.78
3	0.00000	0.00001	0.65

1960–1962 [Roberts, 1966] and in 1971–1974 [Abraham et al., 1979]. The data for males 25 to 34 years of age are given in Tables 11.32 and 11.33. In this problem we use only the 1960–1962 data. Both data sets are used in Problem 11.36. The weight categories were coded as values 124.5, 134.5, ... , 204.5, 214.5 and the height categories as 62, 63, ... , 72, 73. The contingency table was replaced by 675 "observations." As before, we present some of the results from a BMDP computer output. The height was regressed upon weight.

(a) *Goodness-of-Fit Test:* For the polynomial of each degree, a test is made for additional information in the orthogonal polynomials of higher degree. The numerator sum of squares attributed to all orthogonal polynomials of higher degree and the denominator sum of squares is the residual sum of squares from the fit to the highest-degree polynomial (fit to all polynomials). A significant F-statistic thus indicates that a higher-degree polynomial should be considered.

Degree	**SS**	**d.f.**	**MS**	**F-Ratio**	**Tail Probability**
0	900.86747	3	300.28916	54.23	0.00
1	41.69944	2	20.84972	3.77	0.02
2	2.33486	1	2.33486	0.42	0.52
Residual	3715.83771	671	5.53776		

Which degree polynomial appears most satisfactory?

(b) Coefficients with corresponding t-statistics are given in Table 11.34 for the first-, second-, and third-degree polynomials. Does this confirm the results of part (a)? How can the second-order term be significant for the second-degree polynomial, but neither the second or third power has a statistically significant coefficient when a third-order polynomial is used?

(c) The normal probability plot of residuals for the second-degree polynomials is shown in Figure 11.20. What does the tail behavior indicate (as compared to normal tails)? Think about how we obtained those data and how they were generated. Can you explain this phenomenon? This may account for the findings. The original data would be needed to evaluate the extent of this problem.

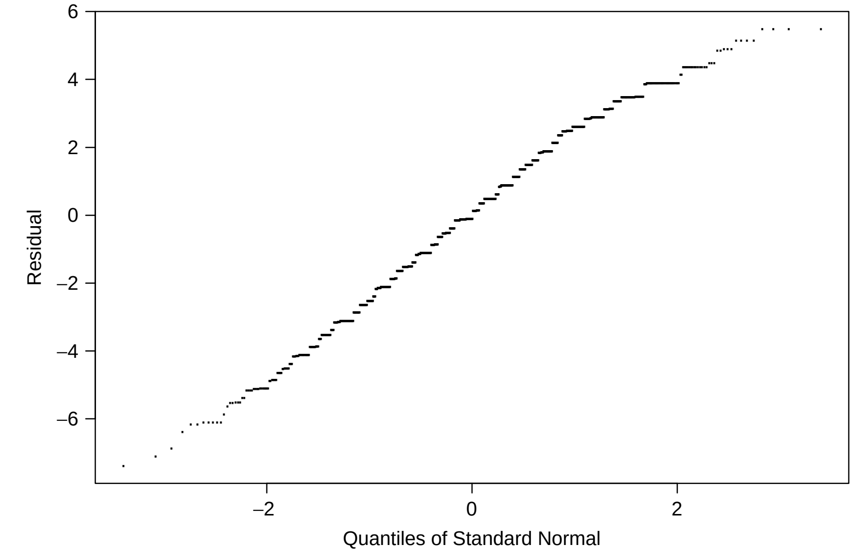

Figure 11.20 Normal probability plot of residuals of degree 2. Figure for Problem 11.23.

Table 11.35 Data for Problems 11.24 to 11.29

Indices of Variables in the Multiple Regression Equation (SS_{TOTAL})	Regression Sum of Squares SS_{REG} ($SS_{\text{TOTAL}} = 32513.75$)	Indices of Variables in the Multiple Regression Equation (SS_{TOTAL})	Regression Sum of Squares SS_{REG} ($SS_{\text{TOTAL}} = 32513.75$)
1	671.04	1,5	2,397.10
2	926.11	2,3	2,547.67
3	1,366.28	2,4	12,619.61
4	12,619.27	2,5	1,145.53
5	658.21	3,4	13,090.47
1,2	1,607.06	3,5	2,066.16
1,3	1,620.17	4,5	21,631.66
1,4	14,973.55		

Most multiple regression analyses (other than examining fit and model assumptions) use sums of squares rather than the original data. Problems 11.24 to 11.29 illustrate this point. The problems and the data in Table 11.35 are based on the 20 aortic valve surgery cases of Chapter 9 (see the introduction to Problems 9.30 to 9.33); Problem 11.3 uses these data. We consider the regression sums of squares for all possible subsets of five predictor variables. Here Y = EDVI postoperative, X_1 = age in years, X_2 = heart rate, X_3 = systolic blood pressure, X_4 = EDVI preoperative, X_5 = SVI preoperative.

11.24 From the regression sums of squares, compute and plot C_p-values for the smallest C_p-value for each p (i.e., for the largest SS_{REG}). Plot these values. Which model appears best?

11.25 From the regression sums of squares, perform a step-up stepwise regression. Use the 0.05 significance level to stop adding variables. Which variables are in the final model?

***11.26** From the regression sums of squares, perform a *stepdown* stepwise regression. Use the 0.10 significance level to stop removing variables. What is your final model?

11.27 Compute the following multiple correlation coefficients:

$$R_{Y(X_4,X_5)}, \qquad R_{Y(X_1,X_2,X_3,X_4,X_5)}, \qquad R_{Y(X_1,X_2,X_3)}$$

Which are statistically significant at the 0.05 significance level?

11.28 Compute the following squared partial correlation coefficients and test their statistical significance at the 1% level.

$$r^2_{Y,X_4 \cdot X_1,X_2,X_3,X_5}, \qquad r^2_{Y,X_5 \cdot X_1,X_2,X_3,X_4}$$

11.29 Compute the following partial multiple correlation coefficients and test their statistical significance at the 5% significance level.

$$R_{Y(X_4,X_5) \cdot X_1,X_2,X_3}, \qquad R_{Y(X_1,X_2,X_3,X_4) \cdot X_5}$$

Data on the 94 sedentary males of Problems 9.9 to 9.12 are used here. The dependent variable was age. The idea is to find an equation that predicted age; this equation might give an approximation to an "exercise age." Subjects might be encouraged, or convinced, to exercise if they heard a statement such as "Mr. Jones, although you are 28, your exercise performance is that of a 43-year-old sedentary man." The potential predictor variables with the regression sum of squares is given below for all combinations.

$$Y = \text{age in years}, \qquad X_1 = \text{duration in seconds}$$

$$X_2 = \text{VO}_2\ \text{MAX}, \qquad X_3 = \text{heart rate in beats/minute}$$

$$X_4 = \text{height in centimeters}, \qquad X_5 = \text{weight in kilograms}$$

$$\text{SS}_{\text{TOTAL}} = 11,395.74$$

Problems 11.30 to 11.35 are based on the data listed in Table 11.36.

11.30 Compute and plot for each p, the smallest C_p-value. Which predictive model would you choose?

11.31 At the 10% significance level, perform stepwise regression (do not compute the regression coefficients) selecting variables. Which variables are in the final model? How does this compare to the answer to Problem 11.30?

***11.32** At the 0.01 significance level, select variables using a *step-down* regression equation (no coefficients computed).

11.33 What are the values of the following correlation and multiple coefficients? Are they significantly nonzero at the 5% significance level?

$$R_{Y(X_1,X_2)}, \qquad R_{Y(X_3,X_4,X_5)},$$

$$R_{YX_1}, \qquad R_{YX_2}, \qquad R_{Y(X_4,X_5)}$$

Table 11.36 Data for Problems 11.30 to 11.35

Indexes of Variables in Multiple Regression Equation	Regression Sum of Squares SS_{REG}	Indexes of Variables in Multiple Regression Equation	Regression Sum of Squares SS_{REG}
1	5382.81	1,2,4	5658.66
2	4900.82	1,2,5	5777.12
3	4527.51	1,3,4	6097.58
4	295.26	1,3,5	6151.91
5	54.80	1,4,5	5723.50
1,2	5454.48	2,3,4	5851.44
1,3	5953.18	2,3,5	5923.41
1,4	5597.08	2,4,5	5243.27
1,5	5685.88	3,4,5	4630.28
2,3	5731.40	1,2,3,4	6128.27
2,4	5089.15	1,2,3,5	6201.39
2,5	5221.73	1,2,4,5	5805.06
3,4	4628.83	1,3,4,5	6179.52
3,5	4568.73	2,3,4,5	5940.03
4,5	299.81	1,2,3,4,5	6223.12
1,2,3	5988.09		

11.34 Compute the following squares of partial correlation coefficients. Are they statistically significant at the 0.10 level?

$$r^2_{Y,X_1\cdot X_2}, \qquad r^2_{Y,X_2\cdot X_1}, \qquad r^2_{Y,X_3X_1\cdot X_2}$$

Describe these quantities in words.

11.35 Compute the following partial multiple correlation coefficients. Are they significant at the 5% level?

$$R_{Y(X_1,X_2,X_3)X_4\cdot X_5}, \qquad R_{Y(X_1,X_3)\cdot X_2},$$
$$R_{Y(X_2,X_3)\cdot X_1}, \qquad R_{Y(X_1,X_2)\cdot X_3}$$

Problems 11.36 and 11.38 are analysis of covariance problems. They use BMDP computer output, which is addressed in more detail in the first problem. This problem should be done before Problem 11.38.

11.36 This problem uses the height and weight data of 25 to 34-year-old men as measured in 1960–1962 and 1971–1974 samples of the U.S. populations. These data are described and presented in Problem 11.23.

(a) The groups are defined by a year variable taking on the value 1 for the 1960 survey and the value 2 for the 1971 survey. Means for the data are:

		Estimates of Means		
		1960	**1971**	**Total**
Height	1	68.5081	68.9353	68.7403
Weight	2	169.3890	171.4030	170.4838

Which survey had the heaviest men? The tallest men? There are at least two possible explanations for weight gain: (1) the weight is increasing due to more overweight and/or building of body muscle; (2) the taller population naturally weighs more.

(b) To distinguish between two hypotheses, an analysis of covariance adjusting for height is performed. The analysis produced the following output, where the dependent variable is weight.

Covariate	**Regression Coefficient**	**Standard Error**	***t*-Value**
Height	4.22646	0.22742	18.58450

Group	***N***	**Group Mean**	**Adjusted Group Mean**	**Standard Error**
1960	675	169.38904	170.37045	0.89258
1971	804	171.40295	170.57901	0.91761

The ANOVA table is as follows:

Source	**d.f.**	**SS**	**MS**	***F*-Ratio**	**Tail Area Probability**
Equality of adjusted cell means	1	15.7500	15.7500	0.0294	0.8639
Zero slope	1	185,086.0000	185,086.0000	345.3833	0.0000
Error	1475	790,967.3750	535.8857		
Equality of slopes	1	0.1250	0.1250	0.0002	0.9878
Error	1475	790,967.2500	536.2490		

Data for the slope within each group:

		1960	**1971**
Height	1	4.2223	4.2298

The t-test matrix for adjusted group means on 1476 degrees of freedom looks as follows:

		1960	**1971**
1960	1	0.0000	
1971	2	0.1720	0.0000

The probabilities for the t-values above are:

	1960_1	**1971_2**
1960_1	1.0000	
1971_2	0.8634	1.0000

(i) Note the "equality of slopes" line of output. This gives the F-test for the equality of the slopes with the corresponding p-value. Is the hypothesis of the equality of the slopes feasible? If estimated separately, what are the two slopes?

(ii) The test for equal (rather than just parallel) regression lines in the groups corresponds to the line labeled "equality of adjusted cell means." Is there a statistically significant difference between the groups? What are the adjusted cell means? By how many pounds do the adjusted cell means differ? Does hypothesis (1) or (2) seem more plausible with these data?

(iii) A t-test for comparing each pair of groups is presented. The p-value 0.8643 is the same (to round off) as the F-statistic. This occurs because only two groups are compared.

11.37 The cases of Bruce et al. [1973] are used. We are interested in comparing $VO_{2,MAX}$, after adjusting for duration and age, in three groups: active males, sedentary males, and active females. The analysis gives the following results:

Number of Cases per Group	
ACTMALE	44
SEDMALE	94
ACTFEM	43
Total	181

The estimates of means is as follows:

		ACTMALE	**SEDMALE**	**ACTFEM**	**Total**
$VO_{2,\text{ MAX}}$	1	40.8046	35.6330	29.0535	35.3271
Duration	2	647.3864	577.1067	514.8837	579.4091
Age	3	47.2046	49.7872	45.1395	48.0553

Data are as follows when the dependent variable is $VO_{2,\text{ MAX}}$:

Covariate	**Regression Coefficient**	**Standard Error**	***t*-Value**
Duration	0.05242	0.00292	17.94199
Age	−0.06872	0.03160	−2.17507

Group	***N***	**Group Mean**	**Adjusted Group Mean**	**Standard Error**
ACTMALE	44	40.80456	37.18298	0.52933
SEDMALE	94	35.63297	35.87268	0.34391
ACTFEM	43	29.05349	32.23531	0.56614

The ANOVA table is:

Source	DF	SS	MS	F-Ratio	Tail Area Probability
Equality of adjusted cell means	2	422.8359	211.4180	19.4336	0.0000
Zero slope	2	7612.9980	3806.4990	349.6947	0.0000
Error	176	1914.7012	10.8790		
Equality of slopes	4	72.7058	18.1765	1.6973	0.1528
Error	172	1841.9954	10.7093		

Values of the slopes within each group are:

		ACTMALE	SEDMALE	ACTFEM
Duration	2	0.0552	0.0522	0.0411
Age	3	−0.1439	−0.0434	−0.1007

The t-test matrix for adjusted group means on 176 degrees of freedom looks as follows:

		ACTMALE	SEDMALE	ACTFEM
ACTMALE	1	0.0000		
SEDMALE	2	−2.1005	0.0000	
ACTFEM	3	−5.9627	−5.3662	0.0000

The probabilities for the t-values above are:

		ACTMALE	SEDMALE	ACTFEM
ACTMALE	1	1.0000		
SEDMALE	2	0.0371	1.0000	
ACTFEM	3	0.0000	0.0000	1.0000

(a) Are the slopes of the adjusting variables (covariates) statistically significant?

(b) Is the hypothesis of parallel regression equations (equal β's in the groups) tenable?

(c) Does the adjustment bring the group means closer together?

(d) After adjustment, is there a statistically significant difference between the groups?

(e) If the answer to part (d) is yes, which groups differ at the 10%, 5%, and 1% significance level?

11.38 This problem deals with the data of Example 10.7 presented in Tables 10.20, 10.21, and 10.22.

(a) Using the quadratic term of Table 10.21 correlate this term with height, weight, and age for the group of females and for the group of males. Are the correlations comparable?

(b) Do part (a) by setting up an appropriate regression analysis with dummy variables.

(c) Test whether gender makes a significant contribution to the regression model of part (b).

(d) Repeat the analyses for the linear and constant terms of Table 10.21.

(e) Do your conclusions differ from those of Example 10.7?

11.39 This problem examines the heart rate response in normal males and females as reported in Hossack et al. [1980, 1981]. As heart rate is related to age and the males were older, this was used as an adjustment covariate. The data are:

Number of Cases per Group	
Male	11
Female	10
Total	21

The estimates of means are:

		Male	Female	Total
Heart rate	1	180.9091	172.2000	176.7619
Age	2	50.4546	45.5000	48.0952

The dependent variable is heart rate:

Covariate	Regression Coefficient	Standard Error	*t*-Value
Age	−0.75515	0.17335	−4.35610

Group	*N*	Group Mean	Adjusted Group Mean	Standard Error
Male	11	180.90909	182.69070	3.12758
Female	10	172.19998	170.24017	3.28303

The ANOVA table:

Source	d.f.	SS	MS	*F*-Ratio	Tail Area Probability
Equality of adjusted cell means	1	783.3650	783.3650	7.4071	0.0140
Zero slope	1	2006.8464	2006.8464	18.9756	0.0004
Error	18	1903.6638	105.7591		
Equality of slopes	1	81.5415	81.5415	0.7608	0.3952
Error	17	1822.1223	107.1837		

The slopes within each group are:

Age	Male	Female
2	−1.0231	−0.6687

(a) Is it reasonable to assume equal age response in the two groups?

(b) Are the adjusted cell means closer or farther apart than the unadjusted cell means? Why?

(c) After adjustment what is the p-value for a difference between the two groups? Do men or women have a higher heart rate on maximal exercise (after age adjustment) in these data?

REFERENCES

Abraham, S., Johnson, C. L., and Najjar, M. F. [1979]. *Weight by Height and Age for Adults 18–74 Years: United States, 1971–1974.* Data from the National Health Survey, Series 11, No. 208. DHEW Publication (PHS) 79-1656. U.S. Government Printing Office, Washington, DC.

Blalock, H. M., Jr. (ed.) [1985]. *Causal Inferences in Nonexperimental Research.* de Gruyter, Aldine, Inc.

Boucher, C. A., Bingham, J. B., Osbakken, M. D., Okada, R. D., Strauss, H. W., Block, P. C., Levine, R. B., Phillips, H. R., and Pohost, G. B. [1981]. Early changes in left ventricular size and function after correction of left ventricular volume overload. *American Journal of Cardiology*, **47**: 991–1004.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. [1994]. *Time Series Analysis, Forecasting and Control.* Holden-Day, San Francisco, CA.

Bruce, R. A., Kusumi, F., and Hosmer, D. [1973]. Maximal oxygen intake and nomographic assessment of functional aerobic impairment in cardiovascular disease. *American Heart Journal*, **85**: 546–562.

Cook, T. D., Campbell, D. T., Stanley, J. C., and Shadish, W. [2001]. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Miflin, New York.

Cullen, B. F., and van Belle, G. [1975]. Lymphocyte transformation and changes in leukocyte count: effects of anesthesia and operation. *Anesthesiology*, **43**: 577–583. Used with permission of J. B. Lippincott Company.

Daniel, C., and Wood, F. S. [1999]. *Fitting Equations to Data*, 2nd ed. Wiley, New York.

Dixon, W. J. (chief ed.) [1988]. *BMDP-81 Statistical Software Manual*, BMDP 1988, Vols. 1 and 2. University of California Press, Berkeley, CA.

Draper, N. R., and Smith, H. [1998]. *Applied Regression Analysis*, 3rd ed. Wiley, New York.

Efron, B., and Tibshirani, R. [1986]. The bootstrap (with discussion), *Statistical Science*, **1**: 54–77.

Efron, B., and Tibshirani, R. [1994]. An Introduction to the Bootstrap. CRC Press, Boca Raton, FL.

Florey, C. du V., and Acheson, R. M. [1969]. Blood pressure as it relates to physique, blood glucose and cholesterol. Vital and Health Statistics. Data from the National Health Survey. Public Health Service Publication 1000, Ser. 11, No. 34. Washington, DC.

Gardner, M. J. [1973]. Using the environment to explain and predict mortality. *Journal of the Royal Statistical Society, Series A*, **136**: 421–440.

Goldberger, A. S., and Duncan, O. D. [1973]. *Structural Equation Models in the Social Sciences.* Elsevier, New York.

Graybill, F. A. [2000]. *Theory and Application of the Linear Model.* Brooks/Cole, Pacific Grove, CA.

Haynes, S. G., Levine, S., Scotch, N., Feinleib, M., and Kannel, W. B. [1978]. The relationship of psychosocial factors to coronary heart disease in the Framingham study. *American Journal of Epidemiology*, **107**: 362–283.

Hocking, R. R. [1976]. The analysis and selection of variables in linear regression. *Biometrics*, **32**: 1–50.

Hossack, K. F., Bruce, R. A., Green, B., Kusumi, F., DeRouen, T. A., and Trimble, S. [1980]. Maximal cardiac output during upright exercise: approximate normal standards and variations with coronary heart disease. *American Journal of Cardiology*, **46**: 204–212.

Hossack, K. F., Kusumi, F., and Bruce, R. A. [1981]. Approximate normal standards of maximal cardiac output during upright exercise in women. *American Journal of Cardiology*, **47**: 1080–1086.

Hurvich, C. M., and Tsai, C.-L. [1990]. The impact of model selection on inference in linear regression. *American Statistician*, **44**: 214–217.

Jensen, D., Atwood, J. E., Frolicher, V., McKirnan, M. D., Battler, A., Ashburn, W., and Ross, J., Jr. [1980]. Improvement in ventricular function during exercise studied with radionuclide ventriculography after cardiac rehabilitation. *American Journal of Cardiology*, **46**: 770–777.

Kaplan, D. [2000]. *Structural Equations Modeling*. Sage Publications.

Keller, R. B., Atlas, S. J., Singer, D. E., Chapin, A. M., Mooney, N. A., Patrick, D. L., and Deyo, R. A. [1996]. The Maine lumbar spine study: I. Background and concepts. *Spine*, **21**: 1769–1776.

Kleinbaum, D. G. [1994]. *Logistic Regression: A Self-Learning Text*. Springer-Verlag, New York.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., and Nizam A. [1998]. *Applied Regression Analysis and Other Multivariate Methods*, 3rd ed. Duxbury Press, North Scituate, MA.

Li, C. C. [1975]. *Path Analysis: A Primer*. Boxwood Press, Pacific Grove, CA.

Little, R. J., and Rubin, D. B. [2000]. Causal effects in clinical and epidemiologic studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, **21**: 121–145.

Maldonado, G., and Greenland, S. [1993]. Simulation study of confounder-selection strategies. *American Journal of Epidemiology*, **138**: 923–936.

Mason, R. L. [1975]. Regression analysis and problems of multicollinearity. *Communications in Statistics*, **4**: 277–292.

Mehta, J., Mehta, P., Pepine, C. J., and Conti, C. R. [1981]. Platelet function studies in coronary artery disease: X. Effects of dipyridamole. *American Journal of Cardiology*, **47**: 1111–1114.

Mickey, R. M., and Greenland, S. [1989]. The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology*, **129**: 125–137.

Morrison, D. F. [1990]. *Multivariate Statistical Methods*, 3rd ed. McGraw-Hill, New York.

Neyman, J. [1923]. On the application of probability theory to agricultural experiments: essay on principles, section 9. Translated in *Statistical Science*, 1990, **5**: 65–80.

Pearl, J. [2000]. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. [2002]. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, **21**: 2917–2930.

Raab, G. M., Day, S., and Sales, J. [2000]. How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials*, **21**: 330–342.

Roberts, J. [1966]. *Weight by Height and Age of Adults: United States, 1960–1962*. Vital and Health Statistics. Data from the National Health Survey. Public Health Service Publication 1000, Series 11, No. 14. U.S. Government Printing Office, Washington, DC.

Robins, J. M. [1986]. A new approach to causal inference in mortality studies with sustained exposure periods: application to the control of the healthy worker survivor effect. *Mathematical Modelling*, **7**: 1393–1512.

Rosenbaum, P. R., and Rubin, D. R. [1983]. The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**: 41–55.

Rothman, K. J., and Greenland, S. [1998]. *Modern Epidemiology*. Lippincott-Raven, Philadelphia.

Rubin, D. B. [1974]. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Education Psychology*, **66**: 688–701.

Stoudt, H. W., Damon, A., and McFarland, R. A. [1970]. *Skinfolds, Body Girths, Biacromial Diameter, and Selected Anthropometric Indices of Adults: United States, 1960–62*. Vital and Health Statistics. Data from the National Health Survey. Public Health Service Publication 1000, Series 11, No. 35. U.S. Government Printing Office, Washington, DC.

Sun, G.-W., Shook, T. L., and Kay, G. L. [1996]. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology*, **8**: 907–916.

Timm, N. H. [2001]. *Applied Multivariate Analysis*. Springer-Verlag, New York.

van Belle, G., Leurgans, S., Friel, P., Guo, S., and Yerby, M. [1989]. Determination of enzyme binding constants using generalized linear models, with particular reference to Michaelis–Menten models. *Journal of Pharmaceutical Science*, **78**: 413–416.

CHAPTER 12

Multiple Comparisons

12.1 INTRODUCTION

Most of us are aware of the large number of coincidences that appear in our lives. “Imagine meeting you here!” “The ticket number is the same as our street address.” One explanation of such phenomena is statistical. There are so many different things going on in our lives that a few events of small probability (the coincidences) are likely to happen at the same time. See Diaconis and Mosteller [1989] for methods for studying coincidences.

In a more formal setting, the same phenomenon can occur. If many tests or comparisons are carried out at the 0.05 significance level (with the null hypothesis holding in all cases), the probability of deciding that the null hypothesis may be rejected in one or more of the tests is considerably larger. If *many* 95% confidence intervals are set up, there is not 95% confidence that *all* parameters are “in” their confidence intervals. If many treatments are compared, each comparison at a given significance level, the overall probability of a mistake is much larger. If significance tests are done continually while data accumulate, stopping when statistical significance is reached, the significance level is much larger than the nominal “fixed sample size” significance level. The category of problems being discussed is called the *multiple comparison* problem: Many (or multiple) statistical procedures are being applied to the same data. We note that one of the most important practical cases of multiple comparisons, the interim monitoring of randomized trials, is discussed in Chapter 19.

This chapter provides a quantitative feeling for the problem. Statistical methods to handle the situation are also described. We first describe the multiple testing or multiple comparison problem in Section 12.2. In Section 12.3 we present three very common methods for obtaining simultaneous confidence intervals for the regression coefficients of a linear model. In Section 12.4 we discuss how to choose between them. The chapter concludes with notes and problems.

12.2 MULTIPLE COMPARISON PROBLEM

Suppose that n statistically independent tests are being considered in an experiment. Each test is evaluated at significance level α. Suppose that the null hypothesis holds in each case. What is the probability, α^*, of incorrectly rejecting the null hypothesis in one or more of the tests? For $n = 1$, the probability is α, by definition. Table 12.1 gives the probabilities for several values of α and n. Note that if each test is carried out at a 0.05 level, then for 20 tests, the probability is 0.64 of incorrectly rejecting at least one of the null hypotheses.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

Table 12.1 Probability, α*, of Rejecting One or More Null Hypotheses When n independent Tests Are Carried Out at Significance Level α and Each Null Hypothesis Is True

Number	α		
of Tests, n	0.01	0.05	0.10
1	0.01	0.05	0.10
2	0.02	0.10	0.19
3	0.03	0.14	0.27
4	0.04	0.19	0.34
5	0.05	0.23	0.41
6	0.06	0.26	0.47
7	0.07	0.30	0.52
8	0.08	0.34	0.57
9	0.09	0.37	0.61
10	0.10	0.40	0.65
20	0.18	0.64	0.88
50	0.39	0.92	0.99
100	0.63	0.99	1.00
1000	1.00	1.00	1.00

The table may also be related to confidence intervals. Suppose that each of $n100(1-\alpha)\%$ confidence intervals comes from an independent data set. The table gives the probability that one or more of the estimated parameters is not straddled by its confidence interval. For example, among five 90% confidence intervals, the probability is 0.41 that at least one of the confidence intervals does not straddle the parameter being estimated.

Now that we see the magnitude of the problem, what shall we do about it? One solution is to use a smaller α level for each test or confidence interval so that the probability of one or more mistakes over all n tests is the desired (nominal) significance level. Table 12.2 shows the α level needed for each test in order that the combined significance level, α^*, be as given at the column heading.

The values of α and α^* are related to each other by the equation

$$\alpha^* = 1 - (1-\alpha)^n \quad \text{or} \quad \alpha = 1 - (1-\alpha^*)^{1/n} \tag{1}$$

where $(1-\alpha)^{1/n}$ is the nth root of $1-\alpha$.

If p-values are being used without a formal significance level, the p-value from an individual test is adjusted by the opposite of equation (1). That is, p^*, the overall p-value, taking into account the fact that there are n tests, is given by

$$p^* = 1 - (1-p)^n \tag{2}$$

For example, if there are two tests and the p-value of each test is 0.05, the overall p-value is $p^* = 1 - (1-0.05)^2 = 0.0975$. For small values of α (or p) and n by the binominal expansion $\alpha^* = 1/n\alpha$ (and $p^* = np$), a relationship that will also be derived in the context of the Bonferroni inequality.

Before giving an example, we introduce some terminology and make a few comments. We consider an "experiment" in which n tests or comparisons are made.

Definition 12.1. The significance level at which each test or comparison is carried out in an experiment is called the *per comparison* error rate.

Table 12.2 Significance Level, α, Needed for Each Test or Confidence Interval So That the Overall Significance Level (Probability of One or More Mistakes) Is α^* When Each Null Hypothesis Is True

Number	α^*		
of Tests, n	0.01	0.05	0.10
1	0.010	0.05	0.10
2	0.005	0.0253	0.0513
3	0.00334	0.0170	0.0345
4	0.00251	0.0127	0.0260
5	0.00201	0.0102	0.0209
6	0.00167	0.00851	0.0174
7	0.00143	0.00730	0.0150
8	0.00126	0.00639	0.0131
9	0.00112	0.00568	0.0116
10	0.00100	0.00512	0.0105
20	0.00050	0.00256	0.00525
50	0.00020	0.00103	0.00210
100	0.00010	0.00051	0.00105
1000	0.00001	0.00005	0.00011

Definition 12.2. The probability of incorrectly rejecting at least one of the true null hypotheses in an experiment involving one or more tests or comparisons is called the *per experiment error rate*.

The terminology is less transparent than it seems. In particular, what defines an "experiment"? You could think of your life as an experiment involving many comparisons. If you wanted to restrict your "per experiment" error level to, say, $\alpha^* = 0.05$, you would need to carry out each of the comparisons at ridiculously low values of α. This has led some to question the entire idea of multiple comparison adjustment [Rothman, 1990; O'Brien, 1983; Proschan and Follman, 1995]. Frequently, groups of tests or comparisons form a natural unit and a suitable adjustment can be made. In some cases it is reasonable to control the total error rate only over tests that in some sense ask the same question.

Example 12.1. The liver carries out many complex biochemical tasks in the body. In particular, it modifies substances in the blood to make them easier to excrete. Because of this, it is very susceptible to damage by foreign substances that become more toxic as they are metabolized. As liver damage often causes no noticeable symptoms until far too late, biochemical tests for liver damage are very important in investigating new drugs or monitoring patients with liver disease. These include measuring substances produced by the healthy liver (e.g., albumin), substances removed by the healthy liver (e.g., bilirubin), and substances that are confined inside liver cells and so not found in the blood when the liver is healthy (e.g., transaminases).

It is easy to end up with half a dozen or more indicators of liver function, creating a multiple comparison problem if they are to be tested. Appropriate solutions to the problem vary with the intentions of the analyst. They might include:

1. *Controlling the Type I error rate.* If a deterioration in any of the indicators leads to the same qualitative conclusion — liver damage — they form a single hypothesis that deserves a single α.

2. *Controlling the Type II error rate*. When a new drug is first being tested, it is important not to miss even fairly rare liver damage. The safety monitoring program must have a low Type II error rate.
3. *Controlling Type I error over smaller groups*. Different indicators are sensitive to various types of liver damage. For a researcher interested in the mechanism of the toxicity, separating the indicators into these groups would be more appropriate.
4. *Combining the indicators*. In some cases the multiple comparison problem can be avoided by creating a composite outcome such as some sort of weighted sum of the indicators. This will typically increase power for alternatives where more than one indicator is expected to be affected.

The fact that different strategies are appropriate for different people suggests that it is useful to report p-values and confidence intervals without adjustment, perhaps in addition to adjusted versions.

Two of the key assumptions in the derivation of equations (1) and (2) are (1) statistical independence and (2) the null hypothesis being true for each comparison. In the next two sections we discuss their relevance and ways of dealing with these assumptions when controlling Type I error rates.

Example 12.2. To illustrate the methods, consider responses to maximal exercise testing within eight groups by Bruce et al. [1974]. The subjects were all males. An indication of exercise performance is functional aerobic impairment (FAI). This index is age- and gender-adjusted to compare the duration of the maximal treadmill test to that expected for a healthy person of the subject's age and gender. A larger score indicates more exercise impairment. Working at a 5% significance level, it is desired to compare the average levels in the eight groups. The data are shown in Table 12.3.

Because it was expected that the healthy group would have a smaller variance, a one-way ANOVA was not performed (in the next section you will see how to handle such problems). Instead, we construct eight *simultaneous* 95% confidence intervals. Hence, $\alpha = 1 - (1 - 0.05)^{1/8} \doteq 0.0064$ is to be the α-level for each interval. The intervals are given by

$$\overline{Y} \pm \frac{\text{SD}}{\sqrt{n}} t_{n-1,1-(0.0064/2)}$$

The t-values are estimated by interpolation from the table of t-critical values and the normal table ($n > 120$). The eight confidence intervals work out to be as shown in Table 12.4. Displaying these intervals graphically and indicating which group each interval belongs to gives Figure 12.1.

Table 12.3 Functional Aerobic Impairment Data for Example 12.2

	Group	N	Mean	Standard Deviation
1	Healthy individuals	1275	0.6	11
2	Hypertensive subjects (HT)	193	8.5	19
3	Postmyocardial infarction (PMI)	97	24.5	21
4	Angina pectoris, chest pain (AP)	306	30.3	24
5	PMI + AP	228	36.9	26
6	HT + AP	138	36.6	23
7	HT + PMI	20	27.6	18
8	PMI + AP + HT	75	44.9	22

Table 12.4 FAI Confidence Intervals by Group for Example 12.2

		Limits	
Group	Critical t-Value	Lower	Upper
1	2.73	−0.2	1.4
2	2.73	4.8	12.2
3	2.79	18.5	30.5
4	2.73	26.6	34.0
5	2.73	32.2	41.6
6	2.77	31.2	42.0
7	3.06	15.3	39.9
8	2.81	37.7	52.1

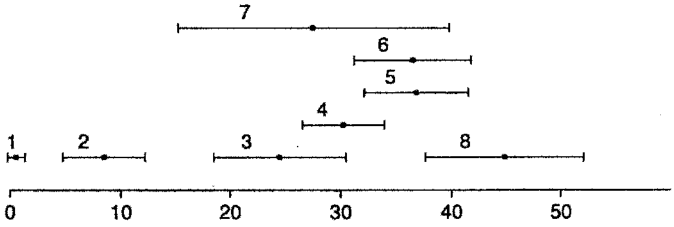

Figure 12.1 Functional aerobic impairment level.

Since all eight groups have a simultaneous 95% confidence interval, it is sufficient (but not necessary) to decide that any two means whose confidence intervals do not overlap are significantly different. Let $\mu_1, \mu_2, \ldots, \mu_8$, be the population means associated with groups $1, 2, \ldots, 8$, respectively. The following conclusions are in order:

1. μ_1 has the smallest mean ($\mu_1 < \mu_i, i = 2, \ldots, 8$).
2. μ_2 is the second smallest mean ($\mu_1 < \mu_2 < \mu_i, i = 3, \ldots, 8$).
3. $\mu_3 < \mu_5$, $\mu_3 < \mu_6$, $\mu_3 < \mu_8$.
4. $\mu_4 < \mu_8$.

There are seeming paradoxes. We know that $\mu_3 < \mu_5$, but we cannot decide whether μ_7 is larger or smaller than those two means.

Restating the conclusions in words: The healthy group had the best exercise performance, followed by the hypertensive subjects, who were better than the rest. The postmyocardial infarction group performed better than the PMI + AP, PMI + AP + HT, and HT + AR groups. The angina pectoris group had better performance than angina pectoris plus an MI and hypertension. The other orderings were not clear from this data set.

12.3 SIMULTANEOUS CONFIDENCE INTERVALS AND TESTS FOR LINEAR MODELS

12.3.1 Linear Combinations and Contrasts

In the linear models, the estimates of the parameters are usually not independent. Even when the estimates of the parameters are independent, the same error mean square, MS_e, is used for each test or confidence interval. Thus, the method of Section 12.2 does not apply. In this section, several techniques dealing with the linear model are considered.

Before introducing the Scheffé method, we need additional concepts of linear combinations and contrasts.

Definition 12.3. A *linear combination of the parameters* $\beta_1, \beta_2, \ldots, \beta_p$ is a sum $\theta = c_1\beta_1 + c_2\beta_2 + \cdots + c_p\beta_p$, where $c_1, c_2, \ldots, c_p$ are known constants.

Associated with any parameter set $\beta_1, \beta_2, \ldots, \beta_p$ is a number that is equal to the number of linearly estimated independent parameters. In ANOVA tables, this is the number of degrees of freedom associated with a particular sum of squares.

A linear combination is a parameter. An estimate of such a parameter is a statistic, a random variable. Let $b_1, b_2, \ldots, b_p$ be unbiased estimates of $\beta_1, \beta_2, \ldots, \beta_p$; then $\widehat{\theta} = c_1b_1 + c_2b_2 + \cdots + c_pb_p$ is an unbiased estimate of θ. If $b_1, b_2, \ldots, b_p$ are jointly normally distributed, $\widehat{\theta}$ will be normally distributed with mean θ and variance $\sigma_{\widehat{\theta}^2}$. The standard error of $\widehat{\theta}$ is usually quite complex and depends on possible relationships among the β's as well as correlations among the estimates of the β's. It will be of the form

$$\text{constant}\sqrt{\text{MS}_e}$$

where MS_e is the residual mean square from either the regression analysis or the analysis of variance. A simple set of linear combinations can be obtained by having only one of the c_i take on the value 1 and all others the value 0.

A particular class of linear combinations that will be very useful is given by:

Definition 12.4. A linear combination $\theta = c_1\beta_1 + c_2\beta_2 + \cdots + c_p\beta_p$ is a *contrast* if $c_1 + c_2 + \cdots + c_p = 0$. The contrast is *simple* if exactly two constants are nonzero and equal to 1 and -1.

The following are examples of linear combinations that are contrasts: $\beta_1 - \beta_2$ (a simple contrast); $\beta_1 - \frac{1}{2}(\beta_2 + \beta_3) = \beta_1 - \frac{1}{2}\beta_2 - \frac{1}{2}\beta_3$, and $(\beta_1 + \beta_8) - (\beta_2 + \beta_4) = \beta_1 + \beta_8 - \beta_2 - \beta_4$. The following are linear combinations that are not contrasts: β_1, $\beta_1 + \beta_6$, and $\beta_1 + \frac{1}{2}\beta_2 + \frac{1}{2}\beta_3$. The linear combinations and contrasts have been defined and illustrated using regression notation. They are also applicable to analysis of variance models (which are special regression models), so that the examples can be rewritten as $\mu_1 - \mu_2$, $\mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$, and so on. The interpretation is now a bit more transparent: $\mu_1 - \mu_2$ is a comparison of treatment 1 and treatment 2; $\mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$ is a comparison of treatment 1 with the average of treatment 2 and treatment 3.

Since hypothesis testing and estimation are equivalent, we state most results in terms of simultaneous confidence intervals.

12.3.2 Scheffé Method (S-Method)

A very general method for protecting against a large per experiment error rate is provided by the Scheffé method. It allows unlimited "fishing," at a price.

Result 12.1. Given a set of parameters $\beta_1, \beta_2, \ldots, \beta_p$, the probability is $1 - \alpha$ that simultaneously *all* linear combinations of $\beta_1, \beta_2, \ldots, \beta_p$, say, $\theta = c_1\beta_1 + c_2\beta_2 + \cdots + c_p\beta_p$, are in the confidence intervals

$$\widehat{\theta} \pm \sqrt{dF_{d,m,1-\alpha}}\widehat{\sigma_{\widehat{\theta}}}$$

where the estimate of θ is $\widehat{\theta} = c_1b_1 + c_2b_2 + \cdots + c_pb_p$ with estimated standard error $\widehat{\sigma}_{\widehat{\theta}}$, F is the usual F-statistic with (d, m) degrees of freedom, d is the number of linearly independent parameters, and m is the number of degrees of freedom associated with MS_e.

Note that these confidence intervals are of the usual form, "statistic ± constant × standard error of statistic," the only difference being the constant, which now depends on the number of parameters involved as well as the degrees of freedom for the error sum of squares. When $d = 1$, for any α,

$$\sqrt{dF_{d,m,1-\alpha}} = \sqrt{F_{1,m,1-\alpha}} = t_{m,1-\alpha}$$

That is, the constant reduces to the usual t-statistic with m degrees of freedom. After discussing some examples, we assess the price paid for the unlimited number of comparisons that can be made.

The easiest way to understand the S-method is to work through some examples.

Example 12.3. In Table 12.5 we present part of the computer output from Cullen and van Belle [1975] discussed in Chapters 9 and 11. We construct simultaneous 95% confidence intervals for the slopes β_i. In this case, the first linear combination is

$$\theta_1 = 1 \times \beta_1 + 0 \times \beta_2 + 0 \times \beta_3 + 0 \times \beta_4 + 0 \times \beta_5$$

the second linear combination is

$$\theta_2 = 0 \times \beta_1 + 1 \times \beta_2 + 0 \times \beta_3 + 0 \times \beta_4 + 0 \times \beta_5$$

and so on.

The standard errors of these linear combinations are simply the standard errors of the slopes. There are five slopes $\beta_1, \beta_2, \ldots, \beta_5$, which are linearly independent, but their estimates $b_1, b_2, \ldots, b_5$ are correlated. The MS_e upon which the standard errors of the slopes are based has 29 degrees of freedom. The F-statistic has value $F_{5,29,0.95} = 2.55$.

The 95% *simultaneous* confidence intervals will be of the form

$$b_i \pm \sqrt{(5)(2.55)}s_{b_i}$$

Table 12.5 Analysis of Variance, Regression Coefficients, and Confidence Intervals

	Analysis of Variance				
Source	d.f.	SS	MS	F-Ratio	Significance
Regression	5.0	95,827	18,965	12.9	0.000
Residual	29.0	42,772	1,474		

				95% Limits	
Variable	b	Standard-Error b	t	Lower	Upper
DPMB	0.575	0.0834	6.89	0.404	0.746
Trauma	−9.21	11.6	−0.792	−33.0	14.6
Lymph B	−8.56	10.2	−0.843	−29.3	12.2
Time	−4.66	5.68	−0.821	−16.3	6.96
Lymph A	−4.55	6.72	−0.677	−18.3	9.19
Constant	−96.3	36.4	2.65	22.0	171

or

$$b_i \pm 3.57 s_{b_i}, \qquad i = 1, 2, \ldots, 5$$

For the regression coefficient of DPMB the interval is

$$0.575 \pm (3.57)(0.0834)$$

resulting in 95% confidence limits of (0.277, 0.873).

Computing these values, the confidence intervals are as follows:

	Limits			Limits	
Variable	**Lower**	**Upper**	**Variable**	**Lower**	**Upper**
DPMB	0.277	0.873	Time	−24.9	15.6
Trauma	−50.8	32.3	Lymph A	−28.5	19.4
Lymph B	−44.8	27.7			

These limits are much wider than those based on a per comparison t-statistic. This is due solely to the replacement of $t_{29,0.975} = 2.05$ by $\sqrt{5F_{5,29,0.95}} = 3.57$. Hence, the confidence interval width is increased by a factor of $3.57/2.05 = 1.74$ or 74%.

Example 12.4. In a one-way ANOVA situation, using the notation of Section 10.2.2, if we wish simultaneous confidence intervals for all I means, then $d = I$, $m = n. - I$, and the standard error of the estimate of μ_i is

$$\sqrt{\frac{\text{MS}_e}{n_i}}, \qquad i = 1, \ldots, I$$

Thus, the confidence intervals are of the form

$$\overline{Y}_i. \pm \sqrt{IF_{I,n.-I,1-\alpha}}\sqrt{\frac{\text{MS}_e}{n_i}}, \qquad i = 1, \ldots, I$$

Suppose that we want simultaneous 99% confidence intervals for the morphine binding data of Problem 10.1. The confidence interval for the chronic group is

$$31.9 \pm \sqrt{(4)\,\underbrace{(4.22)}_{F_{4,24,0.99}}}\sqrt{\frac{9.825}{18}} = 31.9 \pm 3.0$$

or

$$31.9 \pm 3.0$$

The four simultaneous 99% confidence intervals are:

	Limits			Limits	
Group	**Lower**	**Upper**	**Group**	**Lower**	**Upper**
μ_1 = Chronic	28.9	34.9	μ_3 = Dialysis	22.0	36.8
μ_2 = Acute	21.0	39.2	μ_4 = Anephric	19.2	30.8

As all four intervals overlap, we cannot conclude immediately from this approach that the means differ (at the 0.01 level). To compare two means we can also consider confidence intervals for $\mu_i - \mu'_i$. As the Scheffé method allows us to look at all linear combinations, we may also consider the confidence interval for $\mu_i - \mu'_i$.

The formula for the simultaneous confidence intervals is

$$\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot} \pm \sqrt{IF_{I,n.-I,1-\alpha}}\sqrt{\text{MS}_e\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}, \qquad i, i' = 1, \ldots, I, i \neq i'$$

In this case, the confidence intervals are:

	Limits			Limits	
Contrast	**Lower**	**Upper**	**Contrast**	**Lower**	**Upper**
$\mu_1 - \mu_2$	−7.8	11.4	$\mu_2 - \mu_3$	−11.1	12.5
$\mu_1 - \mu_3$	−5.5	10.5	$\mu_2 - \mu_4$	−5.7	15.9
$\mu_1 - \mu_4$	0.4	13.4	$\mu_3 - \mu_4$	−5.0	13.8

As the interval for $\mu_1 - \mu_4$ does not contain zero, we conclude that $\mu_1 - \mu_4 > 0$ or $\mu_1 > \mu_4$. This example is typical in that comparison of the linear combination of interest is best done through a confidence interval for that combination.

The comparisons are in the form of contrasts but were not considered so explicitly. Suppose that we restrict ourselves to contrasts. This is equivalent to deciding which mean values differ, so that we are no longer considering confidence intervals for a particular mean. This approach gives smaller confidence intervals.

Contrast comparisons among the means μ_i, $i = 1, \ldots, I$ are equivalent to comparisons of α_i, $i = 1, \ldots, I$ in the one-way ANOVA model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$; for example, $\mu_1 - \mu_2 = \alpha_1 - \alpha_2$. There are only $(I - 1)$ linearly independent values of α_i since we have the constraint $\sum_i \alpha_i = 0$. This is, therefore, the first example in which the parameters are not linearly independent. (In fact, the main effects are contrasts.) Here, we set up confidence intervals for the simple contrasts $\mu_i - \mu'_i$. Here $d = 3$ and the simultaneous confidence intervals are given by

$$\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot} \pm \sqrt{(I-1)F_{I-1,n.-I,1-\alpha}}\sqrt{\text{MS}_e\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}, \qquad i, i' = 1, \ldots, I, i \neq i'$$

In the case at hand, the intervals are:

	Limits			Limits	
Contrast	**Lower**	**Upper**	**Contrast**	**Lower**	**Upper**
$\mu_1 - \mu_2$	−7.0	10.6	$\mu_2 - \mu_3$	−10.1	11.5
$\mu_1 - \mu_3$	−4.9	9.9	$\mu_2 - \mu_4$	−4.8	15.0
$\mu_1 - \mu_4$	0.9	12.9	$\mu_3 - \mu_4$	−1.9	10.7

As the $\mu_1 - \mu_4$ interval does not contain zero, we conclude that $\mu_1 > \mu_4$. Note that these intervals are shorter then in the first illustration. If you are interested in comparing each pair of means, this method will occasionally detect differences not found if we require confidence intervals for the mean as well.

Example 12.5.

1. *Main effects*. In two-way ANOVA situations there are many possible sets or linear combinations that may be studied; here we consider a few. To study all cell means, consider the IJ cells to be part of a one-way ANOVA and use the approach of Example 12.2 or 12.4.

   Now consider Example 10.5 in Section 10.3.1. Suppose that we want to compare the differences between the means for the different days at a 10% significance level. In this case we are working with the β_j main effects. The intervals for $\overline{\mu}_{\cdot j} - \overline{\mu}_{\cdot j'} = \beta_j - \beta_{j'}$ are given by

$$\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdot j'\cdot} \pm \sqrt{(J-1)F_{J-1,n_{\cdot\cdot}-IJ,1-\alpha}}\sqrt{\text{MS}_e\left(\frac{1}{n_{\cdot j}} + \frac{1}{n_{\cdot j'}}\right)}$$

   The means are 120.4, 158.1, and 118.4, respectively. The following contrasts are of interest:

		90% Limits	
Contrast	**Estimate**	**Lower**	**Upper**
$\beta_1 - \beta_2$	−37.7	−70.7	−4.7
$\beta_2 - \beta_3$	39.7	5.5	73.9
$\beta_1 - \beta_3$	2.0	−31.0	35.0

   At the 10% significance level, we conclude that $\mu_{\cdot 1} - \mu_{\cdot 2} < 0$ or $\mu_{\cdot 1} < \mu_{\cdot 2}$, and that $\mu_{\cdot 3} < \mu_{\cdot 2}$. Thus, the means (combining cases and controls) of days 10 and 14 are less than the means of day 12.

2. *Main effects assuming no interaction*. We illustrate the procedure using Problem 10.12 as an example. This example discussed the effect of histamine shock on the medullary blood vessel surface of the guinea pig thymus.

   The sex of the animal was used as a covariate. The ANOVA table is shown in Table 12.6. There is little evidence of interaction. Suppose that we want to fit the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \qquad \begin{array}{l} i = 1, \ldots, I \\ j = 1, \ldots, J \\ k = 1, \ldots, n_{ij} \end{array}$$

   That is, we ignore the interaction term. It can be shown that the appropriate estimates in the balanced model for the cell means $\mu + \alpha_i + \beta_j$ are

$$\overline{Y}_{\ldots} + a_i + b_j, \qquad \begin{array}{l} i = 1, \ldots, I \\ j = 1, \ldots, J \end{array}$$

Table 12.6 ANOVA Table for Control vs. Histamine Shock

Source	d.f.	Mean Square	F-Ratio	p-Value
Treatment	1	11.56	5.20	<0.05
Sex	1	1.26	0.57	>0.05
Treatment by sex	1	5.40	2.43	>0.05
Error	36	2.225		
Total	39			

or

$$\overline{Y}\ldots + (\overline{Y}_{i}.. - \overline{Y}\ldots) + (\overline{Y}._{j}. - \overline{Y}\ldots) = \overline{Y}_{i}.. + \overline{Y}._{j}. - \overline{Y}\ldots$$

The estimates are $\overline{Y}\ldots = 6.53$, $\overline{Y}_{1}.. = 6.71$, $\overline{Y}_{2}.. = 6.35$, $\overline{Y}._{1}. = 5.99$, $\overline{Y}._{2}. = 7.07$. The estimated cell means fitted to the model $E(Y_{ijk}) = \mu + \alpha_i + \beta_j$ by $\overline{Y}\ldots + a_i + b_j$ are:

	Treatment	
Sex	**Control**	**Shock**
Male	6.17	7.25
Female	5.81	6.89

For multiple comparisons the appropriate formula for simultaneous confidence intervals for each cell mean assuming that the interaction term is zero is given by the formula

$$\overline{Y}_{i}.. + \overline{Y}._{j}. - \overline{Y}\ldots \pm \sqrt{(I + J - 1)F_{I+J-1,n..-IJ+1,1-\alpha}}\sqrt{\text{MS}_e\left(\frac{1}{n_{i}.} + \frac{1}{n._{j}} - \frac{1}{n..}\right)}$$

The degrees of freedom for the F-statistic are $(I + J - 1)$ and $(n.. - IJ + 1)$ because there are $I + J - 1$ linearly independent cell means and the residual MS_e has $(n.. - IJ + 1)$ degrees of freedom. This MS_e can be obtained by pooling the $\text{SS}_{\text{INTERACTION}}$ and $\text{SS}_{\text{RESIDUAL}}$ in the ANOVA table. For our example,

$$\text{MS}_e = \frac{1 \times 5.40 + 36 \times 2.225}{37} = 2.311$$

We will construct the 95% confidence intervals for the four cell means. The confidence interval for the first cell is given by

$$6.17 \pm \sqrt{(2 + 2 - 1)\underbrace{F_{3,37,0.95}}_{2.86}}\sqrt{2.311\left(\frac{1}{20} + \frac{1}{20} - \frac{1}{40}\right)}$$

yielding 6.17 ± 1.22 for limits (4.95, 7.39). The four simultaneous 95% confidence limits are:

	Treatment	
Sex	**Control**	**Shock**
Male	(4.95, 7.39)	(6.03, 8.47)
Female	(4.59, 7.03)	(5.67, 8.11)

Requiring this degree of confidence gives intervals that overlap. However, using the Scheffé method, all linear combinations can be examined. With the same 95% confidence, let us examine the sex and treatment differences. The intervals for sex are defined by

$$\overline{Y}_{1}.. - \overline{Y}_{2}.. \pm \sqrt{3F_{3,37,0.95}}\sqrt{\text{MS}_e\left(\frac{1}{n_{1}.} + \frac{1}{n_{2}.}\right)}$$

or 0.36 ± 1.41 for limits $(-1.05, 1.77)$. Thus, in these data there is no reason to reject the null hypothesis of no difference in sex. The simultaneous 95% confidence interval for treatment is -1.08 ± 1.41 or $(-2.49, 0.33)$. This confidence interval also straddles zero, and at the 95% simultaneous confidence level we conclude that there is no difference in the treatment. This result nicely illustrates a dilemma. The two-way analysis of variance did indicate a significant treatment effect. Is this a contradiction? Not really, we are "protecting" ourselves against an increased Type I error. Since the results are "borderline" even with the analysis of variance, it may be best to conclude that the results are suggestive but not clearly significant. A more substantial point may be made by asking why we should test the effect of sex anyway? It is merely a covariate or blocking factor. This argument raises the question of the appropriate set of comparisons. What do you think?

3. *Randomized block designs*. Usually, we are interested in the treatment means only and not the block means. The confidence interval for the contrast $\tau_j - \tau_j'$ has the form

$$\overline{Y}_{\cdot j} - \overline{Y}_{\cdot j'} \pm \sqrt{(J-1)F_{J-1,IJ-I-J+1,1-\alpha}}\sqrt{\mathrm{MS}_e\frac{2}{I}}$$

The treatment effect τ_j has confidence interval

$$\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot} \pm \sqrt{(J-1)F_{J-1,IJ-I-J+1,1-\alpha}}\sqrt{\mathrm{MS}_e\left(1-\frac{1}{J}\right)\frac{1}{I}}$$

Problem 12.16 uses these formulas in a randomized block analysis.

12.3.3 Tukey Method (T-Method)

Another method that holds in nicely balanced ANOVA situations is the Tukey method, which is based on an extension of the Student t-test. Recall that in the two-sample t-test, we use

$$t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\frac{\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}}{s}$$

where $\overline{Y}_{1\cdot}$ is the mean of the first sample, $\overline{Y}_{2\cdot}$ is the mean of the second sample, and $s = \sqrt{\mathrm{MS}_e}$ is the pooled standard deviation. The process of dividing by s is called *studentizing* the range.

For more than two means, we are interested in the sampling distribution of the (largest–smallest) mean.

Definition 12.5. Let $Y_1, Y_2, \ldots, Y_k$ be independent and identically distributed (iid) $N(\mu, \sigma^2)$. Let s^2 be an estimate of σ^2 with m degrees of freedom, which is independent of the Y_i's. Then the quantity

$$Q_{k,m} = \frac{\textsc{max}(Y_1, Y_2, \ldots, Y_k) - \textsc{min}(Y_1, Y_2, \ldots, Y_k)}{s}$$

is called the *studentized range*.

Tukey derived the distribution of $Q_{k,m}$ and showed that it does not depend on μ or σ; a description is given in Miller [1981]. The distribution of the studentized range is given by some statistical packages and is tabulated in the Web appendix. Let $q_{k,m,1-\alpha}$ denote the upper critical value; that is,

$$P[Q_{k,m} \geq q_{k,m,1-\alpha}] = 1 - \alpha$$

You can verify from the table that for $k = 2$, two groups,

$$q_{2,m,1-\alpha} = \sqrt{2}t_{2,m,1-\alpha/2}$$

We now state the main result for using the T-method of multiple comparisons, which will then be specialized and illustrated with some examples.

The result is stated in the analysis of variance context since it is the most common application.

Result 12.2. Given a set of p population means $\mu_1, \mu_2, \ldots, \mu_p$ estimated by p independent sample means $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_p$ each based on n observations and residual error s^2 based on m degrees of freedom, the probability is $1 - \alpha$ that simultaneously all *contrasts* of $\mu_1, \mu_2, \ldots, \mu_p$, say, $\theta = c_1\mu_1 + c_2\mu_2 + \cdots + c_p\mu_p$, are in the confidence intervals

$$\widehat{\theta} \pm q_{p,m,1-\alpha}\widehat{\sigma_{\widehat{\theta}}}$$

where

$$\widehat{\theta} = c_1\overline{Y}_1 + c_2\overline{Y}_2 + \cdots + c_p\overline{Y}_p \quad \text{and} \quad \widehat{\sigma_{\widehat{\theta}}} = \frac{s}{\sqrt{n}} \sum_{i=1}^{p} \frac{|c_i|}{2}$$

The Tukey method is used primarily with pairwise comparisons. In this case, $\widehat{\sigma_{\widehat{\theta}}}$ reduces to $s/\sqrt{n}$, the standard error of a mean. A requirement is that there be equal numbers of observations in each mean; this implies a balanced design. However, reasonably good approximations can be obtained for some unbalanced situations, as illustrated next.

One-Way Analysis of Variance

Suppose that there are I groups with n observations per group and means $\mu_1, \mu_2, \ldots, \mu_I$. We are interested in all pairwise comparisons of these means. The estimate of $\mu_i - \mu_i'$ is $\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot}$, the variance of each sample mean estimated by $\text{MS}_e(1/n)$ with $m = I(n-1)$ degrees of freedom. The $100(1-\alpha)\%$ simultaneous confidence intervals are given by

$$\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot} \pm q_{I,I(n-1),1-\alpha}\frac{1}{\sqrt{n}}\sqrt{\text{MS}_e}, \qquad i, i' = 1, \ldots, I, i \neq i'$$

This result cannot be applied to the example of Section 12.3.2 since the sample sizes are not equal. However, Dunnett [1980] has shown that the $100(1-\alpha)\%$ simultaneous confidence intervals can be reasonably approximated by replacing

$$\sqrt{\frac{\text{MS}_e}{n}} \quad \text{by} \quad \sqrt{\text{MS}_e\left(\frac{1}{2}\right)\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}$$

where n_i and $n_{i'}$ are the sample sizes in groups i and i', respectively, and the degrees of freedom associated with MS_e are the usual ones from the analysis of variance.

We now apply this approximation to the morphine binding data in Section 12.3.2. For this example, $1 - \alpha = 0.99$, $I = 4$, and the $\text{MS}_e = 9.825$ has 24 d.f., resulting in $q_{4,24,0.99} = 4.907$. Simultaneous 99% confidence intervals are listed in Table 12.7.

Table 12.7 Morphine Binding Data

Contrast	n_i	n'_i	$\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot}$	Estimated Standard Error	99% Limits Lower	99% Limits Upper
$\mu_1 - \mu_2$	18	2	1.7833	1.6520	−6.32	9.98
$\mu_1 - \mu_3$	18	3	2.4500	1.3822	−4.33	9.23
$\mu_1 - \mu_4$	18	5	6.8833	1.1205	1.39	12.4
$\mu_2 - \mu_3$	2	3	0.6167	2.0233	−9.31	10.5
$\mu_2 - \mu_4$	2	5	5.0500	1.8544	−4.05	14.1
$\mu_3 - \mu_4$	3	5	4.4333	1.6186	−3.51	12.4

We conclude, at a somewhat stringent 99% confidence level, that simultaneously, only one of the pairwise contrasts is significantly different: group 1 (normal) differing significantly from group 4 (anephric).

Two-Way ANOVA with Equal Numbers of Observations per Cell

Suppose that in the two-way ANOVA of Section 10.3.1, there are n observations for each cell. The T-method may then be used to find intervals for either set of main effects (but not both simultaneously). For example, to find intervals for the α_i's, the intervals are:

Contrast	**Interval**
α_i	$\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot} \pm \frac{1}{\sqrt{Jn}} q_{I,IJ(n-1),1-\alpha} \sqrt{\text{MS}_e \left(1 - \frac{1}{I}\right)}$
$\alpha_i - \alpha_{i'}$	$\overline{Y}_{i\cdot\cdot} - \overline{Y}_{i'\cdot\cdot} \pm \frac{1}{\sqrt{Jn}} q_{I,IJ(n-1),1-\alpha} \sqrt{\text{MS}_e}$

We again consider the last example of Section 12.3.2 and want to set up 95% confidence intervals for α_1, α_2, and $\alpha_1 - \alpha_2$. In this example $I = 2$, $J = 2$, and $n = 10$. Using $q_{2,36,0.95} = 2.87$ (by interpolation), the intervals are:

Contrast	**Estimate**	**Standard Error**	**95% Limits Lower**	**95% Limits Upper**
α_1	−0.54	0.2358	−1.22	0.68
α_2	0.54	0.2358	−0.68	1.22
$\alpha_1 - \alpha_2$	−1.08	0.3335	−2.04	−0.12

We have used the MS_e with 36 degrees of freedom; that is, we have fitted a model with interaction. The interpretation of the results is that treatment effects do differ significantly at the 0.05 level; even though there is not enough evidence to reject the null hypothesis that the treatment effects differ from zero.

Randomized Block Designs

Using the notation of Section 12.3.2, suppose that we want to compare contrasts among the treatment means (the $\mu + \tau_j$). The τ_j themselves are contrasts among the means. In this case, $m = (I - 1)(J - 1)$. The intervals are:

Table 12.8 Confidence Intervals for the Six Comparisons

		95% Limits	
Contrast	Estimate	Upper	Lower
$\mu_1 - \mu_2$	21.6	4.4	38.8
$\mu_1 - \mu_3$	20.7	3.5	37.9
$\mu_1 - \mu_4$	7.0	−10.2	24.2
$\mu_2 - \mu_3$	−0.9	−18.1	16.3
$\mu_2 - \mu_4$	−14.6	−31.8	2.6
$\mu_3 - \mu_4$	−13.7	−30.9	3.5

Contrast	**Interval**
τ_j	$\overline{Y}._j - \overline{Y}.. \pm \frac{1}{\sqrt{I}} q_{J,(I-1)(J-1),1-\alpha}\sqrt{\text{MS}_e\left(1 - \frac{1}{J}\right)}$
$\tau_j - \tau_{j'}$	$\overline{Y}._j - \overline{Y}._{j'} \pm \frac{1}{\sqrt{2I}} q_{J,(I-1)(J-1),1-\alpha}\sqrt{\text{MS}_e}$

Consider Example 10.6. We want to compare the effectiveness of pancreatic supplements on fat absorption. The treatment means are

$$\overline{Y}._1 = 38.1, \qquad \overline{Y}._2 = 16.5, \qquad \overline{Y}._3 = 17.4, \qquad \overline{Y}._4 = 31.1$$

The estimate of σ^2 is $\text{MS}_e = 107.03$ with 15 degrees of freedom. To construct simultaneous 95% T-confidence intervals, we need $q_{4,15,0.95} = 4.076$. The simultaneous 95% confidence interval for $\tau_1 - \tau_2$ is

$$(38.1 - 16.5) \pm \frac{1}{\sqrt{6}}(4.076)\sqrt{107.03}$$

or

$$21.6 \pm 17.2$$

yielding (4.4, 38.8).

Proceeding similarly, we obtain simultaneous 95% confidence intervals for the six pairwise comparisons (Table 12.8). From this analysis we conclude that treatment 1 differs from treatments 2 and 3 but has not been shown to differ from treatment 4. All other contrasts are not significant.

12.3.4 Bonferroni Method (B-Method)

In this section a method is presented that may be used in all situations. The method is conservative and is based on Bonferroni's inequality. Called the Bonferroni method, it states that the probability of occurrence of one or more of a set of events occurring is less that or equal to the sum of the probabilities. That is, the Bonferroni inequality states that

$$P(A_1 U \cdots U A_n) \leq \sum_{i=1}^{n} P(A_i)$$

We know that for disjoint events, the probability of one or more of $A_1, \dots, A_n$ is equal to the sum of probabilities. If the events are not disjoint, part of the probability is counted twice or more and there is strict inequality.

Suppose now that n simultaneous tests are to be performed. It is desired to have an overall significance level α. That is, if the null hypothesis is true in all n situations, the probability of incorrectly rejecting one or more of the null hypothesis is less than or equal to α. *Perform each test at significance level* α/n; *then the overall significance level is less that or equal to* α. Let A_i be the event of incorrectly rejecting in the ith test. Bonferroni's inequality shows that the probability of rejecting one or more of the null hypotheses is less than or equal to $(\alpha/n + \cdots + \alpha/n)$ (n terms), which is equal to α.

We now state a result that makes use of this inequality:

Result 12.3. Given a set of parameters $\beta_1, \beta_2, \dots, \beta_p$ and N linear combinations of these parameters, the probability is greater than or equal to $1 - \alpha$ that simultaneously these linear combinations are in the intervals

$$\widehat{\theta} \pm t_{m,1-\alpha/2N}\widehat{\sigma_{\widehat{\theta}}}$$

The quantity $\widehat{\theta}$ is $c_1b_1 + c_2b_2 + \cdots + c_pb_p$, $t_{m,1-\alpha/2N}$ is the $100(1 - \alpha/2N)$th percentile of a t-statistic with m degrees of freedom, and $\widehat{\sigma_{\widehat{\theta}}}$ is the estimated standard error of the estimate of the linear combination based on m degrees of freedom.

The value of N will vary with the application. In the one-way ANOVA with all the pairwise comparisons among the I treatment means $N = \binom{I}{2}$. Simultaneous confidence intervals, in this case, are of the form

$$\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot} \pm t_{m,1-\alpha/2\binom{I}{2}} \sqrt{\text{MS}_e\left(\frac{1}{n_i} + \frac{1}{n_i'}\right)}, \qquad i, i' = 1, \dots, I, i \neq i'$$

The value of α need not be partitioned into equal multiples. The simplest is $\alpha = \alpha/N+\alpha/N+\cdots+\alpha/N$, but any partitions of $\alpha = \alpha_1+\alpha_2+\cdots+\alpha_N$ is permissible, yielding a per experiment error rate of at most α. However, any such decision must be made a priori—obviously, one cannot decide after seeing one p-value of 0.04 and 14 larger ones to allow all the Type I error to the 0.04 and declare it significant. Partly for this reason, unequal allocation is very unusual outside group sequential clinical trials (where it is routine but does not use the Bonferroni inequality).

When presenting p-values, when N simultaneous tests are being done, multiplication of the p-value for each test by N gives p-values allowing simultaneous consideration of all N tests.

An example of the use of Bonferroni's inequality is given in a paper by Gey et al. [1974]. This paper considers heartbeats that have an irregular rhythm (or arrythmia). The study examined the administration of the drug procainamide and evaluated variables associated with the maximal exercise test with and without the drug. Fifteen variables were examined using paired t-tests. All the tests came from data on the *same* 23 patients, so the test statistics were not independent. To correct for the multiple comparison values, the p-values were multiplied by 15. Table 12.9 presents 14 of the 15 comparisons. The table shows that even taking the multiple comparisons into account, many of the variables differed when the subject was on the procainamide medication. In particular, the frequency of arrythmic beats was decreased by administration of the drug.

Improved Bonferroni Methods

The Bonferroni adjustment is often regarded as too drastic, causing too great a loss of power. In fact, the adjustment is fairly close to optimal in any situation where only one of the null hypotheses is false. When many of the null hypotheses are false, however, there are better corrections. A number of these are described by Wright [1992]; we discuss two here.

Table 12.9 Variables at Rest and Exercise before and after Oral Procainamide[a]

		Rest						Exercise							
	Procainamide Plasma Level, 1 h	HR		SP		DP		HR Maximum		SP Maximum		DP Maximum		Arrhythmia Frequency	
		Control	1 h	Control	1 h	Control	Hr	Control	1 h	Control	1 h	Control	1 h	Control	1 h
Number of patients	23	23		23		23		23		23		23		23	
Mean	5.99	73	87	129	118	81	81	171	170	187	168	85	76	105	38
±SD	±1.33	±11	±13	±17	±11.8	±11	±9.2	±13.5	±14	±20.6	±20	±12	±10	±108	±69
t		5.053		4.183		0.3796		0.9599		5.225		5.005		3.422	
P^b		<0.0015		<0.0060		NS		NS		<0.0015		<0.015		<0.0360	

	Severity Index		VO_2 MAX		Computer ST_B								Zero Recovery	
					FAI(%)		Rest		Maximum		Slope			
	Control	1 h	Control	1 h	Control	1 h	Control	1 h	Control	1 h	Control	1 h	Control	1 h
Number of patients	23		22		23		22		22		22		23	
Mean	12.9	4.9	33.2	33.0	12.9	13.5	0.036	0.044	−0.190	−0.122	−2.31	−2.05	−0.065	−0.0302
±SD	±3.0	±4.67	±5.8	±6.0	±12.5	±11.5	±0.044	±0.051	±0.126	±0.095	±1.401	±1.29	±0.0003	±0.077
t	5.870		0.3852		0.5253		0.8861		3.915		1.132		4.320	
P^b	<0.0015		NS		NS		NS		<0.0120		NS		<0.0045	

[a] Dose, 15 mg per kilogram body weight; HR, heart rate; SP, systolic pressure (mmHg); DP, diastolic pressure (mmHg); VO_{2MAX}, maximal oxygen consumption (mL/min); FAI, functional aerobic impairment; ST_B, 100-beat averaged S-T depression, from monitored CB, lead, taken 50 to 69 ms after nadir of S-wave; slope, $\delta HR/\delta ST_B$; *t*, paired *t*-test; NS, not significant; h, hour.

[b] Probability multiplied by 15 to correct for multiple comparisons (Bonferroni's inequality correction).

Table 12.10 Application of the Three Methods

Original p	×	=	Hochberg	Holm	Bonferroni
0.001	6	0.006	0.006	0.006	0.006
0.01	5	0.05	0.04	0.05	0.06
0.02	4	0.08	0.04	0.08	0.12
0.025	3	0.075	0.04	0.08	0.15
0.03	2	0.06	0.04	0.08	0.18
0.04	1	0.04	0.04	0.08	0.24

Consider a situation where you perform six tests and obtain p-values of 0.001, 0.01, 0.02, 0.025, 0.03, and 0.04, and you wish to use $\alpha = 0.05$. All the p-values are below 0.05, something that is very unlikely to occur by chance, but the Bonferroni adjustment declares only one of them significant.

Given n p-values, the Bonferroni adjustment multiplies each by n. The Hochberg and Holm adjustments multiply the smallest by n, the next smallest by $n - 1$, and so on (Table 12.10).

This may change the relative ordering of p-values, so they are then restored to the original order. For the Hochberg method this is done by decreasing them where necessary; for the Holm method it is done by increasing them. The Holm adjustment guarantees control of Type I error; the Hochberg adjustment controls Type I error in most but not all circumstances.

Although there is little reason other than tradition to prefer the Bonferroni adjustment over the Holm adjustment, there is often not much difference.

12.4 COMPARISON OF THE THREE PROCEDURES

Of the three methods presented, which should be used? In many situations there is not sufficient balance in the data (e.g., equal numbers in each group in a one-way analysis of variance) to use the T-method; the Scheffé method procedure or the Bonferroni inequality should be used. For paired comparisons, the T-method is preferable. For more complex contrasts, the S-method is preferable. A comparison between the B-method and the S-method is more complicated, depending heavily on the type of application. The Bonferroni method is easier to carry out, and in many situations the critical value will be less than that for the Scheffé method.

In Table 12.11 we compare the critical values for the three methods for the case of one-way ANOVA with k treatments and 20 degrees of freedom for error MS. With two treatments ($k = 2$ and therefore $\nu = 1$) the three methods give identical multipliers (the q statistic has to be divided by $\sqrt{2}$ to have the same scale as the other two statistics).

Table 12.11 Comparison of the Critical Values for One-Way ANOVA with k Treatments[a]

Number of Treatments, k	Degrees of Freedom, $\nu = k - 1$	$\sqrt{\nu F_{\nu,20,0.95}}$	$\frac{1}{\sqrt{2}} q_{\nu,20,0.95}$	$t_{20,1-\alpha/2\binom{k}{2}}$
2	1	2.09	2.09	2.09
3	2	2.64	2.53	2.61
4	3	3.05	2.80	2.93
5	4	3.39	2.99	3.15
11	10	4.85	3.61	3.89
21	20	6.52	4.07	4.46

[a] Assume $\binom{k}{2}$ comparisons for the Tukey and Bonferroni procedures. Based on 20 degrees of freedom for error mean square.

Hence, if pairwise comparisons are carried out, the Tukey procedure will produce the shortest simultaneous confidence intervals. For the type of situation illustrated in the table, the B-method is always preferable to the S-method. It assumes, of course, that the total, N, of comparisons to be made is known. If this is not the case, as in "fishing expeditions," the Scheffé method provides more adequate protection.

For an informative discussion of the issues in multiple comparisons, see comments by O'Brien [1983] in *Biometrics*.

12.5 FALSE DISCOVERY RATE

With the rise of high-throughput genomics in recent years there has been renewed concern about the problem of very large numbers of multiple comparisons. An RNA expression array (gene chip) can measure the activity of several thousand genes simultaneously, and scientists often want to ask which genes differ in their expression between two samples. In such a situation it may be infeasible, but also unnecessary, to design a procedure that prevents a single Type I error out of thousands of comparisons. If we reject a few hundred null hypotheses, we might still be content if a dozen of them were actually Type I errors. This motivates a definition:

Definition 12.6. The *positive false discovery rate* (pFDR) is the expected proportion of rejected hypotheses that are actually true given that at least some null hypotheses are rejected. The *false discovery rate* (FDR) is the positive false discovery rate times the probability that no null hypotheses are rejected.

Example 12.6. Consider an experiment comparing the expression levels of 12,625 RNA sequences on an Affymetrix HG-u95A chip, to see which genes had different expression in benign and malignant colon polyps. Controlling the Type I error rate at 5% means that if we declare 100 sequences to be significantly different, we are not prepared to take more than a 5% chance of even 1 of these 100 being a false positive.

Controlling the positive false discovery rate at 5% means that if we declare 100 sequences to be significantly different, we are not prepared to have, on average, more than 5 of these 100 being false positives.

The pFDR and FDR apparently require knowledge of which hypotheses are true, but we will see that, in fact, it is possible to control the pFDR and FDR without this knowledge and that such control is more effective when we are testing a very large number of hypotheses.

Although like many others, we discuss the FDR and pFDR under the general heading of multiple comparisons, they are very different quantities from the Type I error rates in the rest of this chapter. The Type I error rate is the probability of making a certain decision (rejecting the null hypothesis) conditional on the state of nature (the null hypothesis is actually true). The simplest interpretation of the pFDR is the probability of a state of nature (the null hypothesis is true) given a decision (we reject it). This should cause some concern, as we have not said what we might mean by the probability that a hypothesis is true.

Although it is possible to define probabilities for states of nature, leading to the interesting and productive field of Bayesian statistics, this is not necessary in understanding the false discovery rates. Given a large number N of tests, we know that in the worst case, when all the null hypotheses are true, there will be approximately αN hypotheses (falsely) rejected. In general, fewer that N of the null hypotheses will be true, and there will be fewer than N false discoveries. If we reject R of the null hypotheses and $R > \alpha N$, we would conclude that at least roughly $R - \alpha N$ of the discoveries were correct, and so would estimate the positive false discovery rate as

$$\text{pFDR} \approx \frac{R - \alpha N}{R}$$

This is similar to a graphical diagnostic proposed by Schweder and Spjøtvoll [1982], which involves plotting R/N against the p-value, with a line showing the expected relationship. As it stands, this estimator is not a very good one. The argument can be improved to produce fairly simple estimators of FDR and pFDR that are only slightly conservative [Storey, 002].

As the FDR and pFDR are primarily useful when N is very large (at least hundreds of tests), hand computation is not feasible. We defer the computational details to the Web appendix of this chapter, where the reader will find links to programs for computing the FDR and pFDR.

12.6 POST HOC ANALYSIS

12.6.1 The Setting

A particular form of the multiple comparison problem is post hoc *analysis*. Such an analysis is not explicitly planned at the start of the study but suggested by the data. Other terms associated with such analyses are *data driven* and *subgroup analysis*. Aside from the assignment of appropriate p-values, there is the more important question of the scientific status of such an analysis. Is the study to be considered exploratory, confirmatory, or both? That is, can the post hoc analysis only suggest possible connections and associations that have to be confirmed in future studies, or can it be considered as confirming them as well? Unfortunately, no rigid lines can be drawn here. Every experimenter does, and should do, post hoc analyses to ensure that all aspects of the observations are utilized. There is no room for rigid adherence to artificial schema of hypothesis which are laid out row upon boring row. But what is the status of these analyses? Cox [1977] remarks:

> Some philosophies of science distinguish between exploratory experiments and confirmatory experiments and regard an effect as well established only when it has been demonstrated in a confirmatory experiment. There are undoubtedly good reasons, not specifically concerned with statistical technique, for proceeding this way; but there are many fields of study, especially outside the physical sciences, where mounting confirmatory investigations may take a long time and therefore where it is desirable to aim at drawing reasonably firm conclusions from the same data as used in exploratory analysis.

What statistical approaches and principles can be used? In the following discussion we follow closely suggestions of Cox and Snell [1981] and Pocock [1982, 1984].

12.6.2 Statistical Approaches and Principles

Analyses Must Be Planned

At the start of the study, specific analyses must be planned and agreed to. These may be broadly outlined but must be detailed enough to, at least theoretically, answer the questions being asked. Every practicing statistician has met the researcher who has a filing cabinet full of crucial data "just waiting to be analyzed" (by the statistician, who may also feel free to suggest appropriate questions that can be answered by the data).

Planned Analyses Must Be Carried Out and Reported

This appears obvious but is not always followed. At worst it becomes a question of scientific integrity and honesty. At best it is potentially misleading to omit reporting such analyses. If the planned analysis is amplified by other analyses which begin to take on more importance, a justification must be provided, together with suggested adjustments to the significance level of the tests. The researcher may be compared to the novelist whose minor character develops a life of his own as the novel is written. The development must be rational and believable.

Adjustment for Selection

A post hoc analysis is part of a multiple-comparison procedure, and appropriate adjustments can be made if the family of comparisons is known. Use of the Bonferroni adjustment or other methods can have a dramatic effect. It may be sufficient, and is clearly necessary, to report analyses in enough detail that readers know how much testing was done.

Split-Sample Approach

In the split-sample approach, the data are randomly divided into two parts. The first part is used to generate the exploratory analyses, which are then "confirmed" by the second part. Cox [1977] says that there are "strong objections on general grounds to procedures where different people analyzing the same data by the same method get different answers." An additional aspect of such analyses is that it does not provide a solution to the problem of subgroup analysis.

Interaction Analysis

The number of comparisons is frequently not defined, and most of the foregoing approaches will not work very well. Interaction analysis of subgroups provides valid protection in such post hoc analyses. Suppose that a treatment effect has been shown for a particular subgroup. To assess the validity of this effect, analyze all subgroups jointly and test for an interaction of subgroup and treatment. This procedure embeds the subgroup in a meaningful larger family. If the global test for interaction is significant, it is warranted to focus on the subgroup suggested by the data. Pocock [1984] illustrates this approach with data from the Multiple Risks Factor Intervention Trial Research Group [1982] "MR. FIT". This randomized trial of "12,866 men at high risk of coronary heart disease compared to special intervention (SI) aimed at affecting major risk factors (e.g., hypertension, smoking, diet) and usual care (UC). The overall rates of coronary mortality after an average seven year follow-up (1.79% on SI and 1.93% on UC) are not significantly different." The paper presented four subgroups. The extreme right-hand column in Table 12.12 lists the odds ratio comparing mortality in the special intervention and usual care groups. The first three subgroups appear homogeneous, suggesting a beneficial effect of special intervention. The fourth subgroup (with hypertension and ECG abnormality) appears different. The average odds ratio for the first three subgroups differs significantly from the odds ratio for the fourth group ($p < 0.05$). However, this is a post hoc analysis, and a test for the homogeneity of the odds ratios over all four subgroups shows no significant differences, and furthermore, the average of the odds ratio does not differ significantly from 1. Thus, on the basis of the global interaction test there are no significant differences in mortality among the eight groups. (A chi-square analysis of the 2×8 contingency table formed by the two treatment groups and the eight subgroups shows a value of $\chi^2 = 8.65$ with 7 d.f.) Pocock concludes: "Taking into account the fact that this was not the only subgroup analysis performed, one should feel confident that there are inadequate grounds for supposing that the special intervention did harm to those with hypertension and ECG abnormalities."

If the overall test of interaction had been significant, or if the comparison had been suggested before the study was started, the "significant" p-value would have had clinical implications.

12.6.3 Simultaneous Tests in Contingency Tables

In $r \times c$ contingency tables, there is frequently interest in comparing subsets of the tables. Goodman [1964a,b] derived the large sample form for $100(1-\alpha)\%$ simultaneous contrasts for

Table 12.12 Interaction Analysis: Data for Four MR. FIT Subgroups

		No. of Coronary Death/No. of Men				
Hypertension	ECG Abnormality	Special Intervention	(%)	Usual Care	(%)	Odds Ratio
No	No	24/1817	(1.3)	30/1882	(1.6)	0.83
No	Yes	11/592	(1.9)	15/583	(2.6)	0.72
Yes	No	44/2785	(1.6)	58/2808	(2.1)	0.76
Yes	Yes	36/1233	(2.9)	21/1185	(1.8)	1.67

all 2 × 2 comparisons. This is equivalent to examining all $\binom{r}{2}\binom{c}{2}$ possible odds ratios. The intervals are constructed in terms of the logarithms of the ratio. Let

$$\widehat{\omega} = \log n_{ij} + \log n_{i'j'} - \log n_{i'j} - \log n_{ij}$$

be the log odds associated with the frequencies indicated. In Chapter 7 we showed that the approximate variance of this statistic is

$$\widehat{\sigma}^2_{\widehat{\omega}} \doteq \frac{1}{n_{ij}} + \frac{1}{n_{i'j'}} + \frac{1}{n_{i'j}} + \frac{1}{n_{ij'}}$$

Simultaneous $100(1-\alpha)\%$ confidence intervals are of the form

$$\widehat{\omega} \pm \sqrt{\chi^2_{(r-1)(c-1),(1-\alpha)}}\widehat{\sigma}_{\widehat{\omega}}$$

This again is of the same form as the Scheffé approach, but now based on the chi-square distribution rather that the F-distribution. The price, again, is fairly steep. At the 0.05 level and a 6 × 6 contingency table, the critical value of the chi-square statistic is

$$\sqrt{\chi^2_{25,0.95}} = \sqrt{37.65} = 6.14$$

Of course, there are $\binom{6}{2}\binom{6}{2} = 225$ such tables. It may be more efficient to use the Bonferroni inequality. In the example above, the corresponding Z-value using the Bonferroni inequality is

$$Z_{1-0.025/225} = Z_{0.999889} \doteq 3.69$$

So if only 2 × 2 tables are to be examined, the Bonferroni approach will be more economical.

However, the Goodman approach works and is valid for *all* linear contrasts. See Goodman [1964a,b] for additional details.

12.6.4 Regulatory Statistics and Game Theory

In reviewing newly developed pharmaceuticals, the Food and Drug Administration, takes a very strong view on multiple comparisons and on control of Type I error, much stronger than we have taken in this chapter. Regulatory decision making, however, is a special case because it is in part adversarial. Statistical decision theory deals with decision making under uncertainty and is appropriate for scientific research, but is insufficient as a basis for regulation.

The study of decision making when dealing with multiple rational actors who do not have identical interests is called game theory. Unfortunately, it is much more complex than statistical decision theory. It is clear that FDA policies affect the supply of new treatments not only through their approval of specific products but also through the resulting economic incentives for various sorts of research and development, but it is not clear how to go from this to an assessment of the appropriate p-values.

12.6.5 Summary

Post hoc comparisons should usually be considered exploratory rather than confirmatory, but this rule should not be followed slavishly. It is clear that some adjustment to the significance level must be made to maintain the validity of the statistical procedure. In each instance the p-value will be adjusted upward. The question is whether this should be done by a formal adjustment, and if so, what groups of hypotheses should the fixed Type I error be divided over. One important difficulty in specifying how to divide up the Type I error is that different readers may group hypotheses differently. It is also important to remember that controlling the total Type I error unavoidably increases the Type II error. If your conclusions are that an exposure makes no difference, these conclusions are weakened, rather than strengthened, by controlling Type I error.

When reading research reports that include post hoc analyses, it is prudent to keep in mind that in all likelihood, many such analyses were tried by the authors but not reported. Thus, scientific caution must be the rule. To be confirmatory, results from such analyses must not only make excellent biological sense but must also satisfy the principle of Occam's razor. That is, there must not be a simpler explanation that is also consistent with the data.

NOTES

12.1 Orthogonal Contrasts

Orthogonal contrasts form a special group of contrasts. Consider two contrasts:

$$\theta_1 = c_{11}\beta_1 + \cdots + c_{1p}\beta_p$$

and

$$\theta_2 = c_{21}\beta_1 + \cdots + c_{2p}\beta_p$$

The two contrasts are said to be *orthogonal* if

$$\sum_{j=1}^{p} c_{1j}c_{2j} = 0$$

Clearly, if θ_1, θ_2 are orthogonal, then $\widehat{\theta_1}, \widehat{\theta_2}$ will be orthogonal since orthogonality is a property of the coefficients. Two orthogonal contrasts are *orthonormal* if, in addition,

$$\sum c_{1j}^2 = \sum c_{2j}^2 = 1$$

The advantage to considering orthogonal (and orthonormal) contrasts is that they are uncorrelated, and hence, if the observations are normally distributed, the contrasts are statistically independent. Hence, the Bonferroni inequality becomes an equality. But there are other advantages. To see those we extend the orthogonality to more than two contrasts. A set of contrasts is orthogonal (orthonormal) if all pairs of contrasts are orthogonal (orthonormal).

Now consider the one-way analysis of variance with I treatments. There are $I - 1$ degrees of freedom associated with the treatment effect. It can be shown that there are precisely $I - 1$ orthogonal contrasts to compare the treatment means. The set is not unique; let $\theta_1, \theta_2, \ldots, \theta_{I-1}$

form a set of such contrasts. Assume that they are orthonormal, and let $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_{I-1}$ be the estimate of the orthonormal contrasts. Then it can be shown that

$$\mathrm{SS}_{\mathrm{TREATMENTS}} = \widehat{\theta}_1^2 + \widehat{\theta}_2^2 + \cdots + \widehat{\theta}_{I-1}^2$$

We have thus partitioned the $\mathrm{SS}_{\mathrm{TREATMENTS}}$ into $I - 1$ components (each with one degree of freedom, it turns out) and uncorrelated as well. This is a very nice summary of the data. To illustrate this approach, assume an experiment with four treatments. Let the means be μ_1, μ_2, μ_3, μ_4. A possible set of contrasts is given by the following pattern:

Contrast	μ_1	μ_2	μ_3	μ_4
θ_1	$1/\sqrt{2}$	$-1/\sqrt{2}$	0	0
θ_2	$1/\sqrt{6}$	$1/\sqrt{6}$	$-2/\sqrt{6}$	0
θ_3	$1/\sqrt{12}$	$1/\sqrt{12}$	$1/\sqrt{12}$	$-3/\sqrt{12}$

You can verify that:

- These contrasts are orthonormal.
- There are no additional *orthogonal contrasts*.
- $\theta_1^2 + \theta_2^2 + \theta_3^2 = \sum(\mu_i - \mu)^2$.

The pattern can clearly be extended to any number of means (it is known as the *Gram-Schmidt orthogonalization process*).

The nonuniqueness of this decomposition becomes obvious from starting the first contrast, say, with

$$\theta_1^* = \frac{1}{\sqrt{2}}\mu_1 - \frac{1}{\sqrt{2}}\mu_4$$

Sometimes a meaningful set of orthogonal contrasts can be used to summarize an experiment. This approach, using the statistical independence to determine the significance level, will minimize the cost of multiple testing. Of course, if these contrasts were carefully specified beforehand, you might argue that each one should be tested at level α!

12.2 Tukey Test

The assumptions underlying the Tukey test include that the variances of the means are equal; this translates into equal sample sizes in the analysis of variance situation. Although the procedure is commonly associated with pairwise comparisons among independent means, it can be applied to arbitrary linear combinations and even allows for a common correlation among the means. For further discussion, see Miller [1981, pp. 37–48]. There are extensions of the Tukey test similar in principle to the Holm extension of the Bonferroni adjustment. These are built on the idea of sequential testing. Suppose that we have tested the most extreme pair of means and rejected the hypothesis that they are the same. There are two possibilities:

1. The null hypothesis is actually false, in which case we have not used any Type I error.
2. The null hypothesis is actually true, which happens with probability less than α.

In either case, if we now perform the next-most extreme test we can ignore the fact that we have already done one test without affecting the per experiment Type I error. The resulting procedure is called the *Newman–Keuls* or *Student–Newman–Keuls test* and is available in many statistical packages.

12.3 Likelihood Principle

The likelihood principle is a philosophical principle in statistics which says that all the evidence for or against a hypothesis is contained in the likelihood ratio. It can be derived in various ways from intuitively plausible assumptions. The likelihood principle implies that the evidence about one hypothesis does not depend on what other hypotheses were investigated. One view of this is that it shows that multiple comparison adjustment is undesirable; another is that it shows the that likelihood principle is undesirable. A fairly balanced discussion of these issues can be found in Stuart et al. [1999].

There is no entirely satisfactory resolution to this conflict, which is closely related to the question of what counts as an experiment for the per experiment error rate. One possible resolution is to conclude that the main danger in the multiple comparison problem comes from incomplete publication. That is, the danger is more that other people will be misled than that you yourself will be misled (see also Problem 12.13). In this case the argument from the likelihood principle does not hold in any simple form. The relevant likelihood would now be the likelihood of seeing the results given the selective reporting process as well as the randomness in the data, and this likelihood does depend on what one does with multiple comparisons. This intermediate position suggests that multiple comparison adjustments are critical primarily when only selected results of an exploratory analysis are reported.

PROBLEMS

For the problems in this chapter, the following tasks are defined. Additional tasks are indicated in each problem. Unless otherwise indicated, assume that $\alpha^* = 0.05$.

(a) Calculate simultaneous confidence intervals as discussed in Section 12.2. Graph the intervals and state your conclusions.

(b) Apply the Scheffé method. State your conclusions.

(c) Apply the Tukey method. State your conclusions.

(d) Apply the Bonferroni method. State your conclusions.

(e) Compare the methods indicated. Which result is the most reasonable?

12.1 This problem deals with Problem 10.1. Use a 99% confidence level.

(a) Carry out task (a).

(b) Compare your results with those obtained in Section 12.3.2.

(c) A more powerful test can be obtained by considering the groups to be ranked in order of increasingly severe disorder. A test for trend can be carried out by coding the groups 1, 2, 3, and 4 and regressing the percentage morphine bound on the regressor variable and testing for significance of the slope. Carry out this test and describe its pros and cons.

(d) Carry out task (c) using the approximation recommended in Section 12.3.3.

(e) Carry out task (e).

12.2 This problem deals with Problem 10.2.

(a) Do tasks (a) through (e) for pairwise comparisons of all treatment effects.

12.3 This problem deals with Problem 10.3.

(a) Do tasks (a) through (d) for all pairwise comparisons.

(b) Do task (c) defined in Problem 12.1.

(c) Do task (e).

12.4 This problem deals with Problem 10.4.

- **(a)** Do tasks (a) through (e) setting up simultaneous confidence intervals on both main effects and all pairwise comparisons.
- **(b)** A further comparison of interest is control vs. shock. Using the Scheffé approach, test this effect.
- **(c)** Summarize the results from this experiment in a short paragraph.

12.5 Sometimes we are interested in comparing several treatments against a standard treatment. Dunnett [1954] has considered this problem. If there are I groups, and group 1 is the standard group, $I - 1$ comparisons can be made at level $1 - \alpha/2(I - 1)$ to maintain a per experiment error rate of α. Apply this approach to the data of Bruce et al. [1974] in Section 12.2 by comparing groups 2, ... , 8 with group 1, the healthy individuals. How do your conclusions compare with those of Section 12.2?

12.6 This problem deals with Problem 10.6.

- **(a)** Carry out tasks (a) through (e).
- **(b)** Suppose that we treat these data as a regression problem (as suggested in Chapter 10). Does it still make sense to test the significance of the difference of adjacent means? Why or why not? What if the trend was nonlinear?

12.7 This problem deals with Problem 10.7.

- **(a)** Carry out tasks (a) through (e).

12.8 This problem deals with Problem 10.8.

- **(a)** Carry out tasks (b), (c), and (d).
- **(b)** Of particular interest are the comparisons of each of the test preparations A through D with the standard insulin. The "medium" treatment is not relevant for this analysis. How does this alter task (d)?
- **(c)** Why would it not be very wise to ignore the "medium" treatment totally? What aspect of the data for this treatment can be usefully incorporated into the analysis in part (b)?

12.9 This problem deals with Problem 10.9.

- **(a)** Compare each of the means of the schizophrenic group with the control group using S, T, and B methods.
- **(b)** Which method is preferred?

12.10 This problem deals with Problem 10.10.

- **(a)** Carry out tasks (b) through (e) on the plasma concentration of 45 minutes, comparing the two treatments with controls.
- **(b)** Carry out tasks (b) through (d) on the difference in the plasma concentration at 90 minutes and 45 minutes (subtract the 45-minute reading from the 90-minute reading). Again, compare the two treatments with controls.
- **(c)** Synthesize the conclusions of parts (a) and (b).
- **(d)** Can you think of a "nice" graphical way of presenting part (c)?

(e) Consider parts (a) and (b) combined. From a multiple-comparison point of view, what criticism could you level at this combination? How would you resolve it?

12.11 Data for this problem are from a paper by Winick et al. [1975]. The paper examines the development of adopted Korean children differing greatly in early nutritional status. The study was a retrospective study of children admitted to the Holt Adoption Service and ultimately placed in homes in the United States. The children were divided into three groups on the basis of how their height, at the time of admission to Holt, related to a reference standard of normal Korean children of the same age:

- *Group 1.* designated "malnourished"—below the third percentile for both height and weight.
- *Group 2.* "moderately nourished"—from the third to the twenty-fourth percentile for both height and weight.
- *Group 3.*"well-nourished or control"—at or above the twenty-fifth percentile for both height and weight.

Table 12.13 has data from this paper.

Table 12.13 Current Height (Percentiles, Korean Reference Standard) Comparison of Three Nutrition Groups[a]

Group	N	Mean Percentile	SD	F Probability	Contrast Group	t-Test t	P
1	41	71.32	24.98	0.068	1 vs. 2	−1.25	0.264
2	50	76.86	21.25		1 vs. 3	−2.22	0.029[b]
3	47	82.81	23.26		2 vs. 3	−1.31	0.194
Total	138	77.24	23.41				

[a] F probability is the probability that the F calculated from the one-way ANOVA ratio would occur by chance
[b] Statistically significant.

(a) Carry out tasks (a) through (e) for all pairwise comparisons and state your conclusions.

(b) Read the paper, then compare your results with that of the authors.

(c) A philosophical point may be raised about the procedure of the paper. Since the overall F-test is not significant at the 0.05 level (see Table 12.13), it would seem inappropriate to "fish" further into the data. Discuss the pros and cons of this argument.

(d) Can you suggest alternative, more powerful analyses? (What is meant by "more powerful"?)

12.12 Derive equation (1). Indicate clearly how the independence assumption and the null hypotheses are crucial to this result.

12.13 A somewhat amusing—but also serious—example of the multiple comparison problem is the following. Suppose that a journal tends to accept only papers that show "significant" results. Now imagine multiple groups of independent researchers (say, 20 universities in the United States and Canada) all working on roughly the same topic and hence testing the same null hypothesis. If the null hypothesis is true, we would expect only one of the researchers to come up with a "significant" result. Knowing the editorial policy of the journal, the 19 researchers with nonsignificant results do not bother to write up their research, but the remaining researcher does. The paper is well written, challenging, and provocative. The editor accepts the paper and it is published.

(a) What is the per experiment error rate? Assume 20 independent researchers.

(b) Define an appropriate editorial policy in view of an unknown number of comparisons.

12.14 This problem deals with the data of Problem 10.13. The primary interest in these data involves comparisons of three treatments; that is, the experiments represent blocks. Carry out tasks (a) through (e) focusing on comparison of the means for tasks (b) through (d).

12.15 This problem deals with the data of Problem 10.14.

(a) Carry out the Tukey test for pairwise comparisons on the total analgesia score presented in part (b) of that question. Translate your answers to obtain confidence intervals applicable to single readings.

***(b)** The sum of squares for analgesia can be partitioned into three orthogonal contrasts as follows:

	μ_1	μ_2	μ_3	μ_4	**Divisor**
θ_1	-1	-1	-1	3	$\sqrt{12}$
θ_2	1	-1	-1	1	$\sqrt{4}$
θ_3	-1	3	-3	1	$\sqrt{20}$

(c) Verify that these contrasts are orthogonal. If the coefficients are divided by the divisors at the right, verify that the contrasts are orthonormal.

***(d)** Interpret the contrasts θ_1, θ_2, θ_3 defined in part (b).

***(e)** Let $\widehat{\theta_1}$, $\widehat{\theta_2}$, $\widehat{\theta_3}$ be the estimates of the orthonormal contrasts. Verify that

$$\text{SS}_{\text{TREATMENTS}} = \widehat{\theta_1^2} + \widehat{\theta_2^2} + \widehat{\theta_3^2}$$

Test the significance of each of these contrasts and state your conclusion.

12.16 This problem deals with Problem 10.15.

(a) Carry out tasks (b) through (e) on all pairwise comparisons of treatment means.

***(b)** How would the results in part (a) be altered if the Tukey test for additivity is used? Is it worth reanalyzing the data?

12.17 This problem deals with Problem 10.16.

(a) Carry out tasks (b) through (e) on the treatment effects and on all pairwise comparisons of treatment means.

***(b)** Partition the sums of squares of treatments into two pieces, a part attributable to linear regression and the remainder. Test the significance of the regression, adjusting for the multiple comparison problem.

***12.18** This problem deals with the data of Problem 10.18.

(a) We are going to "mold" these data into a regression problem as follows; define six dummy variables I_1 to I_6.

$$I_i = \begin{cases} 1, & \text{data from subject } i, i = 1, \dots, 6 \\ 0, & \text{otherwise} \end{cases}$$

In addition, define three further dummy variables:

$$I_7 = \begin{cases} 1, & \text{recumbent position} \\ 0, & \text{otherwise} \end{cases}$$

$$I_8 = \begin{cases} 1, & \text{placebo} \\ 0, & \text{otherwise} \end{cases}$$

$$I_9 = I_7 \times I_8$$

(b) Carry out the regression analyses of part (a) forcing in the dummy variables I_1 to I_6 first. Group those into one SS with six degrees of freedom. Test the significance of the regression coefficients of I_7, I_8, I_9 using the Scheffé procedure.

(c) Compare the results of part (c) of Problem 10.18 with the analysis of part (b). How can the two analyses be reconciled?

12.19 This problem deals with the data of Example 10.5 and Problem 10.19.

(a) Carry out tasks (c) and (d) on pairwise comparisons.

(b) In the context of the Friedman test, suggest a multiple-comparison approach.

12.20 This problem deals with Problem 10.4.

(a) Set up simultaneous 95% confidence intervals on the three regression coefficients using the Scheffé method.

(b) Use the Bonferroni method to construct comparable 95% confidence intervals.

(c) Which method is preferred?

(d) In regression models, the usual tests involve null hypotheses of the form H_0: $\beta_i = 0$, $i = 1, \dots, p$. In general, how do you expect the Scheffé method to behave as compared with the Bonferroni method?

(e) Suppose that we have another kind of null hypothesis, for example, H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. Does this create a multiple-comparison problem? How would you test this null hypothesis?

(f) Suppose that we wanted to test, simultaneously, two null hypotheses, H_0: $\beta_1 = \beta_2 = 0$ and H_0: $\beta_3 = 0$. Carry out this test using the Scheffé procedure. State your conclusion. Also use nested hypotheses; how do the two tests compare?

***12.21** **(a)** Verify that the contrasts defined in Problem 10.18, parts (c), (d), and (e) are orthogonal.

(b) Define another set of orthogonal contrasts that is also meaningful. Verify that $\text{SS}_{\text{TREATMENTS}}$ can be partitioned into three sums of squares associated with this set. How do you interpret these contrasts?

REFERENCES

Bruce, R. A., Gey, G. O., Jr., Fisher, L. D., and Peterson, D. R. [1974]. Seattle heart watch: initial clinical, circulatory and electrocardiographic responses to maximal exercise. *American Journal of Cardiology*, **33**: 459–469.

Cox, D. R. [1977]. The role of significance tests. *Scandinavian Journal of Statistics*, **4**: 49–62.

Cox, D. R., and Snell, E. J. [1981]. *Applied Statistics*. Chapman & Hall, London.

Cullen, B. F., and van Belle, G. [1975]. Lymphocyte transformation and changes in leukocyte count: effects of anesthesia and operation. *Anesthesiology*, **43**: 577–583.

Diaconis, P., and Mosteller, F. [1989]. Methods for studying coincidences. *Journal of the American Statistical Association*, **84**: 853–861.

Dunnett, C. W. [1954]. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, **50**: 1096–1121.

Dunnett, C. W. [1980]. Pairwise multiple comparison in the homogeneous variance, unequal sample size case. *Journal of the American Statistical Association*, **75**: 789–795.

Gey, G. D., Levy, R. H., Fisher, L. D., Pettet, G., and Bruce, R. A. [1974]. Plasma concentration of procainamide and prevalence of exertional arrythmias. *Annals of Internal Medicine*, **80**: 718–722.

Goodman, L. A. [1964a]. Simultaneous confidence intervals for contrasts among multinomial populations. *Annals of Mathematical Statistics*, **35**: 716–725.

Goodman, L. A. [1964b]. Simultaneous confidence limits for cross-product ratios in contingency tables. *Journal of the Royal Statistical Society, Series B*, **26**: 86–102.

Miller, R. G. [1981]. *Simultaneous Statistical Inference*, 2nd ed. Springer-Verlag, New York.

Multiple Risks Factor Intervention Trial Research Group [1982]. Multiple risk factor intervention trial: risk factor changes and mortality results. *Journal of the American Medical Association*, **248**: 1465–1477.

O'Brien, P. C. [1983]. The appropriateness of analysis of variance and multiple comparison procedures. *Biometrics*, **39**: 787–794.

Pocock, S. J. [1982]. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics*, **36**: 153–162.

Pocock, S. J. [1984]. Current issues in design and interpretation of clinical trials. *Proceedings of the 12th International Biometric Conference*, Tokyo, pp. 31–39.

Proschan, M., and Follman, D. [1995]. Multiple comparisons with control in a single experiment versus separate experiments: Why do we feel differently? *American Statistician*, **49**:144–149.

Rothman, K. [1990]. No adjustments are needed for multiple comparisons. *Epidemiology*, **1**: 43–46

Schweder, T., and Spjøtvoll, E. [1982]. Plots of P-values to evaluate many tests simultaneously. *Biometrika*, **69**: 493–502.

Storey, J. D. [2002]. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**: 479-498.

Stuart, A., Ord, K., and Arnold, S. [1999]. *Kendall's Advanced Theory of Statistics*, Vol. 2A, *Classical Inference and the Linear Model*. Edward Arnold, London.

Winick, M., Meyer, K. K., and Harris, R. C. [1975]. Malnutrition and environmental enrichment by early adoption. *Science*, **190**: 1173–1175.

Wright, S. P. [1992] Adjusted p-values for simultaneous inference. *Biometrics*, **48**: 1005–1013.

CHAPTER 13

Discrimination and Classification

13.1 INTRODUCTION

Discrimination or classification methods attempt to use measured characteristics to divide people or objects into prespecified groups. As in regression modeling for prediction in Chapter 11, the criteria for assessing classification models are accuracy of prediction and possibly cost of measuring the relevant characteristics. There need not be any relationship between the model and the actual causal processes involved. The computer science literature refers to classification as *supervised learning*, as distinguished from *cluster analysis* or *unsupervised learning*, in which groups are not prespecified and must be discovered as part of the analysis. We discuss cluster analysis briefly in Note 13.5.

In this chapter we discuss the general problem of classification. We present two simple techniques, logistic and linear discrimination, and discuss how to choose and evaluate classification models. Finally, we describe briefly a number of more modern classification methods and give references for further study.

13.2 CLASSIFICATION PROBLEM

In the classification problem we have a group variable Y for each individual, taking values $1, 2, \ldots, K$, called *classes*, and a set of characteristics $X_1, X_2, \ldots, X_p$. Both X and Y are observed for a *training set* of data, and the goal is to create a rule to predict Y from X for new observations and to estimate the accuracy of these predictions.

The most common examples of classification problems in biostatistics have just two classes: with and without a given disease. In screening and diagnostic testing, the classes are based on whether the disease is currently present; in prognostic models, the classes are those who will and will not develop the disease over some time frame.

For example, the Framingham risk score [Wilson et al., 1998] is used widely to determine the probability of having a heart attack over the next 10 years based on blood pressure, age, gender, cholesterol levels, and smoking. It is a prognostic model used in screening for heart disease risk, to help choose interventions and motivate patients. Various diagnostic classification rules also exist for coronary heart disease. A person presenting at a hospital with chest pain may be having a heart attack, in which case prompt treatment is needed, or may have muscle strain or indigestion-related pain, in which case the clot-dissolving treatments used for heart attacks would be unnecessary and dangerous. The decision can be based on characteristics of the pain,

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

blood enzyme levels, and electrocardiogram abnormalities. Finally, for research purposes it is often necessary to find cases of heart attack from medical records. This retrospective diagnosis can use the same information as the initial diagnosis and later follow-up information, including the doctors' conclusions at the time of discharge from a hospital.

It is useful to separate the classification problem into two steps:

1. Estimate the probability p_k that $Y = k$.
2. Choose a predicted class based on these probabilities.

It might appear that the second step is simply a matter of choosing the most probable class, but this need not be the case when the consequences of making incorrect decisions depend on the decision. For example, in cancer screening a *false positive*, calling for more investigation of what turns out not to be cancer, is less serious than a *false negative*, missing a real case of cancer. About 10% of women are recalled for further testing after a mammogram [Health Canada, 2001], but the great majority of these are false positives and only 6 to 7% of these women are diagnosed with cancer.

The consequences of misclassification can be summarized by a *loss function* $L(j, k)$, which gives the relative seriousness of choosing class j when in fact class k is the correct one. The loss function is defined to be zero for a correct decision and positive for incorrect decisions. If $L(j, k)$ has the same value for all incorrect decisions, the correct strategy is to choose the most likely class. In some cases these losses might be actual monetary costs; in others the losses might be probabilities of dying as a result of the decision, or something less concrete. What the theory requires is that a loss of 2 is twice as bad as a loss of 1. In Note 13.3 we discuss some of the practical and philosophical issues involved in assigning loss functions.

Finally, the expected proportion in each class may not be the same in actual use as in training data. This imbalance may be deliberate: If some classes are very rare, it will be more efficient if they are overrepresented in the training data. The imbalance may also be due to a variation in frequency of classes between different times or places; for example, the relative frequency of common cold and influenza will depend on the season. We will write π_k for the expected proportion in class k if it is specified separately from the training data. These are called *prior probabilities*.

Given a large enough training set, the classification problem is straightforward (assume initially that we do not have separately specified proportions π_k). For any new observations with characteristics $x_1, \dots, x_p$, we find all the observations in the training set that have exactly the same characteristics and estimate p_k, the probability of being in class k, as the proportion of these observations that are in class k.

Now that we have probabilities for each class k, we can compute the expected loss for each possible decision. Suppose that there are two classes and we decide on class 1. The probability that we are correct is p_1, in which case there is no loss. The probability that we are incorrect is p_2, in which case the loss is $L(1, 2)$. So the expected loss is $0 \times p_1 + L(1, 2) \times p_2$. Conversely, if we decide on class 2, the expected loss is $L(2, 1) \times p_1 + 0 \times p_2$. We should choose whichever class has the lower expected loss. Even though we are assuming unlimited amounts of training data, the expected loss will typically not be zero. Problems where the loss can be reduced to zero are called *noiseless*. Medical prediction problems are typically very noisy.

Bayes' theorem, discussed in Chapter 6, now tells us how to incorporate separately specified expected proportions (*prior probabilities*) into this calculation: We simply multiply p_1 by π_1, p_2 by π_2, and so on. The expected loss from choosing class 1 is $0 \times p_1 \times \pi_1 + L(1, 2) \times p_2 \times \pi_2$.

Classification is more difficult when we do not have enough training data to use this simple approach to estimation, or when it is not feasible to keep the entire training set available for making predictions. Unfortunately, at least one of these limitations is almost always present. In this chapter we consider only the first problem, the most important in biostatistical applications. It is addressed by building regression models to estimate the probabilities p_k and then following the same strategy as if p_k were known. The accuracy of prediction, and thus the actual average loss, will be greater than in our ideal setting. The error rates in the ideal setting give a lower bound on the error rates attainable by any model; if these are low, improving a model may have a large payoff; if they are high, no model can predict well and improvements in the model may provide little benefit in error rates.

13.3 SIMPLE CLASSIFICATION MODELS

Linear and logistic models for classification have a long history and often perform reasonably well in clinical and epidemiologic classification problems. We describe them for the case of two classes, although versions for more than two classes are available. Linear and logistic discrimination have one important restriction in common: They separate the classes using a linear combination of the characteristics.

13.3.1 Logistic Regression

Example 13.1. Pine et al. [1983] followed patients with intraabdominal sepsis (blood poisoning) severe enough to warrant surgery to determine the incidence of organ failure or death (from sepsis). Those outcomes were correlated with age and preexisting conditions such as alcoholism and malnutrition. Table 13.1 lists the patients with the values of the associated variables. There are 21 deaths in the set of 106 patients. Survival status is indicated by the variable Y. Five potential predictor variables: shock, malnutrition, alcoholism, age, and bowel infarction were labeled X_1, X_2, X_3, X_4, and X_5, respectively. The four variables X_1, X_2, X_3, and X_5 were binary variables, coded 1 if the symptom was present and 0 if absent. The variable $X_4 =$ age in years, was retained as a continuous variable. Consider for now just variables Y and X_1; a 2×2 table could be formed as shown in Table 13.2.

With this single variable we can use the simple approach of matching new observations exactly to the training set. For a patient with shock, we would estimate a probability of death of $7/10 = 0.70$; for a patient without shock, we would estimate a probability of $14/96 = 0.15$.

Once we start to incorporate the other variables, this simple approach will break down. Using all four binary variables would lead to a table with 2^5 cells, and each cell would have too few observations for reliable estimates. The problem would be enormously worse when age is added to the model—there might be no patient in our training set who was an exact match on age.

We clearly need a way to simplify the model. One approach is to assume that to a reasonable approximation, the effect of one variable does not depend on the values of other variables, leading to a linear regression model:

$$P(\text{death}) = \pi = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_5 X_5$$

This model is unlikely to be ideal: If having shock increases the risk of death by 0.55, and the probability can be no larger than 1, the effects of other variables are severely limited. For this reason it is usual to transform the probability to a scale that is not limited by 0 and 1.

The most common reexpression of π leads to the logistic model

$$\log_e \frac{\pi}{1-\pi} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_5 X_5 \qquad (1)$$

commonly written as

$$\text{logit}(\pi) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_5 X_5 \qquad (2)$$

Table 13.1 Survival Status of 106 Patients Following Surgery and Associated Preoperative Variables[a]

ID	Y	X_1	X_2	X_3	X_4	X_5	ID	Y	X_1	X_2	X_3	X_4	X_5
1	0	0	0	0	56	0	301	1	0	1	0	50	1
2	0	0	0	0	80	0	302	0	0	0	0	20	0
3	0	0	0	0	61	0	303	0	0	0	0	74	1
4	0	0	0	0	26	0	304	0	0	0	0	54	0
5	0	0	0	0	53	0	305	1	0	1	0	68	0
6	1	0	1	0	87	0	306	0	0	0	0	25	0
7	0	0	0	0	21	0	307	0	0	0	0	27	0
8	1	0	0	1	69	0	308	0	0	0	0	77	0
9	0	0	0	0	57	0	309	0	0	1	0	54	0
10	0	0	1	0	76	0	401	0	0	0	0	43	0
11	1	0	0	1	66	1	402	0	0	1	0	27	0
12	0	0	0	0	48	0	501	1	0	1	1	66	1
13	0	0	0	0	18	0	502	0	0	1	1	47	0
14	0	0	0	0	46	0	503	0	0	0	1	37	0
15	0	0	1	0	22	0	504	0	0	1	0	36	1
16	0	0	1	0	33	0	505	1	1	1	0	76	0
17	0	0	0	0	38	0	506	0	0	0	0	33	0
19	0	0	0	0	27	0	507	0	0	0	0	40	0
20	1	1	1	0	60	1	508	0	0	1	0	90	0
22	0	0	0	0	31	0	510	0	0	0	1	45	0
102	0	0	0	0	59	1	511	0	0	0	0	75	0
103	0	0	0	0	29	0	512	1	0	0	1	70	1
104	0	1	0	0	60	0	513	0	0	0	0	36	0
105	1	1	0	0	63	1	514	0	0	0	1	57	0
106	0	0	0	0	80	0	515	0	0	1	0	22	0
107	0	0	0	0	23	0	516	0	0	0	0	33	0
108	0	0	0	0	71	0	518	0	0	1	0	75	0
110	0	0	0	0	87	0	519	0	0	0	0	22	0
111	1	1	1	0	70	0	520	0	0	1	0	80	0
112	0	0	0	0	22	0	521	1	0	1	0	85	0
113	0	0	0	0	17	0	523	0	0	1	0	90	0
114	1	0	0	1	49	0	524	1	0	0	1	71	0
115	0	1	0	0	50	0	525	0	0	0	1	51	0
116	0	0	0	0	51	0	526	1	0	1	1	67	0
117	0	0	1	1	37	0	527	0	0	1	0	77	0
118	0	0	0	0	76	0	529	0	0	0	0	20	0
119	0	0	0	1	60	0	531	0	0	0	0	52	1
120	1	1	0	0	78	1	532	1	1	0	1	60	0
122	0	0	1	1	60	0	534	0	0	0	0	29	0
123	1	1	1	0	57	0	535	0	0	0	0	30	1
202	0	0	0	0	28	1	536	0	0	0	0	20	0
203	0	0	0	0	94	0	537	0	0	0	0	36	0
204	0	0	0	0	43	0	538	0	0	1	1	54	0
205	0	0	0	0	70	0	539	0	0	0	0	65	0
206	0	0	0	0	70	0	540	1	0	0	0	47	0
207	0	0	0	0	26	0	541	0	0	0	0	22	0
208	0	0	0	0	19	0	542	1	0	0	1	69	0
209	0	0	0	0	80	0	543	1	0	1	1	68	0
210	0	0	1	0	66	0	544	0	0	1	1	49	0
211	0	0	1	0	55	0	545	0	0	0	0	25	0
214	0	0	0	0	36	0	546	0	1	1	0	44	0
215	0	0	0	0	28	0	549	0	0	0	1	56	0
217	0	0	0	0	59	1	550	0	0	1	1	42	0

Source: Data from Pine et al. [1983].

[a] See the text for labels.

Table 13.2 2 x 2 Table for Survival by Shock Status

		Y		
		Death	Survive	
X_1		1	0	
Shock	1	7	3	10
No Shock	0	14	82	96
		21	85	106

Four comments are in order:

1. The logit of p has range $(-\infty, \infty)$. The following values can easily be calculated:

$$\text{logit}(1) = +\infty$$

$$\text{logit}(0) = -\infty$$

$$\text{logit}(0.5) = 0$$

2. If we solve for π, the expression that results is

$$\pi = \frac{e^{\alpha+\beta_1 X_1+\cdots+\beta_5 X_5}}{1+e^{\alpha+\beta_1 X_1+\cdots+\beta_5 X_5}} = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\cdots+\beta_5 X_5)}} \tag{3}$$

3. We will write a for the estimate of α, b_1 for the estimate of β_1, and so on. Our estimated probability of death is obtained by inserting these values into equation (3) to get

$$\widehat{P}(\text{death}) = a + b_1 X_1 + b_2 X_2 + \cdots + b_5 X_5$$

4. The estimates are obtained by *maximum likelihood*. That is, we choose the values of a, b_1, b_2, ..., b_5 that maximize the probability of getting the death and survival values that we observed. In the simple situation where we can estimate a probability for each possible combination of characteristics, maximum likelihood gives the same answer as our rule of using the observed proportions. Note 13.1 gives the mathematical details. Any general-purpose statistical program will perform logistic regression.

We can check that with a single variable, logistic regression gives the same results as our previous analysis. In the previous analysis we used only the variable X_1, the presence of shock. If we fit this model to the data, we get

$$\text{logit}(\widehat{\pi}) = -1.768 + 2.615X_1$$

If $X_1 = 0$ (i.e., there is no shock),

$$\text{logit}(\widehat{\pi}) = -1.768$$

or

$$\widehat{\pi} = \frac{1}{1+e^{-(-1.768)}} = 0.146$$

If $X_1 = 1$ (i.e., there is shock),

$$\text{logit}(\widehat{\pi}) = -1.768 + 2.615 = 0.847$$

$$\widehat{\pi} = \frac{1}{1 + e^{-0.847}} = 0.700$$

This is precisely the probability of death given no preoperative shock. The coefficient of X_1, 2.615, also has a special interpretation: It is the logarithm of the odds ratio and the quantity $e^{b_1} = e^{2.615} = 13.7$ is the odds ratio associated with shock (as compared to no shock). This can be shown algebraically to be the case (see Problem 13.1).

Example 13.1. (*continued*) We now continue the analysis of the data of Pine et al. listed in Table 13.1. The output and calculations shown in Table 13.3 can be generated for all the variables. We would interpret these results as showing that in the presence of the remaining variables, malnutrition, is not an important predictor of survival status. All the other variables are significant predictors of survival status. All but variable X_4 are discrete binary variables. If malnutrition is dropped from the analysis, the estimates and standard errors are as given in Table 13.4.

If $\widehat{\pi}$ is the predicted probability of death, the equation is

$$\text{logit}(\widehat{\pi}) = -8.895 + 3.701X_1 + 3.186X_3 + 0.08983X_4 + 2.386X_5$$

For each of the values of X_1, X_3, X_5 (a total of eight possible combinations), a regression curve can be drawn for logit($\widehat{\pi}$) vs. age. In Figure 13.1 the lines are drawn for each of the eight combinations. For example, corresponding to $X_1 = 1$ (shock present), $X_3 = 0$ (no alcoholism), and $X_5 = 0$ (no infarction), the line

Table 13.3 Logistic Regression for Example 13.1

Variable	Regression Coefficient	Standard Error	Z-Value	p-Value
Intercept	−9.754	2.534	—	—
X_1 (shock)	3.674	1.162	3.16	0.0016
X_2 (malnutrition)	1.217	0.7274	1.67	0.095
X_3 (alcoholism)	3.355	0.9797	3.43	0.0006
X_4 (age)	0.09215	0.03025	3.04	0.0023
X_5 (infarction)	2.798	1.161	2.41	0.016

Table 13.4 Estimates and Standard Errors for Example 13.1

Variable	Regression Coefficient	Standard Error
Intercept	−8.895	2.314
X_1 (shock)	3.701	1.103
X_3 (alcoholism)	3.186	0.9163
X_4 (age)	0.08983	0.02918
X_5 (infarction)	2.386	1.071

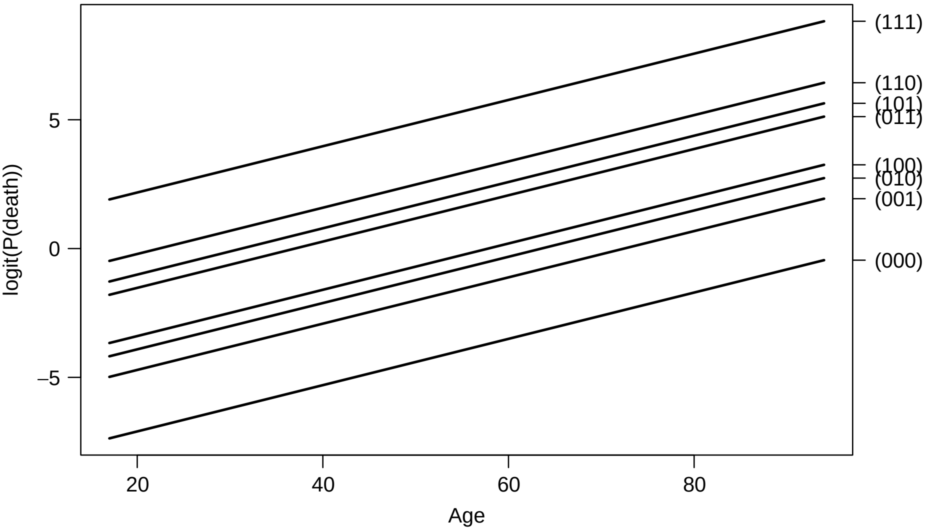

Figure 13.1 Logit of estimated probability of death as a function of age in years and category of status of (X_1, X_3, X_5). (Data from Pine et al. [1983].)

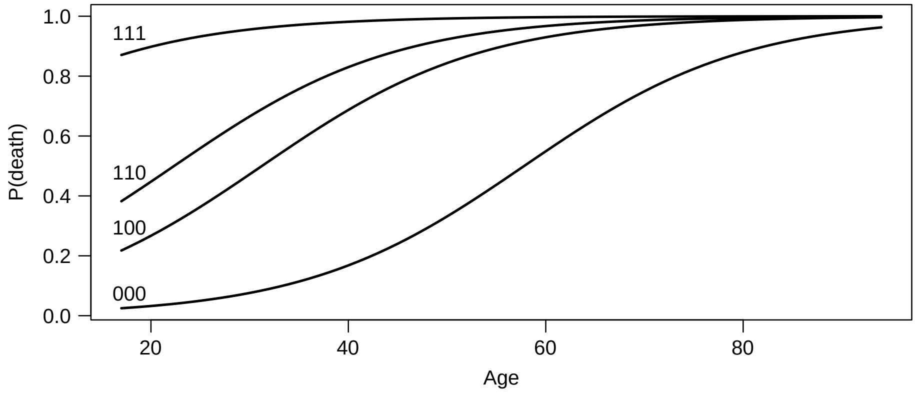

Figure 13.2 Estimated probability of death as a function of age in years and selected values of (X_1, X_3, X_5). (Data from Pine et al. [1983].)

$$\begin{aligned}\text{logit}(\widehat{\pi}) &= -8.895 + 3.701 + 0.08983X_4\\ &= -5.194 + 0.08983X_4\end{aligned}$$

is drawn.

This line is indicated by "(100)" as a shorthand way of writing $(X_1 = 1, X_3 = 0, X_5 = 0)$. The eight lines seem to group themselves into four groups: the top line representing all three symptoms present; the next three lines, groups with two symptoms present; the next three lines, groups with one symptom present; and finally, the group at lowest risk with no symptoms present. In Figure 13.2 the probability of death is plotted on the original probability scale; only four of the eight groups have been graphed. The group at highest risk is the one with all three binary risk factors present. One of the advantages of the model is that we can draw a curve for the situation with all three risk factors present even though there are no patients in that category; but the estimate depends on the model. The curve is drawn on the assumption that the risks are additive in the logistic scale (that is what we *mean* by a linear model). This assumption can be partially tested by including interaction terms involving these three covariates in the model and testing their significance. When this was done, none of the interaction terms were significant, suggesting that the additive model is a reasonable one. Of course, as there are no patients with all three risk factors present, there is no way to perform a complete test of the model.

13.3.2 Linear Discrimination

The first statistical approach to classification, as with so many other problems, was invented by R. A. Fisher. Fisher's linear discriminant analysis is designed for continuous characteristics that have a normal distribution (in fact, a multivariate normal distribution; any sums or differences of multiples of the variables should be normally distributed).

Definition 13.1. A set of random variables $X_1, \dots, X_k$ is *multivariate normal* if every linear combination of $X_1, \dots, X_k$ has a normal distribution.

In addition, we assume that the variances and covariances of the characteristics are the same in the two groups. Under these assumptions, Fisher's method finds a combination of variables (a *discriminant function*) for distinguishing the classes:

$$\Delta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Assuming equal losses for different errors, an observation is assigned to class 1 if $\Delta > 0$ and class 2 if $\Delta < 0$. Estimation of the parameters β again uses maximum likelihood. It is also possible to compute probabilities p_k for membership of each class using the normal cumulative distribution function: $p_1 = \Phi(\Delta)$, $p_2 = 1 - \Phi(\Delta)$, where Φ is the symbol for the cumulative normal distribution.

Because linear discrimination makes more assumptions about the structure of the X's than logistic regression does, it gives more precise estimates of its parameters and more precise predictions [Efron, 1975]. However, in most medical examples the uncertainty in the parameters is a relatively small component of the overall prediction error, compared to model uncertainty and to the inherent unpredictability of human disease. In addition to requiring extra assumptions to hold, linear discrimination is likely to give substantial improvements only when the characteristics determine the classes very accurately so that the main limitation is the accuracy of statistical estimation of the parameters (i.e., a nearly "noiseless" problem).

The robustness can be explained by considering another equivalent way to define Δ. Let D_1 and D_2 be the mean of Δ in groups 1 and 2, respectively, and V be the variance of Δ within each group (assumed to be the same). Δ is the linear combination that maximizes

$$\frac{(D_1 - D_2)^2}{V}$$

the ratio of the between-group and within-group variances.

Truett et al. [1967] applied discriminant analysis to the data of the Framingham study. This was a longitudinal study of the incidence of coronary heart disease in Framingham, Massachusetts. In their prediction model the authors used continuous variables such as age (years) and serum cholesterol (mg/100 mL) as well as discrete or categorical variables such as cigarettes per day (0 = never smoked, 1 = less than one pack a day, 2 = one pack a day, 3 = more than a pack a day) and ECG (0 = normal, 1 = certain kinds of abnormality). It was found that the linear discriminant model gave reasonable predictions. Halperin [1971] came to five conclusions, which have stood the test of time. If the logistic model holds but the normality assumptions for the predictor variables are violated, they concluded that:

1. β_i that are zero will tend to be estimated as zero for large samples by the method of maximum likelihood but not necessarily by the discrimination function method.
2. If any β_i are nonzero, they will tend to be estimated as nonzero by either method, but the discriminant function approach will give asymptotically biased estimates for those β_i and for α.
3. Empirically, the assessment of significance for a variable, as measured by the ratio of the estimated coefficient to its estimated standard error, is apt to be about the same whichever method is used.
4. Empirically, the maximum likelihood method usually gives slightly better fits to the model as evaluated from observed and expected numbers of cases per decile of risk.
5. There is a theoretical basis for the possibility that the discriminant function will give a very poor fit even if the logistic model holds.

Some of these empirical conclusions are supported theoretically by Li and Duan [1989] and Hall and Li [1993], who considered situations similar to this one, where a linear combination

$$\Delta = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

is to be estimated under either of two models. They showed that under some assumptions about the distribution of variables X, using the wrong model would typically lead to estimating

$$\Delta = c\beta_1 X_1 + c\beta_2 X_2 + \cdots + c\beta_p X_p$$

for some constant c. When these conditions apply, using linear discrimination would tend to lead to a similar discriminant function Δ but to poor estimation of the actual class probabilities. See also Knoke [1982]. Problems 13.4, 13.6, and 13.7 address some of these issues.

In the absence of software specifically designed for this method, linear discrimination can be performed with software for linear regression. The details, which are of largely historical interest, are given in Note 13.4.

13.4 ESTIMATING AND SUMMARIZING ACCURACY

When choosing between classification models or describing the performance of a model, it is necessary to have some convenient summaries of the error rates. It is usually important to distinguish between different kinds of errors, although occasionally a simple estimate of the expected loss will suffice.

Statistical methodology is most developed for the case of two classes. In biostatistics, these are typically presence and absence of disease.

13.4.1 Sensitivity and Specificity

In assigning people to two classes (disease and no disease) we can make two different types of error:

1. Detecting disease when none is present
2. Missing disease when it is there

As in Chapter 6, we define the *sensitivity* as the probability of detecting disease given that disease is present (avoiding an error of the first kind) and *specificity* as the probability of not detecting disease given that no disease is present (avoiding an error of the second kind).

The sensitivity and specificity are useful because they can be estimated from separate samples of persons with and without disease, and because they often generalize well between populations. However, in actual use of a classification rule, we care about the probability that a person has disease given that disease was detected (the *positive predictive value*) and the probability that a person is free of disease given that no disease was detected (the *negative predictive value*).

It is a common and serious error to confuse the sensitivity and the positive predictive value. In fact, for a reasonably good test and a rare disease, the positive predictive value depends almost entirely on the disease prevalence and on the specificity. Consider the mammography example mentioned in Section 13.2. Of 1000 women who have a mammogram, about 100 will be recalled for further testing and 7 of those will have cancer. The positive predictive value is 7%, which is quite low, not because the sensitivity of the mammogram is poor but because 93 of those 1000 women are falsely testing positive. Because breast cancer is rare, false positives greatly outnumber true positives, regardless of how sensitive the test is.

When a single binary characteristic is all that is available, the sensitivity and specificity describe the properties of the classification rule completely. When classification is based on a summary criterion such as the linear discriminant function, it is useful to consider the sensitivity and specificity based on a range of possible thresholds.

Example 13.2. Tuberculosis testing is important in attempts to control the disease, which can be quite contagious but in most countries is still readily treatable with a long course of antibiotics. Tests for tuberculosis involve injecting a small amount of antigen under the skin and looking for an inflamed red area that appears a few days later, representing an active T-cell response to the antigen. The size of this indurated area varies from person to person both because of variations in disease severity and because of other individual factors. Some people with HIV infection have no reaction even with active tuberculosis (a state called *anergy*). At the other extreme, migrants from countries where the BCG vaccine is used will have a large response irrespective of their actual disease status (and since the vaccine is incompletely effective, they may or may not have disease).

The diameter of the indurated area is used to classify people as disease-free or possibly infected. It is important to detect most cases of TB (high sensitivity) without too many false positives being subjected to further investigation and unnecessary treatment (high positive predictive value). The diameter used to make the classification varies depending on characteristics of the patient. A 5-mm induration is regarded as positive for close contacts of people with active TB infection or those with chest x-rays suggestive of infection because the prior probability of risk is high. A 5-mm induration is also regarded as positive for people with compromised immune systems due to HIV infection or organ transplant, partly because they are likely to have weaker T-cell responses (so a lower threshold is needed to maintain sensitivity) and partly because TB is much more serious in these people (so the loss for a false negative is higher).

For people at moderately high risk because they are occupationally at higher risk or because they come from countries where TB is common, a 10-mm induration is regarded as positive (their prior probability is moderately elevated). The 10-mm rule is also used for people with poor access to health care or those with diseases that make TB more likely to become active (again, the loss for a false negative is higher in these groups).

Finally, for everyone else, a 15-mm threshold is used. In fact, the recommendation is that they typically not even be screened, implicitly classifying everyone as negative.

Given a continuous variable predicting disease (whether an observed characteristic or a summary produced by logistic or linear discrimination), we would like to display the sensitivity and specificity not just for one threshold but for all possible thresholds. The *receiver operating characteristic* (ROC) *curve* is such a display. It is a graph with "sensitivity" on the y-axis and "1 − specificity" on the x-axis, evaluated for each possible threshold.

If the variable is completely independent of disease, the probability of detecting disease will be the same for people with and without disease, so "sensitivity" and "1 − specificity"

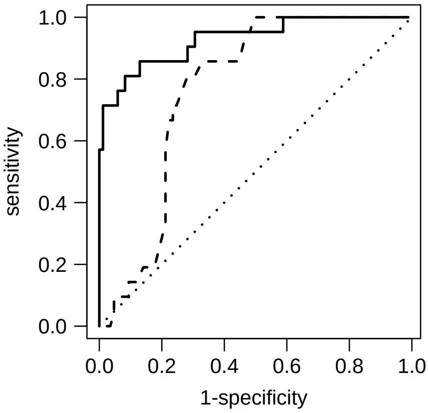

Figure 13.3 Receiver operating characteristic curve for data of Pine et al. [1983]. The solid line is the prediction from all five variables; the dashed line is the prediction from age alone.

will be the same. This is indicated by a diagonal line in Figure 13.3. If higher values of the variable are associated with higher risks of disease, the curve will lie above the diagonal line. By convention, if lower values of the variable are associated with higher risks of disease, the variable is transformed to reverse this, so ROC curves should always lie above the diagonal line.

The area under the ROC curve is a measure of how well the variable discriminates a disease state: If you are given one randomly chosen person with disease and one randomly chosen person without disease, the area under the ROC curve is the probability that the person with disease has the higher value of the variable. The area under the ROC curve is a good analog for binary data of the r^2 value for linear models.

Drawing the ROC curve for two classification rules allows you to compare their accuracy at a range of different thresholds. It might be, for example, that two rules have very different sensitivity when their specificity is low but very similar sensitivity when their specificity is high. In that case, the rules would be equivalently useful in screening low-risk populations, where specificity must be high, but might be very different in clinical diagnostic use.

13.4.2 Internal and External Error Rates

The *internal* or *apparent* or *training* or *in-sample error rates* are those obtained on the same data as those used to fit the model. These always underestimate the true error rate, sometimes very severely. The underestimation becomes more severe when many characteristics are available for modeling, when the model is very flexible in form, and when the data are relatively sparse.

An extreme case is given by a result from computer science called the *perceptron capacity bound* [Cover, 1965]. Suppose that there are d continuous characteristics and n observations from two classes in the training set, and suppose that the characteristics are purely random, having no real association whatsoever with the classes. The probability of obtaining an in-sample error rate of zero for some classification rule based on a single linear combination of characteristics is then approximately

$$1 - \Phi\left(\frac{n-2d}{\sqrt{n}}\right)$$

If d is large and $n/d < 2$, this probability will be close to 1. Even without considering nonlinear models and interactions between characteristics, it is quite possible to obtain an apparent error rate of zero for a model containing no information whatsoever. Note that $n/d > 2$ does not guarantee a good in-sample estimate of the error rate; it merely rules out this worst possible case.

Estimates of error rates are needed for model selection and in guiding the use of classification models, so this is a serious problem. The only completely reliable solution is to compute the error rate on a completely new sample of data, which is often not feasible.

When no separate set of data will be available, there are two options:

1. Use only part of the data for building the model, saving out some data for testing.
2. Use all the data for model building and attempt to estimate the true error rate statistically.

Experts differ on which of these is the best strategy, although the majority probably leans toward the second strategy. The first strategy has the merit of simplicity and requires less programming expertise. We discuss one way to estimate the true error rate, cross-validation, and one way to choose between models without a direct error estimate, the Akaike information criterion.

13.4.3 Cross-Validation

Statistical methods to estimate true error rate are generally based on the idea of refitting a model to part of the data and using the refitted model to estimate the error rate on the rest of the data. Refitting the model is critical so that the data left out are genuinely independent of the model fit. It is important to note that refitting ideally means redoing the entire model selection process, although this is feasible only when the process was automated in some way.

In *10-fold cross-validation*, the most commonly used variant, the data are randomly divided into 10 equal pieces. The model is then refitted 10 times, each time with one of the 10 pieces left out and the other nine used to fit the model. The classification errors (either the expected loss or the false positive and false negative rates) are estimated for the left-out data from the refitted model. The result is an estimate of the true error rate, since each observation has been classified using a model fitted to data not including that observation. Clearly, 10-fold cross-validation takes 10 times as much computer time as a single model selection, but with modern computers this is usually negligible. Cross-validation gives an approximately unbiased estimate of the true error rate, but a relatively noisy one.

13.4.4 Akaike's Information Criterion

Akaike's information criterion (AIC) [Akaike, 1973] is an asymptotic estimate of expected loss for a particular loss function, one that is proportional to the logarithm of the likelihood. It is extremely simple to compute but can only be used for models fitted by maximum likelihood and requires great caution when used to compare models fitted by different modeling techniques. In the case of linear regression, model selection with AIC is equivalent to model selection with Mallow's C_p, discussed in Chapter 11, so it can be seen as a generalization of Mallow's C_p to nonlinear models.

The primary difficulty in model selection is that increasing the number of variables always decreases the apparent error rate even if the variables contain no useful information. The AIC is based on the observation that for one particular loss function, the log likelihood, the decrease depends only on the number of variables added to the model. If a variable is uninformative, it will on average increase the log likelihood by 1 unit. When comparing model A to model B, we can compute

$$\begin{gathered}\log(\text{likelihood of A}) - \log(\text{likelihood of B})\\ -(\text{no. parameters in A} - \text{no. parameters in B})\end{gathered} \tag{4}$$

If this is positive, we choose model A, if it is negative we choose model B. The AIC is most often defined as

$$\text{AIC} = -2\,\log(\text{likelihood of model}) + 2(\text{no. parameters in model}) \tag{5}$$

so that choosing the model with the lower AIC is equivalent to our strategy based on equation (4). Sometimes the AIC is defined without the factor of -2, in which case the largest value indicates the best model: It is important to check which definition is being used.

Akaike showed that given two fixed models and increasing amounts of data, this criterion would eventually pick the best model. When the number of candidate models is very large, like the 2^p models in logistic regression with p characteristics, AIC still tends to overfit to some extent. That is, the model chosen by the AIC tends to have more variables than the best model.

In principle, the AIC can be used to compare models fitted by different techniques, but caution is needed. The log likelihood is only defined up to adding or subtracting an arbitrary constant, and different programs or different procedures within the same program may use different constants for computational convenience. When comparing models fitted by the same procedure, the choice of constant is unimportant, as it cancels out of the comparison. When comparing models fitted by different procedures, the constant does matter, and it may be difficult to find out what constant has been used.

13.4.5 Automated Stepwise Model Selection

Automated stepwise model selection has a deservedly poor reputation when the purpose of a model is causal inference, as model choice should then be based on a consideration of the probable cause-and-effect relationships between variables. When modeling for prediction, however, this is unimportant: We do not need to know *why* a variable is predictive to know that it *is* predictive.

Most statistical packages provide tools that will automatically consider a set of variables and attempt to find the model that gives the best prediction. Some of these use AIC, but more commonly they use significance testing of predictors. Stepwise model selection based on AIC can be approximated by significance-testing selection using a critical p-value of 0.15.

Example 13.3. We return to the data of Pine et al. [1983] and fit a logistic model by stepwise search, optimizing the AIC. We begin with a model using none of the characteristics and giving the same classification for everyone. Each of the five characteristics is considered for adding to the model, and the one optimizing the AIC is chosen. At subsequent steps, every variable is considered either for adding to the model or for removal from the model. The procedure stops when no change improves the AIC.

This procedure is not guaranteed to find the best possible model but can be carried out much more quickly than an exhaustive search of all possible models. It is at least as good as, and often better than, forward or backward stepwise procedures that only add or only remove variables.

Starting with an empty model the possible changes were as follows:

	d.f.	**Deviance**	**AIC**		**d.f.**	**Deviance**	**AIC**
+ X4	1	90.341	94.341	+ X5	1	97.877	101.877
+ X1	1	91.977	95.977	+ X2	1	99.796	103.796
+ X3	1	95.533	99.533	<none>		105.528	107.528

The d.f. column counts the number of degrees of freedom for each variable (in this case, one for each variable, but more than one if a variable had multiple categories). The deviance is -2 log likelihood. The best (lowest AIC) choice was to add X4 (age). In the second step, X1 (shock) was added, and then X3 (alcoholism). The possible changes in the fourth step were:

	d.f.	**Deviance**	**AIC**		**d.f.**	**Deviance**	**AIC**
+ X5	1	56.073	66.073	− X4	1	76.970	82.970
<none>		61.907	69.907	− X3	1	79.088	85.088
+ X2	1	60.304	70.304	− X1	1	79.925	85.925

Table 13.5 Step 1 Using Linear Discrimination

	d.f.	SS	RSS	AIC
+ X1	1	2.781	14.058	−210.144
+ X4	1	2.244	14.596	−206.165
+ X3	1	1.826	15.014	−203.172
+ X5	1	1.470	15.370	−200.691
+ X2	1	0.972	15.867	−197.312
<none>			16.840	−193.009

Table 13.6 Subsequent Steps Using Linear Discrimination

	d.f.	SS	RSS	AIC
<none>			10.031	−239.922
+ X2	1	0.164	9.867	−239.673
− X5	1	0.733	10.764	−234.447
− X4	1	0.919	10.950	−232.627
− X3	1	1.733	11.764	−225.029
− X1	1	2.063	12.094	−222.095

and the lowest AIC came with adding X5 (infarction) to the model. Finally, adding X2 also reduced the AIC, and no improvement could be obtained by deleting a variable, so the procedure terminated. The model minimizing AIC uses all five characteristics.

We can perform the same classification using linear discrimination. The characteristics clearly do not have a multivariate normal distribution, but it will be interesting to see how well the robustness of the methods stands up in this example.

At the first step we have the data shown in Table 13.5.

For this linear model the residual sum of squares and the change in residual sum of squares are given and used to compute the AIC. The first variable added is X1. In subsequent steps X3, X4, and X5 are added, and then we have the data shown in Table 13.6.

The procedure ends with a model using the four variables X1, X3, X4, and X5. The fifth variable (malnutrition) is not used. We can now compare the fitted values from the two models shown in Figure 13.4. It is clear that both discriminant functions separate the surviving and dying patients very well and that the two functions classify primarily the same people as being at high risk. Looking at the ROC curves suggests that the logistic discriminant function is very slightly better, but this conclusion could not be made reliably without independent data.

13.5 MODERN CLASSIFICATION TECHNIQUES

Most modern classification techniques are similar in spirit to automated stepwise logistic regression. A computer search is made through a very large number of possible models for p_k, and a criterion similar to AIC or an error estimate similar to cross-validation is used to choose a model. All these techniques are capable of approximating any relationship between p_k and X arbitrarily well, and as a consequence will give very good prediction if n is large enough in relation to p.

Modern classification techniques often produce "black-box" classifiers whose internal structure can be difficult to understand. This need not be a drawback: As the models are designed for prediction rather than inference about associations, the opaqueness of the model reduces the

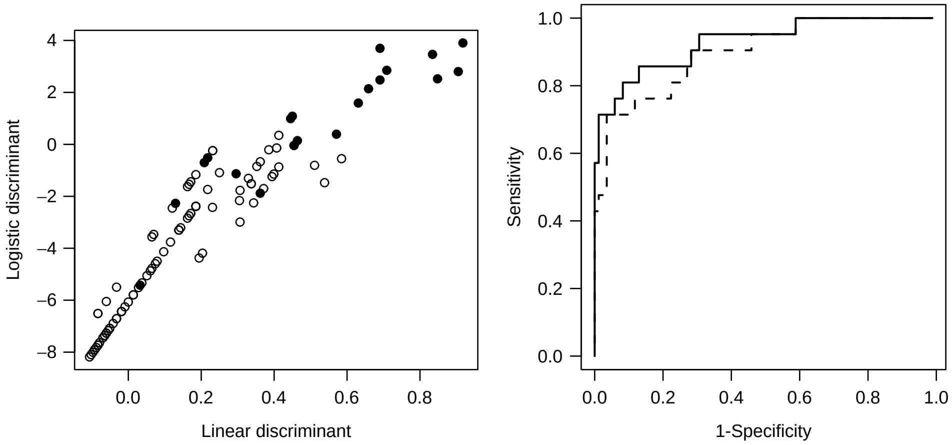

Figure 13.4 Comparison of discriminant functions and ROC curves from logistic and linear models for data of Pine et al. [1983]. Solid circles are deaths; open circles are survival. The solid line is the logistic model; the dashed line is the linear model.

temptation to leap to unjustified causal conclusions. On the other hand, it can be difficult to decide which variables are important in the classification and how strongly the predictions have been affected by outliers. There is some current statistical research into ways of opening up the black box, and techniques may become available over the next few years.

At the time of writing, general-purpose statistical packages often have little classification functionality beyond logistic and linear discrimination. It is still useful for the nonspecialist to understand the concepts behind some of these techniques; we describe two samples.

13.5.1 Recursive Partitioning

Recursive partitioning is based on the idea of classifying by making repeated binary decisions. A *classification tree* such as the left side of Figure 13.5 is constructed step by step:

1. Search every value c of every variable X for the best possible prediction by $X > c$ vs. $X \leq c$.
2. For each of the two resulting subsets of the data, repeat step 1.

In the tree displayed, each split is represented by a logical expression, with cases where the expression is true going left and others going right, so in the first split in Figure 13.5 the cases with white blood cell counts below 391.5 mL^{-1} go to the left.

An exhaustive search procedure such as this is sure to lead to overfitting, so the tree is then *pruned* by snipping off branches. The pruning is done to minimize a criterion similar to AIC:

$$\text{loss} + \text{CP} \times \text{number of splits}$$

The value of CP, called the *cost-complexity penalty*, is most often chosen by 10-fold cross-validation (Section 13.4.3). Leaving out 10% of the data, a tree is grown and pruned with many different values of CP. For each tree pruned, the error rate is computed on the 10% of data left out. This is repeated for each of the ten 10% subsets of the data. The result is a cross-validation estimate of the loss (error rate) for each value of CP, as in the right-hand side of Figure 13.5.

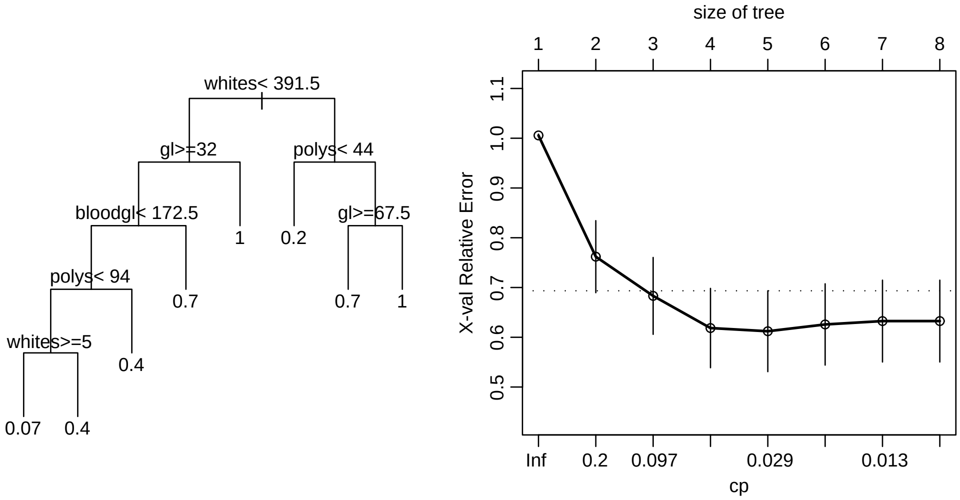

Figure 13.5 Classification tree and cross-validated error rates for differential diagnosis of acute meningitis.

Because cross-validation is relatively noisy (see the standard error bars on the graph), we choose the largest CP (smallest tree) that gives an error estimate within one standard error of the minimum, represented by the horizontal dotted line on the graph.

Example 13.4. In examining these methods we use data from Spanos et al. [1989], made available by Frank Harrell at a site linked from the Web appendix to the chapter. The classification problem is to distinguish viral from bacterial meningitis, based on a series of 581 patients treated at Duke University Medical Center. As immediate antibiotic treatment for acute bacterial meningitis is often life-saving, it is important to have a rapid and accurate initial classification. The definitive classification based on culturing bacteria from cerebrospinal fluid samples will take a few days to arrive. In some cases bacteria can be seen in the cerebrospinal fluid, providing an easy decision in favor of bacterial meningitis with good specificity but inadequate sensitivity.

The initial analysis used logistic regression together with transformations of the variables, but we will explore other possibilities. We will use the following variables:

- *AGE*: in years
- *SEX*
- *BLOODGL*: glucose concentration in blood
- *GL*: glucose concentration in cerebrospinal fluid
- *PR*: protein concentration in cerebrospinal fluid
- *WHITES*: white blood cells per milliliter of cerebrospinal fluid
- *POLYS*: % of white blood cells that are polymorphonuclear leukocytes
- *GRAM*: result of Gram smear (bacteria seen under microscope): 0 negative, $>$ 0 positive
- *ABM*: 1 for bacterial, 0 for viral meningitis

The original analysis left GRAM out of the model and used it only to override the predicted classification if GRAM $>$ 0. This is helpful because the variable is missing in many cases, and because the decision to take a Gram smear appears to be related to suspicion of bacterial meningitis.

In the resulting tree, each *leaf* is labeled with the probability of bacterial meningitis for cases ending up in that leaf. Note that they range from 1 down to 0.07, so that in some cases bacterial meningitis is almost certain, but it is harder to be certain of viral meningitis.

It is interesting to note what happens when Gram smear status is added to the variable list for growing a tree. It is by far the most important variable, and prediction error is distinctly reduced. On the other hand, bacterial meningitis is predicted not only in those whose Gram smear is positive, but also in those whose Gram smear is negative. Viral meningitis is predicted only in a subset of those whose Gram smear is missing. If the goal of the model were to classify the cases retrospectively from hospital records, this would not be a problem. However, the original goal was to construct a diagnostic tool, where it is undesirable to have the prediction strongly dependent on another physician choice. Presumably, the Gram smear was being ordered based on other information available to the physician but not to the investigators.

Classification trees are particularly useful where there are strong interactions between characteristics. Completely different variables can be used to split each subset of the data. In our example tree, blood glucose is used only for those with high white cell counts and high glucose in the cerebrospinal fluid. This ability is particularly useful when there are missing data.

On the other hand, classification trees do not perform particularly well when there are smooth gradients in risk with a few characteristics. For example, the prediction of acute bacterial meningitis can be improved by adding a new variable with the ratio of blood glucose to cerebrospinal fluid glucose.

The best known version of recursive partitioning, and arguably the first to handle overfitting carefully, is the CART algorithm of Breiman et al. [1984]. Our analysis used the free "rpart" package [Therneau, 2002], which automates both fitting and the cross-validation analysis. It follows the prescriptions of Breiman et al. [1984] quite closely.

A relatively nontechnical overview of recursive partitioning in biostatistics is given by Zhang and Singer [1999]. More recently, techniques using multiple classification trees (*bagging*, *boosting*, and *random forests*) have become popular and appear to work better with very large numbers of characteristics than do other methods.

13.5.2 Neural Networks

The terminology *neural network* and the original motivation were based on a model for the behavior of biological neurons in the brain. It is now clear that real neurons are much more complicated, and that the fitting algorithms for neural networks bear no detailed relationship to anything happening in the brain. Neural networks are still very useful black-box classification tools, although they lack the miraculous powers sometimes attributed to them.

A computational neuron in a neural net is very similar to a logistic discrimination function. It takes a list of inputs $Z_1, Z_2, \ldots, Z_m$ and computes an output that is a function of a weighted combination of the inputs, such as

$$\operatorname{logit}(\alpha + \beta_1 Z_1 + \cdots + \beta_m Z_m) \tag{6}$$

There are many variations on the exact form of the output function, but this is one widely used variation. It is clear from equation (6) that even a single neuron can reproduce any classification from logistic regression.

The real power of neural network models comes from connecting multiple neurons together in at least two layers, as shown in Figure 13.6. In the first layer the inputs are the characteristics $X_1, \ldots, X_p$. The outputs of these neurons form a "hidden layer" and are used as inputs to the second layer, which actually produces the classification probability p_k.

Example 13.5. A neural net fitted to the acute meningitis data has problems because of missing observations. Some form of imputation or variable selection would be necessary for a

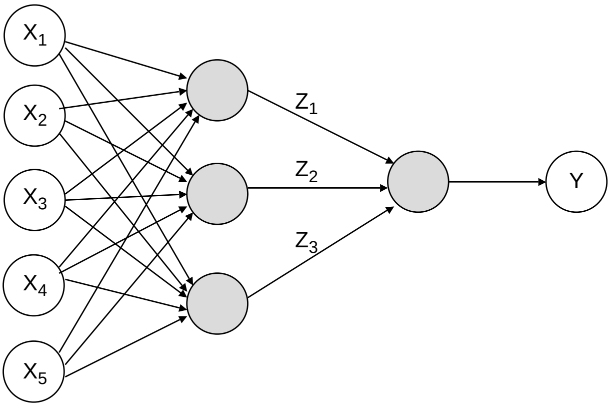

Figure 13.6 Simple neural network with three hidden nodes.

serious analysis of these data. We used the neural network package that accompanies Venables and Ripley [2002], choosing a logistic output function and two hidden nodes (Z_1 and Z_2). That is, the model was

$$\begin{aligned}
\text{logit}(p) &= -0.52 + 2.46Z_1 - 2.31Z_2 \\
\text{logit}(Z_1) &= 0.35 + 0.11\text{POLYS} + 0.58\text{WHITES} - 0.31\text{SEX} + 0.39\text{AGE} \\
&\quad - 0.47\text{GL} - 2.02\text{BLOODGL} - 2.31\text{PR} \\
\text{logit}(Z_2) &= 0.22 + 0.66\text{POLYS} + 0.25\text{WHITES} - 0.06\text{SEX} + 0.31\text{AGE} \\
&\quad + 0.03\text{GL} + 0.33\text{BLOODGL} - 0.02\text{PR}
\end{aligned}$$

The sensitivity of the classification was approximately 50% and the specificity nearly 90%.

Two hidden nodes is the minimum interesting number (one hidden node just provides a transformation of a logistic regression model), and we did not want to use more than this because of the relatively small size of the data set.

NOTES

13.1 Maximum Likelihood for Logistic Regression

The regression coefficients in the logistic regression model are estimated using the maximum likelihood criterion. A full discussion of this topic is beyond the scope of this book, but in this note we outline the procedure for the situation involving one covariate. Suppose first that we have a Bernoulli random variable, Y, with probability function

$$\begin{aligned}
P[Y = 1] &= p \\
P[Y = 0] &= 1 - p
\end{aligned}$$

A mathematical trick allows us to combine these into one expression:

$$P[Y = y] = p^y(1 - p)^{(1-y)}$$

using the fact that any number to the zero power is 1. We observe n values of Y, $y_1, y_2, \ldots, y_n$ (a sequence of zeros and ones). The probability of observing this sequence is proportional to

$$\prod_{j=1}^{n} p^{y_j}(1-p)^{1-y_j} = p^{\Sigma y_j}(1-p)^{n-\Sigma y_j} \tag{7}$$

This quantity is now considered as a function of p and defined to be the likelihood. To emphasize the dependence on p, we write

$$L\left(p|\sum y_j, n\right) = p^{\Sigma y_j}(1-p)^{n-\Sigma y_j} \tag{8}$$

Given the value of $\sum y_j$, what is the "best" choice for a value for p? The maximum likelihood principle states that the value of p that maximizes $L(p|\sum y_j, n)$ should be chosen. It can be shown by elementary calculus that the value of p that maximizes $L(p|\sum y_j, n)$ is equal to $\sum y_j/n$. You will recognize this as the proportion of the n values of Y that have the value 1. This can also be shown graphically; Figure 13.7 is a graph of $L(p|\sum y_j, n)$ as a function of p for the situation $\sum y = 6$ and $n = 10$. Note that the graph has one maximum and that it is not quite symmetrical.

In the logistic regression model the probability p is assumed to be a function of an underlying covariate, X; that is, we model

$$\text{logit}(p) = \alpha + \beta X$$

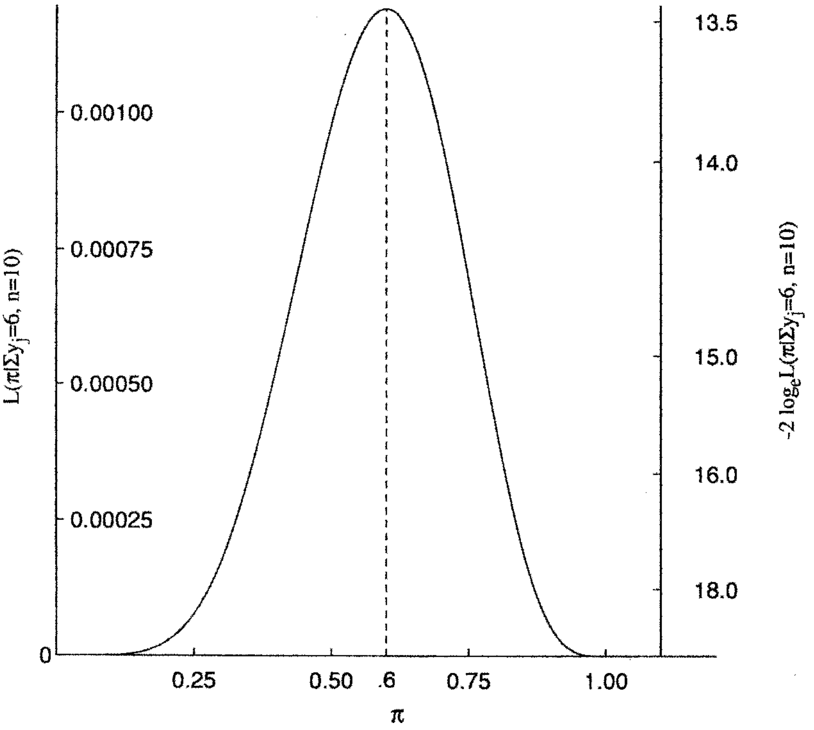

Figure 13.7 Likelihood function, $L(\pi|6, 10)$.

where α and β are constants. Conversely,

$$p = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} = \frac{1}{1+e^{-(\alpha+\beta X)}} \tag{9}$$

For fixed values of X the probability p is determined (since α and β are parameters to be estimated from the data). A set of data now consists of *pairs* of observations: (y_j, x_j), $j = 1, \ldots, n$, where y_j is again a zero–one variable and x_j is an observed value of X for set j. For each outcome, indexed by set j, there is now a probability $p(j)$ determined by the value of x_j. The likelihood function is written

$$L(p(1), \ldots, p(n)|y_1, \ldots, y_n, x_1, \ldots, x_n, n) = \prod_{j=1}^{n} p(j)^{y_j}[1 - p(j)]^{1-y_j} \tag{10}$$

but $p(j)$ can be expressed as

$$p(j) = \frac{e^{\alpha+\beta X_j}}{1+e^{\alpha+\beta X_j}} \tag{11}$$

where x_j is the value of the covariate for subject j. The likelihood function can then be written and expressed as a function of α and β as follows:

$$\begin{aligned} L(\alpha, \beta|y_1, \ldots, y_n; x_1, \ldots, x_n; n) &= \prod_{j=1}^{n} \left(\frac{e^{\alpha+\beta x_j}}{1+e^{\alpha+\beta x_j}}\right)^{y_j} \left(\frac{1}{1+e^{\alpha+\beta x_j}}\right)^{1-y_j} \\ &= \prod_{j=1}^{n} \frac{(e^{\alpha+\beta x_j})^{y_j}}{1+e^{\alpha+\beta x_j}} \\ &= \frac{e^{\Sigma_{j=1}^{n} y_j(\alpha+\beta x_j)}}{\prod_{j=1}^{n}(1+e^{\alpha+\beta x_j})} \end{aligned} \tag{12}$$

The maximum likelihood criterion then requires values for α and β to be chosen so that the likelihood function above is maximized. For more than one covariate, the likelihood function can be deduced similarly.

13.2 Logistic Discrimination with More Than Two Groups

Anderson [1972] and Jones [1975], among others, have considered the case of logistic discrimination with more than two groups. Following Anderson [1972], let for two groups

$$P(G_1|X) = \frac{\exp(\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p)}{1 + \exp(\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p)}$$

Then

$$P(G_2|X) = \frac{1}{1 + \exp(\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p)}$$

This must be so because $P(G_1|X) + P(G_2|X) = 1$; that is, the observation X belongs to either the G_1 or G_2. For k groups, define

$$P(G_s|X) = \frac{\exp(\alpha_{s0} + \alpha_{s1} X_1 + \cdots + \alpha_{sp} X_p)}{1 + \sum_{j=1}^{k-1} \exp(\alpha_{j0} + \alpha_{j1} X_1 + \cdots + \alpha_{jp} X_p)}$$

for groups $s = 1, \ldots, k-1$, and for group G_k, let

$$P(G_k|X) = \frac{1}{1 + \sum_{j=1}^{k-1} \exp(\alpha_{j0} + \alpha_{j1}X_1 + \cdots + \alpha_{jp}X_p)} \tag{13}$$

Most statistical packages provide this analysis, which is often called *polytomous logistic regression* (or occasionally and incorrectly, "polychotomous" logistic regression).

13.3 Defining Losses

In order to say that one prediction is better than another, we need some way to compare the relative importance of false positive and false negative errors. Even looking at total error rate implicitly assigns a relative importance. When the main adverse or beneficial effects are directly comparable, this is straightforward. We can compare the monetary costs of false negatives and false positives, or the probability of death caused by a false positive or false negative. In most cases, however, there will not be direct comparability. When evaluating a cancer screening program, the cost of false negatives is an increase in the risk of death, due to untreated cancer. The cost of a false positive includes the emotional effects and health risks of further testing needed to rule out disease. Even without weighing monetary costs against health costs we can see that it is not clear how many false negatives are worth one false positive. The problem is much more controversial, although perhaps no more difficult when monetary costs are important, as they usually are.

It can be shown [Savage, 1954] that the ability to make consistent choices between courses of action whose outcome is uncertain implies the ability to rate all the possible outcomes on the same scale, so this problem cannot be avoided. Perhaps the most important general guidance we can give is that it is important to recognize that different people will assign different losses and so prefer different classification rules.

13.4 Linear Discrimination Using Linear Regression Software

Given two groups of size n_1 and n_2, it has been shown by Fisher [1936] that the discriminant analysis is equivalent to a multiple regression on the dummy variable Y defined as follows:

$$\begin{aligned} Y &= \frac{n_2}{n_1 + n_2} \text{members of group 1} \\ &= \frac{-n_1}{n_1 + n_2} \text{members of group 2} \end{aligned} \tag{14}$$

We can now treat this as a regression analysis problem. The multiple regression equation obtained will define the regions in the sample space identical to these defined by the discriminant analysis model.

13.5 Cluster Analysis

Cluster analysis is a set of techniques for dividing observations into classes based on a set of characteristics, without the classes being specified in advance. Cluster analysis may be carried out in an attempt to discover classes that are hypothesized to exist but whose structure is unknown, but may also be used simply to create relatively homogeneous subsets of the data.

One application of cluster analysis to clinical epidemiology is in refining the definition of a new syndrome. The controversial *Gulf War syndrome* has been analyzed this way by various authors. Everitt et al. [2002] found five clusters: one *healthy* cluster and four with different distributions of symptoms. On the other hand, Hallman et al. [2003] found only two clusters: healthy and not. Cherry et al. [2001] found six clusters, three of which were relatively healthy and three representing distinct clusters of symptoms. This lack of agreement suggests that there is little evidence for genuine, strongly differentiated clusters.

Cluster analysis has become more visible in biostatistics in recent years with the rise of genomic data. A popular analysis for RNA expression data is to cluster genes based on their patterns of expression across tissue samples or experimental conditions, following Eisen et al. [1998]. The goal of these analyses is intermediate: The clusters are definitely not biologically meaningful in themselves, but are likely to contain higher concentrations of related genes, thus providing a useful starting point for further searches.

Another very visible example of cluster analysis is given by the Google News service (*http://news.google.com*). Google News extracts news stories from a very large number of traditional newspapers and other sources on the Web and finds clusters that indicate popular topics. The most prominent clusters are then displayed on the Web page.

Cluster analysis has a number of similarities to both factor analysis and principal components analysis, discussed in Chapter 14.

13.6 Predicting Categories of a Continuous Variable

In some cases the categorical outcome being predicted is defined in terms of a continuous variable. For example, low birthweight is defined as birthweight below 2500 g, diabetes may be diagnosed by a fasting blood glucose concentration over 140 mg/dL on two separate occasions, hypertension is defined as blood pressure greater than 140/90 mmHg. An obvious question is whether it is better to predict the categorical variable directly or to predict the continuous variable and then divide into categories.

In contrast to the question of whether a predictor should be dichotomized, to which we can give a clear "no!," categorizing an outcome variable may be helpful or harmful. Using the continuous variable has the advantage of making more information available, but the disadvantage of requiring the model to fit well over the entire range of the response. For example, when fitting a model to (continuous) birthweight, the parameter values are chosen by giving equal weight to a 100-g error at a weight of 4000 g as at 2450 g. When fitting a model to (binary) low birthweight, more weight is placed on errors near 2500 g, where they are more important. See also Problem 13.5.

13.7 Further Reading

Harrell [2001] discusses regression modeling for prediction, including binary outcomes, in a medical context. This is a good reference for semiautomatic modeling that uses the available features of statistical software and incorporates background knowledge about the scientific problem. Lachenbruch [1977] covers discriminant analysis, and Hosmer and Lemeshow [2000] discuss logistic regression for prediction (as well as for inference). Excellent but very technical summaries of modern classification methods are given by Ripley [1996] and Hastie et al. [2001]. Venables and Ripley [2002] describe how to use many of these methods in widely available software. As already mentioned, Zhang and Singer [1999] describe recursive partitioning and its use in health sciences. Two excellent texts on screening are Pepe [2003] and Zhou et al. [2002].

PROBLEMS

13.1 For the logistic regression model $\text{logit}(\pi) = \alpha + \beta X$, where X is a dichotomous 0–1 variable, show that e^{β} is the odds ratio associated with the exposure to X.

13.2 For the data of Table 13.7, the logistic regression model using only the variable X_1, malnutrition, is

$$\text{logit}(\widehat{\pi}) = -0.646 + 1.210X_1$$

Table 13.7 **Comparison of Logistic Regression and Linear Regression (One Predictor Variable)**

	Logistic Regression	Normal Regression
Dependent variable	Y discrete (binary)	Y continuous
Covariates	X categorical or continuous	X categorical or continuous
Distribution of Y (given X)	Binomial$(n\pi)$	Normal(μ, σ^2)
Model	$E(Y) = \pi$	$E(Y) = \mu$
Link to X	$\text{logit}(\pi_j) = \alpha + \beta X_j$	$\mu_j = \alpha + \beta X_j$
Data	$y_1, y_2, \ldots, y_n; x_1, x_2, \ldots, x_n$	$y_1, y_2, \ldots, y_n; x_1, x_2, \ldots, x_n$
Likelihood function (LF)	$\prod_{j=1}^{n} \pi_j^{y_j}(1-\pi_j)^{1-y_j}$ $= \prod_{j=1}^{n} \left(\frac{e^{\alpha+\beta x_j}}{1+e^{\alpha+\beta x_j}}\right)^{y_j} \left(\frac{1}{1+e^{\alpha+\beta x_j}}\right)^{1-y_j}$	$\prod_{j=1}^{n} \left(\frac{1}{\sqrt{2\sigma\pi}}\right)^{n} \exp\left(-1/2 \sum \left(\frac{y_j - \mu_j}{\sigma}\right)\right)$ $= \prod_{j=1}^{n} \left(\frac{1}{\sqrt{2\sigma\pi}}\right)^{n} \exp\left(-1/2 \sum \left(\frac{y_j - \alpha - \beta x_j}{\sigma}\right)^2\right)$
Fitting criterion (for choosing estimates of α, β)	Maximize LF	Maximize LF
-2 log LF (is proportional to)	$-2\sum y_j(\alpha + \beta X_j) + 2\sum \ln(1 + e^{\alpha+\beta X_j})$	$\frac{1}{\sigma^2} \sum (y_j - \alpha - \beta X_j)^2$
Equivalent fitting criterion	Minimize -2 log LF (*not* least squares)	Minimize -2 log LF (least squares)
Notation	$D(X) = \underset{\text{over } \alpha,\, \beta}{\text{minimum}} \ (-2 \log LF) = \text{deviance}$	$D(X) = \underset{\text{over } \alpha,\, \beta}{\text{minimum}} \ (-2 \log LF) = \text{deviance}$
Testing: $H_0 : \beta = 0$ in model	$D - D(X)$ is approximately chi-square	$D - D(X)$ is chi-square
Alternative test $H_0 : \beta = 0$ in model	$\frac{D - D(X)}{D(X)/(n-2)}$ is approximately $F_{1,n-2}$	$\frac{D - D(X)}{D(X)/(n-2)} = F_{1,n-2}$

Table 13.8 2 x 2 Table for Vital Status vs. Nutritional Status

		Y		
		Death	Survive	
X_1		1	0	
Malnutrition	1	11	21	32
No malnutrition	0	10	64	74
		21	85	106

The 2 × 2 table associated with these data is shown in Table 13.8.

(a) Verify that the coefficient of X_1 is equal to the logarithm of the odds ratio for malnutrition.

(b) Calculate the probability of death given malnutrition using the model above and compare it with the probability observed.

(c) The standard error of the regression coefficient is 0.5035; test the significance of the observed value, 1.210. Set up 95% confidence limits on the population value and translate these limits into limits for the population odds ratio.

(d) Calculate the standard error of the logarithm of the odds ratio from the 2 × 2 table and compare it with the value in part (c).

13.3 The full model for the data of Table 13.2 is given in Section 13.2.

(a) Calculate the logit line for $X_2 = 0$, $X_3 = 1$, and $X_5 = 1$. Plot logit($\widehat{\pi}$) vs. age in years.

(b) Plot $\widehat{\pi}$ vs. age in years for part (a).

(c) What is the probability of death for a 60-year-old patient with no evidence of shock, but with symptoms of alcoholism and prior bowel infarction?

13.4 One of the problems in the treatment of acute appendicitis is that perforation of the appendix cannot be predicted accurately. Since the consequences of perforation are serious, surgeons tend to be conservative by removing the appendix. Koepsell et al. [1981] attempted to relate the occurrence (or absence) of perforation to a variety of risk factors to enable better assessment of the risk of perforation. A consecutive series of 281 surgery patients was selected initially; of these, 192 were appropriate for analysis, 41 of whom had demonstrable perforated appendices according to the pathology report. The data are listed in Table 13.9. Of the 12 covariates studied, six are listed here, with the group indicator Y.

Y = perforation status (1 = yes; 0 = no)

X_1 = gender (1 = male; 0 = female)

X_2 = age (in years)

X_3 = duration of symptoms in hours prior to physician contact

X_4 = time from physician contact to operation (in hours)

X_5 = white blood count (in thousands)

X_6 = gangrene (1 = yes; 0 = no)

Table 13.9 Data for Problem 13.4

	Y	X_1	X_2	X_3	X_4	X_5	X_6		Y	X_1	X_2	X_3	X_4	X_5	X_6
1	0	0	41	19	1	16	0	49	0	1	15	6	6	19	0
2	1	1	42	48	0	24	1	50	0	0	17	10	4	9	0
3	0	0	11	24	5	14	0	51	0	0	10	72	6	17	0
4	0	1	17	12	2	9	0	52	0	1	9	8	999	15	0
5	1	1	45	36	3	99	1	53	1	1	3	4	2	18	1
6	0	0	15	24	5	14	0	54	0	0	7	16	1	24	0
7	0	1	17	11	24	8	0	55	0	1	60	14	2	11	0
8	0	1	52	30	1	13	0	56	0	1	11	48	3	8	0
9	0	1	15	26	6	13	0	57	0	1	8	48	24	14	0
10	1	1	18	48	2	20	1	58	0	1	9	12	1	12	0
11	0	0	23	48	5	14	0	59	0	1	19	36	1	99	0
12	1	1	9	336	11	13	1	60	1	0	44	24	1	11	1
13	0	0	18	24	3	13	0	61	0	0	46	9	4	12	0
14	0	0	30	8	15	11	0	62	0	1	11	36	2	13	0
15	0	0	16	19	9	10	0	63	0	1	18	8	2	19	0
16	0	1	9	8	2	15	0	64	0	0	21	24	5	12	0
17	0	1	15	48	4	12	0	65	0	0	31	24	8	16	0
18	1	1	25	120	4	8	1	66	0	0	14	7	4	12	0
19	0	0	17	7	17	14	0	67	0	1	17	6	6	19	0
20	0	1	17	12	2	14	0	68	0	0	15	24	1	9	0
21	1	0	63	72	7	11	1	69	0	0	18	24	4	9	0
22	0	0	19	8	1	15	0	70	0	0	38	48	2	99	0
23	0	1	9	48	24	9	0	71	0	1	13	18	4	18	0
24	1	0	9	48	12	14	1	72	1	0	23	168	4	18	0
25	0	0	17	5	1	14	0	73	0	0	15	3	2	14	0
26	0	0	12	48	3	15	0	74	1	0	34	48	3	16	1
27	0	1	6	48	1	26	0	75	0	1	21	24	47	8	1
28	0	0	8	48	3	99	0	76	0	1	50	8	4	12	0
29	1	1	17	30	6	12	1	77	0	0	10	23	6	16	1
30	0	0	11	8	7	15	0	78	0	0	14	48	12	15	0
31	0	1	16	48	2	11	0	79	0	1	26	48	12	13	0
32	0	1	15	10	12	12	0	80	1	0	16	22	1	14	1
33	0	1	13	24	11	15	1	81	1	0	9	24	12	16	1
34	1	1	26	48	4	11	1	82	0	1	26	5	1	16	0
35	0	1	14	7	4	16	0	83	0	1	29	24	1	30	0
36	0	0	44	20	2	13	0	84	0	1	35	408	72	6	0
37	1	1	13	168	999	10	1	85	0	0	18	168	16	12	0
38	0	0	13	14	22	13	0	86	0	1	12	18	4	12	0
39	0	1	24	10	2	19	0	87	0	1	14	7	3	21	0
40	1	0	12	72	2	16	1	88	1	1	45	24	3	18	1
41	0	1	18	15	1	16	0	89	0	1	16	5	21	12	0
42	0	0	19	15	0	9	0	90	0	0	19	240	163	6	0
43	0	0	11	336	20	8	0	91	1	1	9	48	7	23	1
44	0	1	13	14	1	99	0	92	1	1	50	30	5	15	1
45	0	1	25	10	10	11	0	93	0	0	18	2	10	15	0
46	0	1	16	72	5	7	0	94	0	0	27	2	24	17	1
47	0	1	25	72	45	7	0	95	0	1	48	27	5	16	0
48	0	1	42	12	33	19	1	96	0	1	7	18	5	14	0
97	0	1	16	13	1	11	0	145	0	1	41	24	4	14	0
98	0	1	29	5	24	19	1	146	0	0	28	6	1	15	0
99	0	1	18	48	3	11	0	147	1	0	13	48	9	15	1

Table 13.9 (*continued*)

	Y	X_1	X_2	X_3	X_4	X_5	X_6		Y	X_1	X_2	X_3	X_4	X_5	X_6
100	0	1	18	9	2	14	0	148	0	1	10	15	1	99	0
101	1	1	14	14	1	15	1	149	0	1	16	18	4	14	0
102	0	1	32	240	24	7	0	150	0	1	17	18	10	17	0
103	0	1	23	18	2	17	1	151	0	1	38	9	7	11	0
104	0	1	26	16	2	13	0	152	0	1	12	18	2	13	0
105	0	0	30	24	4	20	0	153	0	0	12	72	3	15	0
106	0	1	44	39	15	11	0	154	0	0	27	16	0	14	1
107	1	1	17	24	4	16	1	155	0	1	31	7	8	14	0
108	0	1	30	36	3	15	1	156	0	0	45	20	4	27	0
109	0	1	18	24	2	11	1	157	1	1	52	48	3	15	1
110	0	1	34	96	1	10	0	158	1	1	26	48	13	16	1
111	0	1	15	12	2	10	0	159	0	0	38	15	1	16	0
112	0	1	10	24	4	99	0	160	0	0	19	24	5	99	0
113	0	1	12	14	13	5	0	161	0	1	14	20	2	15	0
114	0	1	10	12	17	17	0	162	0	0	27	22	8	18	0
115	0	1	28	24	2	15	0	163	0	1	20	21	1	99	0
116	0	1	10	96	8	8	0	164	1	1	11	24	8	10	1
117	0	0	22	12	2	12	0	165	0	1	17	72	20	10	0
118	0	0	30	15	5	12	0	166	0	0	27	24	3	9	0
119	0	1	16	36	3	12	0	167	1	0	52	16	4	13	1
120	0	0	16	30	4	15	0	168	1	1	38	48	2	13	1
121	0	1	9	12	12	15	0	169	0	1	16	19	3	12	0
122	1	1	16	144	4	15	1	170	0	1	19	9	4	17	0
123	0	1	17	36	13	6	0	171	0	0	24	24	2	11	0
124	1	1	12	120	2	11	1	172	0	1	12	17	20	6	1
125	0	1	28	17	26	10	0	173	1	1	51	72	2	16	1
126	1	0	13	48	3	21	1	174	1	1	50	72	6	11	1
127	0	0	23	72	3	13	0	175	0	0	28	12	3	13	0
128	1	0	62	72	2	12	1	176	0	0	19	48	8	14	1
129	0	1	17	24	4	14	0	177	0	1	9	24	999	99	0
130	0	0	12	24	12	15	0	178	0	0	40	48	7	14	0
131	0	1	10	12	10	11	0	179	0	0	17	504	7	99	0
132	0	1	47	48	8	9	0	180	0	1	51	24	1	9	1
133	0	1	43	11	8	13	0	181	0	1	31	24	2	10	0
134	1	1	18	36	2	15	1	182	0	0	25	8	9	8	0
135	0	0	6	24	1	9	0	183	0	0	14	24	8	10	0
136	0	0	24	2	22	10	0	184	0	1	7	24	4	15	0
137	0	0	22	11	24	7	0	185	0	1	27	7	2	14	0
138	1	1	39	36	3	15	1	186	0	1	35	72	3	19	1
139	1	1	43	48	2	11	1	187	0	0	11	12	9	11	0
140	0	1	12	7	1	14	0	188	0	1	20	8	6	12	0
141	0	1	14	48	6	16	0	189	0	1	50	48	27	19	0
142	0	1	21	24	1	17	0	190	0	1	16	6	7	7	0
143	1	1	34	48	12	9	1	191	0	1	45	24	4	20	0
144	1	0	60	24	3	14	1	192	1	1	47	336	4	9	1

For X_4 the code 999 is for unknown; for X_5 the code 99 is an unknown code.

(a) Compare the means of the continuous variables (X_2, X_3, X_4, X_5) in the two outcome groups ($Y = 0, 1$) by some appropriate test. Make an appropriate comparison of the association of X_5 and Y. State your conclusion at this point.

(b) Carry out a stepwise discriminant analysis. Which variables are useful predictors? How much improvement in prediction is there in using the discriminant procedure? How appropriate is the procedure?

(c) Carry out a stepwise logistic regression and compare your results with those of part (b).

(d) The authors introduced two additional variables in their analysis: $X_7 = \log(X_2)$ and $X_8 = \log(X_3)$. Test whether these variables improve the prediction scheme. Interpret your findings.

(e) Plot the probability of perforation as a function of the duration of symptoms; using the logistic model, generate a separate curve for subjects aged 10, 20, 30, 40, and 50 years. Interpret your findings.

13.5 The Web appendix to this chapter has a data set with daily concentrations of particulate air pollution in Seattle, Washington. The air quality index for fine particulate pollution below 2.5 μm in diameter (PM2.5) will be "unhealthy for sensitive groups" at 40 μg/m^3 and "moderate" at 20 μg/m^3. The Puget Sound Clean Air Agency is interested in predicting high air pollution days so that it can issue burn bans to reduce fireplace use. Using information on weather and pollution from previous days and the time of year, build logistic models to predict when PM2.5 will exceed 20 or 40 μg/m^3. Also build a linear regression model for predicting PM2.5 or log(PM2.5). Summarize the predictive accuracy of these models. Do you get more accurate prediction using the logistic model or categorizing the prediction from the linear model? Does the answer depend on what losses you assign to false positive and false negative predictions?

13.6 A classic in the use of discriminant analysis is the paper by Truett et al. [1967], in which the authors attempted to predict the risk of coronary heart disease using data from the Framingham study, a longitudinal study of the incidence of coronary heart disease in Framingham, Massachusetts. The two groups under consideration were those who did and did not develop coronary heart disease (CHD) in a 12-year follow-up period. There were 2669 women and 2187 men, aged 30 to 62, involved in the study and free from CHD at their first examination. The variables considered were:

- Age (years)
- Serum cholesterol (mg/100 mL)
- Systolic blood pressure (mmHg)
- Relative weight (100 × actual weight ÷ median for sex–height group)
- Hemoglobin (g/100 mL)
- Cigarettes per day, coded as 0 = never smoked, 1 = less than a pack a day, 2 = one pack a day, and 3 = more than a pack a day
- ECG, coded as 0 = for normal, and 1 = for definite or possible left ventricular hypertrophy, definite nonspecific abnormality, and intraventricular block

Note that the variables "cigarettes" and "ECG" cannot be distributed normally, as they are discrete variables. Nevertheless, the linear discriminant function model was tried. It was found that the predictions (in terms of the risk or estimated probability of being in the coronary heart disease groups) fitted the data well. The coefficients of the linear discriminant functions for men and women, including the standard errors, are shown in Table 13.10.

Table 13.10 Coefficients and Standard Errors for Predicting Coronary Heart Disease.

Risk Factors	Women	Men	Standard Errors of Estimated Coefficients	
Constant ($\widehat{\alpha}$)	−12.5933	−10.8986		
Age (years)	0.0765	0.0708	0.0133	0.0083
Cholesterol (mg %)	0.0061	0.0105	0.0021	0.0016
Systolic blood pressure (mmHg)	0.0221	0.0166	0.0043	0.0036
Relative weight	0.0053	0.0138	0.0054	0.0051
Hemoglobin (g %)	0.0355	−0.0837	0.0844	0.0542
Cigarettes smoked (*see code*)	0.0766	0.3610	0.1158	0.0587
ECG abnormality (*see code*)	1.4338	1.0459	0.4342	0.2706

(a) Determine for both women and men in terms of the p-value the most significant risk factor for CHD in terms of the p-value.

(b) Calculate the probability of CHD for a male with the following characteristics: age = 35 years; cholesterol = 220 mg %; systolic blood pressure = 110 mmHg; relative weight = 110; hemoglobin = 130 g%; cigarette code = 3; and ECG code = 0.

(c) Calculate the probability of CHD for a female with the foregoing characteristics.

(d) How much is the probability in part (b) changed for a male with all the characteristics above except that he does not smoke (i.e., cigarette code = 0)?

(e) Calculate and plot the probability of CHD for the woman in part (c) as a function of age.

13.7 In a paper that appeared four years later, Halperin et al. [1971] reexamined the Framingham data analysis (see Problem 13.6) by Truett et al. [1967] using a logistic model. Halperin et al. analyzed several subsets of the data; for this problem we abstract the data for men aged 29 to 39 years, and three variables: cholesterol, systolic blood pressure, and cigarette smoking (0 = never smoked; 1 = smoker); cholesterol and systolic blood pressure are measured as in Problem 13.6. The following coefficients for the logistic and discriminant models (with standard errors in parentheses) were obtained:

	Intercept	**Cholesterol (mg/100 mL)**	**Systolic Blood Pressure**	**Cigarettes**
Logistic	−11.6246	0.0179(0.0036)	0.0277(0.0085)	1.7346(0.6236)
Discriminant	−13.5300	0.0236(0.0039)	0.0302(0.0100)	1.1191(0.3549)

(a) Calculate the probability of CHD for a male with relevant characteristics defined in Problem 13.6, part (b), for both the logistic and discriminant models.

(b) Interpret the regression coefficients of the logistic model.

(c) In comparing the two methods, the authors state: “Empirically, the assessment of significance of a variable, as measured by the ratio of the estimated coefficient to its estimated standard error, is apt to be about the same whichever method is used.” Verify that this is so for this problem. (However, see also the discussion in Section 13.3.2.)

13.8 In a paper in *American Statistician*, Hauck [1983] derived confidence bands for the logistic response curve. He illustrated the method with data from the Ontario Exercise Heart Collaborative Study. The logistic model dealt with the risk of myocardial infarction (MI) during a study period of four years. A logistic model based on the two most important variables, smoking (X_1) and serum triglyceride level (X_2), was calculated to be

$$\text{logit}(P) = -2.2791 + 0.7682X_1 + 0.001952(X_2 - 100)$$

where P is the probability of an MI during the four-year observation period. The variable X_1 had values $X_1 = 0$ (nonsmoker) and $X_1 = 1$ (smoker). As in ordinary regression, the confidence band for the entire line is narrowest at the means of X_1 and $(X_2 - 100)$ and spreads out the farther you go from the means. (See the paper for more details.)

(a) The range of values of triglyceride levels is assumed to be from 0 to 550. Graph the probability of MI for smokers and nonsmokers separately.

(b) The standard errors of regression coefficients for smoking and serum triglyceride are 0.3137 and 0.001608, respectively. Test their significance.

13.9 One of the earliest applications of the logistic model to medical screening by Anderson et al. [1972] involved the diagnosis of keratoconjunctivitis sicca (KCS), also known as "dry eyes." It is known that rheumatoid arthritic patients are at greater risk, but the definitive diagnosis requires an ophthalmologist; hence it would be advantageous to be able to predict the presence of KCS on the basis of symptoms such as a burning sensation in the eye. In this study, 40 rheumatoid patients with KCS and 37 patients without KCS were assessed with respect to the presence (scored as 1) or absence (scored as 0) of each of the following symptoms: (1) foreign body sensation; (2) burning; (3) tiredness; (4) dry feeling; (5) redness; (6) difficulty in seeing; (7) itchiness; (8) aches; (9) soreness or pain; and (10) photosensitivity and excess of secretion. The data are reproduced in Table 13.11.

(a) Fit a stepwise logistic model to the data. Test the significance of the coefficients.

(b) On the basis of the proportions of positive symptoms displayed at the bottom of the table, select that variable that should enter the regression model first.

(c) Estimate the probability of misclassification.

(d) It is known that approximately 12% of patients suffering from rheumatoid arthritis have KCS. On the basis of this information, calculate the appropriate logistic scoring function.

(e) Define X = number of symptoms reported (out of 10). Do a logistic regression using this variable. Test the significance of the regression coefficient. Now do a t-test on the X variable comparing the two groups. Discuss and compare your results.

13.10 This problem deals with the data of Pine et al. [1983]. Calculate the posterior probabilities of survival for a patient in the fourth decade arriving at the hospital in shock and history of myocardial infarction and without other risk factors:

(a) Using the logistic model.

(b) Using the discriminant model.

Table 13.11 Data for Problem 13.8

	KCS Patients											Patients Without KCS									
Patient	1	2	3	4	5	6	7	8	9	10	Patient	1	2	3	4	5	6	7	8	9	10
1	1	1	1		1		1			1	1	1						1			
2	1	1	1	1	1			1			2										
3	1	1		1	1	1			1		3		1	1				1			
4	1	1		1	1			1	1		4										
5	1	1	1	1			1			1	5								1		
6	1	1			1					1	6										
7	1	1	1			1		1			7			1				1			
8	1	1		1		1			1		8										
9	1	1	1	1	1		1	1			9										
10	1										10									1	
11	1	1	1	1		1		1			11		1					1			
12	1	1					1	1	1		12										
13	1	1	1	1	1	1	1			1	13							1			
14	1	1	1	1	1		1	1		1	14										
15			1	1			1	1			15							1			
16	1	1		1					1		16										
17		1	1			1		1		1	17							1			
18	1		1	1	1			1			18										
19	1		1	1	1		1				19					1					
20	1	1	1	1			1			1	20										
21					1						21										
22	1	1	1	1	1	1	1				22						1				

continued overleaf

Table 13.11 (*continued*)

	KCS Patients											Patients Without KCS									
Patient	1	2	3	4	5	6	7	8	9	10	Patient	1	2	3	4	5	6	7	8	9	10
23	1	1	1		1					1	23										
24	1	1		1			1	1	1		24					1					
25	1	1	1	1				1			25	1								1	
26				1			1				26										
27	1	1		1	1			1	1	1	27										
28	1	1	1	1						1	28										
29	1		1		1				1		29							1			
30	1	1		1		1				1	30										
31		1	1	1						1	31				1						
32	1	1	1	1	1	1	1			1	32										
33			1	1	1		1		1		33							1			
34	1	1	1	1		1	1			1	34										
35	1		1		1			1			35										1
36	1	1	1	1				1			36										
37	1	1	1	1	1						37							1			1
38	1	1																			
39																					
40		1	1	1			1			1											
Proportion position	$\frac{32}{40}$	$\frac{30}{40}$	$\frac{26}{40}$	$\frac{28}{40}$	$\frac{19}{40}$	$\frac{10}{40}$	$\frac{16}{40}$	$\frac{15}{40}$	$\frac{9}{40}$	$\frac{15}{40}$	Proportion position	$\frac{2}{37}$	$\frac{2}{37}$	$\frac{2}{37}$	$\frac{1}{37}$	$\frac{2}{37}$	$\frac{1}{37}$	$\frac{10}{37}$	$\frac{1}{37}$	$\frac{2}{37}$	$\frac{2}{37}$

- **(c)** Graph the two survival curves as a function of age. Use the values 5, 15, 25, ... for the ages in the discriminant model.
- **(d)** Assume that the prior probabilities are $\pi_1 = P[\text{survival}] = 0.60$ and $\pi_2 = 1 - 0.60 = 0.40$. Recalculate the probabilities in parts (a) and (b).
- **(e)** Define a new variable for the data of Table 13.2 as follows: $X_6 = X_1 + X_2 + X_3 + X_5$. Interpret this variable.
- **(f)** Do a logistic regression and discriminant analysis using variables X_4 and X_6 (defined above). Interpret your results.
- **(g)** Is any information "lost" using the approach of parts (e) and (f)? If so, what is lost? When is this likely to be important?

13.11 This problem requires some programming. Create 100 observations of 20 independent random characteristics (e.g., from a uniform distribution) and one random 0–1 variable. Fit a logistic discrimination model using 1, 2, 5, 10, 15, or 20 of your characteristics, and 20, 40, 60, 80, and 100 of the observations. Compute the in-sample error rate and compare it to the true error rate (1/2).

13.12 This problem deals with the data of Problem 5.14, comparing the effect of the drug nifedipine on vasospasm attacks in patients suffering from Raynaud's phenomenon. We want to make a multivariate comparison of the seven patients with a history of digital ulcers ("yes" in column 4) with the eight patients without a history of digital ulcers ("no" in column 4). Variables to be used are age, gender, duration of phenomenon, total number of attacks on placebo, and total number of attacks on nifedipine.

- **(a)** Carry out a stepwise logistic regression on these data.
- **(b)** Which variable entered first?
- **(c)** State your conclusion.
- **(d)** Make a scatter plot of the logistic scores and indicate the dividing point.

***13.13** This problem deals with the data of Problem 10.10, comparing metabolic clearance rates in three groups of subjects.

- **(a)** Use a discriminant analysis on the *three* groups.
- **(b)** Interpret your results.
- **(c)** Graph the data using different symbols to denote the three groups.
- **(d)** Suppose you "create" a third variable: concentration at 90 minutes minus concentration at 45 minutes. Will this improve the discrimination? Why or why not?

***13.14** Consider two groups, G_1 and G_2 (e.g., "death," "survive"; "disease," "no disease"), and a binary covariate, X, with values 0 or 1 (e.g., "don't smoke," "smoke"; "symptom absent," "symptom present"). The data can be arranged in a 2 × 2 table:

	Group	
X	G_1	G_2
1		
0		
	π_1	π_2

Here π_1 is the prior probability of group G_1 membership; $P(X = i|G_1)$ the likelihood of $X = i$ given G_1 membership, $i = 0, 1$; and $P(G_1|X = i)$ the posterior probability of G_1 membership given that $X = i, i = 0, 1$.

(a) Show that

$$\frac{P(G_1|X = i)}{P(G_2|X = i)} = \frac{\pi_1}{\pi_2} \frac{P(X = i|G_1)}{P(X = i|G_2)}$$

Hint: Use Bayes' theorem.

(b) The expression in part (a) can be written as

$$\frac{P(G_1|X = i)}{1 - P(G_1|X = i)} = \frac{\pi_1}{1 - \pi_1} \frac{P(X = i|G_1)}{P(X = i|G_2)}$$

In words:

posterior odds of group 1 membership = prior odds of group 1 membership × ratio of likelihoods of observed values of X.

Relate the ratio of likelihoods to the sensitivity and specificity of the procedure.

(c) Take logarithms of both sides of the equation in part (b). Relate your result to Note 6.7.

(d) The result in part (b) can be shown to hold for X continuous or multivariate. What are the assumptions [go back to the simple set-up of part (a)].

REFERENCES

Akaike, H. [1973]. Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory.*

Anderson, J. A. [1972]. Separate sample logistic regression. *Biometrika*, **59**: 19–35.

Anderson, J. A., Whaley, K., Williamson, J., and Buchanan, W. W. [1972]. A statistical aid to the diagnosis of keratoconjunctivitis sicca. *Quarterly Journal of Medicine, New Series*, **41**: 175–189. Used with permission from Oxford University Press.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. [1984]. *Classification and Regression Trees.* Wadsworth Press, Belmont, CA.

Cherry, N., Creed, F., Silman, A., Dunn, G., Baxter, D., Smedley, J., Taylor, S., and Macfarlane, G. J. [2001]. Health and exposure of United Kingdom Gulf War veterans: I. The pattern and extent of ill health. *Occupational and Environmental Medicine*, **58**: 291–298.

Cover, T. M. [1965]. Geometrical and statistical properties of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computing*, **14**: 326–334.

Efron, B. [1975]. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, **70**: 892–898.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. [1998]. Cluster analysis and display of genome-wise expression patterns. *Proceedings of the National Academy of Sciences*, **95**(25): 14863–14868.

Everitt, B., Ismail, K., David, A. S., and Wessely, S. [2002]. Searching for a Gulf War syndrome using cluster analysis. *Psychological Medicine*, **32**(8): 1335–1337.

Fisher, R. A. [1936]. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. **7**: 179–188.

Hall, P. and Li, K-C. [1993]. On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, **21**: 867–889.

Hallman, W. K., Kipen, H. M., Diefenbach, M., Boyd, K., Kang, H., Leventhal, H., and Wartenberg, D. [2003]. Symptom patterns among Gulf War registry veterans. *American Journal of Public Health*, **93**(4): 624–630.

Halperin, M. and Gurian, J. [1971]. A note on estimation in straight line regression when both variables are subject to error. *Journal of the American Statistical Association*, **66**: 587–589.

Halperin, M., Blockwelder, W. C., and Verter, J. I. [1971]. *Estimation* of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. *Journal of Chronic Diseases*, **24**: 125–158.

Harrell, F. E. [2001]. *Regression Modeling Strategies*. SpringerVerlag, New York.

Hastie, T., Tibshirani, R., and Friedman, J. H. [2001]. *The Elements of Statistical Learning*. SpringerVerlag, New York.

Hauck, W. W. [1983]. A note on confidence bands for the logistic response curve. *American Statistician*, **37**: 158–160.

Health Canada [2001]. *Organized Breast Cancer Screening Programs in Canada: 1997 and 1998 report*. Downloaded from *http://www.hc-sc.gc.ca/pphb-dgspsp/publications_e.html.*

Hosmer, D. W., and Lemeshow, S. [2000]. *Applied Logistic Regression*, 2nd ed. Wiley, New York.

Jones, R. H. [1975]. Probability estimation using a multinomial logistic function. *Journal of Statistical Computation and Simulation*, **3**: 315–329.

Knoke, J. D. [1982]. Discriminant analysis with discrete and continuous variables. *Biometrics*, **38**: 191–200. See also correction in *Biometrics*, **38**: 1143.

Koepsell, T. D., Inui, T. S., and Farewell, V. T. [1981]. Factors affecting perforation in acute appendicitis. *Surgery, Gynecology and Obstetrics*, **153**: 508–510. Used with permission.

Lachenbruch, P. A. [1977]. *Discriminant Analysis*. Hafner Press, New York.

Li, K-C. and Duan, N. [1989]. Regression analysis under link violation, *The Annals of Statistics*, **17**: 1009–1052.

Pepe, M. S. [2003]. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford.

Pine, R. W. Wertz, M. J., Lennard, E. S., Dellinger, E. P., Carrico, C. J., and Minshew, H. [1983]. Determinants of organ malfunction or death in patients with intra-abdominal sepsis. *Archives of Surgery*, **118**: 242–249. Copyright © 1983 by the American Medical Association.

Ripley, B. D. [1996]. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Savage, L. J. [1954]. *The Foundations of Statistics*. Wiley, New York.

Spanos, A., Harrell, F. E. and Durack, F. T. [1989]. Differential diagnosis of acute meningitis. An analysis of the predictive value of initial observations. *Journal of the American Medical Association*, **262**(19): 2700–2707.

Therneau, T. M. [2002]. *Rpart Software*. Mayo Foundation for Medical Research, Rochester, MN.

Truett, J., Cornfield, J., and Kannel, W. [1967]. A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Diseases*, **20**: 511–524.

Venables, W. N., and Ripley, B. D. [2002]. *Modern Applied Statistics with S*, 4th ed. SpringerVerlag, New York.

Wilson P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. [1998]. Prediction of coronary heart disease using risk factor categories. *Circulation*, **97**: 1837–1847.

Zhang, H., and Singer, B. [1999]. *Recursive Partitioning in the Health Sciences*. SpringerVerlag, New York.

Zhou, X.-H., McClish, D. K., and Obuchowski, A. [2002]. *Statistical Methods in Diagnostic Medicine*. Wiley, New York.

CHAPTER 14

Principal Component Analysis and Factor Analysis

14.1 INTRODUCTION

In Chapters 10 and 11 we considered the dependence of a specified response variable on other variables. The response variable identified played a special role among the variables being considered. This is appropriate in many situations because of the scientific question and/or experimental design. What do you do, however, if you have a variety of variables and desire to examine the relationships between them without identifying a specific response variable?

In this chapter we present two methods of examining the relationships among a set of variables without identifying a specific response variable. For these methods, no single variable has a more distinguished role or importance than any other variable. The first technique we examine, principal component analysis, explains as much variability as possible in terms of a few linear combinations of the variables. The second technique, factor analysis, explains the relationships between variables by a few unobserved factors. Both methods depend on the covariances, or correlations, between variables.

14.2 VARIABILITY IN A GIVEN DIRECTION

Consider the 20 observations on two variables X and Y listed in Table 14.1. These data are such that the original observations had their means subtracted, so that the means of the points are zero. Figure 14.1 plots these points, that is, plots the data points about their common mean.

Rather than thinking of the data points as X and Y values, think of the data points as a point in a plane. Consider Figure 14.2(a); when an origin is identified, each point in the plane is identified with a pair of numbers x and y. The x value is found by dropping a line perpendicular to the horizontal axis; the y value is found by dropping a line perpendicular to the vertical axis. These axes are shown in Figure 14.2(b). It is not necessary, however, to use the horizontal and vertical directions to locate our points, although this is traditional. Lines at any angle θ from the horizontal and vertical, as shown in Figure 14.2(c), might be used. In terms of these two lines, the data point has values found by dropping perpendicular lines to these two directions; Figure 14.2(d) shows the two values. We will call the new values x' and y' and the old values x and y. It can be shown that x' and y' are linear combinations of x and y. This idea of lines in different directions with perpendiculars to describe the position of points is used in principal component analysis.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

Table 14.1 Twenty Biometric Observations

Observation	*X*	*Y*	Observation	*X*	*Y*
1	−0.52	0.60	11	0.08	0.23
2	0.04	−0.51	12	−0.06	−0.59
3	1.29	−1.19	13	1.25	−1.25
4	−1.12	1.90	14	0.53	−0.45
5	−1.02	0.31	15	0.14	0.47
6	0.10	−1.15	16	0.48	−0.11
7	−0.32	−0.13	17	−0.61	1.04
8	0.08	−0.17	18	−0.47	0.34
9	0.49	0.18	19	0.41	0.29
10	−0.54	0.20	20	−0.22	0.00

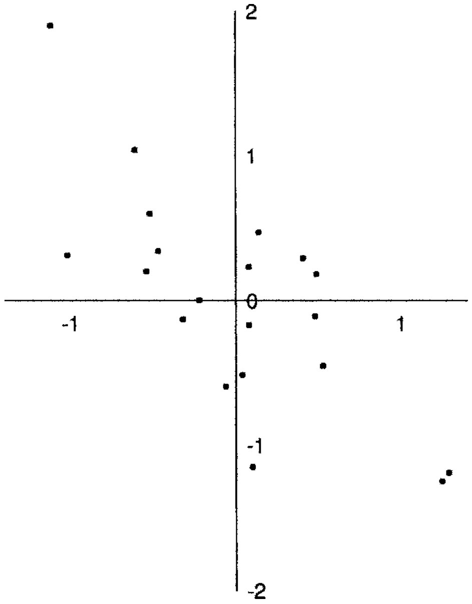

Figure 14.1 Plot of the 20 data points of Table 14.1.

For our data set, the variability in x and y may be summarized by the standard deviation of the x and y values, respectively, as well as the covariance, or equivalently, the correlation between them. Consider now the data of Figure 14.1 and Table 14.1. Suppose that we draw a line in a direction of 30° to the horizontal. The 20 observations give 20 x' values in the X' direction when the perpendicular lines are dropped. Figure 14.3 shows the values in the x' direction. Consider now the points along the line in the x' direction corresponding to the feet of the perpendicular lines. We may summarize the variability among these points by our usual measure of variability, the standard deviation. This would be computed in our usual manner from the 20 values x'. The variability of the data may be summarized by plotting the standard deviation, say $s(\theta)$, in each direction θ at a distance s from the origin. When we look at the standard deviation in all directions, this results in an egg-shaped curve with dents in the side; or a symmetric curve in the shape of a violin or cello body. For the data at hand, this curve is shown in Figure 14.4; the curve is identified as the standard deviation curve. Note that the standard deviation is not the same in all directions. For our data set, the data are spread out more

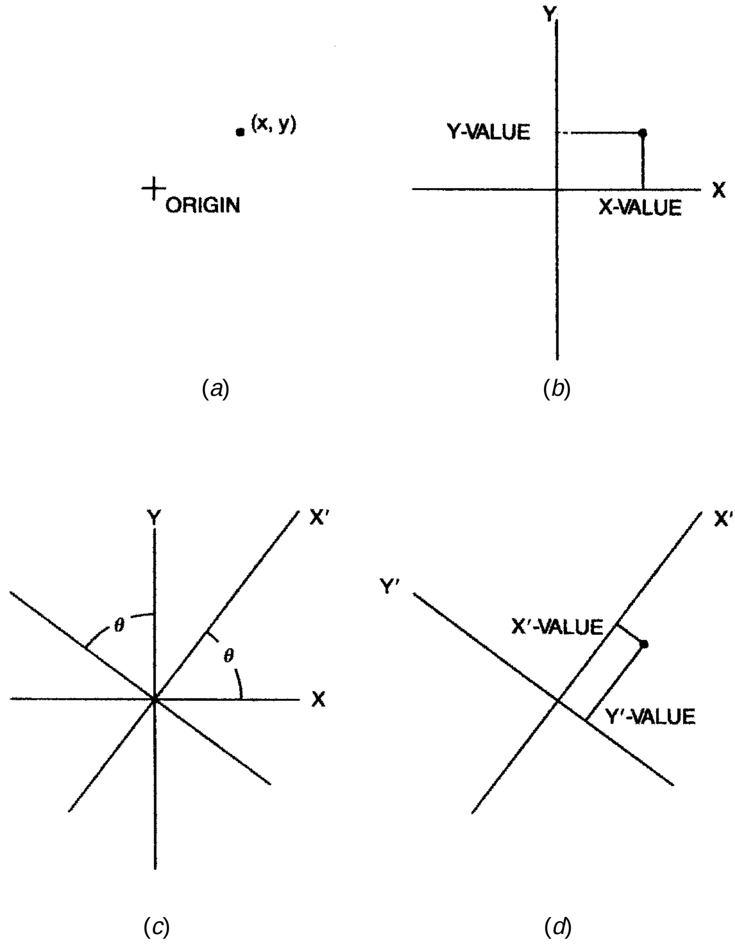

Figure 14.2 Points in the plane, coordinates, and rotation of axes.

along a northwest–southeast direction than in the southwest–northeast direction. The standard deviation curve has a minimum distance at about 38°. The standard deviation increases steadily to a maximum; the maximum is positioned along the line in Figure 14.4, running from the upper left to the lower right. These two directions are labeled directions 1 and 2. If we want to pick one direction that contains as much variability as possible, we would choose direction 1, because the standard deviation is largest in that direction. If all the data points lie on a line, the variability will be a maximum in the direction of the line that contains all the data.

There is some terminology used in finding the value of a data point in a particular direction. The process of dropping a line perpendicular to a direction is called *projecting* the point onto the direction. The value in the particular direction [x' in Figure 14.2(*d*) or Figure 14.3] is called the *projection of the point*. If we know the values x and y, or if we know the values x' and y', we know where the point is in the plane. Two such variables x and y, or equivalently, x' and y', which allow us to find the values of the data, are called a *basis for the variables*.

These concepts may be generalized when there are more than two variables. If we observe three variables x, y, and z, the points may be thought of as points in three dimensions. Suppose that we subtract the means from all the data so that the data are centered about the origin of a three-dimensional plot. As you sit reading this material, picture the points suspended about the

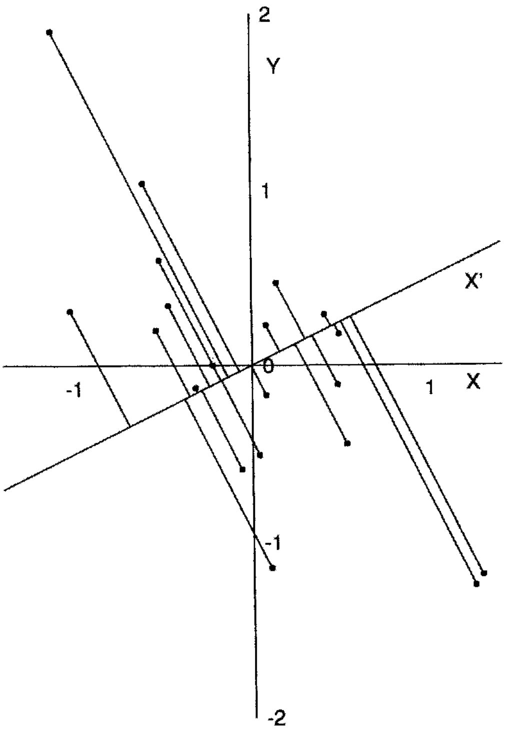

Figure 14.3 Values in the X-direction. X' axis at 30° to the x-axis.

room. Pick an origin. You may draw a line through the origin in any direction. For any point that you have picked in the room, you may drop a perpendicular to the line. Given a line, the point on the line where the perpendicular meets the line is the projection of the point onto the line. We may then calculate the standard deviation for this direction. If the standard deviations are plotted in all directions, a dented egg-shaped surface results. There will be one direction with the greatest variability. When more than three variables are observed, although we cannot picture the situation mentally, mathematically the ideas may be extended; the concept of a direction may be extended in a natural manner. In fact, mathematical statistics is one part of mathematics that heavily uses the geometry of n-dimensional space when there are n variables observed. Fortunately, to understand the statistical methods, we do not need to understand the mathematics!

Let us turn our attention again to Figure 14.4. Rather than plotting the standard deviation curve, it is traditional to summarize the variability in the data by an ellipse. The two perpendicular axes of the ellipse lie along the directions of the greatest variability and the least variability. The ellipse, called the *ellipsoid of concentration*, meets the standard deviation curve along its axes at the points of greatest and least variation. In other directions the standard deviation curve will be larger, that is, farther removed from the origin. In three dimensions, rather than plotting an ellipse we plot an egg-shaped surface, the ellipsoid. (One reason the ellipsoid is used: If you have a bivariate normal distribution in the plane, take a very large sample, divide the plane up into small squares as on graph paper, and place columns whose height is proportional to the number of points; the columns of constant height would lie on an ellipsoid.)

Out of the technical discussion above, we want to remember the following ideas:

1. If we observe a set of variables, we may think of each data point as a point in a space. In this space, when the points are centered about their mean, there is variability in each direction.

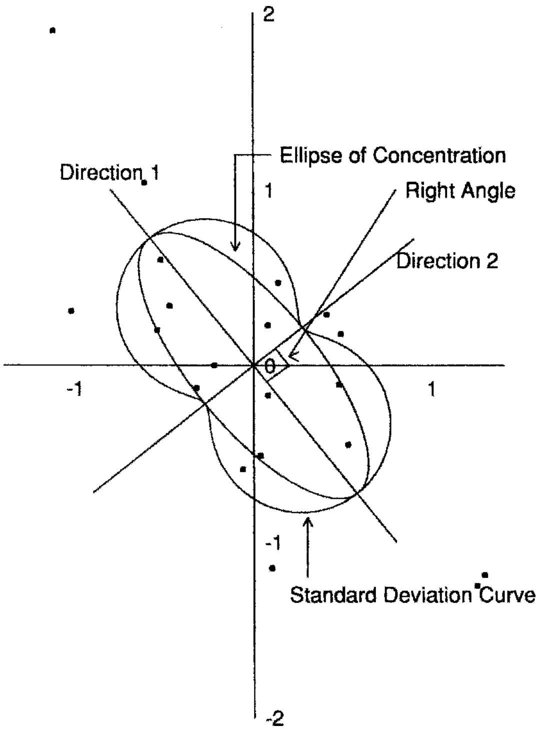

Figure 14.4 Standard deviation in each direction and the ellipse of concentration.

2. The variability is a maximum in one direction. In two dimensions (or more) the minimum lies in a perpendicular direction.
3. The variability is symmetric about each of the particular directions identified.

It is possible to identify the various directions with linear combinations of the variables or coordinates. Each direction for $X_1, \dots, X_p$ is associated with a sum

$$Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p \tag{1}$$

where

$$a_1^2 + a_2^2 + \cdots + a_p^2 = 1$$

The constants $a_1, a_2, \dots, a_p$ are uniquely associated with the direction, except that we may multiply each a by -1. The sum that is given in equation (1) is the value of the projection of the points x_1 to x_p corresponding to the given direction.

14.3 PRINCIPAL COMPONENTS

The motivation behind principal component analysis is to find a direction, or a few directions, that explain as much of the variability as possible. Since each direction is associated with a linear sum of the variables, we may say that we want to find a few new variables, which are linear sums of the old variables, which explain as much of the variability as possible. Thus, the first principal component is the linear sum corresponding to the direction of greatest variability:

Definition 14.1. The *first principal component* is the sum

$$Y = a_1 X_1 + \cdots + a_p X_p, \qquad a_1^2 + \cdots + a_p^2 = 1 \tag{2}$$

corresponding to the direction of greatest variability when variables $X_1, \ldots, X_p$ are under consideration.

Usually, the first principal component will leave much of the variability unexplained. (In the next section, we discuss a method of quantifying the amount of variability explained.) For this reason we wish to search for a second principal component that explains much of the remaining variability. You might think we would take the next linear combination of variables that explains as much of the variability as possible. But when you examine Figure 14.4, you see that the closer the direction gets to the first principal component (which would be direction 1 in Figure 14.4), the more variability one would have. Thus, essentially, we would be driven to the same variable. Therefore, the search for the second principal component is restricted to variables that are uncorrelated with the first principal component. Geometrically, it can be shown that this is equivalent to considering directions that are perpendicular to the direction of the first principal component. In two dimensions such as Figure 14.4, direction 2 would be the direction of the second principal component. However, in three dimensions, when we have the line corresponding to the direction of the first principal component, the set of all directions perpendicular to it correspond to a plane, and there are a variety of possible directions in which to search for the second principal component. This leads to the following definition:

Definition 14.2. Suppose that we have the first $k - 1$ principal components for variables $X_1, \ldots, X_p$. The kth *principal component* corresponds to the variable or direction that is uncorrelated with the first $k - 1$ principal components and has the largest possible variance.

As a summary of these difficult ideas, you should remember the following:

1. Each principal component is chosen to explain as much of the remaining variability as possible after the preceding principal components have been chosen.
2. Each principal component is uncorrelated to the other principal components. In the case of a multivariate normal distribution, the principal components are statistically independent.
3. Although it is not clear from the above, the following is true: For each k, the first k principal components explain as much of the variability in a sample as may be explained by any k directions, or equivalently, k variables.

14.4 AMOUNT OF VARIABILITY EXPLAINED BY THE PRINCIPAL COMPONENTS

Suppose that we want to perform a principal component analysis upon variables $X_1, \ldots, X_p$. If we were dealing with only one variable, say variable X_j, we summarize its variability by the variance. Suppose that there are a total of n observations, so that for each of the p variables, we have n values. Let X_{ij} be the ith observation on the jth variable. Let $\overline{X}_j$ be the mean of the n observations on the jth variable. Then we estimate the variability, that is, the variance, of the variable X_j by

$$\widehat{\text{var}}(X_j) = \sum_{i=1}^{n} \frac{(X_{ij} - \overline{X}_j)^2}{n-1} \tag{3}$$

A reasonable summary of the variability in the p variables is the sum of the individual variances. This leads us to the next definition.

Definition 14.3. The *total variance, denoted by V*, for variables $X_1, \ldots, X_p$ is the sum of the individual variances. That is,

$$\text{total variance} = V = \sum_{j=1}^{p} \text{var}(X_j) \tag{4}$$

The sample total variance, which we will also denote by V since that is the only type of total variance used in this section, is

$$\text{sample total variance} = V = \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{(X_{ij} - \overline{X}_j)^2}{n-1}$$

We now characterize the amount of variability explained by the principal components. Recall that the principal components are themselves variables; they are linear combinations of the X_j variables. Each principal component has a variance itself. It is natural, therefore, to compare the variance of the principal components with the variance of the X_j's. This leads us to the following definitions.

Definition 14.4. Let $Y_1, Y_2, \ldots$ be the first, second, and subsequent principal components for the variables $X_1, \ldots, X_p$. In a sample the variance of each Y_k is estimated by

$$\text{var}(Y_k) = \sum_{i=1}^{n} \frac{(Y_{ik} - \overline{Y}_k)^2}{n-1} = V_k \tag{5}$$

where Y_{ik} is the value of the kth principal component for the ith observation. That is, we first estimate the coefficients for the kth principal component. The value for the ith observation uses those coefficients and the observed values of the X_j's to compute the value of Y_{ik}. The variance for the kth principal component in the sample is then given by the sample variance for Y_{ik}, $i = 1, 2, \ldots, n$. We denote this variance as seen above by V_k. Using this notation, we have the following two definitions:

1. *The percent of variability explained by the* kth *principal component is*

$$\frac{100V_k}{V}$$

2. *The percent of the variability explained by the first m principal components is*

$$100 \sum_{k=1}^{m} \frac{V_k}{V} \tag{6}$$

The following facts about the principal components can be stated:

1. There are exactly p principal components, where p is the number of X variables considered. This is because with p uncorrelated variables, there is a one-to-one correspondence between the values of the principal components and the values of the original data; that is, we can go back and forth so that all of the variability is accounted for; the percent of variability explained by the p principal components is 100%.
2. Because we chose the principal components successively to explain more and more of the variance, we have

$$V_1 \geq V_2 \geq \cdots \geq V_p \geq 0$$

3. The first m principal components explain as much of the total variability as it is possible to explain by m linear functions of the X_j variables.

We now proceed to a geometric interpretation of the principal components. Consider the case where $p = 2$. That is, we observe two variables X_1 and X_2. Plot, as previously in this chapter, the ith data point in the coordinate system that is centered about the means for the X_1 and X_2 variables. Draw a line in the direction of the first principal component and project the data point onto the line. This is done in Figure 14.5.

The square of the distance of the data point from the new origin, which is the sample mean, is given by the following equation, using the Pythagorean theorem:

$$d_i^2 = (X_{i1} - \overline{X}_1)^2 + (X_{i2} - \overline{X}_2)^2 = \sum_{j=1}^{2}(X_{ij} - \overline{X}_j)^2$$

The square of the distance f_i of the projection turns out to be the difference between the value of the first principal component for the ith observation and the mean of the first principal component squared. That is,

$$f_i^2 = (Y_{i1} - \overline{Y}_1)^2$$

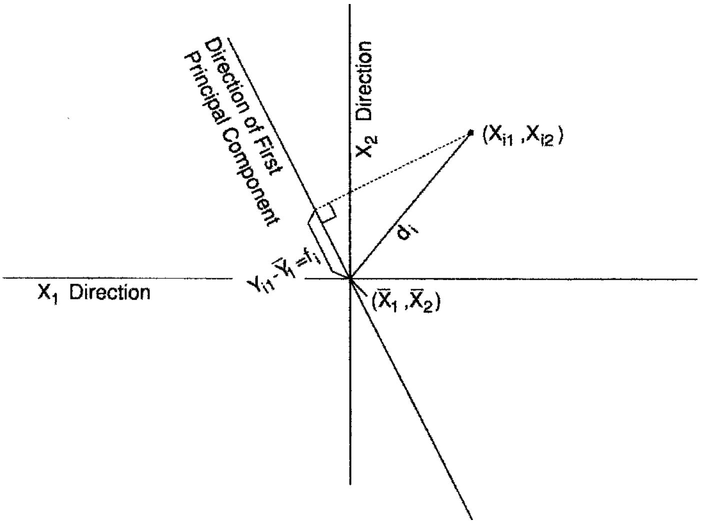
Direction of First Principal Component
X_2 Direction
(X_{i1}, X_{i2})
d_i
$Y_{i1}-\overline{Y}_1=f_i$
X_1 Direction
$(\overline{X}_1, \overline{X}_2)$

Figure 14.5 Projection of a data point onto the first principal component direction.

It is geometrically clear that the distance d_i is larger than f_i. The ith data point will be better represented by its position along the line if it lies closer to the line, that is, if f_i is close to d_i. One way we might judge the adequacy of the variability explained by the first principal component would be to take the ratio of the sum of the lengths of the f_i's squared to the sum of the lengths of the d_i's squared. If we do this, we have

$$\frac{\sum_{i=1}^{n} f_i^2}{\sum_{i=1}^{n} d_i^2} = \frac{\sum_{i=1}^{n}(Y_{i1} - \overline{Y}_1)^2}{\sum_{i=1}^{n}\sum_{j=1}^{2}(X_{ij} - \overline{X}_j)^2} = \frac{V_1}{V} \tag{7}$$

That is, we have the proportion of the variability explained. If we multiplied the equation throughout by 100, we would have the percent of the variability explained by the first principal component. This gives us an alternative way of characterizing the first principal component. The direction of the first principal component is the line for which the following holds: When the data are projected onto this line, the sum of the squares of the projections is as large as possible; equivalently, the sum of squares is as close as possible to the sum of squares of the lengths of the lines to the original data points from the origin (which is also the mean). From this we see that the percent of variability explained by the first principal component will be 100 if and only if the lengths d_i and f_i are all the same; that is, the first principal component will explain all the variability if and only if all of the data points lie on a single line. The closer all the data points come to lie on a single line, the larger the percent of variability explained by the first principal component.

We now proceed to examine the geometric interpretation in three dimensions. In this case we consider a data point plotted not in terms of the original axes X_1, X_2, and X_3 but rather, in terms of the coordinate system given by the principal components Y_1, Y_2, and Y_3. Figure 14.6 presents such a plot for a particular data point. The figure is a two-dimensional representation of a three-dimensional situation; two of the axes are vertical and horizontal on the paper. The third axis recedes into the plane formed by the page in this book. Consider the ith data point,

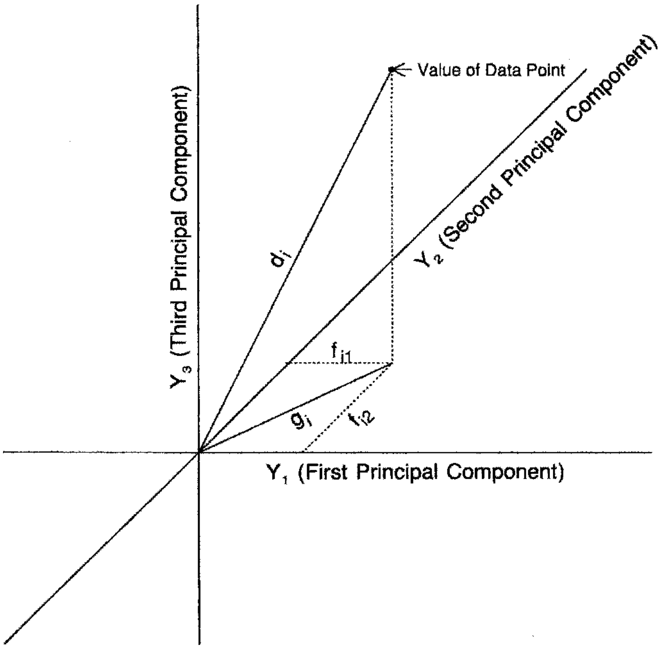
Value of Data Point
Y₂ (Second Principal Component)
Y₃ (Third Principal Component)
dᵢ
fᵢ1
gᵢ
fᵢ2
Y₁ (First Principal Component)

Figure 14.6 Geometric interpretation of principal components for three variables.