CHAPTER 3

Multiple Regression

Multiple linear regression generalizes the simple linear regression model by allowing for many *terms* in a mean function rather than just one intercept and one slope.

3.1 ADDING A TERM TO A SIMPLE LINEAR REGRESSION MODEL

We start with a response Y and the simple linear regression mean function

$$\mathrm{E}(Y|X_1 = x_1) = \beta_0 + \beta_1 x_1$$

Now suppose we have a second variable X_2 with which to predict the response. By adding X_2 to the problem, we will get a mean function that depends on both the value of X_1 and the value of X_2,

$$\mathrm{E}(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{3.1}$$

The main idea in adding X_2 is to explain the part of Y that has not already been explained by X_1.

United Nations Data

We will reconsider the United Nations data discussed in Problem 1.3. To the regression of log(*Fertility*), the base-two log fertility rate on log(*PPgdp*), the base-two log of the per person gross domestic product, we consider adding *Purban*, the percentage of the population that lives in an urban area. The data in the file `UN2.txt` give values for these three variables, as well as the name of the *Locality* for 193 localities, mostly countries, for which the United Nations provides data.

Figure 3.1 presents several graphical views of these data. Figure 3.1a can be viewed as a summary graph for the simple regression of log(*Fertility*) on log(*PPgdp*). The fitted mean function using OLS is

$$\widehat{\mathrm{E}}(\log(Fertility)|\log(PPgdp)) = 2.703 - 0.153\log(PPgdp)$$

Applied Linear Regression, Third Edition, by Sanford Weisberg
ISBN 0-471-66379-4

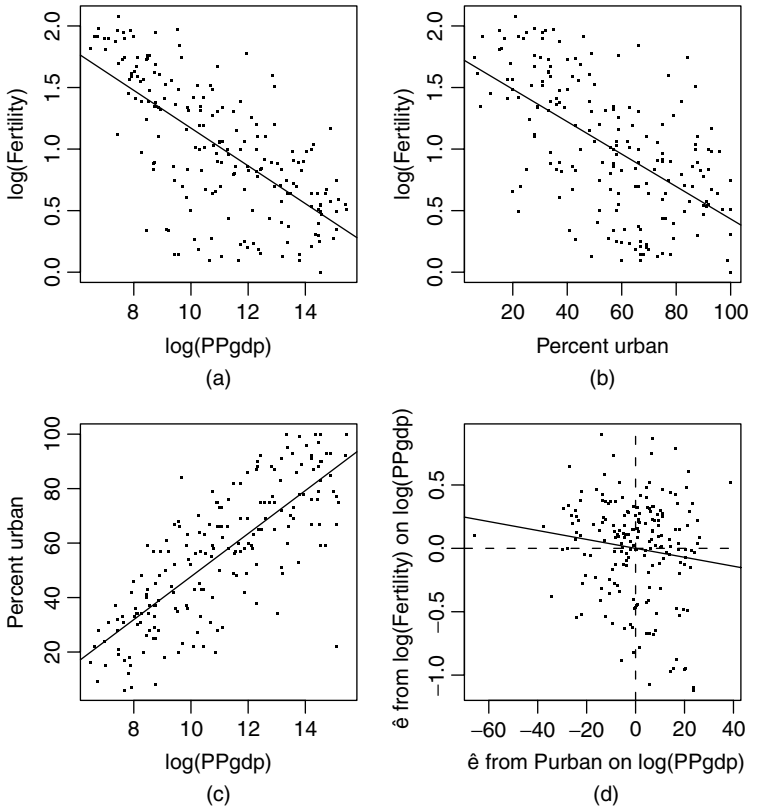

FIG. 3.1 United Nations data on 193 localities, mostly nations. (a) log(*Fertility*) versus log(*PPgdp*); (b) *log*(*Fertility*) versus *Purban*; (c) *Purban* versus log(*PPgdp*); (d) Added-variable plot for *Purban* after log(*PPgdp*).

with $R^2 = 0.459$, so about 46% of the variability in log(*Fertility*) is explained by log(*PPgdp*). An increase of one unit in log(*PPgdp*), which corresponds to a *doubling* of *PPgdp*, is estimated to decrease log(*Fertility*) by 0.153 units.

Similarly, Figure 3.1b is the summary graph for the regression of log(*Fertility*) on *Purban*. This simple regression has fitted mean function

$$\widehat{\mathrm{E}}(\log(Fertility)|Purban) = 1.750 - 0.013\ Purban$$

with $R^2 = 0.348$, so *Purban* explains about 35% of the variability in log(*Fertility*). An increase of one percent urban implies a change on the average in log(*Fertility*) of -0.13.

To get a summary graph of the regression of log(*Fertility*) on *both* log(*PPgdp*) and *Purban* would require a three-dimensional plot of these three variables, with log(*PPgdp*) on one of the horizontal axes, *Purban* on the other horizontal axis, and log(*Fertility*) on the vertical axis. Although such plots are possible by using

either perspective or motion to display the third dimension, using them is much more difficult than using two-dimensional graphics, and their successful use is not widespread. Cook and Weisberg (1999a) discuss using motion to understand three-dimensional graphics for regression.

As a partial substitute for looking at the full three-dimensional plot, we add a third plot to the first two in Figure 3.1, namely, the plot of *Purban* versus log(*PPgdp*) shown in Figure 3.1c. This graph does not include the response, so it only shows the relationship between the two potential predictors. In this problem, these two variables are positively correlated, and the mean function for Figure 3.1c seems to be well approximated by a straight line.

The inference to draw from Figure 3.1c is that to the extent that *Purban* can be predicted by log(*PPgdp*), these two potential predictors are measuring the same thing, and so the role of these two variables in predicting log(*Fertility*) will be overlapping, and they will both, to some extent, be explaining the same variability.

3.1.1 Explaining Variability

Given these graphs, what can be said about the proportion of variability in log(*Fertility*) explained by log(*PPgdp*) and *Purban*? We can say that the total explained variation must exceed 46 percent, the larger of the two values explained by each variable separately, since using both log(*PPgdp*) and *Purban* must surely be at least as informative as using just one of them. The total variation will be additive, $46\% + 35\% = 91\%$, only if the two variables are completely unrelated and measure different things. The total can be less than the sum if the terms are related and are at least in part explaining the same variation. Finally, the total can exceed the sum if the two variables act jointly so that knowing both gives more information than knowing just one of them. For example, the area of a rectangle may be only poorly determined by either the length or width alone, but if both are considered at the same time, area can be determined exactly. It is precisely this inability to predict the joint relationship from the marginal relationships that makes multiple regression rich and complicated.

3.1.2 Added-Variable Plots

The *unique* effect of adding *Purban* to a mean function that already includes log(*PPgdp*) is determined by the relationship between the part of log(*Fertility*) that is not explained by log(*PPgdp*) and the part of *Purban* that is not explained by log(*PPgdp*). The "unexplained parts" are just the residuals from these two simple regressions, and so we need to examine the scatterplot of the residuals from the regression of log(*Fertility*) on log(*PPgdp*) versus the residuals from the regression of *Purban* on log(*PPgdp*). This plot is shown in Figure 3.1d. Figure 3.1b is the summary graph for the relationship between log(*Fertility*) and *Purban ignoring* log(*PPgdp*), while Figure 3.1d shows this relationship, but after *adjusting* for log(*PPgdp*). If Figure 3.1d shows a stronger relationship than does Figure 3.1b, meaning that the points in the plot show less variation about the fitted straight line,

then the two variables act jointly to explain extra variation, while if the relationship is weaker, or the plot exhibits more variation, then the total explained variability is less than the additive amount. The latter seems to be the case here.

If we fit the simple regression mean function to Figure3.1d, the fitted line has zero intercept, since the averages of the two plotted variables are zero, and the estimated slope via OLS is $\hat{\beta}_2 = -0.0035 \approx -0.004$. It turns out that this is exactly the estimate $\hat{\beta}_2$ that would be obtained using OLS to get the estimates using the mean function (3.1). Figure 3.1d is called an *added-variable plot*.

We now have two estimates of the coefficient β_2 for *Purban*:

$$\hat{\beta}_2 = -0.013 \textit{ ignoring } \log(PPgdp)$$
$$\hat{\beta}_2 = -0.004 \textit{ adjusting for } \log(PPgdp)$$

While both of these indicate that more urbanization is associated with lower fertility, adjusting for log(*PPgdp*) suggests that the magnitude of this effect is only about one-fourth as large as one might think if log(*PPgdp*) were ignored. In other problems, slope estimates for the same term but from different mean functions can be even more wildly different, changing signs, magnitude, and significance. This naturally complicates the interpretation of fitted models, and also comparing between studies fit with even slightly different mean functions.

To get the coefficient estimate for log(*PPgdp*) in the regression of log(*Fertility*) on both predictors, we would use the same procedure we used for *Purban* and consider the problem of adding log(*PPgdp*) to a mean function that already includes *Purban*. This would require looking at the graph of the residuals from the regression of log(*Fertility*) on *Purban* versus the residuals from the regression of log(*PPgdp*) on *Purban* (see Problem 3.2).

3.2 THE MULTIPLE LINEAR REGRESSION MODEL

The general multiple linear regression model with response Y and terms $X_1, \ldots, X_p$ will have the form

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{3.2}$$

The symbol X in $\mathrm{E}(Y|X)$ means that we are conditioning on all the terms on the right side of the equation. Similarly, when we are conditioning on specific values for the predictors $x_1, \ldots, x_p$ that we will collectively call $\mathbf{x}$, we write

$$\mathrm{E}(Y|X = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{3.3}$$

As in Chapter 2, the βs are unknown parameters we need to estimate. Equation (3.2) is a *linear function of the parameters*, which is why this is called linear regression. When $p = 1$, X has only one element, and we get the simple regression problem discussed in Chapter 2. When $p = 2$, the mean function (3.2) corresponds

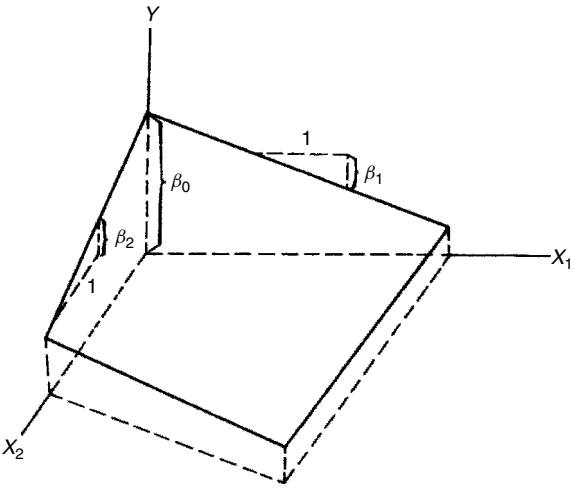

FIG. 3.2 A linear regression surface with $p = 2$ predictors.

to a plane in three dimensions, as shown in Figure 3.2. When $p > 2$, the fitted mean function is a *hyperplane*, the generalization of a p-dimensional plane in a $(p + 1)$-dimensional space. We cannot draw a general p-dimensional plane in our three-dimensional world.

3.3 TERMS AND PREDICTORS

Regression problems start with a collection of potential predictors. Some of these may be continuous measurements, like the height or weight of an object. Some may be discrete but ordered, like a doctor's rating of overall health of a patient on a nine-point scale. Other potential predictors can be categorical, like eye color or an indicator of whether a particular unit received a treatment. All these types of potential predictors can be useful in multiple linear regression.

From the pool of potential predictors, we create a set of *terms* that are the X-variables that appear in (3.2). The terms might include:

The intercept The mean function (3.2) can we rewritten as

$$\mathrm{E}(Y|X) = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where X_0 is a term that is always equal to one. Mean functions without an intercept would not have this term included.

Predictors The simplest type of term is equal to one of the predictors, for example, the variable *Mheight* in the heights data.

Transformations of predictors Sometimes the original predictors need to be transformed in some way to make (3.2) hold to a reasonable approximation. This was the case with the UN data just discussed, in which *PPgdp* was

used in log scale. The willingness to replace predictors by transformations of them greatly expands the range of problems that can be summarized with a linear regression model.

Polynomials Problems with curved mean functions can sometimes be accommodated in the multiple linear regression model by including polynomial terms in the predictor variables. For example, we might include as terms both a predictor X_1 and its square X_1^2 to fit a quadratic polynomial in that predictor. Complex polynomial surfaces in several predictors can be useful in some problems[1].

Interactions and other combinations of predictors Combining several predictors is often useful. An example of this is using body mass index, given by height divided by weight squared, in place of both height and weight, or using a total test score in place of the separate scores from each of several parts. Products of predictors called *interactions* are often included in a mean function along with the original predictors to allow for joint effects of two or more variables.

Dummy variables and factors A categorical predictor with two or more levels is called a *factor*. Factors are included in multiple linear regression using *dummy variables*, which are typically terms that have only two values, often zero and one, indicating which category is present for a particular observation. We will see in Chapter 6 that a categorical predictor with two categories can be represented by one dummy variable, while a categorical predictor with many categories can require several dummy variables.

A regression with say k predictors may combine to give fewer than k terms or expand to require more than k terms. The distinction between predictors and terms can be very helpful in thinking about an appropriate mean function to use in a particular problem, and in using graphs to understand a problem. For example, a regression with one predictor can always be studied using the 2D scatterplot of the response versus the predictor, regardless of the number of terms required in the mean function.

We will use the fuel consumption data introduced in Section 1.6 as the primary example for the rest of this chapter. As discussed earlier, the goal is to understand how fuel consumption varies as a function of state characteristics. The variables are defined in Table 1.2 and are given in the file `fuel2001.txt`. From the six initial predictors, we use a set of four combinations to define terms in the regression mean function.

Basic summary statistics for the relevant variables in the fuel data are given in Table 3.1, and these begin to give us a bit of a picture of these data. First, there is quite a bit of variation in *Fuel*, with values between a minimum of about 626 gallons per year and a maximum of about 843 gallons per year. The gas *Tax* varies

[1]This discussion of polynomials might puzzle some readers because in Section 3.2, we said the linear regression mean function was a hyperplane, but here we have said that it might be curved, seemingly a contradiction. However, *both* of these statements are correct. If we fit a mean function like $\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2$, the mean function is a quadratic curve in the plot of the response versus x but a plane in the three-dimensional plot of the response versus x and x^2.

TABLE 3.1 Summary Statistics for the Fuel Data

Variable	N	Average	Std Dev	Minimum	Median	Maximum
Tax	51	20.155	4.5447	7.5	20.	29.
Dlic	51	903.68	72.858	700.2	909.07	1075.3
Income	51	28.404	4.4516	20.993	27.871	40.64
logMiles	51	15.745	1.4867	10.583	16.268	18.198
Fuel	51	613.13	88.96	317.49	626.02	842.79

from only 7.5 cents per gallon to a high of 29 cents per gallon, so unlike much of the world gasoline taxes account for only a small part of the cost to consumers of gasoline. Also of interest is the range of values in *Dlic*: The number of licensed drivers per 1000 population over the age of 16 is between about 700 and 1075. Some states appear to have more licensed drivers than they have population over age 16. Either these states allow drivers under the age of 16, allow nonresidents to obtain a driver's license, or the data are in error. For this example, we will assume one of the first two reasons.

Of course, these univariate summaries cannot tell us much about how the fuel consumption depends on the other variables. For this, graphs are very helpful. The scatterplot matrix for the fuel data is repeated in Figure 3.3. From our previous

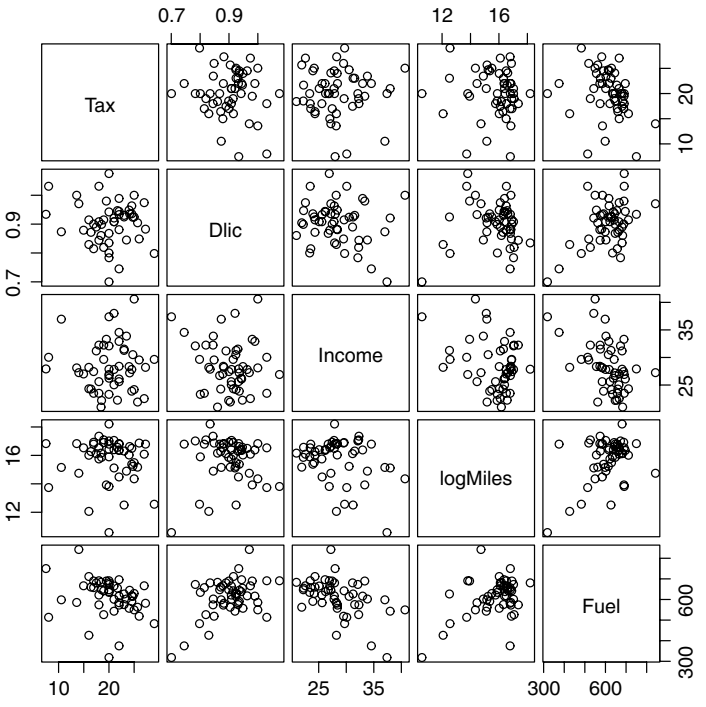

FIG. 3.3 Scatterplot matrix for the fuel data.

TABLE 3.2 Sample Correlations for the Fuel Data

Sample Correlations

	Tax	Dlic	Income	logMiles	Fuel
Tax	1.0000	-0.0858	-0.0107	-0.0437	-0.2594
Dlic	-0.0858	1.0000	-0.1760	0.0306	0.4685
Income	-0.0107	-0.1760	1.0000	-0.2959	-0.4644
logMiles	-0.0437	0.0306	-0.2959	1.0000	0.4220
Fuel	-0.2594	0.4685	-0.4644	0.4220	1.0000

discussion, *Fuel* decreases on the average as *Tax* increases, but there is lot of variation. We can make similar qualitative judgments about each of the regressions of *Fuel* on the other variables. The overall impression is that *Fuel* is at best weakly related to each of the variables in the scatterplot matrix, and in turn these variables are only weakly related to each other.

Does this help us understand how *Fuel* is related to all four predictors simultaneously? We know from the discussion in Section 3.1 that the marginal relationships between the response and each of the variables is *not* sufficient to understand the *joint* relationship between the response and the terms. The interrelationships among the terms are also important. The pairwise relationships between the terms can be viewed in the remaining cells of the scatterplot matrix. In Figure 3.3, the relationships between all pairs of terms appear to be very weak, suggesting that for this problem the marginal plots including *Fuel* are quite informative about the multiple regression problem.

A more traditional, and less informative, summary of the two-variable relationships is the matrix of sample correlations, shown in Table 3.2. In this instance, the correlation matrix helps to reinforce the relationships we see in the scatterplot matrix, with fairly small correlations between the predictors and *Fuel*, and essentially no correlation between the predictors themselves.

3.4 ORDINARY LEAST SQUARES

From the initial collection of potential predictors, we have computed a set of $p+1$ terms, including an intercept, $X = (X_0, X_1, \ldots, X_p)$. The mean function and variance function for multiple linear regression are

$$\begin{aligned} \mathrm{E}(Y|X) &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \\ \mathrm{Var}(Y|X) &= \sigma^2 \end{aligned} \tag{3.4}$$

Both the βs and σ^2 are unknown parameters that need to be estimated.

3.4.1 Data and Matrix Notation

Suppose we have observed data for n cases or units, meaning we have a value of Y and all of the terms for each of the n cases. We have symbols for the response and

the terms using matrices and vectors; see Appendix A.6 for a brief introduction. We define

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \tag{3.5}$$

so $\mathbf{Y}$ is an $n \times 1$ vector and $\mathbf{X}$ is an $n \times (p+1)$ matrix. We also define $\boldsymbol{\beta}$ to be a $(p+1) \times 1$ vector of regression coefficients and $\mathbf{e}$ to be the $n \times 1$ vector of statistical errors,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{and} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

The matrix $\mathbf{X}$ gives all of the observed values of the terms. The ith row of $\mathbf{X}$ will be defined by the symbol $\mathbf{x}_i'$, which is a $(p+1) \times 1$ vector for mean functions that include an intercept. Even though $\mathbf{x}_i$ is a row of $\mathbf{X}$, we use the convention that all vectors are column vectors and therefore need to write $\mathbf{x}_i'$ to represent a row. An equation for the mean function evaluated at $\mathbf{x}_i$ is

$$\begin{aligned} \mathrm{E}(Y|X = \mathbf{x}_i) &= \mathbf{x}_i'\boldsymbol{\beta} \\ &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \end{aligned} \tag{3.6}$$

In matrix notation, we will write the multiple linear regression model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{3.7}$$

The ith row of (3.7) is $y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i$.

For the fuel data, the first few and the last few rows of the matrix $\mathbf{X}$ and the vector $\mathbf{Y}$ are

$$\mathbf{X} = \begin{pmatrix} 1 & 18.00 & 1031.38 & 23.471 & 16.5271 \\ 1 & 8.00 & 1031.64 & 30.064 & 13.7343 \\ 1 & 18.00 & 908.597 & 25.578 & 15.7536 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 25.65 & 904.894 & 21.915 & 15.1751 \\ 1 & 27.30 & 882.329 & 28.232 & 16.7817 \\ 1 & 14.00 & 970.753 & 27.230 & 14.7362 \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} 690.264 \\ 514.279 \\ 621.475 \\ \vdots \\ 562.411 \\ 581.794 \\ 842.792 \end{pmatrix}$$

The terms in $\mathbf{X}$ are in the order intercept, *Tax*, *Dlic*, *Income* and finally log(*Miles*). The matrix $\mathbf{X}$ is 51×5 and $\mathbf{Y}$ is 51×1.

3.4.2 Variance-Covariance Matrix of e

The 51×1 error vector is an unobservable random vector, as in Appendix A.6. The assumptions concerning the e_is given in Chapter 2 are summarized in matrix form as

$$\mathrm{E}(\mathbf{e}) = \mathbf{0} \qquad \mathrm{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$$

where Var(**e**) means the covariance matrix of **e**, $\mathbf{I}_n$ is the $n \times n$ matrix with ones on the diagonal and zeroes everywhere else, and **0** is a matrix or vector of zeroes of appropriate size. If we add the assumption of normality, we can write

$$\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

3.4.3 Ordinary Least Squares Estimators

The least squares estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is chosen to minimize the residual sum of squares function

$$RSS(\boldsymbol{\beta}) = \sum (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.8}$$

The OLS estimates can be found from (3.8) by differentiation in a matrix analog to the development of Appendix A.3. The OLS estimate is given by the formula

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3.9}$$

provided that the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ exists. The estimator $\hat{\boldsymbol{\beta}}$ depends only on the sufficient statistics $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$, which are matrices of uncorrected sums of squares and cross-products.

Do not compute the least squares estimates using (3.9)! Uncorrected sums of squares and cross-products are prone to large rounding error, and so computations can be highly inaccurate. The preferred computational methods are based on matrix decompositions as briefly outlined in Appendix A.8. At the very least, computations should be based on *corrected* sums of squares and cross-products.

Suppose we define $\mathcal{X}$ to be the $n \times p$ matrix

$$\mathcal{X} = \begin{pmatrix} (x_{11} - \overline{x}_1) & \cdots & (x_{1p} - \overline{x}_p) \\ (x_{21} - \overline{x}_1) & \cdots & (x_{2p} - \overline{x}_p) \\ \vdots & \vdots & \vdots \\ (x_{n1} - \overline{x}_1) & \cdots & (x_{np} - \overline{x}_p) \end{pmatrix}$$

This matrix consists of the original **X** matrix, but with the first column removed and the column mean subtracted from each of the remaining columns. Similarly, $\mathcal{Y}$ is the vector with typical elements $y_i - \overline{y}$. Then

$$\mathcal{C} = \frac{1}{n-1} \begin{pmatrix} \mathcal{X}'\mathcal{X} & \mathcal{X}'\mathcal{Y} \\ \mathcal{Y}'\mathcal{X} & \mathcal{Y}'\mathcal{Y} \end{pmatrix} \tag{3.10}$$

is the matrix of sample variances and covariances. When $p = 1$, the matrix $\mathcal{C}$ is given by

$$\mathcal{C} = \frac{1}{n-1}\begin{pmatrix} SXX & SXY \\ SXY & SYY \end{pmatrix}$$

The elements of $\mathcal{C}$ are the summary statistics needed for OLS computations in simple linear regression. If we let $\boldsymbol{\beta}^*$ be the parameter vector excluding the intercept β_0, then for $p \geq 1$,

$$\hat{\boldsymbol{\beta}}^* = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{Y}$$
$$\hat{\beta}_0 = \overline{y} - \hat{\boldsymbol{\beta}}^{*\prime}\overline{\mathbf{x}}$$

where $\overline{\mathbf{x}}$ is the vector of sample means for all the terms except for the intercept.

Once $\hat{\boldsymbol{\beta}}$ is computed, we can define several related quantities. The fitted values are $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and the residuals are $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$. The function (3.8) evaluated at $\hat{\boldsymbol{\beta}}$ is the residual sum of squares, or *RSS*,

$$RSS = \hat{\mathbf{e}}'\hat{\mathbf{e}} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{3.11}$$

3.4.4 Properties of the Estimates

Additional properties of the OLS estimates are derived in Appendix A.8 and are only summarized here. Assuming that $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{e}) = \sigma^2\mathbf{I}_n$, then $\hat{\boldsymbol{\beta}}$ is unbiased, $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{3.12}$$

Excluding the intercept term,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^*) = \sigma^2(\mathcal{X}'\mathcal{X})^{-1} \tag{3.13}$$

and so $(\mathcal{X}'\mathcal{X})^{-1}$ is all but the first row and column of $(\mathbf{X}'\mathbf{X})^{-1}$. An estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{RSS}{n-(p+1)} \tag{3.14}$$

which is the residual sum of squares divided by its $\mathrm{df} = n - (p+1)$. Several formulas for *RSS* can be computed by substituting the value of $\hat{\boldsymbol{\beta}}$ into (3.11) and simplifying:

$$\begin{aligned} RSS &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \\ &= \mathcal{Y}'\mathcal{Y} - \hat{\boldsymbol{\beta}}^{*\prime}(\mathcal{X}'\mathcal{X})\hat{\boldsymbol{\beta}}^* \\ &= \mathcal{Y}'\mathcal{Y} - \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} + n\overline{y}^2 \end{aligned} \tag{3.15}$$

Recognizing that $\mathcal{Y}'\mathcal{Y} = SYY$, (3.15) has the nicest interpretation, as it writes *RSS* as equal to the total sum of squares minus a quantity we will call the *regression sum of squares*, or *SSreg*. In addition, if **e** is normally distributed, then the residual sum of squares has a Chi-squared distribution,

$$(n - (p+1))\hat{\sigma}^2/\sigma^2 \sim \chi^2(n - (p+1))$$

By substituting $\hat{\sigma}^2$ for σ^2 in (3.12), we find the estimated variance of $\hat{\boldsymbol{\beta}}$, $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$, to be

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \tag{3.16}$$

3.4.5 Simple Regression in Matrix Terms

For simple regression, **X** and **Y** are given by

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

and thus

$$(\mathbf{X}'\mathbf{X}) = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \qquad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum y_i \\ \sum y_i^2 \end{pmatrix}$$

By direct multiplication, $(\mathbf{X}'\mathbf{X})^{-1}$ can be shown to be

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{SXX}\begin{pmatrix} \sum x_i^2/n & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix} \tag{3.17}$$

so that

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \frac{1}{SXX}\begin{pmatrix} \sum x_i^2/n & -\overline{x} \\ -\overline{x} & \sum x_i y_i \end{pmatrix}\begin{pmatrix} \sum y_i \\ \sum y_i^2 \end{pmatrix}$$
$$= \begin{pmatrix} \overline{y} - \hat{\beta}_1\overline{x} \\ SXY/SXX \end{pmatrix}$$

as found previously. Also, since $\sum x_i^2/(nSXX) = 1/n + \overline{x}^2/SXX$, the variances and covariances for $\hat{\beta}_0$ and $\hat{\beta}_1$ found in Chapter 2 are identical to those given by $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

The results are simpler in the deviations from the sample average form, since

$$\mathcal{X}'\mathcal{X} = SXX \qquad \mathcal{X}'\mathcal{Y} = SXY$$

and

$$\hat{\beta}_1 = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{Y} = \frac{SXY}{SXX}$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$$

Fuel Consumption Data

We will generally let p equal the number of terms in a mean function excluding the intercept, and $p' = p + 1$ equal if the intercept is included; $p' = p$ if the intercept is not included. We shall now fit the mean function with $p' = 5$ terms, including the intercept for the fuel consumption data. Continuing a practice we have already begun, we will write *Fuel* on *Tax Dlic Income* log(*Miles*) as shorthand for using OLS to fit the multiple linear regression model with mean function

$$\mathrm{E}(Fuel|X) = \beta_0 + \beta_1 Tax + \beta_2 Dlic + \beta_3 Income + \beta_4 \log(Miles)$$

where conditioning on X is short for conditioning on all the terms in the mean function. All the computations are based on the summary statistics, which are the sample means given in Table 3.1 and the sample covariance matrix $\mathcal{C}$ defined at (3.10) and given by

	Tax	Dlic	Income	logMiles	Fuel
Tax	20.6546	-28.4247	-0.2162	-0.2955	-104.8944
Dlic	-28.4247	5308.2591	-57.0705	3.3135	3036.5905
Income	-0.2162	-57.0705	19.8171	-1.9580	-183.9126
logMiles	-0.2955	3.3135	-1.9580	2.2103	55.8172
Fuel	-104.8944	3036.5905	-183.9126	55.8172	7913.8812

Most statistical software will give the sample correlations rather than the covariances. The reader can verify that the correlations in Table 3.2 can be obtained from these covariances. For example, the sample correlation between *Tax* and *Income* is $-0.2162/\sqrt{(20.6546 \times 19.8171)} = -0.0107$ as in Table 3.2. One can convert back from correlations and sample variances to covariances; the square root of the sample variances are given in Table 3.1.

The 5×5 matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is given by

	Intercept	Tax	Dlic	Income	logMiles
Intercept	9.02151	-2.852e-02	-4.080e-03	-5.981e-02	-1.932e-01
Tax	-0.02852	9.788e-04	5.599e-06	4.263e-05	1.602e-04
Dlic	-0.00408	5.599e-06	3.922e-06	1.189e-05	5.402e-06
Income	-0.05981	4.263e-05	1.189e-05	1.143e-03	1.000e-03
logMiles	-0.19315	1.602e-04	5.402e-06	1.000e-03	9.948e-03

The elements of $(\mathbf{X}'\mathbf{X})^{-1}$ often differ by several orders of magnitude, as is the case here, where the smallest element in absolute value is $3.9 \times 10^{-6} = 0.0000039$, and the largest element is 9.02. It is the combining of these numbers of very different magnitude that can lead to numerical inaccuracies in computations.

The lower-right 4×4 sub-matrix of $(\mathbf{X}'\mathbf{X})^{-1}$ is $(\mathcal{X}'\mathcal{X})^{-1}$. Using the formulas based on corrected sums of squares in this chapter, the estimate $\hat{\boldsymbol{\beta}}^*$ is computed to be

$$\hat{\boldsymbol{\beta}}^* = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{Y} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{pmatrix} = \begin{pmatrix} -4.2280 \\ 0.4719 \\ -6.1353 \\ 18.5453 \end{pmatrix}$$

The estimated intercept is

$$\hat{\beta}_0 = \overline{y} - \hat{\boldsymbol{\beta}}^{*\prime}\overline{\mathbf{x}} = 154.193$$

and the residual sum of squares is

$$RSS = \mathcal{Y}'\mathcal{Y} - \hat{\boldsymbol{\beta}}^{*\prime}(\mathcal{X}'\mathcal{X})\hat{\boldsymbol{\beta}}^* = 193{,}700$$

so the estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{RSS}{n - (p+1)} = \frac{193{,}700}{51 - 5} = 4211$$

Standard errors and estimated covariances of the $\hat{\beta}_j$ are found by multiplying $\hat{\sigma}$ by the square roots of elements of $(\mathbf{X}'\mathbf{X})^{-1}$. For example,

$$\text{se}(\hat{\beta}_2) = \hat{\sigma}\sqrt{3.922 \times 10^{-6}} = 0.1285$$

Virtually all statistical software packages include higher-level functions that will fit multiple regression models, but getting intermediate results like $(\mathbf{X}'\mathbf{X})^{-1}$ may be a challenge. Table 3.3 shows typical output from a statistical package. This output gives the estimates $\hat{\boldsymbol{\beta}}$ and their standard errors computed based on $\hat{\sigma}^2$ and the

TABLE 3.3 Edited Output from the `Summary` Method in R for Multiple Regression in the Fuel Data

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.1928   194.9062   0.791 0.432938
Tax          -4.2280     2.0301  -2.083 0.042873
Dlic          0.4719     0.1285   3.672 0.000626
Income       -6.1353     2.1936  -2.797 0.007508
logMiles     18.5453     6.4722   2.865 0.006259

Residual standard error: 64.89 on 46 degrees of freedom
Multiple R-Squared: 0.5105
F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.33e-07
```

diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$. The column marked `t-value` is the ratio of the estimate to its standard error. The column labelled `Pr(>|t|)` will be discussed shortly. Below the table are a number of other summary statistics; at this point only the estimate of σ called the residual standard error and its df are familiar.

3.5 THE ANALYSIS OF VARIANCE

For multiple regression, the analysis of variance is a very rich technique that is used to compare mean functions that include different nested sets of terms. In the *overall analysis of variance*, the mean function with all the terms

$$\mathrm{E}(Y|X = \mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} \tag{3.18}$$

is compared with the mean function that includes only an intercept:

$$\mathrm{E}(Y|X = \mathbf{x}) = \beta_0 \tag{3.19}$$

For simple regression, these correspond to (2.16) and (2.13), respectively. For mean function (3.19), $\hat{\beta}_0 = \overline{y}$ and the residual sum of squares is *SYY*. For mean function (3.18), the estimate of β is given by (3.9) and *RSS* is given in (3.11). We must have $RSS < SYY$, and the difference between these two

$$SSreg = SYY - RSS \tag{3.20}$$

corresponds to the sum of squares of Y explained by the larger mean function that is not explained by the smaller mean function. The number of df associated with *SSreg* is equal to the number of df in *SYY* minus the number of df in *RSS*, which equals p, the number of terms in the mean function excluding the intercept.

These results are summarized in the analysis of variance table in Table 3.4. We can judge the importance of the regression on the terms in the larger model by determining if *SSreg* is sufficiently large by comparing the ratio of the mean square for regression to $\hat{\sigma}^2$ to the $F(p, n - p')$ distribution[2] to get a significance

TABLE 3.4 The Overall Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	p	$SSreg$	$SSreg/1$	$MSreg/\hat{\sigma}^2$	
Residual	$n-(p+1)$	RSS	$\hat{\sigma}^2 = RSS/(n-2)$		
Total	$n-1$	SYY			

[2]Reminder: $p' = p$ for mean functions with no intercept, and $p' = p + 1$ for mean functions with an intercept.

level. If the computed significance level is small enough, then we would judge that the mean function (3.18) provides a significantly better fit than does (3.19). The ratio will have an exact F distribution if the errors are normal and (3.19) is true. The hypothesis tested by this F-test is

$$\begin{array}{ll} \text{NH:} & \mathrm{E}(Y|X=\mathbf{x}) = \beta_0 \\ \text{AH:} & \mathrm{E}(Y|X=\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} \end{array}$$

3.5.1 The Coefficient of Determination

As with simple regression, the ratio

$$R^2 = \frac{SSreg}{SYY} = 1 - \frac{RSS}{SYY} \tag{3.21}$$

gives the proportion of variability in Y explained by regression on the terms. R^2 can also be shown to be the square of the correlation between the observed values Y and the fitted values $\hat{Y}$; we will explore this further in the next chapter. R^2 is also called the *multiple correlation coefficient* because it is the maximum of the correlation between Y and *any* linear combination of the terms in the mean function.

Fuel Consumption Data

The overall analysis of variance table is given by

```
           Df Sum Sq Mean Sq F value    Pr(>F)
Regression 4 201994   50499  11.992  9.33e-07
Residuals 46 193700    4211
Total     50 395694
```

To get a significance level for the test, we would compare $F = 11.992$ with the $F(4, 46)$ distribution. Most computer packages do this automatically, and the result is shown in the column marked `Pr(>F)` to be about 0.0000009, a very small number, leading to very strong evidence against the null hypothesis that the mean of *Fuel* does not depend on any of the terms. The value of $R^2 = 201994/395694 = 0.5105$ indicates that about half the variation in *Fuel* is explained by the terms. The value of F, its significance level, and the value of R^2 are given in Table 3.3.

3.5.2 Hypotheses Concerning One of the Terms

Obtaining information on one of the terms may be of interest. Can we do as well, understanding the mean function for *Fuel* if we delete the *Tax* variable? This amounts to the following hypothesis test of

$$\begin{array}{lll} \text{NH:} & \beta_1 = 0, & \beta_0, \beta_2, \beta_3, \beta_4 \text{ arbitrary} \\ \text{AH:} & \beta_1 \neq 0, & \beta_0, \beta_2, \beta_3, \beta_4 \text{ arbitrary} \end{array} \tag{3.22}$$

The following procedure can be used. First, fit the mean function that excludes the term for *Tax* and get the residual sum of squares for this smaller mean function.

Then fit again, this time including *Tax*, and once again get the residual sum of squares. Subtracting the residual sum of squares for the larger mean function from the residual sum of squares for the smaller mean function will give the sum of squares for regression on *Tax* after adjusting for the terms that are in both mean functions, *Dlic*, *Income* and log(*Miles*). Here is a summary of the computations that are needed:

	Df	SS	MS	F	Pr(>F)
Excluding Tax	47	211964			
Including Tax	46	193700			
Difference	1	18264	18264	4.34	0.043

The row marked "Excluding Tax" gives the df and *RSS* for the mean function without *Tax*, and the next line gives these values for the larger mean function including *Tax*. The difference between these two given on the next line is the sum of squares explained by *Tax* after adjusting for the other terms in the mean function. The F-test is given by $F = (18{,}264/1)/\hat{\sigma}^2 = 4.34$, which, when compared to the F distribution with $(1, 46)$ df gives a significance level of about 0.04. We thus have modest evidence that the coefficient for *Tax* is different from zero. This is called a *partial F-test*. Partial F-tests can be generalized to testing *several* coefficients to be zero, but we delay that generalization to Section 5.4.

3.5.3 Relationship to the t-Statistic

Another reasonable procedure for testing the importance of *Tax* is simply to compare the estimate of the coefficient divided by its standard error to the $t(n - p')$ distribution. One can show that the square of this t-statistic is the same number of the F-statistic just computed, so these two procedures are identical. Therefore, the t-statistic tests hypothesis (3.22) concerning the importance of terms adjusted for all the other terms, not ignoring them.

From Table 3.3, the t-statistic for *Tax* is $t = -2.083$, and $t^2 = (-2.083)^2 = 4.34$, the same as the F-statistic we just computed. The significance level for *Tax* given in Table 3.3 also agrees with the significance level we just obtained for the F-test, and so the significance level reported is for the two-sided test. To test the hypothesis that $\beta_1 = 0$ against the one-sided alternative that $\beta_1 < 0$, we could again use the same t-value, but the significance level would be one-half of the value for the two-sided test.

A t-test that β_j has a specific value versus a two-sided or one-sided alternative (with all other coefficients arbitrary) can be carried out as described in Section 2.8.

3.5.4 t-Tests and Added-Variable Plots

In Section 3.1, we discussed adding a term to a simple regression mean function. The same general procedure can be used to add a term to *any* linear regression mean function. For the added-variable plot for a term, say X_1, plot the residuals from the regression of Y on all the other X's versus the residuals for the regression

of X_1 on all the other Xs. One can show (Problem 3.2) that (1) the slope of the regression in the added-variable plot is the estimated coefficient for X_1 in the regression with all the terms, and (2) the t-test for testing the slope to the zero in the added-variable plot is essentially the same as the t-test for testing $\beta_1 = 0$ in the fit of the larger mean function, the only difference being a correction for degrees of freedom.

3.5.5 Other Tests of Hypotheses

We have obtained a test of a hypothesis concerning the effect of *Tax* adjusted for all the other terms in the mean function. Equally well, we could obtain tests for the effect of *Tax* adjusting for some of the other terms or for none of the other terms. In general, these tests will not be equivalent, and a variable can be judged useful ignoring variables but useless when adjusted for them. Furthermore, a predictor that is useless by itself may become important when considered in concert with the other variables. The outcome of these tests depends on the sample correlations between the terms.

3.5.6 Sequential Analysis of Variance Tables

By separating *Tax* from the other terms, *SSreg* is divided into two pieces, one for fitting the first three terms, and one for fitting *Tax* after the other three. This subdivision can be continued by dividing *SSreg* into a sum of squares "explained" by each term separately. Unless all the terms are uncorrelated, this breakdown is not unique. For example, we could first fit *Dlic*, then *Tax* adjusted for *Dlic*, then *Income* adjusted for both *Dlic* and *Tax*, and finally log(*Miles*) adjusted for the other three. The resulting table is given in Table 3.5a. Alternatively, we could fit in the order log(*Miles*), *Income*, *Dlic* and then *Tax* as in Table 3.5b. The sums of squares can be quite different in the two tables. For example, the sum of squares for *Dlic* ignoring the other terms is about 25% larger than the sum of squares for *Dlic* adjusting for the other terms. In this problem, the terms are nearly uncorrelated, see Table 3.2, so the effect of ordering is relatively minor. In problems with high sample correlations between terms, order can be very important.

TABLE 3.5 Two Analysis of Variance Tables with Different Orders of Fitting

(a) First analysis

	Df	Sum Sq	Mean Sq
Dlic	1	86854	86854
Tax	1	19159	19159
Income	1	61408	61408
logMiles	1	34573	34573
Residuals	46	193700	4211

(b) Second analysis

	Df	Sum Sq	Mean Sq
logMiles	1	70478	70478
Income	1	49996	49996
Dlic	1	63256	63256
Tax	1	18264	18264
Residuals	46	193700	4211

3.6 PREDICTIONS AND FITTED VALUES

Suppose we have observed, or will in the future observe, a new case with its own set of predictors that result in a vector of terms $\mathbf{x}_*$. We would like to predict the value of the response given $\mathbf{x}_*$. In exactly the same way as was done in simple regression, the point prediction is $\tilde{y}_* = \mathbf{x}_*'\hat{\boldsymbol{\beta}}$, and the standard error of prediction, sepred$(\tilde{y}_*|\mathbf{x}_*)$, using Appendix A.8, is

$$\text{sepred}(\tilde{y}_*|\mathbf{x}_*) = \hat{\sigma}\sqrt{1 + \mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*} \tag{3.23}$$

Similarly, the estimated average of all possible units with a value $\mathbf{x}$ for the terms is given by the estimated mean function at $\mathbf{x}$, $\hat{E}(Y|X = \mathbf{x}) = \hat{y} = \mathbf{x}'\hat{\boldsymbol{\beta}}$ with standard error given by

$$\text{sefit}(\hat{y}|\mathbf{x}) = \hat{\sigma}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}} \tag{3.24}$$

Virtually all software packages will give the user access to the fitted values, but getting the standard error of prediction and of the fitted value may be harder. If a program produces sefit but not sepred, the latter can be computed from the former from the result

$$\text{sepred}(\tilde{y}_*|\mathbf{x}_*) = \sqrt{\hat{\sigma}^2 + \text{sefit}(\tilde{y}_*|\mathbf{x}_*)^2}$$

PROBLEMS

3.1. Berkeley Guidance Study The Berkeley Guidance Study enrolled children born in Berkeley, California, between January 1928 and June 1929, and then measured them periodically until age eighteen (Tuddenham and Snyder, 1954). The data we use is described in Table 3.6, and the data is given in the data files `BGSgirls.txt` for girls only, `BGSboys.txt` for boys only, and `BGSall.txt` for boys and girls combined. For this example, use only the data on the girls.

3.1.1. For the girls only, draw the scatterplot matrix of all the age two variables, all the age nine variables and *Soma*. Write a summary of the information in this scatterplot matrix. Also obtain the matrix of sample correlations between the height variables.

3.1.2. Starting with the mean function $\text{E}(Soma|WT9) = \beta_0 + \beta_1 WT9$, use added-variable plots to explore adding *LG9* to get the mean function $\text{E}(Soma|WT9, LG9) = \beta_0 + \beta_1 WT9 + \beta_2 LG9$. In particular, obtain the four plots equivalent to Figure 3.1, and summarize the information in the plots.

3.1.3. Fit the multiple linear regression model with mean function

$$\text{E}(Soma|X) = \beta_0 + \beta_1 HT2 + \beta_2 WT2 + \beta_3 HT9 + \beta_4 WT9 + \beta_5 ST9 \tag{3.25}$$

TABLE 3.6 Variable Definitions for the Berkeley Guidance Study in the Files `BGSgirls.txt`, `BGSboys.txt`, and `BGSall.txt`

Variable	Description
Sex	0 for males, 1 for females
WT2	Age 2 weight, kg
HT2	Age 2 height, cm
WT9	Age 9 weight, kg
HT9	Age 9 height, cm
LG9	Age 9 leg circumference, cm
ST9	Age 9 strength, kg
WT18	Age 18 weight, kg
HT18	Age 18 height, cm
LG18	Age 18 leg circumference, cm
ST18	Age 18 strength, kg
Soma	Somatotype, a scale from 1, very thin, to 7, obese, of body type

Find $\hat{\sigma}$, R^2, the overall analysis of variance table and overall F-test. Compute the t-statistics to be used to test each of the β_j to be zero against two-sided alternatives. Explicitly state the hypotheses tested and the conclusions.

3.1.4. Obtain the sequential analysis of variance table for fitting the variables in the order they are given in (3.25). State the hypotheses tested and the conclusions for each of the tests.

3.1.5. Obtain analysis of variance again, this time fitting with the five terms in the order given from right to left in (3.25). Explain the differences with the table you obtained in Problem 3.1.4. What graphs could help understand the issues?

3.2. Added-variable plots This problem uses the United Nations example in Section 3.1 to demonstrate many of the properties of added-variable plots. This problem is based on the mean function

$$\mathrm{E}(\log(\mathit{Fertility})|\log(\mathit{PPgdp}) = x_1, \mathit{Purban} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

There is nothing special about the two-predictor regression mean function, but we are using this case for simplicity.

3.2.1. Show that the estimated coefficient for $\log(\mathit{PPgdp})$ is the same as the estimated slope in the added-variable plot for $\log(\mathit{PPgdp})$ after *Purban*. This correctly suggests that *all the estimates in a multiple linear regression model are adjusted for all the other terms in the mean function.* Also, show that the residuals in the added-variable plot are identical to the residuals from the mean function with both predictors.

3.2.2. Show that the t-test for the coefficient for log(*PPgdp*) is not quite the same from the added-variable plot and from the regression with both terms, and explain why they are slightly different.

3.3. The following questions all refer to the mean function

$$\mathrm{E}(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{3.26}$$

3.3.1. Suppose we fit (3.26) to data for which $x_1 = 2.2x_2$, with no error. For example, x_1 could be a weight in pounds, and x_2 the weight of the same object in kg. Describe the appearance of the added-variable plot for X_2 after X_1.

3.3.2. Again referring to (3.26), suppose now that Y and X_1 are perfectly correlated, so $Y = 3X_1$, without any error. Describe the appearance of the added-variable plot for X_2 after X_1.

3.3.3. Under what conditions will the added-variable plot for X_2 after X_1 have exactly the same shape as the scatterplot of Y versus X_2?

3.3.4. True or false: The vertical variation in an added-variable plot for X_2 after X_1 is always less than or equal to the vertical variation in a plot of Y versus X_2. Explain.

3.4. Suppose we have a regression in which we want to fit the mean function (3.1). Following the outline in Section 3.1, suppose that the two terms X_1 and X_2 have sample correlation equal to zero. This means that, if $x_{ij}, i = 1, \ldots, n$, and $j = 1, 2$ are the observed values of these two terms for the n cases in the data, $\sum_{i=1}^{n}(x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) = 0$.

3.4.1. Give the formula for the slope of the regression for Y on X_1, and for Y on X_2. Give the value of the slope of the regression for X_2 on X_1.

3.4.2. Give formulas for the residuals for the regressions of Y on X_1 and for X_2 on X_1. The plot of these two sets of residuals corresponds to the added-variable plot in Figure 3.1d.

3.4.3. Compute the slope of the regression corresponding to the added-variable plot for the regression of Y on X_2 after X_1, and show that this slope is exactly the same as the slope for the simple regression of Y on X_2 ignoring X_1. Also find the intercept for the added-variable plot.

3.5. Refer to the data described in Problem 1.5, page 18. For this problem, consider the regression problem with response *BSAAM*, and three predictors as terms given by *OPBPC*, *OPRC* and *OPSLAKE*.

3.5.1. Examine the scatterplot matrix drawn for these three terms and the response. What should the correlation matrix look like (that is, which correlations are large and positive, which are large and negative, and which are small)? Compute the correlation matrix to verify your results. Get the regression summary for the regression of *BSAAM* on these three terms. Explain what the "t-values" column of your output means.

3.5.2. Obtain the overall test if the hypothesis that *BSAAM* is independent of the three terms versus the alternative that it is not independent of them, and summarize your results.

3.5.3. Obtain three analysis of variance tables fitting in the order (*OPBPC*, *OPRC* and *OPSLAKE*), then (*OPBPC*, *OPSLAKE* and *OPRC*), and finally (*OPSLAKE*, *OPRC* and *OPBPC*). Explain the resulting tables, and discuss in particular any apparent inconsistencies. Which F-tests in the Anova tables are equivalent to t-tests in the regression output?

3.5.4. Using the output from the last problem, test the hypothesis that the coefficients for both *OPRC* and *OPBPC* are both zero against the alternative that they are not both zero.