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

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}
⋮	⋮	⋮	⋮
i	Y_i	X_{i1}	X_{i2}
⋮	⋮	⋮	⋮
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 $\widehat{Y}_i = a + b_1 X_{i1} + \cdots + b_k X_{ik}$. 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	$SS_{REG} = \sum_i (\widehat{Y}_i - \overline{Y})^2$	$MS_{REG} = SS_{REG}$	$\frac{MS_{REG}}{MS_{RESID}}$
Residual	$n-2$	$SS_{RESID} = \sum_i (Y_i - \widehat{Y}_i)^2$	$MS_{RESID} = \frac{SS_{RESID}}{n-2}$	
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	$SS_{REG} = \sum_i (\widehat{Y}_i - \overline{Y})^2$	$MS_{REG} = \frac{SS_{REG}}{k}$	$\frac{MS_{REG}}{MS_{RESID}}$
Residual	$n-k-1$	$SS_{RESID} = \sum_i (Y_i - \widehat{Y}_i)^2$	$MS_{RESID} = \frac{SS_{RESID}}{n-k-1}$	
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 $VO_{2\ MAX}$ 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	

Source	d.f.	SS	MS	F-Ratio ($F_{3,39,0.95} \doteq 2.85$)
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 $VO_{2\ MAX}$, 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:

Covariate or Constant	b_j	SE(b_j)	t-Statistic ($t_{39,0.975} \doteq 2.02$)
Duration	−0.0524	0.0268	−1.96
$VO_{2\ MAX}$	−0.633	0.378	−1.67
Maximum heart rate	−0.0884	0.119	−0.74
Constant	106.51	18.63	

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		

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, $VO_{2\ MAX}$, 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 $VO_{2\ MAX}$ 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, $VO_{2\ MAX}$, 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, $VO_{2\ MAX}$, 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

$$\text{var}(b_j) = \frac{\sigma^2}{[x_j^2](1 - R_j^2)} \tag{14}$$

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 $X_1, \ldots, X_k$ is denoted by $\rho_{X,Y.X_1,\ldots,X_k}$. The sample partial correlation coefficient of X and Y adjusting for $X_1, \ldots, X_k$ is denoted by $r_{X,Y.X_1,\ldots,X_k}$. 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 $X_1, \ldots, X_k$ and letting $\widehat{X}$ be the predicted value of X from the multiple linear regression of X on $X_1, \ldots, X_k$, 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 $X_1, \ldots, X_k$ is the correlation of X and Y conditionally upon knowing the values of $X_1, \ldots, X_k$. The conditional correlation of X and Y in this multivariate normal case is the same for each fixed set of the values for $X_1, \ldots, X_k$ 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 $X, X_1, \ldots, X_k$ 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

$$\rho_{X,Y\cdot Z} = \frac{\rho_{X,Y} - \rho_{X,Z}\rho_{Y,Z}}{\sqrt{(1-\rho_{X,Z}^2)(1-\rho_{Y,Z}^2)}}$$

$$r_{X,Y.Z} = \frac{r_{X,Y} - r_{X,Z}r_{Y,Z}}{\sqrt{(1-r_{X,Z}^2)(1-r_{Y,Z}^2)}} \tag{15}$$

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, $\mathrm{VO_2\ MAX}$, 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 $\mathrm{VO_2\ MAX}$ 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 $VO_{2\ MAX}$, 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 $X_1, \ldots, X_k$, while adjusting or controlling for a third set of variables, $Z_1, \ldots, Z_p$. 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 $Z_1, \ldots, Z_p$. The best linear predictor for a sample results from the multiple regression of the variable on the independent variables $Z_1, \ldots, Z_p$. The *partial multiple correlation coefficient* between the variable Y and the variables $X_1, \ldots, X_k$ adjusting for $Z_1, \ldots, Z_p$ is the multiple correlation between the variable $Y - \widehat{Y}$ and the variables $X_1 - \widehat{X}_1, \ldots, X_k - \widehat{X}_k$. The partial multiple correlation coefficient of Y and $X_1, \ldots, X_k$ adjusting for $Z_1, \ldots, Z_p$ is denoted by

$$R_{Y(X_1,\ldots,X_k).Z_1,\ldots,Z_p}$$

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

$$\mathrm{SS}_{\mathrm{REG}}(X_1, \ldots, X_j)$$

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

$$\mathrm{SS}_{\mathrm{RESID}}(X_1, \ldots, X_j)$$

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:

$$\mathrm{SS}_{\mathrm{REG}}(X_{j+1}, \ldots, X_k | X_1, \ldots, X_j) = \mathrm{SS}_{\mathrm{REG}}(X_1, \ldots, X_k) - \mathrm{SS}_{\mathrm{REG}}(X_1, \ldots, X_j)$$

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

$$F = \frac{\text{SS}_{\text{REG}}(X_{j+1}, \ldots, X_k | X_1, \ldots, X_j)/(k-j)}{\text{SS}_{\text{RESID}}(X_1, \ldots, X_k)/(n-k-1)}$$

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 $X_1, \ldots, X_k$.

$$r^2_{X,Y \cdot X_1,\ldots,X_k} = \frac{\text{SS}_{\text{REG}}(X | X_1, \ldots, X_k)}{\text{SS}_{\text{RESID}}(X_1, \ldots, X_k)} \tag{16}$$

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

$$F = \frac{\text{SS}_{\text{REG}}(X | X_1, \ldots, X_k)}{\text{SS}_{\text{RESID}}(X, X_1, \ldots, X_k)/(n-k-2)} \tag{17}$$

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

$$t^2 = \frac{\beta_X^2}{\text{SE}(\beta_X)^2} = F$$

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

$$R^2_{Y(X_1,\ldots,X_k) \cdot Z_1,\ldots,Z_p} = \frac{\text{SS}_{\text{REG}}(X_1, \ldots, X_k | Z_1, \ldots, Z_p)}{\text{SS}_{\text{RESID}}(Z_1, \ldots, Z_p)} \tag{18}$$

The F-test for statistical significance uses the test statistic

$$F = \frac{\text{SS}_{\text{REG}}(X_1, \ldots, X_k | Z_1, \ldots, Z_p)/k}{\text{SS}_{\text{RESID}}(X_1, \ldots, X_k, Z_1, \ldots, Z_p)/(n-k-p-1)} \tag{19}$$

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:

$$r_{\mathrm{VO_2\ MAX,DURATION}\,\cdot\,\mathrm{AGE}}$$

$$R^2_{\mathrm{AGE(VO_2\ MAX,\ HEART\ RATE)}\,\cdot\,\mathrm{DURATION}}$$

To examine the relationship between $\mathrm{VO_2\ MAX}$ 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 $\mathrm{VO_2\ MAX}$ as the predictive variable. These runs give the following data: for $Y =$ duration and $X_1 =$ age:

Covariate or Constant	b_j	$\mathrm{SE}(b_j)$	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 $X_2 = \mathrm{VO_2\ MAX}$:

Covariate or Constant	b_j	$\mathrm{SE}(b_j)$	t-statistic ($t_{40,0.975} \doteq 2.09$)
Age	−2.327	0.901	−2.583
$\mathrm{VO_2\ MAX}$	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 $\mathrm{VO_2\ MAX}$	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:

$$\begin{aligned} r^2_{\mathrm{VO_2\ MAX,\ DURATION\cdot AGE}} &= \frac{168{,}961.48 - 119{,}324.47}{131{,}935.95} \\ &= \frac{49{,}637.01}{131{,}935.95} \\ &= 0.376 \end{aligned}$$

Since the regression coefficient for $VO_{2\ MAX}$ is positive (when regressed with age) having a value of 9.151, the positive square root gives r:

$$r_{VO_2\ MAX,\ DURATION\ \cdot\ AGE} = +\sqrt{0.376} = 0.613$$

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 $t^2_{VO_2MAX} = 24.127 = F$ 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, $VO_{2\ MAX}$, 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, $VO_{2\ MAX}$, 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),

$$R^2_{AGE(VO_2\ MAX,\ HEART\ RATE)\ \cdot\ DURATION} = \frac{2256.97 - 2089.18}{2309.98} = 0.0726$$

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: $VO_{2\ MAX}$ 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

Subject i	Potential Outcome $Y_i(0)$	Potential Outcome $Y_i(1)$	Causal Effect Δ_i	Subject i	Potential Outcome $Y_i(0)$	Potential Outcome $Y_i(1)$	Causal 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:

$$\begin{aligned} E(\overline{Y}_1 - \overline{Y}_0) &= E(\overline{Y}_1) - E(\overline{Y}_1) \\ &= \mu_1 - \mu_0 \end{aligned}$$

$$= \frac{1}{N}\sum_i Y_i(1) - \frac{1}{N}\sum_i Y_i(0)$$

$$= \frac{1}{N}\sum_i [Y_i(1) - Y_i(0)]$$

$$= \frac{1}{N}\sum_i \Delta_i$$

$$= \overline{\Delta}$$

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 i	Assignment	Outcome Observed $Y_i(0)$	Outcome Observed $Y_i(1)$	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

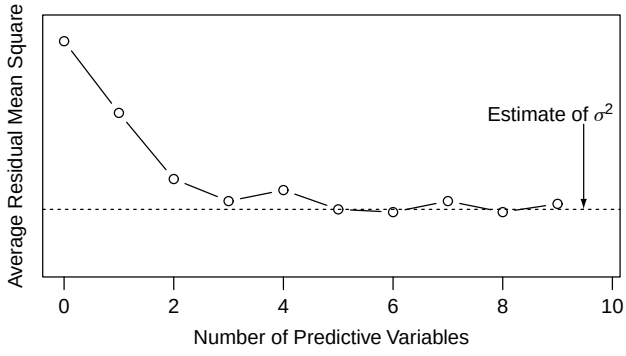

Figure 11.1 Average residual mean square as a function of the number of predictive variables.

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

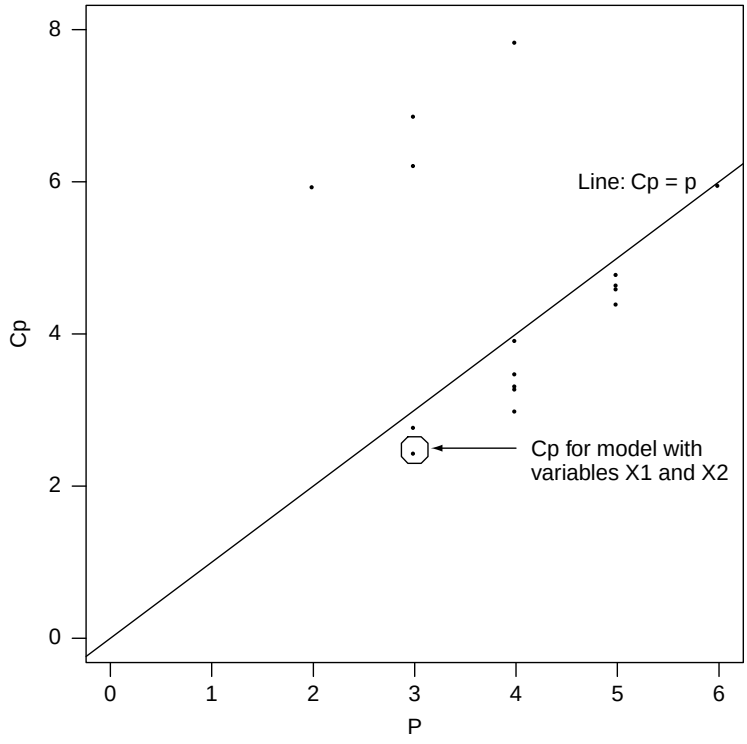

Figure 11.2 Mallow's C_p plot for the data of Cullen and van Belle [1975]. Only points with $C_p < 8$ are plotted.

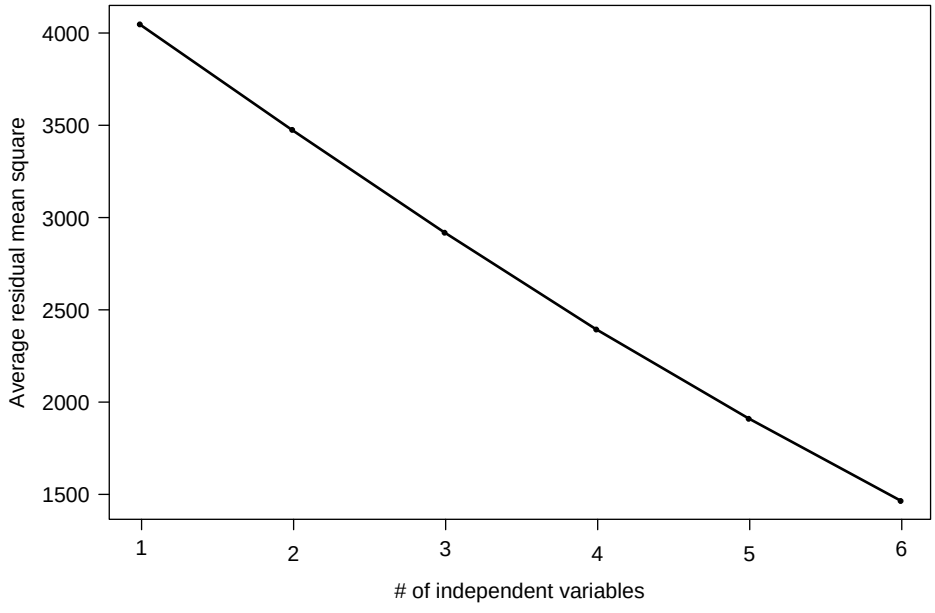

Figure 11.3 Average mean square plot for the Cullen and van Belle data [1975].

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 $\mathrm{SS}_{\mathrm{REG}}(X_i)$.
3. Choose i to minimize $\mathrm{SS}_{\mathrm{RESID}}(X_i)$.

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 $r^2_{Y,X_i \cdot X_1}$.
3. Choose i to maximize $\mathrm{SS}_{\mathrm{REG}}(X_1, X_i)$.
4. Choose i to maximize $\mathrm{SS}_{\mathrm{REG}}(X_i | X_1)$.
5. Choose i to minimize $\mathrm{SS}_{\mathrm{RESID}}(X_1, X_i)$.

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 $\mathrm{SS}_{\mathrm{REG}}(X_1, \ldots, X_j, X_i)$.
4. Choose i to maximize $\mathrm{SS}_{\mathrm{REG}}(X_i | X_1, \ldots, X_j)$.
5. Choose i to minimize $\mathrm{SS}_{\mathrm{RESID}}(X_1, \ldots, X_j, X_i)$.

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 $r^2_{Y,X_i \cdot X_1,\ldots,X_j}$ is greater than a fixed level.
2. The partial F-statistic

$$\frac{\mathrm{SS}_{\mathrm{REG}}(X_i | X_1, \ldots, X_j)}{\mathrm{SS}_{\mathrm{RESID}}(X_1, \ldots, X_j, X_i)/(n - j - 2)}$$

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	$a^{(1)}, b_1^{(1)}$	$\mathrm{SS}_{\mathrm{REG}}(X_1)$	$\mathrm{SS}_{\mathrm{REG}}(X_1)$	$\dfrac{\mathrm{SS}(X_1)(n-2)}{\mathrm{SS}_{\mathrm{RESID}}(X_1)} = F_{1,n-2}$
2	X_2	$a^{(2)}, b_1^{(2)}, b_2^{(2)}$	$\mathrm{SS}_{\mathrm{REG}}(X_1, X_2)$	$\mathrm{SS}_{\mathrm{REG}}(X_2 \mid X_1)$	$\dfrac{\mathrm{SS}(X_2 \mid X_1)(n-3)}{\mathrm{SS}_{\mathrm{RESID}}(X_1, X_2)} = F_{1,n-3}$
3	X_3	$a^{(3)}, b_1^{(3)}, b_2^{(3)}, b_3^{(3)}$	$\mathrm{SS}_{\mathrm{REG}}(X_1, X_2, X_3)$	$\mathrm{SS}_{\mathrm{REG}}(X_3 \mid X_1, X_2)$	$\dfrac{\mathrm{SS}(X_3 \mid X_1, X_2)(n-4)}{\mathrm{SS}_{\mathrm{RESID}}(X_1, X_2, X_3)} = F_{1,n-4}$
⋮	⋮	⋮	⋮	⋮	⋮
j	X_j	$a^{(j)}, b_1^{(j)}, b_2^{(j)}, \ldots, b_j^{(j)}$	$\mathrm{SS}_{\mathrm{REG}}(X_1, X_2, \ldots, X_j)$	$\mathrm{SS}_{\mathrm{REG}}(X_j \mid X_1, \ldots, X_{j-1})$	$\dfrac{\mathrm{SS}(X_j \mid X_1, \ldots, X_{j-1})(n-j-1)}{\mathrm{SS}_{\mathrm{RESID}}(X_1, \ldots, X_j)} = F_{1,n-j-1}$
⋮	⋮	⋮	⋮	⋮	⋮
p	X_p	$a^{(p)}, b_1^{(p)}, b_2^{(p)}, \ldots, b_p^{(p)}$	$\mathrm{SS}_{\mathrm{REG}}(X_1, X_2, \ldots, X_p)$	$\mathrm{SS}_{\mathrm{REG}}(X_p \mid X_1, \ldots, X_{p-1})$	$\dfrac{\mathrm{SS}(X_p \mid X_1, \ldots, X_p)(n-p-1)}{\mathrm{SS}_{\mathrm{RESID}}(X_1, \ldots, X_p)} = F_{1,n-p-1}$

[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

VARIABLE	COEFFICIENT	STD. ERROR OF COEFF	STD REG COEFF	TOLERANCE	F TO REMOVE	LEVEL
(Y-INTERCEPT	29.05349)					

VARIABLES NOT IN EQUATION

	VARIABLE	PARTIAL CORR.	TOLERANCE	F 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

VARIABLE	COEFFICIENT	STD. ERROR OF COEFF	STD REG COEFF	TOLERANCE	F TO REMOVE	LEVEL
(Y-INTERCEPT	3.15880)					
DUR 1	0.05029	0.0062	0.786	1.00000	66.28	1

VARIABLES NOT IN EQUATION

VARIABLE		PARTIAL CORR.	TOLERANCE	F 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

VARIABLE	COEFFICIENT	STD. ERROR OF COEFF	STD REG COEFF	TOLERANCE	F 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

VARIABLE	CORR.	PARTIAL TOLERANCE	TO ENTER	F 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 $VO_{2\ MAX}$.

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 $VO_{2\ MAX}$ 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 $VO_{2\ MAX}$ 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 $VO_{2\ MAX}$. 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 $VO_{2\ MAX}$ 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 $VO_{2\ MAX}$ 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.

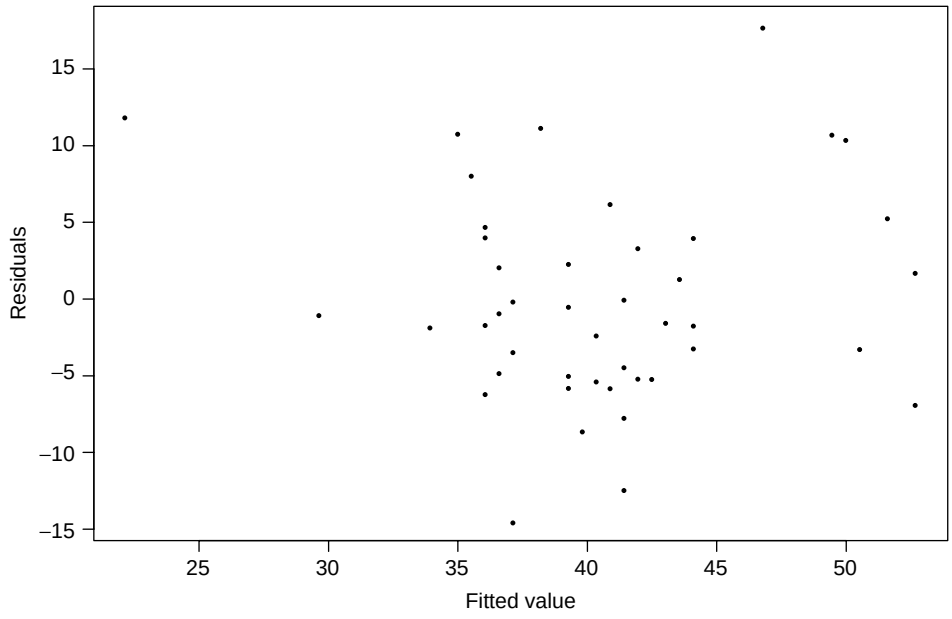

Figure 11.4 Residual plot of the regression of $VO_{2\ MAX}$ on age, active males.

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

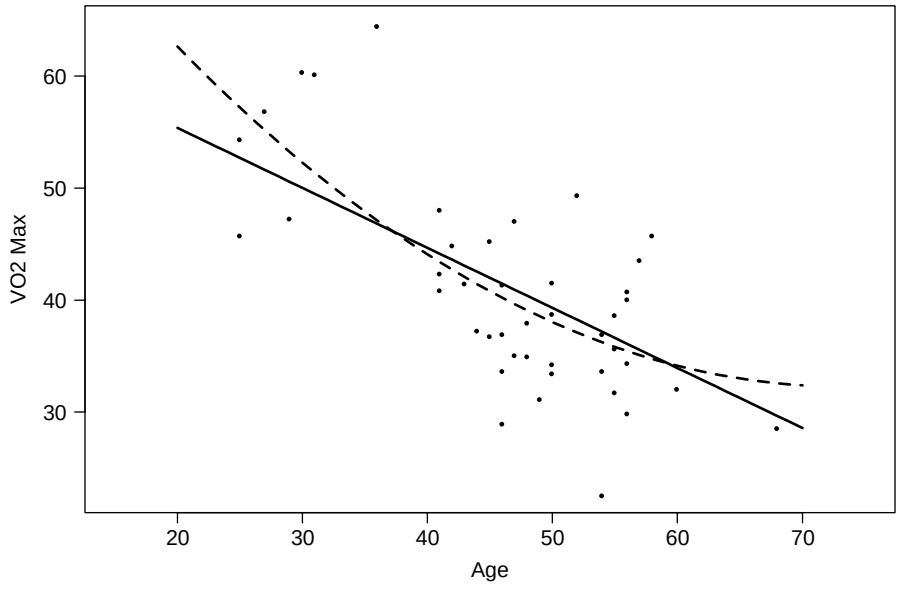

Figure 11.5 Active males with treadmill test: linear (solid line) and quadratic (dashed line) fits. (From Bruce et al. [1973].)

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:

$$Y = a + b_1 X_1 + b_2 X_2 + e$$

(linear in X_1 and X_2), and

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_1^2 + b_4 X_1 X_2 + b_5 X_2^2 + e$$

(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 $VO_{2\ MAX}$ 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

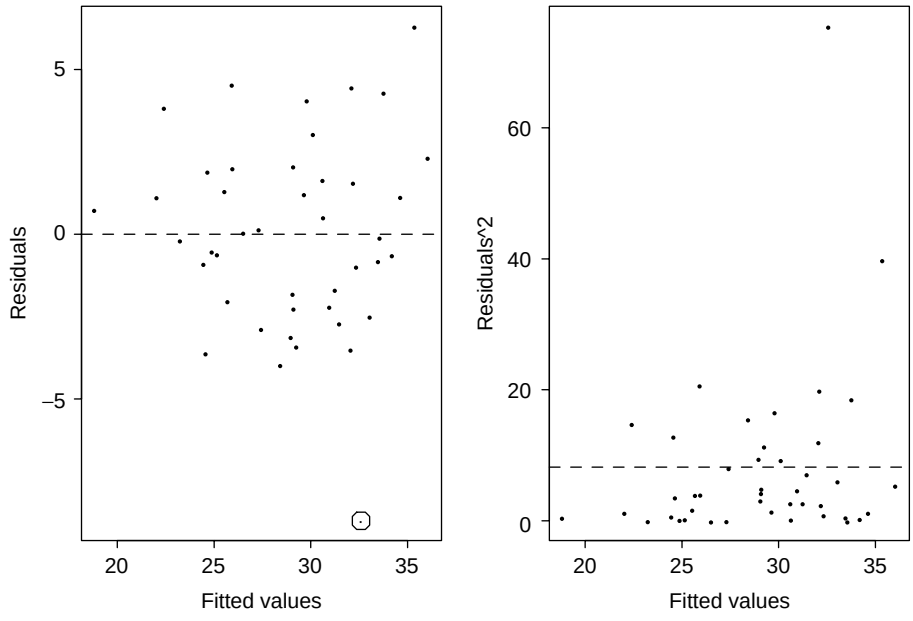

Figure 11.6 Residual plots.

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	

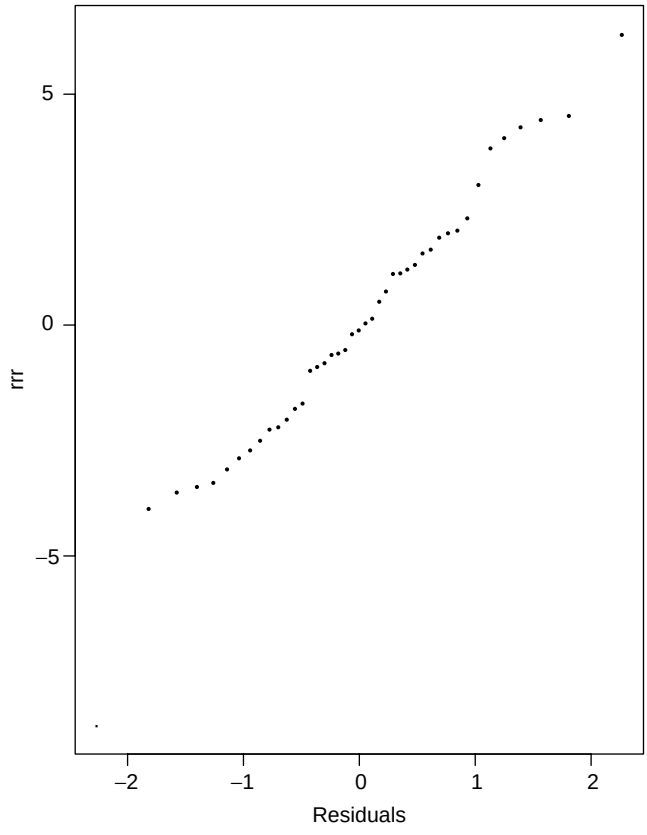

Figure 11.7 Normal residual plot.

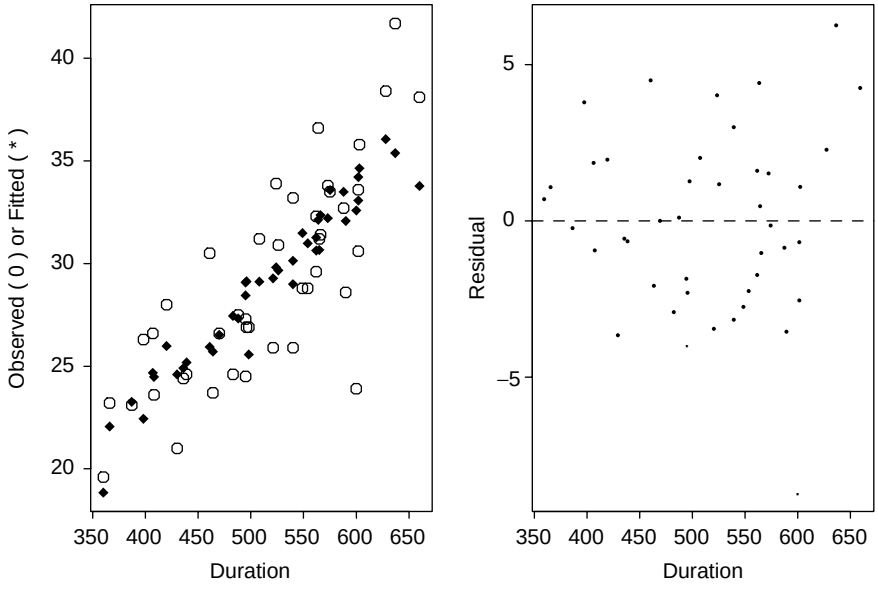

Figure 11.8 Duration vs. residual plots.

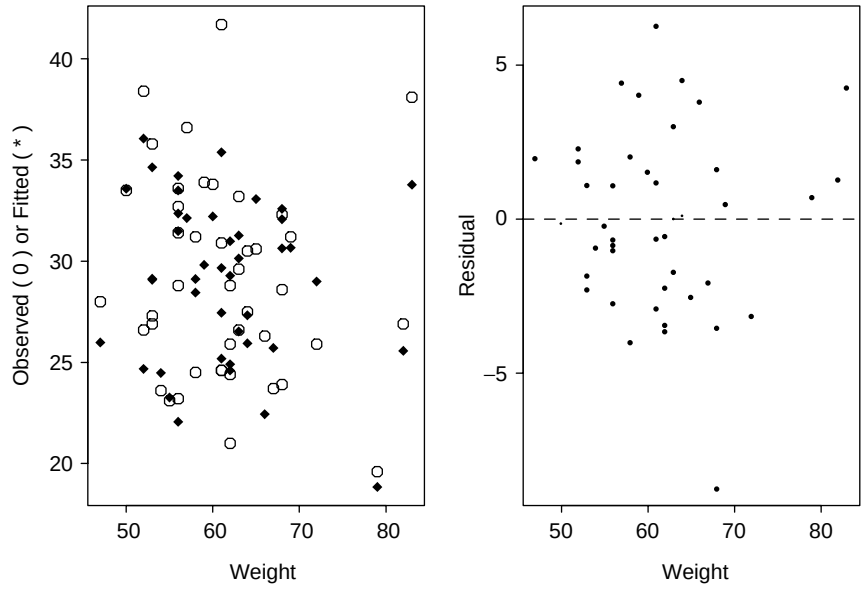

Figure 11.9 Weight vs. residual plots.

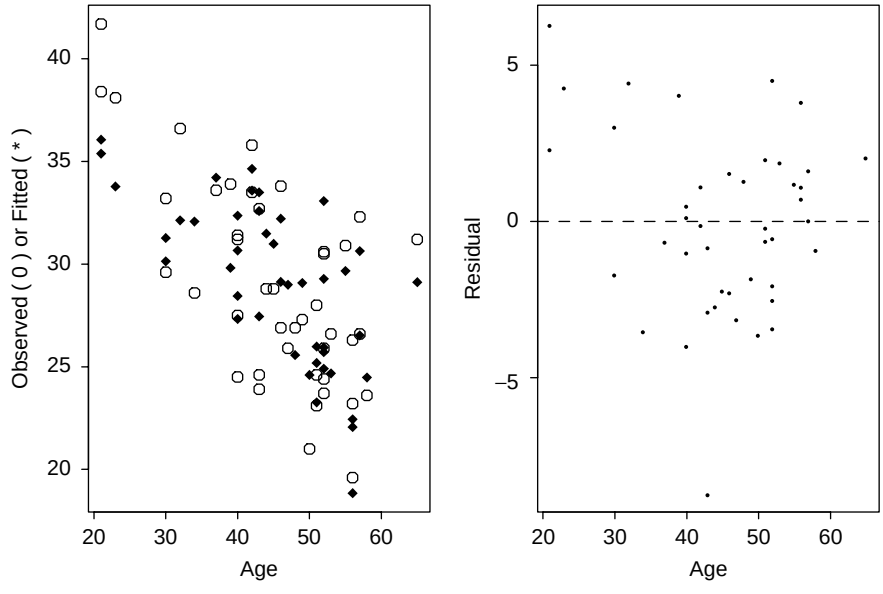

Figure 11.10 Age vs. residual plots.

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

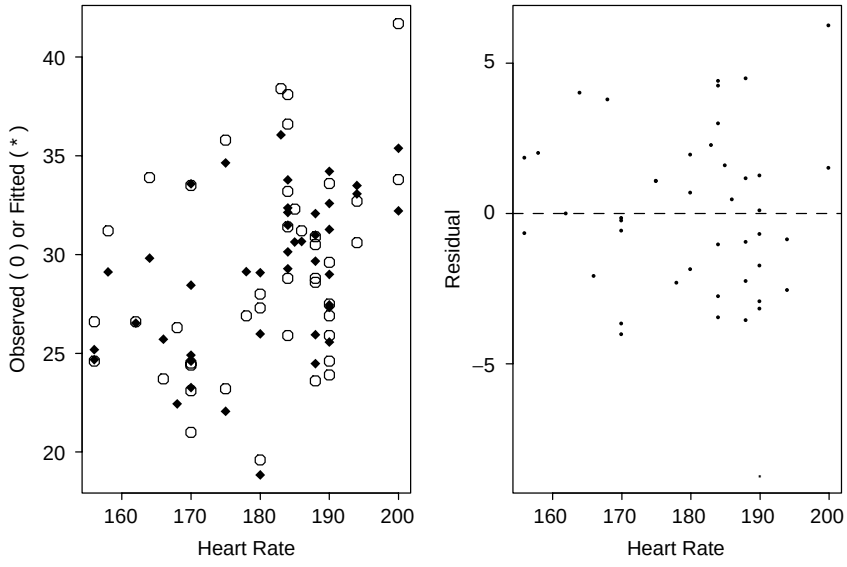

Figure 11.11 Maximum heart rate vs. residual plots.

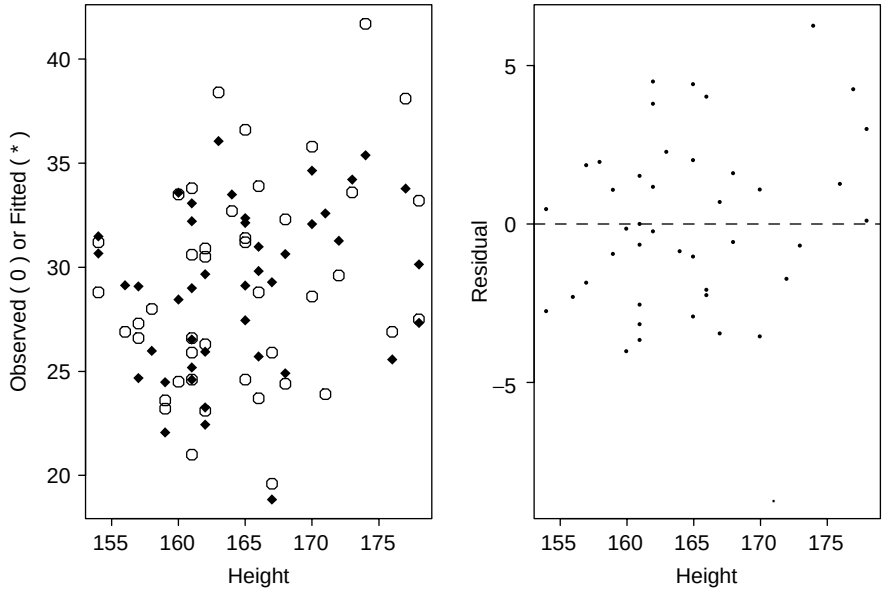

Figure 11.12 Height vs. residual plots.

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.)

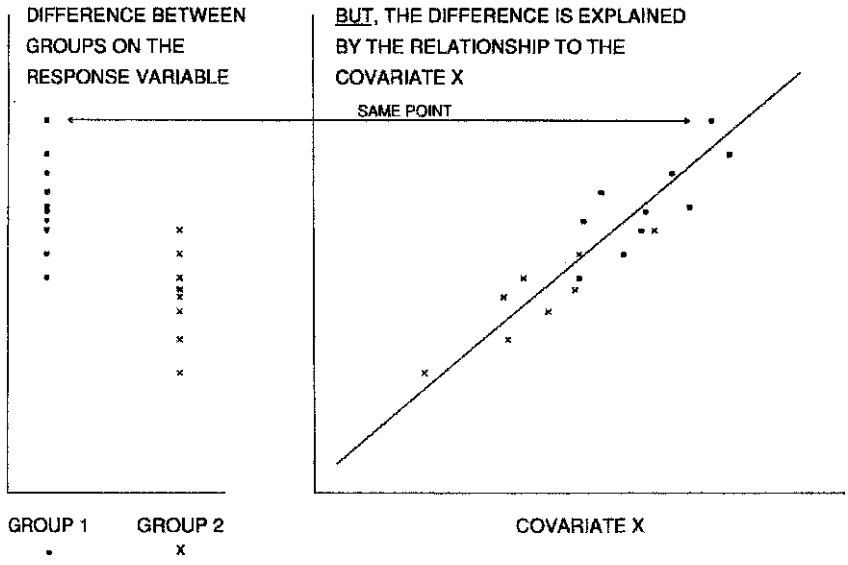

Figure 11.13 One-way analysis of variance with two categories: group difference because of bias due to different distribution on the covariate X.

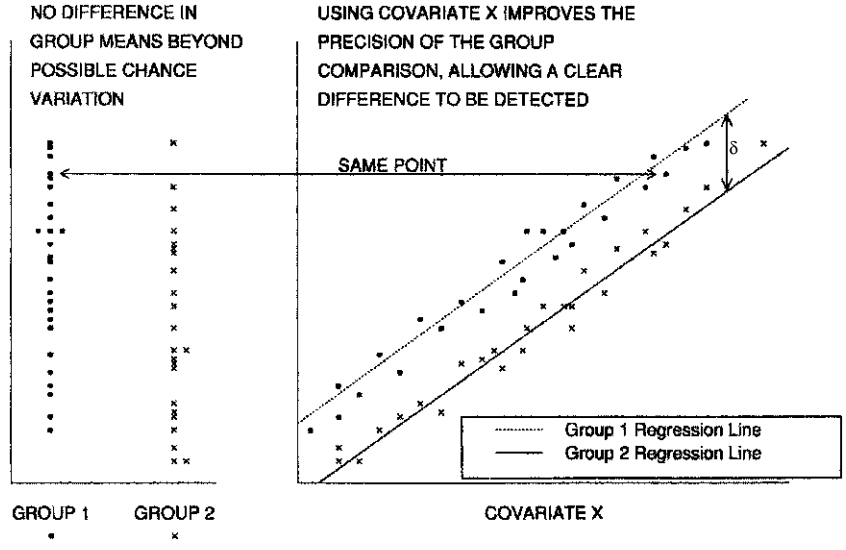

Figure 11.14 Two groups with close distribution on the covariate X. By using the relationship of the response to X separately in each group, a group difference obscured by the variation in X is revealed.

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.*

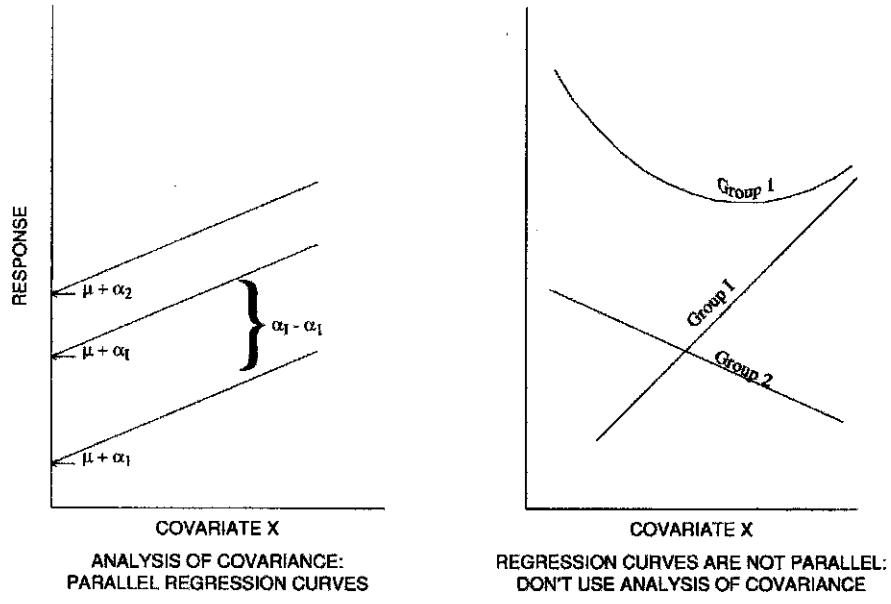

Figure 11.15 Parallel regression curves are assumed in the analysis of covariance.

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 $\mathrm{SS}_{\mathrm{REG}}(X)$ be the regression sum of squares for the model $Y = \delta + \beta X + e$. Let

$$\mathrm{SS}_{\mathrm{REG}}(\gamma|X) = \mathrm{SS}_{\mathrm{REG}}(X_1, \ldots, X_{I-1}, X) - \mathrm{SS}_{\mathrm{REG}}(X)$$

and

$$\mathrm{SS}_{\mathrm{RESID}}(\gamma, X) = \mathrm{SS}_{\mathrm{TOTAL}} - \mathrm{SS}_{\mathrm{REG}}(X_1, \ldots, X_{I-1}, X)$$

The analysis of covariance table is:

Source	d.f.	SS	MS	F-Ratio
Regression on X	1	$\mathrm{SS}_{\mathrm{REG}}(X)$	$\mathrm{MS}_{\mathrm{REG}}(X)$	$\dfrac{\mathrm{MS}_{\mathrm{REG}}(X)}{\mathrm{MS}_{\mathrm{RESID}}}$
Groups adjusted for X	$I-1$	$\mathrm{SS}_{\mathrm{REG}}(\gamma\|X)$	$\mathrm{MS}_{\mathrm{REG}}(\gamma\|X)$	$\dfrac{\mathrm{MS}_{\mathrm{REG}}(\gamma\|X)}{\mathrm{MS}_{\mathrm{RESID}}}$
Residual	$n-I-1$	$\mathrm{SS}_{\mathrm{RESID}}(\gamma\|X)$	$\mathrm{MS}_{\mathrm{RESID}}$	
Total	$n-1$	$\mathrm{SS}_{\mathrm{TOTAL}}$		

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

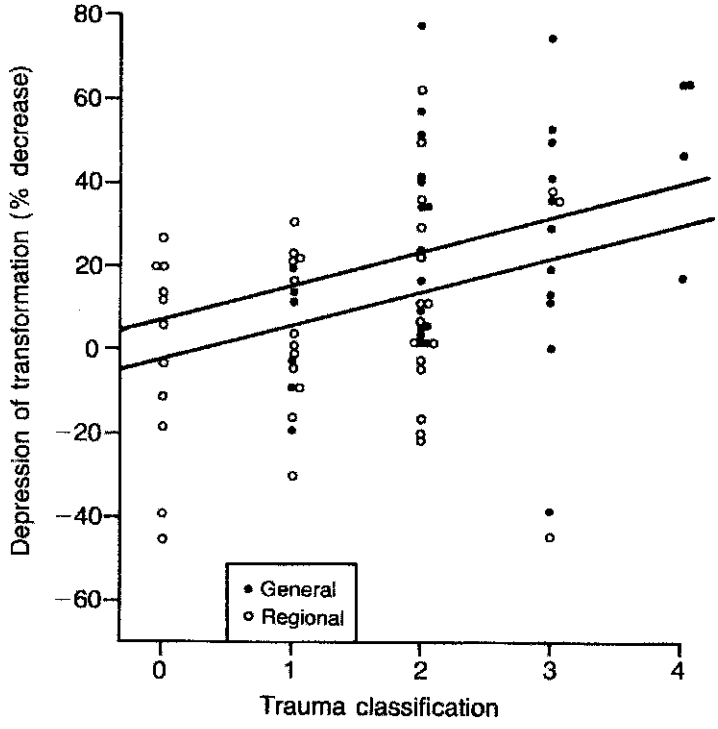

Figure 11.16 Relationship of postoperative depression of lymphocyte transformation to the level of trauma. Each point represents the response of one patient.

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 $\overline{Y}_1. - \overline{Y}_2. = 25.6000 - 6.7381 = 18.86$. 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:

$$\text{Females:} \quad VO_{2\ MAX} = -1.59 + 0.0595 \times \text{duration}$$
$$\text{Males:} \quad VO_{2\ XMAX} = 2.27 + 0.0595 \times \text{duration}$$

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	$SS_{REG}(\gamma_1, \ldots, \gamma_{I-1})$	$MS_{REG}(\gamma's)$	
Model with $\gamma_1, \ldots, \gamma_{I-1}, \beta, \beta_1, \ldots, \beta_I$; extra ss	$I-1$	$SS_{REG}(\beta_1, \ldots, \beta_I \mid \gamma_1, \ldots, \gamma_{I-1}, \beta)$	$MS_{REG}(\beta_i's \mid \gamma_i's, \beta)$	$\dfrac{MS_{REG}(\beta_i's \mid \gamma_i's, \beta)}{MS_{RESID}(\gamma_i's, \beta_i's)}$
Residual	$n-2I$	$SS_{RESID}(\gamma_1, \ldots, \gamma_{I-1}, \beta_1, \ldots, \beta_I)$	$MS_{RESID}(\gamma_i's, \beta_i's)$	
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

$$[x_1^2]b_1 + [x_1x_2]b_2 + \cdots + [x_1x_k]b_k = [x_1y]$$
$$[x_1x_2]b_1 + [x_2^2]b_2 + \cdots + [x_2x_k]b_k = [x_2y]$$
$$\vdots$$
$$[x_1x_k]b_1 + [x_2x_k]b_2 + \cdots + [x_k^2]b_k = [x_ky]$$

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}	$\frac{MS_{REG}}{MS_{RESID}}$
Residual	$n-p-1$	SS_{RESID}	MS_{RESID}	
*Model	*$d.f._{MODEL}$	SS_{MODEL}	MS_{MODEL}	$\frac{MS_{MODEL}}{MS_{PURE\ ERROR}}$
*Pure error	*$d.f._{PURE\ ERROR}$	$SS_{PURE\ ERROR}$	$MS_{PURE\ ERROR}$	
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:

$$SS_{PURE\ ERROR} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (Y_{ik} - \overline{Y}_k)^2$$

$$SS_{MODEL\ FIT} = SS_{RESID} - SS_{PURE\ ERROR}$$

Also,

$$MS_{PURE\ ERROR} = \frac{SS_{PURE\ ERROR}}{d.f._{PURE\ ERROR}}$$

$$MS_{MODEL\ FIT} = \frac{SS_{MODEL}}{d.f._{MODEL}}$$

where

$$d.f._{PURE\ ERROR} = \sum_{k=1}^{K} n_k - K$$

$$d.f._{MODEL} = n + K - \sum_{k=1}^{K} n_k - p - 1$$

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 $MS_{MODEL}/MS_{PURE\ ERROR}$ with $d.f._{MODEL}$ and $d.f._{PURE\ ERROR}$ 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 $Y = VO_{2\ MAX}$, $X_1 =$ 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	$\hat{Y}$	Residual	Y	$\hat{Y}$	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.
- VO_{2MAX}: 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 $= 100 \times (\text{OD} - \text{ED})/\text{ED}$.
- $\dot{Q}_{MAX}$: 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.
- $C\bar{v}O_2$: oxygen content of the venous (vein) system in milliliters of oxygen per liter of blood.
- $a\bar{v}O_2 D_{MAX}$: 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, $a\bar{v}O_2D_{MAX} = CaO_2 - C\bar{v}O_2$.
- $\overline{P}_{SA,\ MAX}$: average pressure in the arterial system at the end of exercise in milliliters of mercury (mmHg).
- $\overline{P}_{PA,\ MAX}$: 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	$\hat{Y}$	Residual	Y	$\hat{Y}$	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	$\hat{Y}$	Residual	Y	$\hat{Y}$	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 $VO_{2,\ MAX}$ is to be regressed upon three variables.

$$Y = VO_{2,\ MAX}$$

$$X_1 = \text{maximal ejection fraction}$$

$$X_2 = \text{maximal heart rate}$$

$$X_3 = \text{maximal systolic blood pressure}$$

The residual mean square with all three variables in the model is 73.40. The residual sums of squares are:

$$SS_{RESID}(X_1, X_2) = 1101.58$$
$$SS_{RESID}(X_1, X_3) = 1839.80$$
$$SS_{RESID}(X_2, X_3) = 1124.78$$
$$SS_{RESID}(X_1) = 1966.32$$
$$SS_{RESID}(X_2) = 1125.98$$
$$SS_{RESID}(X_3) = 1885.98$$

(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]

Step	Variables Entered	Multiple R	Multiple 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]

Step	Variables Entered	Multiple R	Multiple 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]

Step	Variables Entered	Multiple R	Multiple 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]

Step	Variables Entered	Multiple R	Multiple 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

$$R^2_{Y(X_{i_1},\ldots,X_{i_m})} \leq R^2_{Y(X_{j_1},\ldots,X_{j_n})}$$

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,

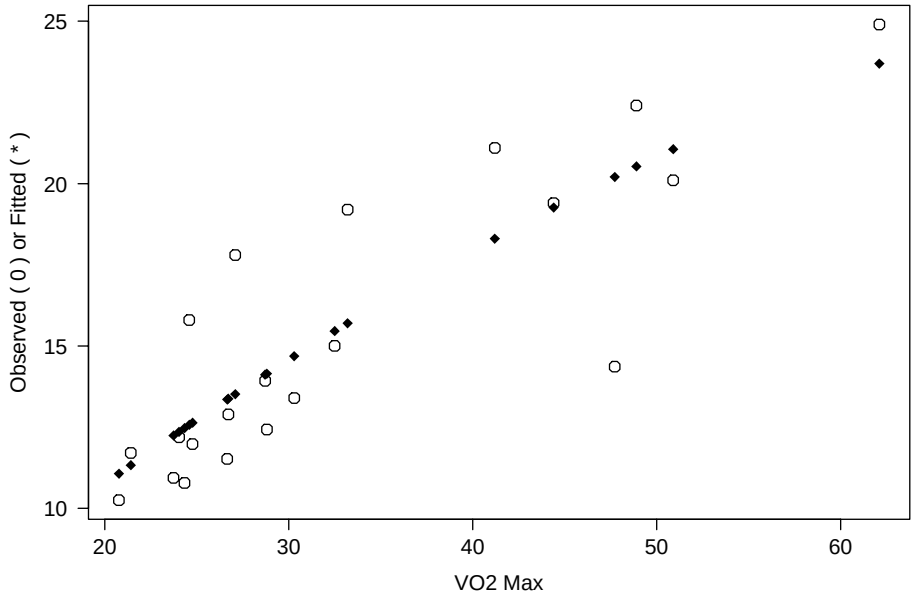

Figure 11.17 Polynomial regression of QDOT on $VO_{2\ MAX}$. Figure for Problem 11.22.

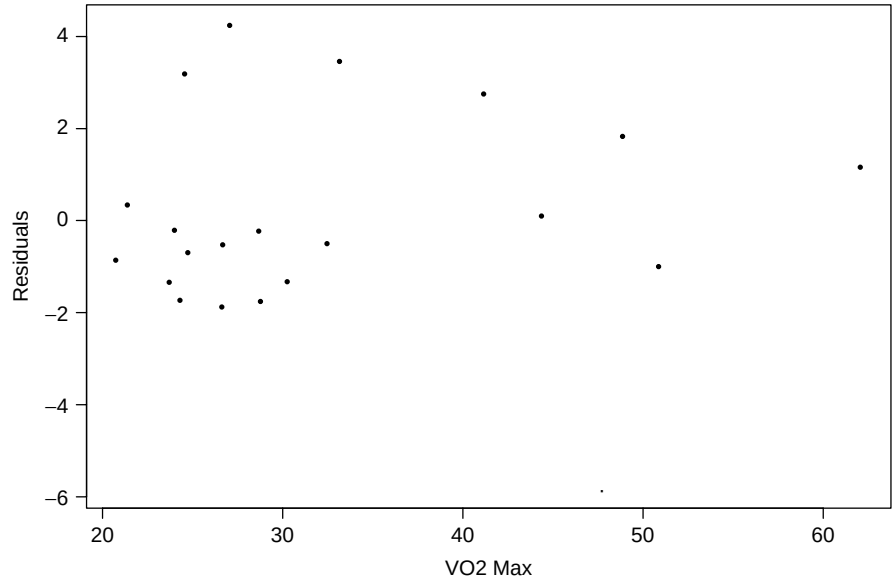

Figure 11.18 Figure for Problem 11.22.

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)?

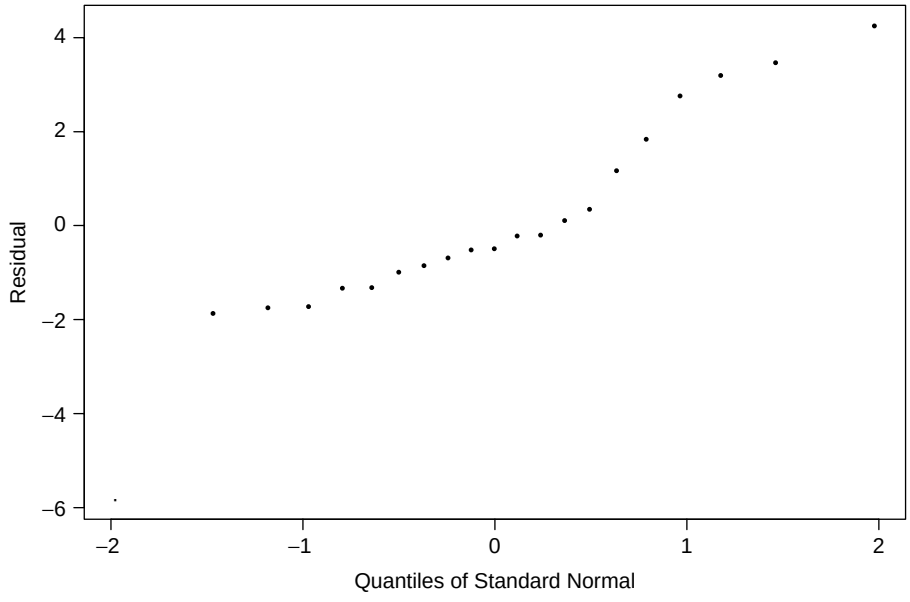

Figure 11.19 Figure for Problem 11.22.

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]

Height (in.)	Number of Examinees at Weight (lb)										
	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.

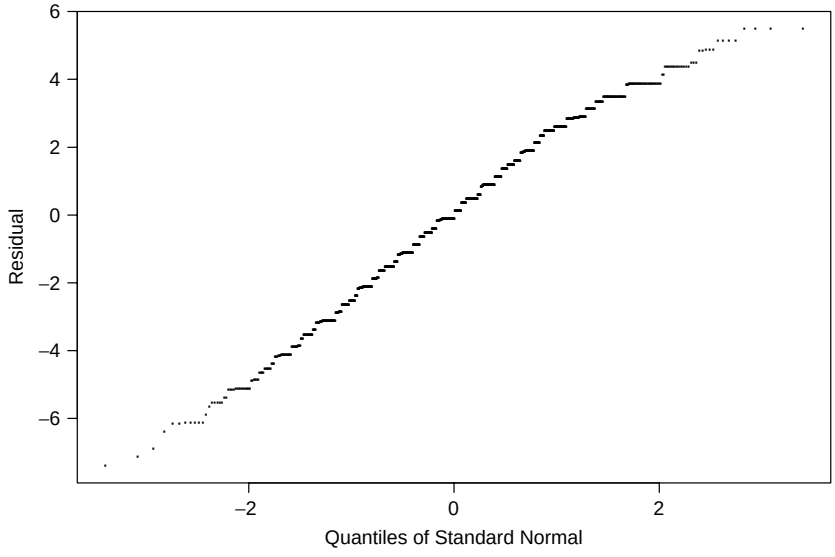

Figure 11.20 Normal probability plot of residuals of degree 2. Figure for Problem 11.23.

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} ($SS_{TOTAL} = 32513.75$)	Indices of Variables in the Multiple Regression Equation (SS_{TOTAL})	Regression Sum of Squares SS_{REG} ($SS_{TOTAL} = 32513.75$)
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 = age in years, X_1 = duration in seconds

$X_2 = \mathrm{VO}_{2\ \mathrm{MAX}}$, X_3 = heart rate in beats/minute

X_4 = height in centimeters, X_5 = weight in kilograms

$\mathrm{SS}_{\mathrm{TOTAL}} = 11,395.74$

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
$VO_{2,\ MAX}$	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 $VO_{2,\ MAX}$:

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

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 of Tests, n	α		
	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 of Tests, n	α^*		
	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

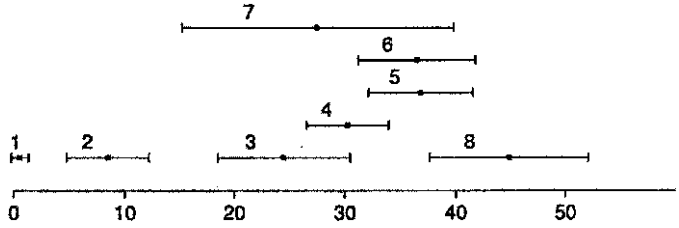

Figure 12.1 Functional aerobic impairment level.

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

$$\widehat{\theta} \pm \sqrt{d\,F_{d,m,1-\alpha}}\,\widehat{\sigma}_{\widehat{\theta}}$$

where the estimate of θ is $\widehat{\theta} = c_1 b_1 + c_2 b_2 + \cdots + c_p b_p$ 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 $\pm$ constant $\times$ 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 $d = I$, $m = n_\cdot - I$, 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

$$\overline{Y}_{i\cdot} \pm \sqrt{I F_{I,n_\cdot - I,1-\alpha}} \sqrt{\frac{\text{MS}_e}{n_i}}, \qquad i = 1, \ldots, I$$

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

$$\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot} \pm \sqrt{I F_{I,n.-I,1-\alpha}} \sqrt{\text{MS}_e \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)}, \qquad i, i' = 1, \ldots, I, i \neq i'$$

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

$$\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdot j'\cdot} \pm \sqrt{(J-1)F_{J-1,n_{\cdot\cdot}-IJ,1-\alpha}}\sqrt{\mathrm{MS}_e\left(\frac{1}{n_{\cdot j}} + \frac{1}{n_{\cdot j'}}\right)}$$

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

$$\overline{Y}_{\cdots} + a_i + b_j, \qquad \begin{array}{l} i = 1, \ldots, I \\ j = 1, \ldots, J \end{array}$$

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

$$\overline{Y}... + (\overline{Y}_{i\cdot\cdot} - \overline{Y}...) + (\overline{Y}_{\cdot j\cdot} - \overline{Y}...) = \overline{Y}_{i\cdot\cdot} + \overline{Y}_{\cdot j\cdot} - \overline{Y}...$$

The estimates are $\overline{Y}... = 6.53$, $\overline{Y}_{1\cdot\cdot} = 6.71$, $\overline{Y}_{2\cdot\cdot} = 6.35$, $\overline{Y}_{\cdot 1\cdot} = 5.99$, $\overline{Y}_{\cdot 2\cdot} = 7.07$. The estimated cell means fitted to the model $E(Y_{ijk}) = \mu + \alpha_i + \beta_j$ by $\overline{Y}... + a_i + b_j$ 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

$$\overline{Y}_{i\cdot\cdot} + \overline{Y}_{\cdot j\cdot} - \overline{Y}... \pm \sqrt{(I+J-1)F_{I+J-1,n..-IJ+1,1-\alpha}}\sqrt{\text{MS}_e\left(\frac{1}{n_{i\cdot}} + \frac{1}{n_{\cdot j}} - \frac{1}{n_{\cdot\cdot}}\right)}$$

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

$$\overline{Y}_{1\cdot\cdot} - \overline{Y}_{2\cdot\cdot} \pm \sqrt{3F_{3,37,0.95}}\sqrt{\text{MS}_e\left(\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}\right)}$$

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

$$Q_{k,m} = \frac{\mathrm{MAX}(Y_1, Y_2, \ldots, Y_k) - \mathrm{MIN}(Y_1, Y_2, \ldots, Y_k)}{s}$$

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	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	$\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot} \pm \frac{1}{\sqrt{Jn}} q_{I,IJ(n-1),1-\alpha} \sqrt{\mathrm{MS}_e \left(1 - \frac{1}{I}\right)}$
$\alpha_i - \alpha_{i'}$	$\overline{Y}_{i\cdot\cdot} - \overline{Y}_{i'\cdot\cdot} \pm \frac{1}{\sqrt{Jn}} q_{I,IJ(n-1),1-\alpha} \sqrt{\mathrm{MS}_e}$

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	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	$\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot} \pm \frac{1}{\sqrt{I}} q_{J,(I-1)(J-1),1-\alpha} \sqrt{\text{MS}_e \left(1 - \frac{1}{J}\right)}$
$\tau_j - \tau_{j'}$	$\overline{Y}_{\cdot j} - \overline{Y}_{\cdot j'} \pm \frac{1}{\sqrt{2I}} q_{J,(I-1)(J-1),1-\alpha} \sqrt{\text{MS}_e}$

Consider Example 10.6. We want to compare the effectiveness of pancreatic supplements on fat absorption. The treatment means are

$$\overline{Y}_{\cdot 1} = 38.1, \qquad \overline{Y}_{\cdot 2} = 16.5, \qquad \overline{Y}_{\cdot 3} = 17.4, \qquad \overline{Y}_{\cdot 4} = 31.1$$

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 $A_1, \ldots, A_n$ 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 $\beta_1, \beta_2, \ldots, \beta_p$ 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 $N = \begin{pmatrix} I \\ 2 \end{pmatrix}$. Simultaneous confidence intervals, in this case, are of the form

$$\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot} \pm t_{m,1-\alpha/2\begin{pmatrix} I \\ 2 \end{pmatrix}} \sqrt{\mathrm{MS}_e \left(\frac{1}{n_i} + \frac{1}{n_i'} \right)}, \qquad i, i' = 1, \ldots, I, i \neq i'$$

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									
					FAI(%)		Rest		Maximum		Slope		Zero Recovery	
	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	$\times$	$=$	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

$$\text{SS}_{\text{TREATMENTS}} = \widehat{\theta}_1^2 + \widehat{\theta}_2^2 + \cdots + \widehat{\theta}_{I-1}^2$$

We have thus partitioned the $\text{SS}_{\text{TREATMENTS}}$ 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 $2, \ldots, 8$ 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

$$\mathrm{SS}_{\mathrm{TREATMENTS}} = \widehat{\theta}_1^2 + \widehat{\theta}_2^2 + \widehat{\theta}_3^2$$

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.

$$I_i = \begin{cases} 1, & \text{data from subject } i,\ i = 1, \ldots, 6 \\ 0, & \text{otherwise} \end{cases}$$

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: $\beta_i = 0$, $i = 1, \ldots, p$. 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 $SS_{TREATMENTS}$ 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.