APPLIED REGRESSION MODELING

Third Edition

IAIN PARDOE
Thompson Rivers University
The Pennsylvania State University

WILEY

CHAPTER 3

MULTIPLE LINEAR REGRESSION

In the preceding chapter, we considered simple linear regression for analyzing two variables measured on a sample of observations, that is, bivariate data. In this chapter, we consider more than two variables measured on a sample of observations, that is, *multivariate data*. In particular, we will learn about *multiple linear regression*, a technique for analyzing certain types of multivariate data. This can help us to understand the association between a response variable and one or more predictor variables, to see how a change in one of the predictor variables is associated with a change in the response variable, and to estimate or predict the value of the response variable knowing the values of the predictor variables.

After reading this chapter, you should be able to

- Define a multiple linear regression model as a linear association between a quantitative response variable and one or more predictor variables.
- Express the value of an observed response variable as the sum of a deterministic linear function of the corresponding predictor values plus a random error.
- Use statistical software to apply the least squares criterion to estimate the sample multiple linear regression equation by minimizing the residual sum of squares.
- Interpret the intercept and slope parameters of an estimated multiple linear regression equation.
- Calculate and interpret the regression standard error in multiple linear regression.
- Calculate and interpret the coefficient of determination in multiple linear regression.

Applied Regression Modeling, Third Edition. Iain Pardoe.

Book companion site: www.wiley.com/go/pardoe/AppliedRegressionModeling3e

- Calculate and know how to use the adjusted coefficient of determination in multiple linear regression.
- Understand the relationship between the coefficient of determination and the correlation between the observed response values and the fitted response values from the model.
- Conduct and draw a conclusion from a global usefulness test (overall F-test) for the population regression parameters in multiple linear regression.
- Conduct and draw a conclusion from a nested model test (general linear F-test) for a subset of population regression parameters in multiple linear regression.
- Conduct and draw conclusions from individual t-tests for the regression parameters in multiple linear regression.
- Calculate and interpret confidence intervals for the regression parameters in multiple linear regression.
- Summarize and assess the four assumptions that underlie the multiple linear regression model.
- Distinguish between estimating a mean response (confidence interval) and predicting an individual observation (prediction interval) in multiple linear regression.
- Calculate and interpret a confidence interval for the population mean response at a specified set of predictor values in multiple linear regression.
- Calculate and interpret a prediction interval for an individual response value at a specified set of predictor values in multiple linear regression.

3.1 PROBABILITY MODEL FOR (X_1, X_2, . . .) AND Y

The *multiple linear regression model* is represented mathematically as an algebraic relationship between a response variable and one or more predictor variables.

- Y is the *response* variable, which can also go by the name dependent, outcome, or output variable. This variable should be quantitative, having meaningful numerical values. In Chapter 7 (www.wiley.com/go/pardoe/AppliedRegressionModeling3e), we introduce some extensions to qualitative (categorical) response variables.
- $(X_1, X_2, \ldots)$ are the *predictor* variables, which can also go by the name independent or input variables, or covariates. For the purposes of this chapter, these variables should also be quantitative. In Section 4.3, we will see how to incorporate qualitative information in predictor variables.

We have a sample of n sets of (X_1, X_2, . . . , Y) values, denoted (X_{1i}, X_{2i}, . . . , Y_i) for $i = 1, 2, \ldots, n$ (the index i keeps track of the sample observations). The simple linear regression model considered in Chapter 2 represents the special case in which there is just one predictor variable. For any particular problem, it is often clear which variable is the response variable, Y; it often "responds" in some way to a change in the values of the predictor variables, $(X_1, X_2, \ldots)$. Similarly, if our model provides a useful approximation to the association between Y and $(X_1, X_2, \ldots)$, then knowing values for the predictor variables can help us to "predict" a corresponding value for the response variable.

The variables therefore take on very different roles in a multiple linear regression analysis, so it is important, for any particular problem with multivariate data, to identify

which is the response variable and which are the predictor variables. For example, consider the problem of quantifying the association between the final exam score of a student taking a course in statistics, and the number of hours spent partying during the last week of the term and the average number of hours per week spent studying for this course. It makes sense to think of the final exam score as responding to time spent partying and time spent studying, so we should set the response variable as exam score (variable $Exam$), and the predictor variables as time spent partying (variable $Party$) and time spent studying (variable $Study$). Furthermore, this will allow us to predict $Exam$ for a student with particular values of $Party$ and $Study$.

As with simple linear regression models, multiple regression does not require there to be a *causal* link between Y and $(X_1, X_2, \ldots)$. The regression modeling described in this book can really only be used to quantify linear associations and to identify whether a change in one variable is associated with a change in another variable, not to establish whether changing one variable "causes" another to change. For further discussion of causality in the context of regression, see Gelman and Hill (2006).

Having identified the response and predictor variables and defined them carefully, we take a random sample of n observations of $(X_1, X_2, \ldots, Y)$. We then use the observed association between Y and $(X_1, X_2, \ldots)$ in this sample to make statistical inferences about the corresponding population association. (We might think of the population in the final exam score example as a conditional probability distribution of possible values of $Exam$ given $Party$ and $Study$ for the particular statistics course that we are considering.)

Before we specify the model algebraically, consider the kind of association between Y and $(X_1, X_2, \ldots)$ that we might expect. It is often useful to think about these matters before analyzing the data. Often, expert knowledge can be tapped to find expected associations between variables. For example, there may be theories in the field of application relating to why certain variables tend to have particular associations, or previous research may suggest how certain variables tend to be associated with one another. In the final exam score example, common sense tells us a lot: $Exam$ probably decreases as Party increases, but increases as Study increases (at least we would hope this is the case).

We can express the multiple linear regression model as

$$\begin{aligned} &Y\text{-value}|X\text{-values} = \text{deterministic part} + \text{random error},\\ &Y_i|(X_{1i}, X_{2i}, \ldots) = \mathrm{E}(Y|(X_{1i}, X_{2i}, \ldots)) + e_i \quad (i = 1, \ldots, n), \end{aligned}$$

where the vertical bar, "|," means "given," so that $\mathrm{E}(Y|(X_{1i}, X_{2i}, \ldots))$ means "the expected value of Y given that X_1 is equal to X_{1i}, X_2 is equal to X_{2i}, and so on." In other words, each sample Y-value is decomposed into two pieces—a deterministic part depending on the X-values, and a random error part varying from observation to observation.

As an example, suppose that for the final exam score example, the exam score is (on average) 70 minus 1.6 times the number of hours spent partying during the last week of the term plus 2.0 times the average number of hours per week spent studying for this course. In other words, for each additional hour spent partying the final exam score tends to decrease by 1.6, and for each additional hour per week spent studying the score tends to increase by 2.0. The deterministic part of the model for such a situation is thus

$$\mathrm{E}(Exam|(Party_i, Study_i)) = 70 - 1.6Party_i + 2.0Study_i \quad (i = 1, \ldots, n).$$

The whole model, including the random error part, is

$$Exam_i|(Party_i, Study_i) = 70 - 1.6Party_i + 2.0Study_i + e_i \quad (i = 1, \ldots, n),$$

although typically only the deterministic part of the model is reported in a regression analysis, usually with the index i suppressed. The random error part of this model is the difference between the value of $Exam$ actually observed with particular observed predictor values, and what we expect $Exam$ to be on average for those particular observed predictor values: that is, $e_i = Exam_i - \mathrm{E}(Exam|(Party_i, Study_i)) = Exam_i - 70 + 1.6Party_i - 2.0Study_i$. This random error represents variation in $Exam$ due to factors other than $Party$ and $Study$ that we have not measured. In this example, these might be factors related to quantitative skills, exam-taking ability, and so on.

Figure 3.1 displays a 3D-scatterplot of some hypothetical students, together with a flat surface, called the *regression plane*, going through the data. The values of $Party$ increase from 0 at the "front" of the plot to 10 on the right, while the values of $Study$ increase from 0 at the front of the plot to 10 on the left. The values of $Exam$ increase from 50 at the bottom of the plot to 90 at the top of the plot. The $Exam$-values are also represented by the shading on the regression plane according to the scale on the right.

The regression plane represents $\mathrm{E}(Exam|(Party, Study)) = 70 - 1.6Party + 2.0Study$, the deterministic part of the model. For example, a student who parties for 7.5 hours during the last week of the term but studies only 1.3 hours per week for this

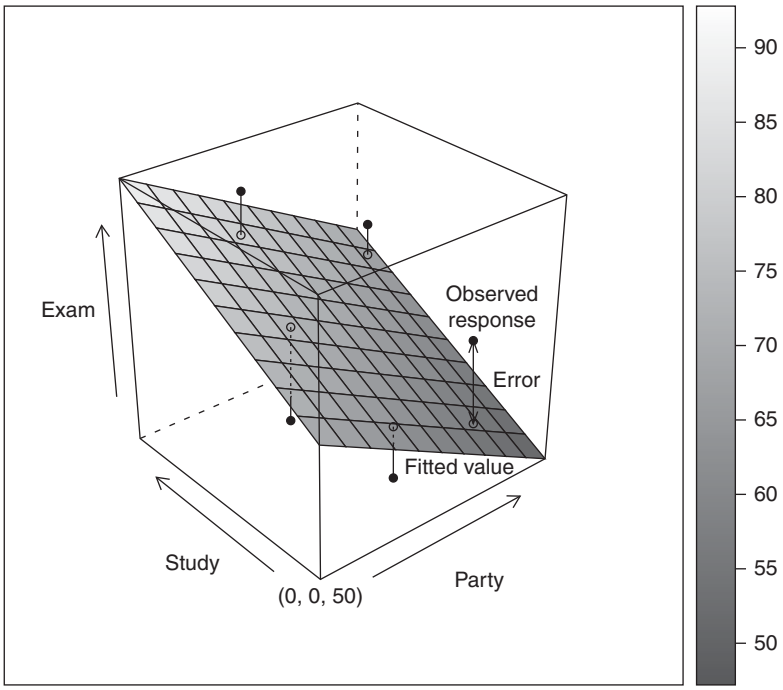

Figure 3.1 Multiple linear regression model with two predictors fitted to a hypothetical population for the final exam score example.

course has an expected final exam score of $\mathrm{E}(Exam|(Party, Study)) = 70 - 1.6(7.5) + 2.0(1.3) = 60.6$. If their observed value of $Exam$ is 65, then their random error (shown in the figure) is $e = 65 - 60.6 = 4.4$. Perhaps for that particular student, exam performance was better than expected because of strong analytical skills, for example.

To estimate the expected $Exam$-values and random errors, we need to know the numbers in the deterministic part of the model, that is, "70," "-1.6," and "2.0." Before we see how to obtain these numbers, we need to formalize the representation of a multiple linear regression model in an algebraic expression:

$$Y_i|(X_{1i}, X_{2i}, \ldots) = \mathrm{E}(Y|(X_{1i}, X_{2i}, \ldots)) + e_i \quad (i = 1, \ldots, n),$$
$$\text{where } \mathrm{E}(Y|(X_{1i}, X_{2i}, \ldots)) = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots \quad (i = 1, \ldots, n).$$

We usually write this more compactly without the i indices as

$$Y|(X_1, X_2, \ldots) = \mathrm{E}(Y|(X_1, X_2, \ldots)) + e,$$
$$\text{where } \mathrm{E}(Y|(X_1, X_2, \ldots)) = b_0 + b_1 X_1 + b_2 X_2 + \cdots .$$

As mentioned earlier, typically only the latter expression—the deterministic part of the model—is reported in a regression analysis. The regression parameter (or regression coefficient) b_0 ("b-zero") is the intercept (the value of Y when $X_1 = 0, X_2 = 0, \ldots$). The regression parameter b_1 ("b-one") is the change in Y for a 1-unit change in X_1 when all the other predictor X-variables are held constant. Similarly, the regression parameter b_2 ("b-two") is the change in Y for a 1-unit change in X_2 when all the other predictor X-variables are held constant, and so on.

For example, in the final exam score model, $b_0 = 70$ represents the expected score for a student who went to no parties during the last week of the term, but who also spent no time studying for this course. If we were to consider two students who spent the same time studying per week, but one spent one more hour than the other partying in the last week of the term, we would expect the former student to score $b_1 = -1.6$ points more on the final exam (in other words, 1.6 points less) than the latter student. On the other hand, if we were to consider two students who spent the same time partying during the last week of the term, but one spent one more hour per week than the other studying, we would expect the former student to score $b_2 = 2.0$ points more on the final exam than the latter student.

Figure 3.1 represents each of these quantities: $b_0 = 70$ is the $Exam$-value for the corner of the regression plane at the front of the graph, where $Party$ and $Study$ are both zero; $b_1 = -1.6$ is the slope of the regression plane in the "$Party$-direction" (i.e., when $Study$ is held constant); $b_2 = 2.0$ is the slope of the regression plane in the "$Study$-direction" (i.e., when $Party$ is held constant).

The expression $\mathrm{E}(Y|(X_1, X_2, \ldots)) = b_0 + b_1 X_1 + b_2 X_2 + \cdots$ is called the *regression equation* and is an algebraic representation of the regression plane. When there are more than two predictors, we cannot visualize this plane in three dimensions as we can in Figure 3.1. Nevertheless, the mathematical theory still works in higher dimensions (the regression plane becomes a regression "hyperplane") and the intuitive interpretations we have been using up until now all carry over.

The (population) multiple linear regression model equation is

$$\mathrm{E}(Y|(X_1, X_2, \dots)) = b_0 + b_1 X_1 + b_2 X_2 + \cdots ,$$

where Y is the response variable, $X_1, X_2, \dots$ are the predictor variables, and $\mathrm{E}(Y|(X_1, X_2, \dots))$ represents the expected value of Y given $(X_1, X_2, \dots)$. The regression parameters (or coefficients) are b_0, the Y-intercept, and $b_1, b_2, \dots$, the parameters that multiply the predictors.

We are now ready to find a systematic, easily calculated way to obtain estimates for $b_0, b_1, b_2, \dots$ We explore this in more detail in the next section.

3.2 LEAST SQUARES CRITERION

Figure 3.1 represents a hypothetical population association between Y and $(X_1, X_2, \dots)$. Usually, we do not get to observe all the values in the population. Rather, we just get to observe the values in the sample, in this case the n observations of $(X_1, X_2, \dots, Y)$. If we can estimate a "best fit" regression equation to go through our sample $(X_1, X_2, \dots, Y)$ values, then we can use probability theory results to make inferences about the corresponding regression equation for the population—we will see how to do that in Section 3.3. In the meantime, how can we estimate a "best fit" regression equation for our sample?

Consider tilting the regression plane in Figure 3.1 from side to side and up and down, until the plane is as close to the data points as possible. One way to do this is to make the vertical distances between the data points and the plane as small as possible: These vertical distances are the random errors in the model, that is, $e_i = Y_i - \mathrm{E}(Y|(X_{1i}, X_{2i}, \dots)) = Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i} - \cdots$. Since some random errors are positive (corresponding to data points above the plane) and some are negative (data points below the plane), a mathematical way to make the "magnitudes" of the random errors as small as possible is to square them, add them up, and then minimize the resulting *error sum of squares*.

Since we observe the sample only (not the population), we can only find an *estimated* regression equation for the sample. Recall the "hat" notation from Chapter 2, such that estimated quantities in the sample have "hats"; for example, $\hat{e}$ ("e-hat") represents an estimated random error (or residual) in the sample. We will again also drop the "$|(X_1, X_2, \dots)$" (given $X_1, X_2, \dots$) notation so that from this point on this concept will be implicit in all expressions relating Y and $(X_1, X_2, \dots)$.

In particular, we write the estimated multiple linear regression model as

$$Y_i = \hat{Y}_i + \hat{e}_i \quad (i = 1, \dots, n),$$
$$\text{where } \hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} + \cdots \quad (i = 1, \dots, n).$$

Again, we usually write this more compactly without the i indices as

$$Y = \hat{Y} + \hat{e},$$
$$\text{where } \hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \cdots .$$

$\hat{Y}$ ("Y-hat") represents an estimated expected value of Y, also known as a *fitted or predicted value* of Y. The expression $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \cdots$ is called the estimated

(or sample) regression equation and is an algebraic representation of the estimated (or sample) regression hyperplane.

The estimated (sample) multiple linear regression equation is

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \cdots ,$$

where $\hat{Y}$ is the fitted (or predicted) value of Y (the response variable) and X_1, X_2, ... are the predictor variables. The estimated regression parameters (or coefficients) are $\hat{b}_0$, the estimated Y-intercept, and $\hat{b}_1$, $\hat{b}_2$, ..., the estimated parameters that multiply the predictors.

To find values for the point estimates, $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$, and so on, we minimize the residual sum of squares (RSS):

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} \hat{e}_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i} - \cdots)^2. \end{aligned}$$

We have already seen this method of *least squares* in Chapter 2. Mathematically, to minimize a function such as this, we can set the partial derivatives with respect to $b_0, b_1, b_2, \ldots$ equal to zero and then solve for $b_0, b_1, b_2, \ldots$, which leads to relatively simple expressions for the regression parameter estimates, $\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots$. Since we will be using statistical software to do the calculations, these expressions are provided for interest only at the end of this section. Again, the intuition behind the expressions is more useful than the formulas themselves.

Recall the home prices–floor size example from Chapter 2, in which we analyzed the association between sale prices and floor sizes for single-family homes in Eugene, Oregon. We continue this example here and extend the analysis to see if an additional predictor, lot size (property land area), helps to further explain variation in sale prices. Again, to keep things simple to aid understanding of the concepts in this chapter, the data for this example consist of a subset of a larger data file containing more extensive information on 76 homes, which is analyzed as a case study in Section 6.1 (www.wiley.com/go/pardoe/AppliedRegressionModeling3e).

The data for the example in this chapter consist of $n = 6$ homes with variables $Price =$ sale price in thousands of dollars, $Floor =$ floor size in thousands of square feet, and $Lot =$ lot size category and are available in the **HOMES3** data file:

$Price =$ Sale price in thousands of dollars	252.5	259.9	259.9	269.9	270.0	285.0
$Floor =$ Floor size in thousands of square feet	1.888	1.683	1.708	1.922	2.053	2.269
$Lot =$ Lot size category (explanation follows)	2	5	4	4	3	3

This dataset contains an additional home to the dataset from Chapter 2 (**HOMES2**), which had a strong linear association between sale price and floor size. This additional home has values of $Price = 252.5$ and $Floor = 1.888$, which do not fit the estimated simple

linear regression model from Chapter 2 nearly as well as the other five homes (this is evident from Figure 3.2 presented later). However, we shall see that a multiple linear regression model between Price and (Floor, Lot) together fits the data for all six homes very well.

The variable *Lot* in this dataset requires some further explanation. It is reasonable to assume that homes built on properties with a large amount of land area command higher sale prices than homes with less land, all else being equal. However, it is also reasonable to suppose that an increase in land area of 2,000 square feet from 4,000 to 6,000 would make a larger difference (to sale price) than going from 24,000 to 26,000. Thus, realtors have constructed lot size "categories," which in their experience correspond to approximately equal-sized increases in sale price. The categories used in this dataset are

Lot size	0–3k	3–5k	5–7k	7–10k	10–15k	15–20k	20k–1ac	1–3ac	3–5ac	5–10ac	10–20ac
Category	1	2	3	4	5	6	7	8	9	10	11

Lot sizes ending in "k" represent thousands of square feet, while "ac" stands for acres (there are 43,560 square feet in an acre).

It makes sense to look at scatterplots of the data before we start to fit any models. With simple linear regression, a scatterplot with the response, Y, on the vertical axis and the predictor, X, on the horizontal axis provides all the information necessary to identify an association between Y and X. However, with a response variable, Y, but more than one predictor variable, $X_1, X_2, \ldots$, we can use scatterplots only to identify bivariate associations between any two variables (e.g., Y and X_1, Y and X_2, or even X_1 and X_2). We cannot identify a multivariate association between Y and $(X_1, X_2, \ldots)$ just from bivariate scatterplots.

Nevertheless, we can use these bivariate scatterplots to see whether the data have any strange patterns or odd-looking values that might warrant further investigation. For example, data entry errors are often easy to spot in bivariate scatterplots when one data point appears isolated a long way from all the other data points. A useful method for looking at all possible bivariate scatterplots in a multivariate data setting is to construct a *scatterplot matrix*, such as the scatterplot matrix for the home prices dataset in Figure 3.2 (see computer help #16 in the software information files available from the book website). Here, the scatterplot of $Price$ versus $Floor$ is shown in the top middle part of the matrix, $Price$ versus Lot is at top right, and $Floor$ versus Lot is just below. Reflections of these three plots are below the diagonal of this matrix. Scatterplot matrices can be challenging to decipher initially. One key to understanding a particular graph in a scatterplot matrix is to look left or right to see the label for the variable plotted on the vertical axis and to look up or down to see the label for the variable plotted on the horizontal axis. In this respect, scatterplot matrices are similar to scatterplots, where the vertical axis label is typically to the left of the graph and the horizontal axis label is typically below the graph.

We can see an increasing pattern between $Price$ and $Floor$, an ambiguous pattern between $Price$ and Lot, and a decreasing pattern between $Floor$ and Lot in the plots, but such patterns *cannot* tell us whether the multiple linear regression model that we consider later can provide a useful mathematical approximation to these bivariate associations. In Section 3.3.2, we shall see examples of how such thinking can be misleading. The scatterplot matrix is useful primarily for identifying any strange patterns or odd-looking values that might warrant further investigation *before* we start modeling. In this case, there

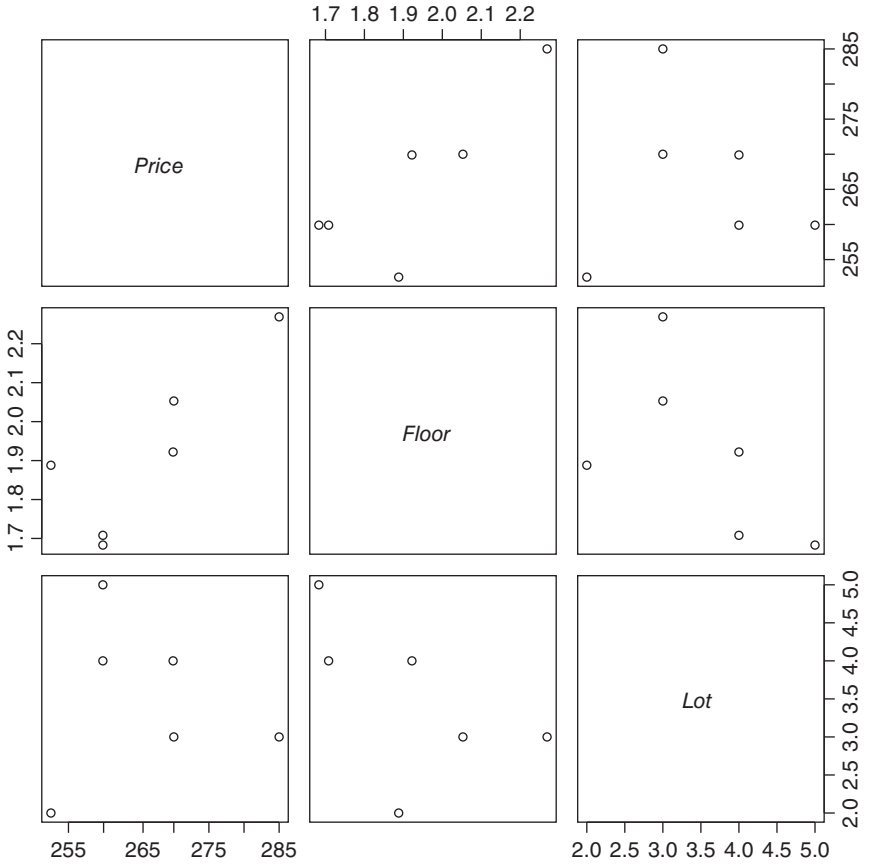

Figure 3.2 Scatterplot matrix for the home prices example.

are no data points that appear isolated a long way from all the other data points, so it seems reasonable to proceed.

We propose the following multiple linear regression model:

$$\begin{aligned} Price &= \mathrm{E}(Price) + e \\ &= b_0 + b_1 Floor + b_2 Lot + e, \end{aligned}$$

with $Price$, $Floor$, and Lot defined as above. The random errors, e, represent variation in $Price$ due to factors other than $Floor$ and Lot that we have not measured. In this example, these might be factors related to numbers of bedrooms/bathrooms, property age, garage size, or nearby schools. We estimate the deterministic part of the model, $\mathrm{E}(Price)$, as

$$\widehat{Price} = \hat{b}_0 + \hat{b}_1 Floor + \hat{b}_2 Lot,$$

by using statistical software to find the values of b_0, b_1, and b_2 that minimize $\mathrm{RSS} = \sum_{i=1}^{n} (Price_i - \hat{b}_0 - \hat{b}_1 Floor_i - \hat{b}_2 Lot_i)^2$.

Here is part of the output produced by statistical software when a multiple linear regression model is fit to this home prices example (see computer help #31):

Parameters[a]

Model		Estimate	Standard error	t-Statistic	Pr(> |t|)
1	(Intercept)	122.357	14.786	8.275	0.004
	Floor	61.976	6.113	10.139	0.002
	Lot	7.091	1.281	5.535	0.012

[a]Response variable: *Price*.

The regression parameter estimates $\hat{b}_0$, $\hat{b}_1$, and $\hat{b}_2$ are in the column headed "Estimate," with $\hat{b}_0$ in the row labeled "(Intercept)," $\hat{b}_1$ in the row labeled with the name of the corresponding predictor, "*Floor*" in this case, and $\hat{b}_2$ in the row labeled "*Lot*." We discuss the other numbers in the output later in the book.

Having obtained these estimates, how can we best interpret these numbers? Overall, we have found that *if* we were to model *Price*, *Floor*, and *Lot* for a housing market population represented by this sample with the multiple linear regression model, $\mathrm{E}(Price) = b_0 + b_1 Floor + b_2 Lot$, then the best-fitting model is $\widehat{Price} = 122.36 + 61.98 Floor + 7.09 Lot$. This association holds only over the range of the sample predictor values, that is, *Floor* from 1,683 to 2,269 square feet and *Lot* from lot size category 2 to 5. In the next section, we discuss whether this multiple linear regression model is appropriate for this example.

What of the particular values for $\hat{b}_0$, $\hat{b}_1$, and $\hat{b}_2$? Since $\hat{b}_0$ is the estimated *Price*-value when $Floor = 0$ and $Lot = 0$, it makes practical sense to interpret this estimate only if predictor value of zero make sense for the particular situation being considered, *and* if we have some data *close* to $Floor = 0$ and $Lot = 0$. In this example, it does not make practical sense to estimate sale price when floor size and lot size category are both zero. Also, we do not have any sample data particularly close to $Floor = 0$ and $Lot = 0$. So, in this case, it is *not* appropriate to interpret $\hat{b}_0 = 122.36$ in practical terms.

The estimate $\hat{b}_1 = 61.98$ represents the change in *Price* for a 1-unit increase in *Floor* when all the other predictor variables are held constant. In particular, we can say that we expect sale price to increase by \$61,980 for each 1,000-square foot increase in floor size when lot size is held constant. A more meaningful interpretation in this example is that we expect sale price to increase by \$6,198 for each 100-square foot increase in floor size when lot size is held constant.

Similarly, the estimate $\hat{b}_2 = 7.09$ represents the change in *Price* for a 1-unit increase in *Lot*, when all the other predictor variables are held constant. In particular, we can say that we expect sale price to increase by \$7,090 for each 1-category increase in lot size when floor size is held constant.

It is important to state the units of measurement for *Price*, *Floor*, and *Lot* when making these interpretations. Again, these interpretations are valid only over the range of the sample predictor values, that is, *Floor* from 1,683 to 2,269 square feet and *Lot* from lot size category 2 to 5.

Interpretations of the estimated regression parameters in the fitted multiple linear regression model, $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \cdots$:

- The estimated intercept, $\hat{b}_0$, is the expected Y-value when $X_1 = X_2 = \cdots = 0$ (if $X_1 = X_2 = \cdots = 0$ make sense for the particular situation being considered and if we have some data with all the predictor values close to zero).

- The regression parameter estimate, $\hat{b}_1$, represents the expected change in Y for each 1-unit increase in X_1, when all the other predictor variables are held constant (sometimes described as, "adjusting for all the other predictor variables").
- The regression parameter estimate, $\hat{b}_2$, represents the expected change in Y for each 1-unit increase in X_2, when all the other predictor variables are held constant (sometimes described as, "adjusting for all the other predictor variables").
- Similar interpretations apply to other regression parameter estimates in the model, $\hat{b}_3, \hat{b}_4, \ldots$

The estimates $\hat{b}_1 = 61.98$ and $\hat{b}_2 = 7.09$ can be combined to find changes in sale price for different changes in floor size and lot size together. For example, since $Price$ is measured in thousands of dollars, we would expect that a 200-square foot increase in floor size coupled with an increase of one lot size category would lead to an increase in sale price of $\$1{,}000 \times (0.2 \times 61.98 + 1 \times 7.09) = \$19{,}500$.

Remember that $\hat{b}_1 = 61.98$ and $\hat{b}_2 = 7.09$ cannot be given *causal* interpretations. The regression modeling described in this book can really only be used to quantify linear associations and to identify whether a change in one variable is associated with a change in another variable, not to establish whether changing one variable "causes" another to change.

Some statistical software will calculate *standardized* regression parameter (or coefficient) estimates in addition to the (unstandardized) estimates considered here. These are the parameter estimates that would result if the response variable and predictor variables were first standardized to have mean 0 and standard deviation 1. Their interpretations are then in terms of "standard deviations." For example, a standardized regression parameter estimate for a particular predictor variable would represent the standard deviation change in the response variable for a 1-standard deviation increase in that predictor variable when all the other predictor variables are held constant. In some circumstances, use of standardized regression parameter estimates can enable comparisons of the relative contributions of each predictor in explaining the overall variation in the response variable. However, this use of standardized estimates does not carry over to models with transformations, interactions, and indicator variables, which we consider in Chapter 4. In Section 5.5, we introduce a graphical method that can be used to compare the relative contributions of predictors for such models.

In the next section, we cover methods for evaluating whether a multiple linear regression model is appropriate for a particular multivariate dataset.

Optional—formula for regression parameter estimates. Suppose that our multiple linear regression model has k predictor variables. Consider all the observations for each predictor variable as column vectors (each length n, the sample size). Then put all these column vectors side by side into a matrix, $\mathbf{X}$ (called the *model matrix*). Note that this matrix has an additional column of ones at the far left representing the intercept term in the model. Thus, $\mathbf{X}$ has n rows and $k + 1$ columns. Also let $\mathbf{Y}$ be a column vector representing all the observations of the response variable. Then calculate the following vector, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, where $\mathbf{X}^{\mathrm{T}}$ is the transpose of $\mathbf{X}$ and $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ is the inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. The $k + 1$ entries of this vector are the regression parameter estimates, $\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k$. In practice, statistical software uses a variation of this method, which centers the variables about their means first.

Further details and examples are provided in Appendix E (www.wiley.com/go/pardoe/AppliedRegressionModeling3e), which gives an informal overview of matrices in the context of multiple linear regression.

3.3 MODEL EVALUATION

Before making the kinds of interpretations discussed at the end of the preceding section, we need to be reasonably confident that our multiple linear regression model provides a useful approximation to the actual association between Y and $(X_1, X_2, \ldots)$. All we have to base that decision on are the results of the fit of the model to the sample data. It is important to be able to present unambiguous numerical justification for whether or not the model provides a good fit.

We used three standard methods to evaluate numerically how well a simple linear regression model fits some sample data. Two of those methods—the regression standard error, s, and the coefficient of determination, R^2—carry over essentially unchanged. The last method, which focused on the slope parameter, b_1, becomes a little more complicated since we now have a series of regression parameters, $b_1, b_2, \ldots$ It turns out that we can tackle this issue globally (looking at all the regression parameters, $b_1, b_2, \ldots$, simultaneously), in subsets (looking at two or more of the regression parameters at a time), or individually (considering just one of the regression parameters at a time). We consider each of these methods—s, R^2, regression parameters globally, regression parameters in subsets, and regression parameters individually—in turn.

3.3.1 Regression standard error

Suppose that our multiple linear regression model has k predictor X-variables. For example, $k = 2$ for the home prices dataset above with predictors $Floor =$ floor size and $Lot =$ lot size. Recall the least squares method used for estimating the regression parameters, $b_0, b_1, b_2, \ldots, b_k$. The estimates $\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k$ are the values that minimize the residual sum of squares,

$$\mathrm{RSS} = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i} - \cdots - \hat{b}_k X_{ki})^2.$$

We can use this minimum value of RSS to say how far (on average) the actual observed values, Y_i, are from the model-based fitted values, $\hat{Y}_i$, by calculating the regression standard error, s:

$$s = \sqrt{\frac{\mathrm{RSS}}{n-k-1}},$$

which is an estimate of the standard deviation of the random errors in the multiple linear regression model. The quantity $\mathrm{RSS}/(n-k-1)$ is called the mean square error, which is often abbreviated MSE. The $n-k-1$ denominator in this expression generalizes from the simple linear regression case when k was 1 so that $n-k-1 = n-2$. The unit of measurement for s is the same as the unit of measurement for Y. For example, the value of the regression standard error for the home prices dataset is $s = 2.48$ (see the following statistical software output). In other words, loosely speaking, the actual observed $Price$-values are, on average, a distance of \$2,480 from the model-based fitted values, $\widehat{Price}$.

Calculation and interpretation of the regression standard error, s, in multiple linear regression:

$$s = \sqrt{\frac{\text{RSS}}{n-k-1}},$$

where RSS is the residual sum of squares, n is the sample size, and k is the number of predictor variables. The value of s estimates the standard deviation of the random errors and tells us approximately how far, on average, the observed response values, Y, are from the predicted values, $\hat{Y}$.

Here is the output produced by statistical software that displays the value of s for the multiple linear regression model fit to the home prices example (see computer help #31 in the software information files available from the book website):

Model summary

Model	Sample size	Multiple R^2	Adjusted R^2	Regression standard error
1[a]	6	0.9717	0.9528	2.475

[a] Predictors: (Intercept), *Floor*, *Lot*.

The value of the regression standard error, s, is in the column headed "Regression standard error." This can go by a different name depending on the statistical software used. For example, in SPSS it is called the "standard error of the estimate," while in SAS it is called the "root mean squared error," and in R it is called the "residual standard error." We discuss the other numbers in the output later in the book.

A multiple linear regression model is more effective the closer the observed Y-values are to the fitted $\hat{Y}$-values. Thus, for a particular dataset, we would prefer a *small* value of s to a large one. How small is small depends on the measurement scale of Y (since Y and s have the same unit of measurement). Thus, s is most useful for comparing one model to another *for the same response variable* Y. For example, suppose that we have alternative possible predictors to use instead of floor size and lot size, say, numbers of bedrooms and property age. We might fit a multiple linear regression model with the response variable, $Price$ = sale price in thousands of dollars, and predictor variables, number of bedrooms and property age, and find that the value of the regression standard error for this model is $s = 9.33$. Thus, the observed $Price$-values are further away (on average) from the fitted $\widehat{Price}$-values for this model than they were for the model that used floor size and lot size (which had $s = 2.48$). In other words, the random errors tend to be larger, and consequently, the deterministic part of the model that uses numbers of bedrooms and property age must be less accurate (on average).

Just as with simple linear regression, another way of interpreting s is to multiply its value by 2 to provide an approximate level of "prediction uncertainty." In particular, approximately 95% of the observed Y-values lie within plus or minus $2s$ of their fitted $\hat{Y}$-values (recall from Section 2.3.1 that this is an approximation derived from the central limit theorem). In other words, if we use a multiple linear regression model to predict an unobserved Y-value from potential X-values, we can expect to be accurate to within

approximately $\pm 2s$ (at a 95% confidence level). Returning to the home prices example, $2s = 4.95$, so if we use a multiple linear regression model to predict an unobserved sale price for an individual home with particular floor size and lot size values, we can expect to be accurate to within approximately $\pm \$4{,}950$ (at a 95% confidence level).

3.3.2 Coefficient of determination—R^2

Another way to evaluate the fit of a multiple linear regression model is to contrast the model with a situation in which the predictor X-variables are not available. If there are no predictors, then all we would have is a list of Y-values; that is, we would be in the situation that we found ourselves in for Chapter 1. Then, when predicting an individual Y-value, we found that the sample mean, m_Y, is the best point estimate in terms of having no bias and relatively small sampling variability. One way to summarize how well this univariate model fits the sample data is to compute the sum of squares of the differences between the Y_i-values and this point estimate m_Y; this is known as the *total sum of squares* (TSS):

$$\mathrm{TSS} = \sum_{i=1}^{n} (Y_i - m_Y)^2.$$

This is similar to the residual sum of squares (RSS) in Section 3.2, but it measures how far off our observed Y-values are from predictions, m_Y, which *ignore* the predictor X-variables.

For the multiple linear regression model, the predictor X-variables should allow us to predict an individual Y-value more accurately. To see how much more accurately the predictors help us to predict an individual Y-value, we can see how much we can reduce the random errors between the observed Y-values and our new predictions, the fitted $\hat{Y}$-values. Recall from Section 3.2 that RSS for the multiple linear regression model is

$$\mathrm{RSS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Just as in simple linear regression, to quantify how much smaller RSS is than TSS, we can calculate the proportional reduction from TSS to RSS, known as the coefficient of determination or R^2 ("R-squared"):

$$\mathrm{R}^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - m_Y)^2}.$$

To fully understand what R^2 measures, think of multiple linear regression as a method for using predictor X-variables to help explain the variation in a response variable, Y. The "total variation" in Y is measured by TSS (which ignores the X-variables) and considers how far the Y-values are from their sample mean, m_Y. The multiple linear regression model predicts Y through the estimated regression equation, $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \cdots$. Any differences between observed Y-values and fitted $\hat{Y}$-values remains "unexplained" and is measured by RSS. The quantity $\mathrm{TSS} - \mathrm{RSS}$, known as the *regression sum of squares*, therefore represents the variation in Y-values (about their sample mean) that has been "explained" by the multiple linear regression model. In other words, R^2 is the proportion

of variation in Y (about its mean) explained by a multiple linear regression association between Y and $(X_1, X_2, \ldots)$.

Calculation and interpretation of the coefficient of determination, R^2, in multiple linear regression:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS},$$

where TSS is the total sum of squares and RSS is the residual sum of squares. R^2 is the proportion of variation in Y (about its mean) "explained" by a linear association between Y and $(X_1, X_2, \ldots)$.

Again, the word "explained" is in quotes here to warn against assuming a *causal* association between the predictor variable and the response variable. To avoid any possible misunderstanding, some analysts prefer to say, "accounted for" instead of "explained" when interpreting R^2.

In practice, we can obtain the value for R^2 directly from statistical software for any particular multiple linear regression model. For example, here is the output that displays the value of R^2 for the home prices dataset (see computer help #31 in the software information files available from the book website):

Model summary

Model	Sample size	Multiple R^2	Adjusted R^2	Regression standard error
1[a]	6	0.9717	0.9528	2.475

[a] Predictors: (Intercept), *Floor*, *Lot*.

The value of the coefficient of determination, $R^2 = 0.9717$, is in the column headed "Multiple R^2." (We have already discussed the "Regression standard error" in Section 3.3.1 and discuss the other numbers in the output later in the book.) To interpret this number, it is standard practice to report the value as a percentage. In this case, we would conclude that 97.2% of the variation in sale price (about its mean) can be explained by a multiple linear regression association between sale price and (floor size, lot size).

Since TSS and RSS are both nonnegative (i.e., greater than or equal to 0), and TSS is always greater than or equal to RSS, the value of R^2 must always be between 0 and 1. Higher values of R^2 correspond to better-fitting multiple linear regression models. However, there is no "reference value" such that R^2 greater than this value suggests a "good model" and R^2 less than this suggests a "poor model." This type of judgment is very context-dependent, and an apparently "low" value of $R^2 = 0.3$ may actually correspond to a useful model in a setting where the response variable is particularly hard to predict with the available predictors. Thus, R^2 by itself cannot tell us whether a particular model is good, but it can tell us something useful (namely, how much variation in Y can be explained).

***Adjusted* R^2.** This chapter is concerned with the mechanics of the multiple linear regression model and how the various concepts from simple linear regression carry over

to this new multivariate setting. For the simple linear regression model of Chapter 2, we discussed just one possible way to model the association between a response variable, Y, and a single predictor, X: using a straight line. By contrast, when there are two or more potential predictors, we have more possibilities for how we model the association between Y and $(X_1, X_2, \ldots)$. We could include just one of the potential predictors (i.e., use a simple linear regression model), or all of the potential predictors, or something in between (i.e., a subset of the potential predictors). In addition, we could "transform" some or all of the predictors, or create "interactions" between predictors. We discuss these topics, part of *model building*, in detail in Chapters 4 and 5.

However, some model building topics arise naturally as we explore how to adapt concepts from simple linear regression to the multiple linear regression setting—R^2 is one such concept. We have seen that R^2 tells us the proportion of variation in Y (about its mean) explained by a multiple linear regression association between Y and $(X_1, X_2, \ldots)$. Since higher values of R^2 are better than lower values of R^2, all other things being equal, we might think that R^2 could be used as a criterion for guiding model building (i.e., out of a collection of possible models, the one with the highest value of R^2 is the "best"). Unfortunately, R^2 *cannot* be used in this way to guide model building because of a particular property that it has. This property dictates that if one model—model A, say—has a value of R^2 equal to $\mathrm{R}^2_{\mathrm{A}}$, then R^2 for a second model with the same predictors as model A *plus* one or more additional predictors will be greater than (or equal to) $\mathrm{R}^2_{\mathrm{A}}$. In other words, as we add predictors to a model, R^2 either increases or stays the same.

While we can justify this result mathematically, a geometrical argument is perhaps more intuitive. Consider a bivariate scatterplot of a response variable, Y, versus a predictor variable, X_1, with a regression line going through the points. Call the model represented by this line "model A" so that the line minimizes the residual sum of squares, $\mathrm{RSS}_{\mathrm{A}}$. If we add a second predictor variable, X_2, to this model, we can think of adding a third axis to this scatterplot (much like in Figure 3.1) and moving the data points out along this axis according to their values for X_2. The regression model, "model B" say, is now represented by a plane rather than a line, with the plane minimizing the residual sum of squares, $\mathrm{RSS}_{\mathrm{B}}$. Whereas for model A, we can only tilt the regression line in two dimensions (represented by the Y-axis and the X_1-axis), for model B we can tilt the regression plane in three dimensions (represented by the Y-axis, the X_1-axis, and the X_2-axis). So, we can always make $\mathrm{RSS}_{\mathrm{B}}$ less than (or at least equal to) $\mathrm{RSS}_{\mathrm{A}}$. This in turn makes $\mathrm{R}^2_{\mathrm{B}}$ always greater than (or equal to) R^2_{A}. This result holds in higher dimensions also, for any model B with the same predictors as model A plus one or more additional predictors.

Consider a collection of *nested models*, that is, a sequence of models with the next model in the sequence containing the same predictor variables as the preceding model in the sequence plus one or more additional predictor variables. Then *if* we were to use R^2 as a criterion for assessing the "best" model, R^2 would pick the last model in the sequence, that is, the one with the most predictor variables. This model certainly does the best job of getting closest to the *sample* data points (since it has the smallest RSS of all the models), but that does not necessarily mean that it does the best job of modeling the *population*. Often, the model with all the potential predictor variables will "overfit" the sample data so that it reacts to every slight twist and turn in the sample associations between the variables. A simpler model with fewer predictor variables will then be preferable *if* it can capture the major, important population associations between the variables without getting distracted by minor, unimportant sample associations.

Since R^2 is inappropriate for finding such a model (one that captures the major, important population associations), we need an alternative criterion, which penalizes models that contain too many unimportant predictor variables. The *adjusted* R^2 measure does just this:

$$\text{adjusted } R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2).$$

As the number of predictors (k) in the model increases, R^2 increases (which also causes adjusted R^2 to increase), but the factor "$-(n-1)/(n-k-1)$" causes adjusted R^2 to decrease. This trade-off penalizes models that contain too many unimportant predictor variables and allows us to use adjusted R^2 to help find models that do a reasonable job of finding the population association between Y and $(X_1, X_2, \ldots)$ without overcomplicating things.

Calculation and use of the adjusted coefficient of determination, *adjusted* R^2, in multiple linear regression:

$$\text{adjusted } R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2),$$

where R^2 is the coefficient of determination, n is the sample size, and k is the number of predictor variables. *Adjusted* R^2 is used as a model selection criterion of how well a model is likely to generalize to the population.

In practice, we can obtain the value for adjusted R^2 directly from statistical software for any particular multiple linear regression model. For example, here is the output that displays adjusted R^2 for the home prices model with predictors, *Floor* and *Lot* (see computer help #31 in the software information files available from the book website):

Model summary

Model	Sample size	Multiple R^2	Adjusted R^2	Regression standard error
1[a]	6	0.9717	0.9528	2.475

[a] Predictors: (Intercept), *Floor*, *Lot*.

The value of adjusted $R^2 = 0.9528$ is in the column headed "Adjusted R^2." Contrast the output if we just use *Floor* as a single predictor:

Model summary

Model	Sample size	Multiple R^2	Adjusted R^2	Regression standard error
2[a]	6	0.6823	0.6029	7.178

[a] Predictors: (Intercept), *Floor*.

In this case, since adjusted R^2 for this single-predictor model is 0.6029 and adjusted R^2 for the two-predictor model is 0.9528, this suggests that the two-predictor model is better

than the single-predictor model (at least according to this criterion). In other words, there is no indication that adding the variable Lot to the model causes overfitting.

As a further example of adjusted R^2 for a multiple linear regression analysis, consider the following example, adapted from McClave et al. (2018) and based on accounting methods discussed in Datar and Rajan (2018). The **SHIPDEPT** data file contains 20 weeks of a firm's accounting and production records on cost information about the firm's shipping department—see Table 3.1.

Suppose that we propose the following multiple linear regression model:

$$E(Lab) = b_0 + b_1 Tws + b_2 Pst + b_3 Asw + b_4 Num.$$

Here is the output produced by statistical software that displays the relevant results for this model (see computer help #31):

Model summary

Model	Sample size	Multiple R^2	Adjusted R^2	Regression standard error
1[a]	20	0.8196	0.7715	9.103

[a]Predictors: (Intercept), Tws, Pst, Asw, Num.

Contrast these results with those for the following two-predictor model:

$$E(Lab) = b_0 + b_1 Tws + b_3 Asw.$$

Table 3.1 Shipping data with response variable Lab = weekly labor hours and four potential predictor variables: Tws = total weight shipped in thousands of pounds, Pst = proportion shipped by truck, Asw = average shipment weight in pounds, and Num = week number.

Lab	Tws	Pst	Asw	Num
100	5.1	0.90	20	1
85	3.8	0.99	22	2
108	5.3	0.58	19	3
116	7.5	0.16	15	4
92	4.5	0.54	20	5
63	3.3	0.42	26	6
79	5.3	0.12	25	7
101	5.9	0.32	21	8
88	4.0	0.56	24	9
71	4.2	0.64	29	10
122	6.8	0.78	10	11
85	3.9	0.90	30	12
50	2.8	0.74	28	13
114	7.5	0.89	14	14
104	4.5	0.90	21	15
111	6.0	0.40	20	16
115	8.1	0.55	16	17
100	7.0	0.64	19	18
82	4.0	0.35	23	19
85	4.8	0.58	25	20

Model summary

Model	Sample size	Multiple R^2	Adjusted R^2	Regression standard error
2[a]	20	0.8082	0.7857	8.815

[a]Predictors: (Intercept), Tws, Asw.

Whereas R^2 decreases from 0.8196 to 0.8082 (from the four-predictor model to the two-predictor model), adjusted R^2 increases from 0.7715 to 0.7857. In other words, although the four-dimensional regression hyperplane can get a little closer to the sample data points, it appears to do so at the expense of overfitting by including apparently redundant predictor variables. The adjusted R^2 criterion suggests that the simpler two-predictor model does a better job than the four-predictor model of finding the population association between Lab and (Tws, Pst, Asw, Num).

Since we have considered all four predictor variables in the analysis, we should mention them all in reporting the results. We found that the population association between Lab and (Tws, Pst, Asw, Num) is summarized well by a multiple linear regression model with just Tws and Asw. But, because we started by including Pst and Num in our analysis, we have essentially averaged over the sample values of Pst and Num to reach this conclusion. So, it is the population association between Lab and (Tws, Pst, Asw, Num) that we have modeled, even though Pst and Num do not appear in our two-predictor model regression equation. If we had ignored Pst and Num all along, we could say that we are modeling the population association between Lab and (Tws, Asw), but that is not what we have done here.

Although R^2 and adjusted R^2 are related, they measure different things and both have their uses when fitting multiple linear regression models.

- R^2 has a clear interpretation since it represents the proportion of variation in Y (about its mean) explained by a multiple linear regression association between Y and $(X_1, X_2, \ldots)$.
- Adjusted R^2 is useful for identifying which models in a sequence of nested models provide a good fit to sample data without overfitting. We can use adjusted R^2 to guide model building since it tends to decrease in value when extra, unimportant predictors have been added to the model. It is not a foolproof measure, however, and should be used with caution, preferably in conjunction with other model building criteria (see Section 5.4).

Another such criterion is the regression standard error, s, which in the shipping example is 9.103 for the four-predictor model but only 8.815 for the two-predictor model. This finding reinforces the conclusion that the two-predictor model may be preferable to the four-predictor model for this dataset. Since $s = \sqrt{\mathrm{RSS}/(n-k-1)}$ increases as k increases (all else being equal), we can see that this criterion also penalizes models that contain too many unimportant predictor variables.

Multiple correlation. In simple linear regression, the concepts of R^2 and correlation are distinct but related. However, whereas the concept of R^2 carries over directly from simple linear regression to multiple linear regression, the concept of correlation does not.

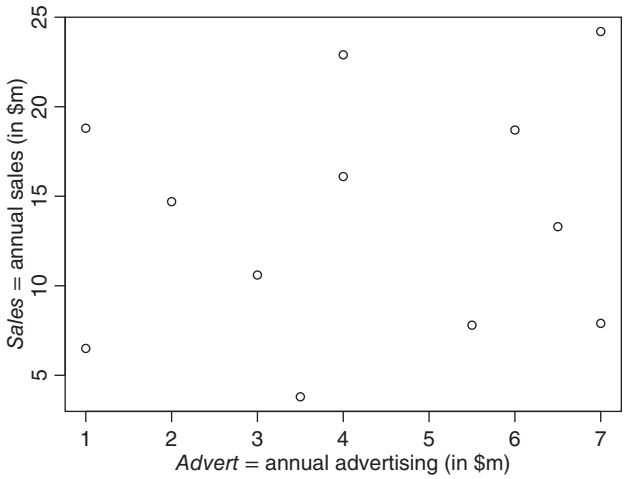

Figure 3.3 Scatterplot of simulated data with low correlation between $Sales$ and $Advert$.

In fact, intuition about correlation can be seriously misleading when it comes to multiple linear regression.

Consider the simulated data represented by the scatterplot of $Sales$ versus $Advert$ in Figure 3.3, where $Sales$ represents annual sales in millions of dollars for a small retail business and X_1 represents total annual spending on advertising in millions of dollars (**SALES2** data file). The correlation between $Sales$ and $Advert$ here is very low (in fact, it is 0.165). This means that $Advert$ is unlikely to be a useful predictor of $Sales$ in a *simple* linear regression model. Nevertheless, it is possible for $Advert$ to be a useful predictor of $Sales$ in a *multiple* linear regression model (if there are other predictors that have a particular association with $Sales$ and $Advert$). For example, there is a second predictor for the dataset represented in Figure 3.3 that produces just such an outcome—in Section 3.3.5 we see exactly how this happens.

This simulated example demonstrates that low correlation between a response variable and a predictor variable does *not* imply that this predictor cannot be useful in a multiple linear regression model. Unfortunately, intuition about correlation can break down in the other direction also: High correlation between a response variable and a predictor variable does *not* imply that this predictor will be useful in a multiple linear regression model. For example, consider a second simulated dataset represented by the scatterplot of $Sales$ versus $Trad$ in Figure 3.4, where $Sales$ represents annual sales in millions of dollars for a small high-tech business and $Trad$ represents annual spending on traditional advertising (TV, print media, etc.) in millions of dollars (**SALES3** data file).

The correlation between $Sales$ and $Trad$ here is very high (in fact, it is 0.986). This means that $Trad$ is likely to be a useful predictor of $Sales$ in a *simple* linear regression model. Nevertheless, it is possible for $Trad$ to apparently be a poor predictor of $Sales$ in a *multiple* linear regression model (if there are other predictors that have a particular association with $Sales$ and $Trad$). For example, there is a second predictor for the dataset represented in Figure 3.4 that produces just such an outcome—in Section 3.3.5 we see exactly how this happens.

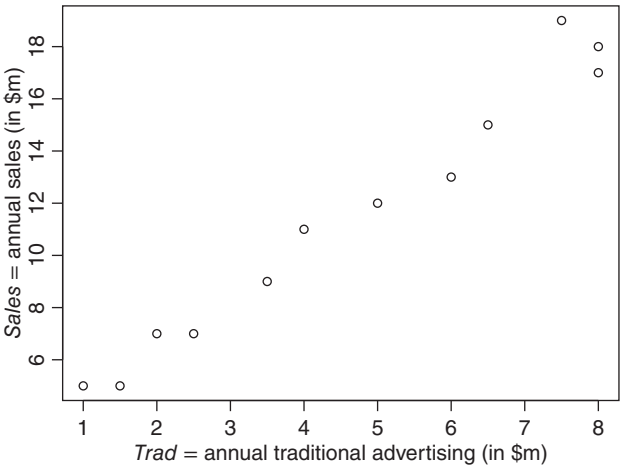

Figure 3.4 Scatterplot of simulated data with high correlation between $Sales$ and $Trad$.

The only correlation coefficient that should not cause confusion when it comes to multiple linear regression is *multiple R*, or the multiple correlation coefficient. It is defined as the correlation between the observed Y-values and the fitted $\hat{Y}$-values from the model. It is related to R^2 in the following way:

$$\text{multiple R} = +\sqrt{\mathrm{R}^2}.$$

If R^2 is high (close to 1), multiple R is also high (close to 1), and there is a strong positive linear association between the observed Y-values and the fitted $\hat{Y}$-values.

In practice, we can find the value for multiple R for any particular multiple linear regression model by calculating the positive square root of R^2. For example, recall that $\mathrm{R}^2 = 0.9717$ for the home prices model with predictors, $Floor$ and Lot. Since the positive square root of 0.9717 is 0.986, multiple R—or the correlation between the observed values of $Price$ and the fitted values of $\widehat{Price}$ from the model—is 0.986. Since there is a direct relationship between multiple R and R^2, and, as we have seen in the two examples above, the concept of correlation can cause problems in multiple linear regression, in this book we tend to prefer the use of R^2 rather than multiple R.

3.3.3 Regression parameters—global usefulness test

Suppose that our population multiple linear regression model has k predictor X-variables:

$$\mathrm{E}(Y) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k,$$

which we estimate from our sample by

$$\hat{Y} = \hat{b}_0 + \hat{b}_1X_1 + \hat{b}_2X_2 + \cdots + \hat{b}_kX_k.$$

Before interpreting the values $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k$, we would like to quantify our uncertainty about the corresponding regression parameters in the population, $b_1, b_2, \ldots, b_k$. For example, is

it possible that all k regression parameters in the population could be zero? If this were the case, it would suggest that our multiple linear regression model contains very little useful information about the population association between Y and $(X_1, X_2, \ldots, X_k)$. So to potentially save a lot of wasted effort (interpreting a model that contains little useful information), we should test this assertion before we do anything else. The *global usefulness test* is a hypothesis test of this assertion.

To see how to apply this test, we need to introduce a new probability distribution, the *F-distribution*. This probability distribution is always positive and is skewed to the right (i.e., has a peak closer to the left side than the right with a long right-hand tail that never quite reaches the horizontal axis at zero). Figure 3.5 shows a typical density curve for an F-distribution. The relative position of the peak and the thickness of the tail are controlled by two degrees of freedom values—the *numerator* degrees of freedom and the *denominator* degrees of freedom. These degrees of freedom values dictate which specific F-distribution gets used for a particular hypothesis test. Since the F-distribution is always positive, hypothesis tests that use the F-distribution are always upper-tail tests.

The critical value, significance level, test statistic, and p-value shown in the figure feature in the following global usefulness hypothesis test (sometimes called an "overall F-test").

- *State null hypothesis*: NH: $b_1 = b_2 = \cdots = b_k = 0$.
- *State alternative hypothesis*: AH: at least one of $b_1, b_2, \ldots, b_k$ is not equal to zero.
- *Calculate test statistic*: Global F-statistic $= \frac{(\mathrm{TSS}-\mathrm{RSS})/k}{\mathrm{RSS}/(n-k-1)} = \frac{\mathrm{R}^2/k}{(1-\mathrm{R}^2)/(n-k-1)}$.

  The first formula provides some insight into how the hypothesis test works. Recall from Section 3.3.2 that the difference between TSS and RSS, $\mathrm{TSS} - \mathrm{RSS}$, is the regression sum of squares. If the regression sum of squares is small (relative to RSS), then the predictors $(X_1, X_2, \ldots, X_k)$ are unable to reduce the random errors

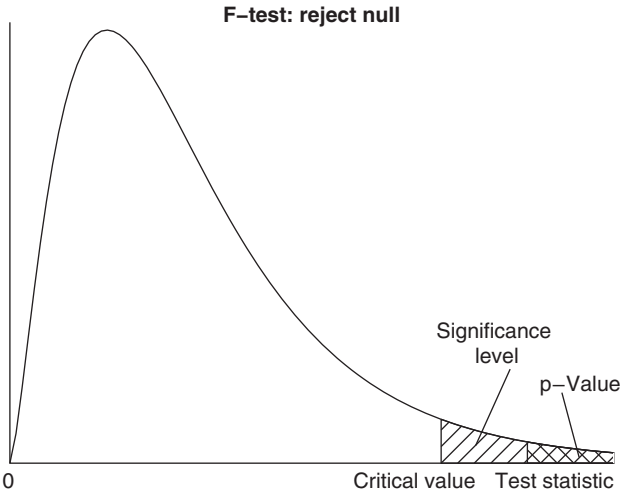

Figure 3.5 Relationships between critical values, significance levels, test statistics, and p-values for hypothesis tests based on the F-distribution. The relative positions of these quantities in this figure would lead to rejecting the null hypothesis. If the positions of the quantities were reversed, then the null hypothesis would not be rejected.

between the Y-values and the fitted $\hat{Y}$-values very much, and we may as well use the sample mean, m_Y, for our model. In such a case, the F-statistic will be small and will probably not be in the rejection region (so the null hypothesis is more plausible). On the other hand, if the regression sum of squares is large (relative to RSS), then the predictors $(X_1, X_2, \ldots, X_k)$ are able to reduce the random errors between the Y-values and the fitted $\hat{Y}$-values sufficiently that we should use at least one of the predictors in our model. In such a case, the F-statistic will be large and will probably be in the rejection region (so the alternative hypothesis is more plausible).

The second formula allows calculation of the global F-statistic using the value of R^2 and shows how a large value for R^2 tends to produce a high value for the global F-statistic (and vice versa).

Statistical software can provide the value of the global F-statistic, as well as the values for R^2, TSS, and RSS to allow calculation by hand.

- *Set significance level*: 5%.
- *Look up a critical value or a p-value using an F-distribution* (use computer help #8 or #9 in the software information files available from the book website):
  - *Critical value*: A particular percentile of the F-distribution with k numerator degrees of freedom and $n - k - 1$ denominator degrees of freedom; for example, the rejection region for a significance level of 5% is any global F-statistic greater than the 95th percentile.
  - *p-value*: The area to the right of the global F-statistic for the F-distribution with k numerator degrees of freedom and $n - k - 1$ denominator degrees of freedom.
- *Make decision*:
  - If the global F-statistic falls in the rejection region, or the p-value is less than the significance level, then we reject the null hypothesis in favor of the alternative (Figure 3.5 provides an illustration of this situation).
  - If the global F-statistic does not fall in the rejection region, or the p-value is greater than or equal to the significance level, then we fail to reject the null hypothesis (it should be clear how Figure 3.5 would need to be redrawn to correspond to this situation).
- *Interpret in the context of the situation*: Rejecting the null hypothesis in favor of the alternative means that at least one of $b_1, b_2, \ldots, b_k$ is not equal to zero (i.e., at least one of the predictors, $X_1, X_2, \ldots, X_k$, is linearly associated with Y); failing to reject the null hypothesis means that we cannot rule out the possibility that $b_1 = b_2 = \cdots = b_k = 0$ (i.e., it is plausible that none of the predictors, $X_1, X_2, \ldots, X_k$, are linearly associated with Y).

To conduct a global usefulness test we need to be able to look up a critical value or a p-value using the F-distribution with k numerator degrees of freedom and $n - k - 1$ denominator degrees of freedom. While it is possible to consult tables similar to the table Univariate Data in Notation and Formulas, we will find it easier from now on to use computer software to find the necessary information (see computer help #8 or #9 in the software information files available from the book website). Alternatively, statistical software provides the p-value for this test directly. For example, here is software output

that displays the results of the global usefulness test for the home prices dataset (see computer help #31):

ANOVA[a]

Model		Sum of squares	df	Mean square	Global F-statistic	Pr(>F)
1	Regression	630.259	2	315.130	51.434	0.005[b]
	Residual	18.381	3	6.127		
	Total	648.640	5			

[a]Response variable: *Price*.
[b]Predictors: (Intercept), *Floor*, *Lot*.

The heading "ANOVA" for this output stands for *analysis of variance* and relates to the comparison of RSS and TSS in the global F-statistic formula. Thus the global usefulness test is an example of an ANOVA test. RSS and TSS are in the column headed "Sum of squares," with RSS in the row labeled "Residual" and TSS in the row labeled "Total." The regression sum of squares in the row labeled "Regression" represents $\mathrm{TSS} - \mathrm{RSS}$. The degrees of freedom values are in the column headed "df," with numerator degrees of freedom, k, in the row labeled "Regression" and denominator degrees of freedom, $n - k - 1$, in the row labeled "Residual." Note that the regression and residual sums of squares and degrees of freedom add to the corresponding totals: $630.259 + 18.381 = 648.640$ and $2 + 3 = 5$.

The column headed "Mean square" divides each sum of squares by its respective degrees of freedom. For example, the "regression mean square" is $(\mathrm{TSS} - \mathrm{RSS})/k$, while the residual mean square is $\mathrm{RSS}/(n - k - 1)$. This residual mean square also goes by the name of "mean square error," which is often abbreviated MSE.

The global F-statistic is the regression mean square divided by the residual mean square and is displayed in the column labeled "Global F-statistic." The p-value for the global usefulness test is displayed in the column labeled "Pr(>F)."

By hand, we can calculate the global F-statistic as follows:

$$\begin{aligned}\text{global F-statistic} &= \frac{(\mathrm{TSS} - \mathrm{RSS})/k}{\mathrm{RSS}/(n-k-1)} = \frac{(648.640 - 18.381)/2}{18.381/(6-2-1)} \\ &= \frac{\mathrm{R}^2/k}{(1-\mathrm{R}^2)/(n-k-1)} = \frac{0.97166/2}{(1-0.97166)/(6-2-1)} \\ &= 51.4.\end{aligned}$$

The value of R^2 for the second formula was obtained from $\mathrm{R}^2 = (\mathrm{TSS} - \mathrm{RSS})/\mathrm{TSS}$.

Suppose that we choose a significance level of 5% for this test. Using computer help #8, the 95th percentile of the F-distribution with $k = 2$ numerator degrees of freedom and $n - k - 1 = 3$ denominator degrees of freedom is 9.55. Since the global F-statistic of 51.4 is larger than this critical value, it is in the rejection region and we reject the null hypothesis in favor of the alternative. Alternatively, using the quicker p-value method, since the p-value of 0.005 (from the preceding statistical software output) is less than our significance level (0.05), we reject the null hypothesis in favor of the alternative. Thus, at least one of b_1 or b_2 is not equal to zero: that is, at least one of the predictors, (X_1, X_2), is linearly associated with Y.

To conduct a global usefulness test (overall F-test) for the population regression parameters in multiple linear regression using the p-value method:

- *State the hypotheses*: NH: $b_1 = b_2 = \cdots = b_k = 0$ versus AH: at least one of $b_1, b_2, \ldots, b_k$ is not equal to zero.
- *Calculate (or simply obtain from statistical software output) the test statistic*: Global F-statistic $= \frac{(\text{TSS}-\text{RSS})/k}{\text{RSS}/(n-k-1)}$.
- *Decide on the significance level* (e.g., 5%).
- *Obtain the p-value* (the area to the right of the F-statistic for an F-distribution with k numerator degrees of freedom and $n-k-1$ denominator degrees of freedom).
- *Make decision*:
  - If the p-value is less than the chosen significance level, reject NH in favor of AH and conclude that at least one of $b_1, b_2, \ldots, b_k$ is not equal to zero (i.e., at least one of the predictors, $X_1, X_2, \ldots, X_k$, is linearly associated with Y).
  - If the p-value is greater than or equal to the chosen significance level, fail to reject NH and conclude that we cannot rule out the possibility that $b_1 = b_2 = \cdots = b_k = 0$ (i.e., it is plausible that none of the predictors, $X_1, X_2, \ldots, X_k$, are linearly associated with Y).

In general, the analysis of variance (ANOVA) table for multiple linear regression has the following form:

ANOVA[a]

Model		Sum of squares	df	Mean square	Global F-statistic	Pr(>F)
1	Regression	$\text{TSS} - \text{RSS}$	k	$(\text{TSS} - \text{RSS})/k$	$\frac{(\text{TSS}-\text{RSS})/k}{\text{RSS}/(n-k-1)}$	p-value
	Residual	RSS	$n-k-1$	$\text{RSS}/(n-k-1)$		
	Total	TSS	$n-1$			

As in the home prices example, the global usefulness test *usually* (but not always) results in concluding that at least one of the predictors, $(X_1, X_2, \ldots, X_k)$, is linearly associated with the response, Y. This is reassuring since it means that we can go on to analyze and interpret the multiple linear regression model confident that we have found something of interest in the association between Y and $(X_1, X_2, \ldots, X_k)$.

We will put the home prices dataset to one side for now, but we shall return to it in some examples and problems in Chapter 4 and again in a case study in Section 6.1 (www.wiley.com/go/pardoe/AppliedRegressionModeling3e). Instead, consider the shipping department example again (data file **SHIPDEPT**), and fit the following multiple linear regression model:

$$\text{E}(Lab) = b_0 + b_1 Tws + b_2 Pst + b_3 Asw + b_4 Num.$$

Here is the output produced by statistical software that displays the results of the global usefulness test for this model (see computer help #31):

ANOVA[a]

Model		Sum of squares	df	Mean square	Global F-statistic	Pr(>F)
1	Regression	5646.052	4	1411.513	17.035	0.000[b]
	Residual	1242.898	15	82.860		
	Total	6888.950	19			

[a]Response variable: Lab.
[b]Predictors: (Intercept), Tws, Pst, Asw, Num.

Suppose that we choose significance level 5% for this test. Since the p-value of 0.000 (from the preceding statistical software output) is less than our significance level (0.05), we reject the null hypothesis (NH: $b_1 = b_2 = b_3 = b_4 = 0$) in favor of the alternative hypothesis AH: at least one of b_1, b_2, b_3, or b_4 is not equal to zero). Thus, at least one of the predictors, (Tws, Pst, Asw, Num), is linearly associated with Lab. We could also manually check that the global F-statistic of 17.035 is larger than the 95th percentile of the F-distribution with $k = 4$ numerator degrees of freedom and $n - k - 1 = 15$ denominator degrees of freedom, but there is no need to do so (remember, there are always two ways to do a hypothesis test, and they will always give the same result if done correctly).

3.3.4 Regression parameters—nested model test

Suppose that we have fit a multiple linear regression model, and a global usefulness test has suggested that at least one of the predictors, $(X_1, X_2, \ldots, X_k)$, is linearly associated with the response, Y. From the application of adjusted R^2 to the shipping department example in Section 3.3.2, we saw that it is possible that a simpler model with fewer than k predictor variables may be preferable to the full k-predictor model. This can occur when, for example, a subset of the predictor variables provides very little information about the response, Y, *beyond* the information provided by the other predictor variables.

A *nested model test* formally investigates such a possibility. Suppose that the full k-predictor model, also known as the *complete model*, has an RSS value equal to $\mathrm{RSS_C}$. Consider removing a subset of the predictor variables that we suspect provides little information about the response, Y, beyond the information provided by the other predictors. Removing this subset leads to a *reduced* model with r predictors (i.e., $k - r$ predictors are removed). Since the reduced model is nested in the complete model (i.e., it contains a subset of the complete model predictors), it will have an RSS value, say, $\mathrm{RSS_R}$, that is greater than or equal to $\mathrm{RSS_C}$; see Section 3.3.2 for a geometrical argument for why this is so.

Intuitively, if the difference between $\mathrm{RSS_R}$ and $\mathrm{RSS_C}$ is small, then the explanatory power of the two models is similar, and we would prefer the simpler reduced model since the complete model seems to be overfitting the sample data. On the other hand, if the difference between $\mathrm{RSS_R}$ and $\mathrm{RSS_C}$ is large, we would prefer the complete model since the $k - r$ extra predictors in the complete model do appear to provide useful information about the response, Y, beyond the information provided by the r reduced model predictors.

To turn this intuition into a formal hypothesis test, we need to find a test statistic proportional to $\mathrm{RSS_R} - \mathrm{RSS_C}$, whose sampling distribution we know under a null hypothesis that states the reduced and complete models are equivalent in the population.

The F-distribution we introduced in Section 3.3.3 serves this purpose in the following nested model test (sometimes called a "general linear F-test").

- Write the reduced model as $\mathrm{E}(Y) = b_0 + b_1X_1 + \cdots + b_rX_r$.
- Write the complete model as $\mathrm{E}(Y) = b_0 + b_1X_1 + \cdots + b_rX_r + b_{r+1}X_{r+1} + \cdots + b_kX_k$. *(Here, the extra predictors in the complete model, $X_{r+1}, \ldots, X_k$, are written after the reduced model predictors $X_1, \ldots, X_r$, but in practice we can test any subset of predictors in the complete model and they do not have to be listed last in the model equation.)*
- *State null hypothesis*: NH: $b_{r+1} = \cdots = b_k = 0$.
- *State alternative hypothesis*: AH: at least one of $b_{r+1}, \ldots, b_k$ is not equal to zero.
- *Calculate test statistic*: Nested F-statistic $= \frac{(\mathrm{RSS_R} - \mathrm{RSS_C})/(k-r)}{\mathrm{RSS_C}/(n-k-1)}$.

  Statistical software can provide the value of the nested F-statistic, as well as the values for $\mathrm{RSS_R}$ and $\mathrm{RSS_C}$ to allow calculation by hand.
- *Set significance level* (e.g., 5%).
- *Look up a critical value or a p-value using an F-distribution* (use computer help #8 or #9 in the software information files available from the book website):
  - *Critical value*: A particular percentile of the F-distribution with $k - r$ numerator degrees of freedom and $n - k - 1$ denominator degrees of freedom; for example, the rejection region for a significance level of 5% is any nested F-statistic greater than the 95th percentile.
  - *p-value*: The area to the right of the nested F-statistic for the F-distribution with $k - r$ numerator degrees of freedom and $n - k - 1$ denominator degrees of freedom.
- *Make decision*:
  - If the nested F-statistic falls in the rejection region, or the p-value is less than the significance level, then we reject the null hypothesis in favor of the alternative (Figure 3.5 provides an illustration of this situation).
  - If the nested F-statistic does not fall in the rejection region, or the p-value is greater than or equal to the significance level, then we fail to reject the null hypothesis (it should be clear how Figure 3.5 would need to be redrawn to correspond to this situation).
- *Interpret in the context of the situation*: Rejecting the null hypothesis in favor of the alternative means that at least one of $b_{r+1}, \ldots, b_k$ is not equal to zero (i.e., at least one of the extra predictors in the complete model, $X_{r+1}, \ldots, X_k$, appears to provide useful information about the response, Y, beyond the information provided by the r predictor variables in the reduced model); failing to reject the null hypothesis means that we cannot rule out the possibility that $b_{r+1} = \cdots = b_k = 0$ (i.e., none of the extra predictors in the complete model, $X_{r+1}, \ldots, X_k$, appear to provide useful information about the response, Y, beyond the information provided by the r predictor variables in the reduced model).

To do a nested model test, we need to look up a critical value or a p-value for the F-distribution with $k - r$ numerator degrees of freedom and $n - k - 1$ denominator

degrees of freedom. As with the global usefulness test, we will find it easier to use computer software to find the necessary information (see computer help #8 or #9).

Recall the shipping department example from Table 3.1. Suppose that we propose the following (complete) multiple linear regression model:

$$\mathrm{E}(Lab) = b_0 + b_1\,Tws + b_2 Pst + b_3 Asw + b_4 Num.$$

Here is part of the output produced by statistical software that displays some results for this model (see computer help #31):

Parameters[a]

Model		Estimate	Standard error	t-Statistic	Pr(> |t|)
C	(Intercept)	95.415	30.036	3.177	0.006
	Tws	6.074	2.662	2.281	0.038
	Pst	8.435	8.870	0.951	0.357
	Asw	−1.746	0.760	−2.297	0.036
	Num	−0.124	0.380	−0.328	0.748

[a]Response variable: Lab.

We will see in Section 3.3.5 that these results suggest that perhaps neither Pst nor Num provides useful information about the response, Lab, beyond the information provided by Tws and Asw. To test this formally, we do a nested model test of the following hypotheses:

- NH: $b_2 = b_4 = 0$.
- AH: at least one of b_2 or b_4 is not equal to zero.

The ANOVA table for the complete model (computer help #31) is

ANOVA[a]

Model		Sum of squares	df	Mean square	Global F-statistic	Pr(>F)
C	Regression	5646.052	4	1411.513	17.035	0.000[b]
	Residual	1242.898	15	82.860		
	Total	6888.950	19			

[a]Response variable: Lab.
[b]Predictors: (Intercept), Tws, Pst, Asw, Num.

$\mathrm{RSS_C}$ is in the "Sum of squares" column, while $n - k - 1$ is in the "df" column, both in the row labeled "Residual," while k is in the "df" column in the row labeled "Regression." Contrast these results with those for the reduced two-predictor model:

$$\mathrm{E}(Lab) = b_0 + b_1\,Tws + b_3 Asw.$$

ANOVA[a]

Model		Sum of squares	df	Mean square	Global F-statistic	Pr(>F)
R	Regression	5567.889	2	2783.945	35.825	0.000[b]
	Residual	1321.061	17	77.709		
	Total	6888.950	19			

[a]Response variable: Lab.
[b]Predictors: (Intercept), Tws, Asw.

$\mathrm{RSS_R}$ is in the "Sum of squares" column in the row labeled "Residual," while r is in the "df" column in the row labeled "Regression." By hand, we can calculate the nested F-statistic as follows:

$$\text{nested F-statistic} = \frac{(\mathrm{RSS_R} - \mathrm{RSS_C})/(k-r)}{\mathrm{RSS_C}/(n-k-1)} = \frac{(1321.061 - 1242.898)/(4-2)}{1242.898/(20-4-1)} = 0.472.$$

Suppose that we choose a significance level of 5% for this test. Using computer help #8, the 95th percentile of the F-distribution with $k - r = 2$ numerator degrees of freedom and $n - k - 1 = 15$ denominator degrees of freedom is 3.68. Since the nested F-statistic of 0.472 is smaller than this critical value (so it is not in the rejection region), we cannot reject the null hypothesis. Thus, it is plausible that both b_2 and b_4 are equal to zero in the population; that is, neither Pst nor Num appears to provide useful information about the response, Lab, beyond the information provided by Tws and Asw.

Intuitively, whereas the four-dimensional regression hyperplane can get a little closer to the sample data points, it appears to do so at the expense of overfitting by including apparently redundant predictor variables. The nested model test suggests that the reduced two-predictor model does a better job of finding the population association between Lab and (Tws, Pst, Asw, Asw) than the complete four-predictor model.

Alternatively, statistical software provides the p-value for this test directly. For example, here is the output that displays the results of the nested model test above for the shipping department dataset (see computer help #34):

Model summary

Model	R^2	Adjusted R^2	Regression standard error	Change statistics			
				F-statistic	df1	df2	Pr(>F)
R[a]	0.8082	0.7857	8.815				
C[b]	0.8196	0.7715	9.103	0.472	2	15	0.633

[a]Predictors: (Intercept), Tws, Asw.
[b]Predictors: (Intercept), Tws, Pst, Asw, Num.

The nested F-statistic is in the second row of the column headed "F-statistic." The associated p-value is in the second row of the column headed "Pr(>F)." (Ignore any numbers in the first rows of these columns.) Since the p-value of 0.633 is more than our significance level (0.05), we cannot reject the null hypothesis—the same conclusion (necessarily) that we made with the rejection region method above.

To conduct a nested model test (general linear F-test) for a subset of population regression parameters in multiple linear regression using the p-value method:

- Write the reduced model as $\mathrm{E}(Y) = b_0 + b_1 X_1 + \cdots + b_r X_r$ and the complete model as $\mathrm{E}(Y) = b_0 + b_1 X_1 + \cdots + b_r X_r + b_{r+1} X_{r+1} + \cdots + b_k X_k$.
- *State the hypotheses*: $b_{r+1} = \cdots = b_k = 0$ versus AH: at least one of $b_{r+1}, \ldots, b_k$ is not equal to zero.
- *Calculate (or simply obtain from statistical software output) the test statistic*: Nested F-statistic $= \frac{(\mathrm{RSS_R} - \mathrm{RSS_C})/(k-r)}{\mathrm{RSS_C}/(n-k-1)}$.
- *Decide on the significance level* (e.g., 5%).
- *Obtain the p-value* (the area to the right of the F-statistic for an F-distribution with $k - r$ numerator degrees of freedom and $n - k - 1$ denominator degrees of freedom).
- *Make decision*:
  - If the p-value is less than the chosen significance level, reject NH in favor of AH and conclude that at least one of $b_{r+1}, \ldots, b_k$ is not equal to zero (i.e., at least one of the extra predictors in the complete model, $X_{r+1}, \ldots, X_k$, appears to provide useful information about the response, Y, beyond the information provided by the r predictor variables in the reduced model).
  - If the p-value is greater than or equal to the chosen significance level, fail to reject NH and conclude that we cannot rule out the possibility that $b_{r+1} = \cdots = b_k = 0$ (i.e., none of the extra predictors in the complete model, $X_{r+1}, \ldots, X_k$, appear to provide useful information about the response, Y, beyond the information provided by the r predictor variables in the reduced model).

When using statistical software, it is possible to calculate the value of $\mathrm{RSS_R} - \mathrm{RSS_C}$ in the numerator of the nested F-statistic by fitting just the complete model, if each predictor can be displayed in a separate row in the output ANOVA table:

- Make sure the extra predictors in the complete model are listed last.
- Select "sequential sum of squares" or "type I sum of squares' to be displayed in the ANOVA table.
- The value of $\mathrm{RSS_R} - \mathrm{RSS_C}$ is the same as the sum of the sequential sums of squares for each regression parameter being tested (i.e., for the extra predictors in the complete model that are listed last).

For example, here is the output that displays the sequential sums of squares for the nested model test above for the shipping department dataset:

ANOVA[a]

Model		Sequential sum of squares	df	Mean square	Global F-statistic	Pr(>F)
C	Regression	5646.052	4	1411.513	17.035	0.000[b]
	Tws	4861.741	1			
	Asw	706.148	1			
	Pst	69.266	1			
	Num	8.897	1			
	Residual	1242.898	15	82.860		
	Total	6888.950	19			

[a]Response variable: Lab.
[b]Predictors: (Intercept), Tws, Asw, Pst, Num.

Using this output, the value of $\mathrm{RSS_R} - \mathrm{RSS_C}$ can be calculated from the sum of the sequential sums of squares for Pst and Num: $69.266 + 8.897 = 78.163$. This matches the value we obtain by explicitly comparing the complete and reduced models, $\mathrm{RSS_R} - \mathrm{RSS_C} = 1321.061 - 1242.898 = 78.163$. Sequential sums of squares take into account the order of the predictors in the model so that each value adjusts for all of the preceding predictors. In particular, each sequential sum of squares measures the *additional* variation in Y-values (about their sample mean) that is "explained" by the corresponding predictor, *beyond* the variation explained by the preceding predictors in the model. Sequential sums of squares also have the property that they sum to the regression sum of squares. For example, in the ANOVA table for the shipping department example, $4861.741 + 706.148 + 69.266 + 8.897 = 5646.052$.

An alternative to the sequential sum of squares is the "adjusted sum of squares" or "type III sum of squares." Adjusted sums of squares adjust for all of the other predictors in the model and therefore do not take into account the order of the predictors in the model. Neither do adjusted sums of squares have the property that they sum to the regression sum of squares. Finally, there is no relationship between sequential sums of squares and adjusted sums of squares, except for the last predictor listed (for which the sequential and adjusted sums of squares are identical—make sure you understand why). There is also a "type II sum of squares," but we do not discuss this further here.

The complete and reduced models considered in this section are "nested" in the sense that the complete model includes all of the predictors in the reduced model as well as some additional predictors unique to the complete model. Equivalently, the reduced model is similar to the complete model except that the regression parameters for these additional predictors are all zero in the reduced model. More generally, one model is nested in another if they each contain the same predictors but the first model constrains some of its regression parameters to be fixed numbers (zero in the examples above). The nested model test then determines whether the second model provides a significant improvement over the first (small p-value); if not, then the constrained values of the regression parameters are plausible (large p-value).

With nested models like this, the RSS for the complete model is always lower than (or the same as) the RSS for the reduced model (see Section 3.3.2). Thus, R^2 for the complete model is always higher than (or the same as) R^2 for the reduced model. However, we can

think of the nested model test as telling us whether the complete model R^2 is *significantly* higher—if not, we prefer the reduced model. Another way to see which model is favored is to consider the regression standard error, s (we would generally prefer the model with the smaller value of s), and adjusted R^2 (we would generally prefer the model with the larger value of adjusted R^2). To summarize:

- R^2 is always higher for the complete model than for the reduced model (so this tells us nothing about which model we prefer).
- The regression standard error, s, may be higher or lower in the reduced model than the complete model, but if it is lower, then we would prefer the reduced model according to this criterion.
- Adjusted R^2 may be higher or lower in the reduced model than in the complete model, but if it is higher, then we would prefer the reduced model according to this criterion.

In many cases, these three methods for comparing nested models—nested model test, comparing regression standard errors, and comparing values of adjusted R^2—will agree, but it is possible for them to conflict. In conflicting cases, a reasonable conclusion might be that the two models are essentially equivalent. Alternatively, other model comparison tools are available—see Section 5.4.

It is easy to confuse the global usefulness and nested model tests since both involve F-statistics and both are examples of ANOVA tests. In summary, the *global usefulness test* considers whether all the regression parameters are equal to zero. In other words, if the test ends up rejecting the null hypothesis in favor of the alternative hypothesis, then we conclude that at least one of the predictor variables has a linear association with the response variable. This test is typically used just once at the beginning of a regression analysis to make sure that there will be something worth modeling (if none of the predictor variables has a linear association with the response variable, then a multiple linear regression model using those predictors is probably not appropriate).

By contrast, the *nested model test* considers whether a subset of the regression parameters are equal to zero. In other words, if the test ends up failing to reject the null hypothesis, then we conclude that the corresponding subset of predictor variables has no linear association with the response variable once the association with the remaining predictors left in the model has been accounted for. This generally means that we can then drop the subset of predictor variables being tested. This test is typically used one or more times during a regression analysis to discover which, if any, predictor variables are redundant given the presence of the other predictor variables and so are better left out of the model.

However, the global usefulness and nested model tests are related to one another. The formulas for the two F-statistics (in Sections 3.3.3 and 3.3.4) show that the global usefulness F-statistic is a special case of the nested model F-statistic, in which the subset of predictors being tested consists of *all* the predictors and the reduced model has no predictor variables at all (i.e., RSS for the reduced model is the same as TSS and $r = 0$).

It is also possible to use the nested model F-test to test whether a subset of regression parameters in a multiple linear regression model could be equal to one another—see the optional section at the end of Section 4.3.2 and also the case study in Section 6.2 (www.wiley.com/go/pardoe/AppliedRegressionModeling3e).

3.3.5 Regression parameters—individual tests

Suppose that we have fit a multiple linear regression model and a global usefulness test has suggested that at least one of the predictors, $(X_1, X_2, \ldots, X_k)$, has a linear association with the response, Y. We have seen in Sections 3.3.2 and 3.3.4 that it is possible that a reduced model with fewer than k predictor variables may be preferable to the complete k-predictor model. This can occur when, for example, a subset of the predictor variables provides very little information about the response, Y, *beyond* the information provided by the other predictor variables. We have seen how to use a nested model test to remove a subset of predictors from the complete model, but how do we identify which predictors should be in this subset? One possible approach is to consider the regression parameters individually. In particular, what do the estimated sample estimates, $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k$, tell us about likely values for the population parameters, $b_1, b_2, \ldots, b_k$?

Since we assume that the sample has been randomly selected from the population, under repeated sampling we would expect the sample estimates and population parameters to match *on average*, but for any particular sample they will probably differ. We cannot be sure how much they will differ, but we can quantify this uncertainty using the sampling distribution of the estimated regression parameters, $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k$. Recall that when analyzing simple linear regression models, we calculated a test statistic:

$$\text{slope t-statistic} = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}}.$$

Under very general conditions, this slope t-statistic has an approximate t-distribution with $n - 2$ degrees of freedom. We used this result to conduct hypothesis tests and construct confidence intervals for the population slope, b_1.

We can use a similar result to conduct hypothesis tests and construct confidence intervals for each of the population regression parameters, $b_1, b_2, \ldots, b_k$, in multiple linear regression. In particular, we can calculate the following test statistic for the pth regression parameter, b_p:

$$\text{regression parameter t-statistic} = \frac{\hat{b}_p - b_p}{s_{\hat{b}_p}},$$

where $s_{\hat{b}_p}$ is the standard error of the regression parameter estimate. The formula for this standard error is provided for interest at the end of this section. Under very general conditions, this regression parameter t-statistic has an approximate t-distribution with $n - k - 1$ degrees of freedom.

Regression parameter hypothesis tests. Suppose that we are interested in a particular value of the pth regression parameter for a multiple linear regression model. Usually, "zero" is an interesting value for the regression parameter since this would be equivalent to there being no linear association between Y and X_p once the linear association between Y and the other $k - 1$ predictors has been accounted for. Another way of saying this is that there is no linear association between Y and X_p when we hold the other $k - 1$ predictors fixed at constant values. One way to test this could be

to literally keep the other $k-1$ predictors fixed at constant values and vary X_p to see if Y changes. This is not usually possible with observational data (see Section 2.1) and even with experimental data would be very time-consuming and expensive to do for each predictor in turn. Alternatively, we can easily do hypothesis tests to see if the information in our sample supports population regression parameters of zero or whether it favors some alternative values.

For the shipping data example, *before* looking at the sample data we might have reason to believe that there is a linear association between weekly labor hours, Lab, and the total weight shipped in thousands of pounds, Tws, once the linear association between Lab and Pst (proportion shipped by truck), Asw (average shipment weight), and Num (week number) has been accounted for (or holding Pst, Asw, and Num constant). To see whether the sample data provide compelling evidence that this is the case, we should conduct a two-tail hypothesis test for the population regression parameter b_1 in the model $\mathrm{E}(Lab) = b_0 + b_1 Tws + b_2 Pst + b_3 Asw + b_4 Asw$.

- *State null hypothesis*: NH: $b_1 = 0$.
- *State alternative hypothesis*: AH: $b_1 \neq 0$.
- *Calculate test statistic*: t-Statistic $= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{6.074-0}{2.662} = 2.28$ ($\hat{b}_1$ and $s_{\hat{b}_1}$ can be obtained using statistical software—see following output—while b_1 is the value in NH).
- *Set significance level*: 5%.
- *Look up t-table*:
  - *Critical value*: The 97.5th percentile of the t-distribution with $20-4-1=15$ degrees of freedom is 2.13 (see computer help #8 in the software information files available from the book website); the rejection region is therefore any t-statistic greater than 2.13 or less than -2.13 (we need the 97.5th percentile in this case because this is a two-tail test, so we need half the significance level in each tail).
  - *p-value*: The sum of the areas to the right of the t-statistic (2.28) and to the left of the negative of the t-statistic (-2.28) for the t-distribution with 15 degrees of freedom is 0.038 (use computer help #9).
- *Make decision*:
  - Since the t-statistic of 2.28 falls in the rejection region, we reject the null hypothesis in favor of the alternative.
  - Since the p-value of 0.038 is less than the significance level of 0.05, we reject the null hypothesis in favor of the alternative.
- *Interpret in the context of the situation*: The 20 sample observations suggest that a population regression parameter, b_1, of zero seems implausible and the sample data favor a nonzero value (at a significance level of 5%); in other words, there does appear to be a linear association between Lab and Tws once Pst, Asw, and Num have been accounted for (or holding Pst, Asw, and Num constant).

Hypothesis tests for the other regression parameters, b_2, b_3, and b_4, are similar. Sometimes, we may have a particular interest in doing an upper- or lower-tail test rather than a two-tail test. The test statistic is the same value for all three flavors of test, but

the significance level represents an area in just one tail rather than getting split evenly between both tails (this affects where the critical values for the rejection regions are), and the p-value also represents an area in just one tail rather than getting split evenly between both tails.

In practice, we can do population regression parameter hypothesis tests in multiple linear regression directly using statistical software. For example, here is the relevant output for the shipping dataset (see computer help #31):

Parameters[a]

Model		Estimate	Standard error	t-Statistic	Pr(> |t|)
1	(Intercept)	95.415	30.036	3.177	0.006
	Tws	6.074	2.662	2.281	0.038
	Pst	8.435	8.870	0.951	0.357
	Asw	−1.746	0.760	−2.297	0.036
	Num	−0.124	0.380	−0.328	0.748

[a]Response variable: Lab.

The regression parameter estimates $\hat{b}_p$ are in the column headed "Estimate" and the row labeled with the name of the predictor. The standard errors of the estimates, $s_{\hat{b}_p}$, are in the column headed "Standard error," while the t-statistics are in the column headed "t-Statistic," and the two-tail p-values are in the column headed "Pr(> |t|)" (meaning "the probability that a t random variable with $n - k - 1$ degrees of freedom could be larger than the absolute value of the t-statistic or smaller than the negative of the absolute value of the t-statistic").

To test an individual regression parameter in multiple linear regression using the p-value method:

- To carry out a two-tail hypothesis test for a zero value for the pth population regression parameter in multiple linear regression, decide on the significance level (e.g., 5%), and check to see whether the two-tail p-value ("Pr(> |t|)" in the statistical software output) is smaller than this significance level. If it is, reject NH: $b_p = 0$ in favor of AH: $b_p \neq 0$ and conclude that the sample data favor a nonzero regression parameter (at the chosen significance level). Otherwise, there is insufficient evidence to reject NH: $b_p = 0$, and conclude that a zero population parameter cannot be ruled out (at the chosen significance level).
- For an upper-tail hypothesis test, decide on the significance level (e.g., 5%) and calculate the upper-tail p-value. The upper-tail p-value is the area to the right of the t-statistic under the appropriate t-distribution density curve. For a positive t-statistic, this area is equal to the two-tail p-value divided by 2. Then, if the upper-tail p-value is smaller than the chosen significance level, reject NH: $b_p = 0$ in favor of AH: $b_p > 0$. Otherwise, there is insufficient evidence to reject NH: $b_p = 0$, and conclude that a zero population parameter cannot be ruled out (at the chosen significance level).
- For a lower-tail hypothesis test, decide on the significance level (e.g., 5%) and calculate the lower-tail p-value. The lower-tail p-value is the area to the left of

the t-statistic under the appropriate t-distribution density curve. For a negative t-statistic, this area is equal to the two-tail p-value divided by 2. Then, if the lower-tail p-value is smaller than the chosen significance level, reject NH: $b_p = 0$ in favor of AH: $b_p < 0$. Otherwise, there is insufficient evidence to reject NH: $b_p = 0$ and conclude that a zero population parameter cannot be ruled out (at the chosen significance level).

However, be careful when an upper-tail hypothesis test has a negative t-statistic or a lower-tail test has a positive t-statistic. In such situations, the p-value must be at least 0.5 (draw a picture to convince yourself of this), so it is also going to be larger than any reasonable significance level that we might have picked. Thus, we will not be able to reject NH: $b_p = 0$ in favor of AH: $b_1 > 0$ (for an upper-tail test) or AH: $b_1 < 0$ (for a lower-tail test).

For the shipping example, since the two-tail p-value for Tws is 0.038, we reject NH: $b_1 = 0$ in favor of AH: $b_1 \neq 0$ and conclude that the sample data favor a nonzero regression parameter (at a significance level of 5%). For an upper-tail test, since the t-statistic is positive, the upper-tail p-value is 0.019, and we reject NH: $b_1 = 0$ in favor of AH: $b_1 > 0$ and conclude that the sample data favor a positive regression parameter (at a significance level of 5%). For a lower-tail test, the positive t-statistic means that the lower-tail p-value is at least 0.5, so we cannot reject NH: $b_1 = 0$ in favor of AH: $b_1 < 0$.

We present all three flavors of hypothesis test here (two-tail, upper-tail, and lower-tail), but in real-life applications, we would usually conduct only one—selected before looking at the data. Remember also that each time we do a hypothesis test, we have a chance of making a mistake (either rejecting NH when we should not have, or failing to reject NH when we should have)—see Section 1.6.3. Thus, when trying to decide which predictors should remain in a particular multiple linear regression model, we should use as few hypothesis tests as possible. One potential strategy is to identify a subset of predictors with relatively high two-tail p-values, and then use the nested model test of Section 3.3.4 to formally decide whether this subset of predictors provides information about the response, Y, beyond the information provided by the other predictor variables. For example, for the shipping data, the relatively high p-values for Pst (0.357) and Num (0.748) suggest that we should do the nested model test we conducted in Section 3.3.4.

Also keep in mind that the p-value of 0.357 for Pst in this example suggests that there is no linear association between Lab and Pst once the linear association between Lab and Tws, Asw, and Num has been accounted for (or holding Tws, Asw, and Num constant). In other words, Pst may be redundant in the model as long as Tws, Asw, and Num remain in the model. Similarly, the p-value of 0.748 for Num suggests that there is no linear association between Lab and Num once Tws, Pst, and Asw have been accounted for (or holding Tws, Pst, and Asw constant). In other words, Num may be redundant in the model as long as Tws, Pst, and Asw remain in the model.

This is *not* quite the same as the conclusion for the nested model test, which was that there does not appear to be a linear association between Lab and (Pst, Num) once Tws and Asw have been accounted for (or holding Tws and Asw constant). In other words, Pst and Num may be redundant in the model as long as Tws and Asw remain in the model.

Thus, we can do individual regression parameter t-tests to remove just one redundant predictor at a time or to identify which predictors to investigate with a nested model F-test.

However, we need the nested model test to actually remove more than one redundant predictor at a time. Using nested model tests allows us to use fewer hypothesis tests overall to help identify redundant predictors (so that the remaining predictors appear to explain the response variable adequately); this also lessens the chance of making any hypothesis test errors.

It is also possible to conduct a hypothesis test for the intercept parameter, b_0; the procedure is exactly the same as for the regression parameters, $b_1, b_2, \ldots, b_k$. For example, if the intercept p-value is greater than a significance level of 5%, then we cannot reject the null hypothesis that the intercept (b_0) is zero (at that significance level). In other words, a zero intercept is quite plausible. Whether this makes sense depends on the practical context. For example, in some physical systems, we might know that when the predictors are zero, the response must necessarily be zero. In contexts such as this it might make sense to drop the intercept from the model and fit what is known as *regression through the origin*. In most practical applications, however, testing a zero intercept (and dropping the intercept from the model if the p-value is quite high) is relatively rare (it generally occurs only in contexts where an intercept exactly equal to zero is expected). Thus, we will not dwell any further on this.

As with simple linear regression, do not confuse being unable to meaningfully interpret the estimated intercept parameter, $\hat{b}_0$, with dropping the intercept from the model—see Section 2.3.3.

In principle, we can also test values other than zero for the population regression parameters, $b_1, b_2, \ldots, b_k$—just plug the appropriate value for b_p into the t-statistic formula earlier in this section and proceed as usual. This is quite rare in practice since testing whether the population regression parameters could be zero is usually of most interest. We can also test values other than zero for b_0—again, this is quite rare in practice.

It turns out that the individual regression parameter t-test considered in this section is related to the nested model F-test considered in the preceding section. It can be shown that if we square a t-distributed random variable with d degrees of freedom, then we obtain an F-distributed random variable with 1 numerator degree of freedom and d denominator degrees of freedom. This relationship leads to the following result. Although the nested model F-test is usually used to test more than one regression parameter, *if* we use it to test just a single parameter, then the F-statistic is equal to the square of the corresponding individual regression parameter t-statistic. Also, the F-test p-value and two-tail t-test p-value are identical. Try this for the shipping example [you should find that the nested model F-statistic for b_1 in the model $\mathrm{E}(Lab) = b_0 + b_1 Tws + b_2 Pst + b_3 Asw + b_4 Asw$ is 5.204, which is the square of the individual regression parameter t-statistic, 2.281; the p-values for both tests are 0.038].

Regression parameter confidence intervals. Another way to express our level of uncertainty about the population regression parameters, $b_1, b_2, \ldots, b_k$, is with confidence intervals. For example, a 95% confidence interval for b_p results from the following:

$$\Pr\left(-97.5\text{th percentile} < t_{n-k-1} < 97.5\text{th percentile}\right) = 0.95,$$

$$\Pr(-97.5\text{th percentile} < (\hat{b}_p - b_p)/s_{\hat{b}_p} < 97.5\text{th percentile}) = 0.95,$$

$$\Pr\left(\hat{b}_p - 97.5\text{th percentile}\ (s_{\hat{b}_p}) < b_p < \hat{b}_p + 97.5\text{th percentile}\ (s_{\hat{b}_p})\right) = 0.95,$$

where the 97.5th percentile is from the t-distribution with $n - k - 1$ degrees of freedom. In other words, the 95% confidence interval is $\hat{b}_p \pm 97.5\text{th percentile } (s_{\hat{b}_p})$. For example, the 95% confidence interval for b_1 in the four-predictor shipping model is

$$\begin{aligned}\hat{b}_1 \pm 97.5\text{th percentile } (s_{\hat{b}_1}) &= 6.074 \pm 2.131 \times 2.662 \\ &= 6.074 \pm 5.673 \\ &= (0.40,\ 11.75).\end{aligned}$$

The values for $\hat{b}_1 = 6.074$ and $s_{\hat{b}_1} = 2.662$ come from the statistical software output in Section 3.3.2, while the 97.5th percentile from the t-distribution with $n - k - 1 = 15$ degrees of freedom is obtained using computer help #8 in the software information files available from the book website. We can calculate the confidence intervals for b_2, b_3, and b_4 similarly. Here is statistical software output displaying 95% confidence intervals for the regression parameters, $b_1, \ldots, b_4$, in the four-predictor shipping model (see computer help #27):

	Parameters[a]				**95% Confidence interval**	
Model	Estimate	Standard error	t-Statistic	Pr(> |t|)	Lower bound	Upper bound
1 (Intercept)	95.415	30.036	3.177	0.006		
Tws	6.074	2.662	2.281	0.038	0.399	11.749
Pst	8.435	8.870	0.951	0.357	−10.472	27.341
Asw	−1.746	0.760	−2.297	0.036	−3.366	−0.126
Num	−0.124	0.380	−0.328	0.748	−0.934	0.685

[a] Response variable: *Lab*.

The confidence intervals are in the columns headed "95% Confidence Interval" and the row labeled with the name of the predictor.

Now that we have calculated some confidence intervals, what exactly do they tell us? Well, for this example, *loosely speaking*, we can say that "we are 95% confident that the population regression parameter, b_1, is between 0.40 and 11.75 in the model $\mathrm{E}(Lab) = b_0 + b_1 Tws + b_2 Pst + b_3 Asw + b_4 Asw$." In other words, "we are 95% confident that labor hours increases by between 0.40 and 11.75 for each 1,000-pound increase in Tws (weight shipped) once Pst (proportion shipped by truck), Asw (average shipment weight), and Num (week number) have been accounted for (or holding Pst, Asw, and Num constant)." A more precise interpretation would be something like "if we were to take a large number of random samples of size 20 from our population of shipping numbers and calculate a 95% confidence interval for b_1 in each, then 95% of those confidence intervals would contain the true (unknown) population regression parameter." We can provide similar interpretations for the confidence intervals for b_2, b_3, and b_4. As further practice, find 95% intervals for the regression parameters, b_1 and b_3, in the two-predictor shipping model—you should find that they are narrower (more precise) than in the four-predictor model (which contains two unimportant predictors).

Thus, in general, we can write a confidence interval for a multiple linear regression slope parameter, b_p, as

$$\hat{b}_p \pm \text{t-percentile } (s_{\hat{b}_p}),$$

where $\hat{b}_p$ is the sample regression parameter estimate, $s_{\hat{b}_p}$ is the standard error of the regression parameter estimate, and the t-percentile comes from a t-distribution with $n - k - 1$ degrees of freedom (n is the sample size, k is the number of predictor variables).

Suppose we have calculated a 95% confidence interval for a multiple linear regression parameter, b_p, to be (a, b). Then we can say that we are 95% confident that b_p is between a and b.

For confidence levels other than 95%, percentiles in the calculations must be changed as appropriate. For example, 90% intervals (5% in each tail) need the 95th percentile, whereas 99% intervals (0.5% in each tail) need the 99.5th percentile. These percentiles can be obtained using computer help #8. As further practice, calculate 90% intervals for b_1 for the four- and two-predictor shipping models (see Problem 3.2)—you should find that they are (1.41, 10.74) and (1.07, 8.94), respectively. Some statistical software automatically calculates 95% intervals only for this type of confidence interval.

Correlation revisited. Recall from Section 3.3.2 the warning about using the concept of correlation in multiple linear regression. We return to the two earlier simulated examples to see how we can be led astray if we are not careful. First, consider the simulated data represented by the scatterplot in Figure 3.3, where $Sales$ is annual sales in millions of dollars for a small retail business, and $Advert$ is total annual spending on advertising in millions of dollars (**SALES2** data file). The correlation between $Sales$ and $Advert$ here is very low (in fact, it is 0.165), but there is a second predictor, $Stores$ (the number of retail stores operated by the company), which enables $Advert$ to be a useful predictor in the multiple linear regression model:

$$\mathrm{E}(Sales) = b_0 + b_1 Advert + b_2 Stores.$$

Statistical software output for this model (see computer help #31 in the software information files available from the book website) is

Parameters[a]

Model		Estimate	Standard error	t-Statistic	Pr(> |t|)
1	(Intercept)	−4.769	0.820	−5.818	0.000
	$Advert$	1.053	0.114	9.221	0.000
	$Stores$	5.645	0.215	26.242	0.000

[a] Response variable: $Sales$.

Since the two-tail p-value for $Advert$ is 0.000, we know that there is a strong linear association between $Sales$ and $Advert$, holding $Stores$ constant. Similarly, since the two-tail p-value for $Stores$ is 0.000, we know that there is a strong linear association between $Sales$ and $Stores$, holding $Advert$ constant. Thus, the low correlation between $Sales$ and $Advert$ is irrelevant to the outcome of the multiple linear regression model. Figure 3.6 shows why this is the case (see computer help #17). When $Stores$ is held constant at 1, the points represented by 1s have an increasing trend. When $Stores$ is held constant at 2, the points represented by 2s have a similarly increasing trend. It is a

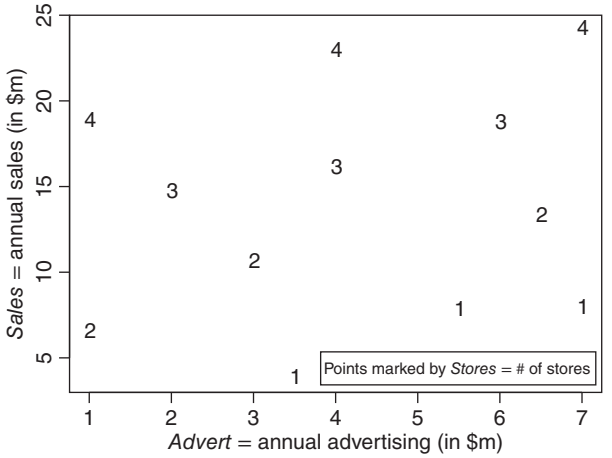

Figure 3.6 Scatterplot of simulated data with low correlation between $Sales$ and $Advert$, but a strong positive linear association between $Sales$ and $Advert$ when $Stores$ is fixed at a constant value.

similar story when $Stores$ is held constant at 3 or 4. The estimated regression parameter, $\hat{b}_1 = 1.0530$, represents the common slope of all these associations.

This simulated example demonstrates that low correlation between a response variable and a predictor variable does *not* imply that this predictor variable cannot be useful in a multiple linear regression model. Unfortunately, intuition about correlation can break down in the other direction also: High correlation between a response variable and a predictor variable does *not* imply that this predictor variable will necessarily be useful in a multiple linear regression model. For example, consider the simulated dataset represented by the scatterplot of $Sales$ versus $Trad$ in Figure 3.4, where $Sales$ represents annual sales in millions of dollars for a small high-tech business and $Trad$ represents annual spending on traditional advertising (TV, print media, etc.) in millions of dollars (**SALES3** data file).

The correlation between $Sales$ and $Trad$ here is very high (in fact, it is 0.986), but there is a second predictor, Int (annual spending on Internet advertising in millions of dollars), which results in $Trad$ apparently being a poor predictor of $Sales$ in the multiple linear regression model:

$$\mathrm{E}(Sales) = b_0 + b_1 Trad + b_2 Int.$$

Statistical software output for this model (see computer help #31) is

Parameters[a]

Model		Estimate	Standard error	t-Statistic	Pr(> |t|)
1	(Intercept)	1.992	0.902	2.210	0.054
	$Trad$	1.275	0.737	1.730	0.118
	Int	0.767	0.868	0.884	0.400

[a]Response variable: $Sales$.

The relatively large two-tail p-value for $Trad$ (0.118) means that any linear association between $Sales$ and $Trad$, holding Int constant, is weak. Similarly, the relatively large

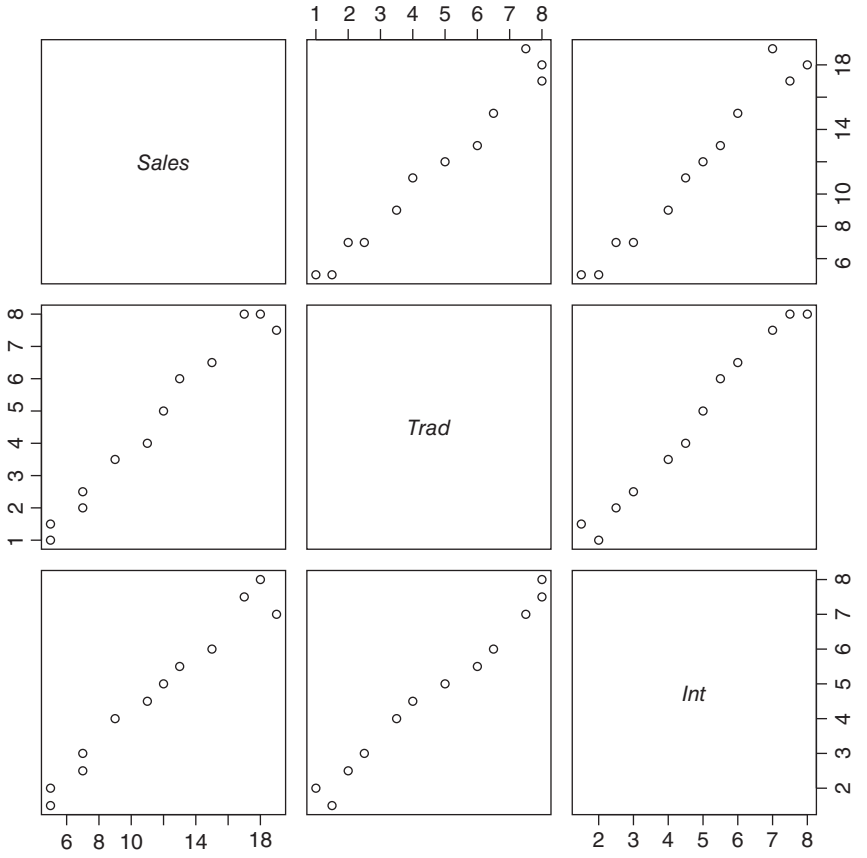

Figure 3.7 Scatterplot matrix for simulated data with high correlation between $Sales$ and $Trad$, but also high correlation between $Sales$ and Int and between $Trad$ and Int.

two-tail p-value for Int (0.400) means that any linear association between $Sales$ and Int, holding $Trad$ constant, is weak. Thus, the high correlation between $Sales$ and $Trad$ is irrelevant to the outcome of the multiple linear regression model. Figure 3.7 shows why this is the case (see computer help #16). Since $Trad$ and Int are highly correlated, when Int is held constant there is little variation in $Trad$ and hence relatively little variation possible in $Sales$. Thus, there is only a weak linear association between $Sales$ and $Trad$, holding Int constant. Similarly, with $Trad$ constant, there is little variation in Int or $Sales$. Thus, there is only a weak linear association between $Sales$ and Int, holding $Trad$ constant. This problem is known as *multicollinearity*, which we shall return to in Section 5.2.3. One way to address this problem is to drop one of the highly correlated predictors, for example, Int (with the larger p-value) in this case. Now, the significant linear association between $Sales$ and $Trad$ reveals itself:

Parameters[a]

Model		Estimate	Standard error	t-Statistic	Pr(> |t|)
2	(Intercept)	2.624	0.542	4.841	0.001
	$Trad$	1.919	0.104	18.540	0.000

[a]Response variable: $Sales$.

Another way to consider the role of an individual predictor, X_1 say, in a multiple linear regression model with response variable Y is to adjust for all the other predictor variables as follows. First regress Y on all the other predictor variables except X_1 and calculate the residuals from that model. Then regress X_1 on all the other predictor variables except X_1 and calculate the residuals from that model. The correlation between these two sets of residuals is called the *partial correlation* between Y and X_1. For example, the partial correlation between $Sales$ and $Advert$ for the **SALES2** data is 0.951 (recall that the ordinary correlation was just 0.165). Conversely, the partial correlation between $Sales$ and $Trad$ for the **SALES3** data is just 0.500 (recall that the ordinary correlation was 0.986). To illustrate partial correlation, we could construct a scatterplot of the two sets of residuals, known as an *added variable plot* or *partial regression plot*. For more details, see Weisberg (2013) and Cook (1996).

Predictor selection. We can use a global usefulness test to determine whether any of the potential predictors in a dataset are useful for modeling the response variable. Assuming that this is the case, we can then use nested model F-tests and individual regression parameter t-tests to identify the most important predictors. We should employ these tests judiciously to avoid conducting too many tests and reduce our chance of making a mistake (by excluding important predictors or failing to exclude unimportant ones). If possible, identification of the important predictors should be guided not just by the results of statistical tests but also by practical considerations and background knowledge about the application.

For some applications, the number of predictors, k, is very large, so determining which are the most important can be a challenging problem. Statistical software provides some automated methods for predictor selection in such cases. Examples include *forward selection* (predictors are added sequentially to an initial zero-predictor model in order of their individual significance), *backward elimination* (predictors are excluded sequentially from the full k-predictor model in order of their individual significance), and a combined *stepwise* method (which can proceed forwards or backwards at each stage). The "final" model selected depends on the particular method used and the model evaluation criterion used at each step. There are also alternative computer-intensive approaches to predictor selection that have been developed in the *machine learning* and *data mining* fields.

While automated predictor selection methods can be quick and easy to use, in applications with a manageable number of potential predictors (say, less than 10), manual selection of the important ones through practical considerations, background knowledge, and judiciously chosen hypothesis tests should usually lead to good results. In Section 5.3, we provide practical guidelines for implementing this approach. In larger applications with tens (or even hundreds) or potential predictors, automated methods can be useful for making an initial pass through the data to identify a smaller, more manageable set of potentially useful predictors. This smaller set can then be evaluated more carefully in the usual way.

We consider some of these ideas about predictor selection further in Section 5.4.

Optional—formula for regression parameter standard errors. Define the model matrix $\mathbf{X}$ as in Section 3.2. Then calculate the matrix $s^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, where s is the regression standard error. The square roots of the $k+1$ diagonal entries of this matrix are

the regression parameter standard errors, $s_{\hat{b}_0}, s_{\hat{b}_1}, s_{\hat{b}_2}, \ldots, s_{\hat{b}_k}$. Further details and examples are provided in Appendix E (www.wiley.com/go/pardoe/AppliedRegressionModeling3e).

3.4 MODEL ASSUMPTIONS

The multiple linear regression model relies on a number of assumptions being satisfied in order for it to provide a reliable approximation to the true association between a response variable, Y, and predictor variables, $(X_1, X_2, \ldots, X_k)$. These assumptions describe the probability distributions of the random errors in the model:

$$\text{random error} = e = Y - \text{E}(Y) = Y - b_0 - b_1 X_1 - \cdots - b_k X_k.$$

In particular, there are four assumptions about these random errors, e:

- The probability distribution of e at each set of values $(X_1, X_2, \ldots, X_k)$ has a **mean of zero** (in other words, the data points are balanced on both sides of the regression "hyperplane" so that the random errors average out to zero at each set of X-values).
- The probability distribution of e at each set of values $(X_1, X_2, \ldots, X_k)$ has **constant variance**, sometimes called *homoscedasticity* (in other words, the data points spread out evenly around the regression hyperplane so that the (vertical) variation of the random errors remains similar at each set of X-values).
- The probability distribution of e at each set of values $(X_1, X_2, \ldots, X_k)$ is **normal** (in other words, the data points are more likely to be closer to the regression hyperplane than further away and have a gradually decreasing chance of being further away).
- The value of e for one observation is **independent** of the value of e for any other observation (in other words, knowing the value of one random error gives us no information about the value of another).

Figure 2.13 illustrates these assumptions for simple linear regression. It is difficult to illustrate them for multiple regression, but we can check them in a similar way.

3.4.1 Checking the model assumptions

The model assumptions relate to the random errors in the population, so we are left with the usual statistical problem. Is there information in the sample data that we can use to ascertain what is likely to be going on in the population? One way to address this is to consider the estimated random errors from the multiple linear regression model fit to the sample data. We can calculate these estimated errors or *residuals* as follows:

$$\text{residual} = \hat{e} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1 X_1 - \cdots - \hat{b}_k X_k.$$

These numbers represent the distances between sample Y-values and fitted $\hat{Y}$-values on the corresponding regression hyperplane. We can construct *residual plots*, which are scatterplots with $\hat{e}$ along the vertical axis and a function of $(X_1, X_2, \ldots, X_k)$ along the horizontal axis. Examples of functions of $(X_1, X_2, \ldots, X_k)$ to put on the horizontal axis include

- the fitted $\hat{Y}$-values, that is, $\hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_k X_k$;

- each predictor variable in the model;
- potential predictor variables that have not been included in the model;
- a variable representing the order in which data values were observed if the data were collected over time (see also Section 5.2.2).

We can construct residual plots for each of these horizontal axis quantities—the more horizontal axis quantities we can assess, the more confidence we can have about whether the model assumptions have been satisfied. For example, suppose that we fit a two-predictor model in a dataset with three predictors. Then, we should construct four residual plots with different quantities on the horizontal axis: one with the fitted $\hat{Y}$-values, two with the predictors in the model, and one with the predictor that is not in the model. The reason for constructing a residual plot for the predictor that is not in the model is that such a plot can sometimes help determine whether that predictor really ought to be included in the model. We illustrate this in the **MLRA** example later in this section. We can then assess each plot by eye to see whether it is plausible that the four model assumptions described in Section 3.4 could hold *in the population*. Since this can be somewhat subjective, to help build intuition refer back to Figure 2.14, which displays residual plots generated from simulated populations in which the four model assumptions hold. By contrast, Figure 2.15 displays residual plots in which the four model assumptions fail. We can use each residual plot to assess the assumptions as follows:

- To assess the **zero mean** assumption, visually divide each residual plot into five or six vertical slices and consider the approximate average value of the residuals in each slice. The five or six within-slice averages should each be "close" to zero (the horizontal lines in Figures 2.14 and 2.15). We should seriously question the zero mean assumption only if some of the within-slice averages are clearly different from zero.
- To assess the **constant variance** assumption, again visually divide each residual plot into five or six vertical slices, but this time consider the spread of the residuals in each slice. The variation should be approximately the same within each of the five or six slices. We should seriously question the constant variance assumption only if there are clear changes in variation between some of the slices. For example, clear "fan/megaphone" or "funnel" shapes can indicate possible violation of the constant variance assumption—see the left-hand and middle residual plots in the second row of Figure 2.15 for examples of these patterns.
- The **normality** assumption is quite difficult to check with the "slicing" technique since there are usually too few residuals within each slice to assess normality for each. Instead, we can use histograms and QQ-plots (normal probability plots).
- To assess the **independence** assumption, take one final quick look at each residual plot. If any nonrandom patterns jump out at you, then the independence assumption may be in doubt. Otherwise, the independence assumption is probably satisfied.

While residual plots work well for assessing the zero mean, constant variance, and independence assumptions, histograms and QQ-plots are more useful for assessing normality. Refer back to Figure 2.16, which displays residual histograms that look sufficiently normal in the upper row of plots but that suggest violation of the normality assumption in the lower

row. Similarly, Figure 2.17 displays reasonably normal residual QQ-plots in the upper row of plots but nonnormal QQ-plots in the lower row.

Checking the multiple linear regression model assumptions graphically:

- Create a series of residual scatterplots with the model residuals on the vertical axis and various quantities on the horizontal axis. (Possible quantities include the fitted $\hat{Y}$-values, each predictor variable in the model, and any potential predictor variables that have not been included in the model.) Moving across each plot from left to right, the residuals should average close to zero (zero mean assumption), remain equally variable (constant variance assumption), and show no clear nonrandom patterns (independence assumption).
- If the data were collected over time, create a residual scatterplot with the model residuals on the vertical axis and a variable representing the order in which data values were observed on the horizontal axis. Moving across the plot from left to right, the residuals should show no clear nonrandom patterns (independence assumption).
- Create a histogram and QQ-plot (normal probability plot) of the model residuals. The histogram should look approximately normal (symmetric, bell-shaped) and in the QQ-plot the points should lie reasonably close to the diagonal line.

Multiple linear regression is reasonably robust to mild violations of the four assumptions. Although we should rely on model results only when we can be reasonably confident that the assumptions check out, we really need to worry only when there is a clear violation of an assumption. Recognizing clear violations can be challenging—one approach is to be concerned only if a pattern we see in a residual plot "jumps off the screen and slaps us in the face." If we find a clear violation, we can fit an alternative model (e.g., with a different subset of available predictors) and recheck the assumptions for this model. In Chapter 4, we introduce some additional strategies for dealing with such a situation (see also the remedies suggested in Section 5.1.1).

A graphical tool in statistical software can make the process of checking the zero mean assumption a little easier. Consider the simulated **MLRA** data file, in which Y depends potentially on three predictor variables: X_1, X_2, and X_3. Consider the following model first:

$$\text{Model 1}: \quad \mathrm{E}(Y) = b_0 + b_1 X_1 + b_2 X_2.$$

We can assess the four multiple linear regression model assumptions with the following:

- a residual plot with the fitted $\hat{Y}$-values on the horizontal axis (check zero mean, constant variance, and independence assumptions);
- residual plots with each predictor in turn (X_1, X_2, and X_3) on the horizontal axis (check zero mean, constant variance, and independence assumptions);
- a histogram and QQ-plot of the residuals (check normality assumption).

Most of these graphs (not shown here) lend support to the four assumptions, but Figure 3.8 shows the residual plot with X_3 on the horizontal axis, which indicates violation of the zero mean assumption. To help make this call, Figure 3.8a adds a *loess fitted line* to

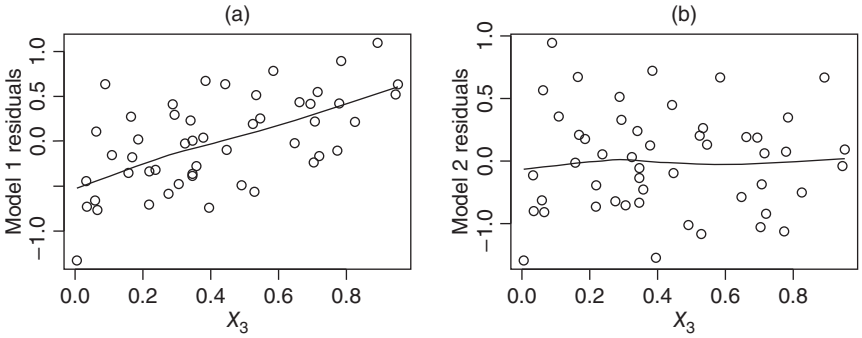

Figure 3.8 Residual plots for the **MLRA** example, with model 1 in (a) and model 2 in (b). X_3 is on the horizontal axes, a predictor that is not in model 1 but that is in model 2. The loess fitted line in plot (a) is sufficiently different from a horizontal line at zero to suggest that the zero mean assumption is violated for model 1. By contrast, the loess fitted line in plot (b) is sufficiently close to a horizontal line at zero to suggest that the zero mean assumption seems reasonable for model 2.

this residual plot (see computer help #36 in the software information files available from the book website). This line, essentially a computational method for applying the "slicing and averaging" smoothing technique described previously, is sufficiently different from a horizontal line at zero to violate the zero mean assumption for this model. Some statistical software uses a *lowess fitted line* instead of (or as an alternative to) a loess fitted line. Both types of line generally produce similar results, although technically they are based on slightly different smoothing methods.

Since X_3 is not in model 1, one possible remedy to try is to include X_3 in a new model:

$$\text{Model 2}: \quad \mathrm{E}(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3.$$

We can then assess the four assumptions for this model using the same set of graphs as for model 1 (but using the residuals for model 2 instead). Figure 3.8b shows the model 2 residual plot with X_3 on the horizontal axis. In contrast with the model 1 residuals, there is no clear positive trend in the model 2 residuals.

While the visual impression of a single graph can raise doubt about a model assumption, to have confidence in all four assumptions we need to consider all the suggested residual plots, histograms, and QQ-plots—see Figures 3.9 and 3.10. The four left-hand residual plots in Figure 3.9 include loess fitted lines, which should be reasonably flat to satisfy the zero mean assumption. For the four right-hand plots (similar to the left-hand plots except without loess fitted lines), the average vertical variation of the points should be reasonably constant across each plot to satisfy the constant variance assumption. The four right-hand plots should also have no clear nonrandom patterns to satisfy the independence assumption. In Figure 3.10, the histogram should be reasonably bell-shaped and symmetric, and the QQ-plot points should lie reasonably close to the line to satisfy the normality assumption.

Note that residual plots with a predictor on the horizontal axis are generally most useful for *quantitative* predictor variables. We will see in Section 4.3 how to incorporate *qualitative* predictor variables into a multiple linear regression model. Such variables have a small number (often two, three, or four) of categories or levels. We can use such variables in residual plots but must adapt our approach slightly. For example, a residual plot with

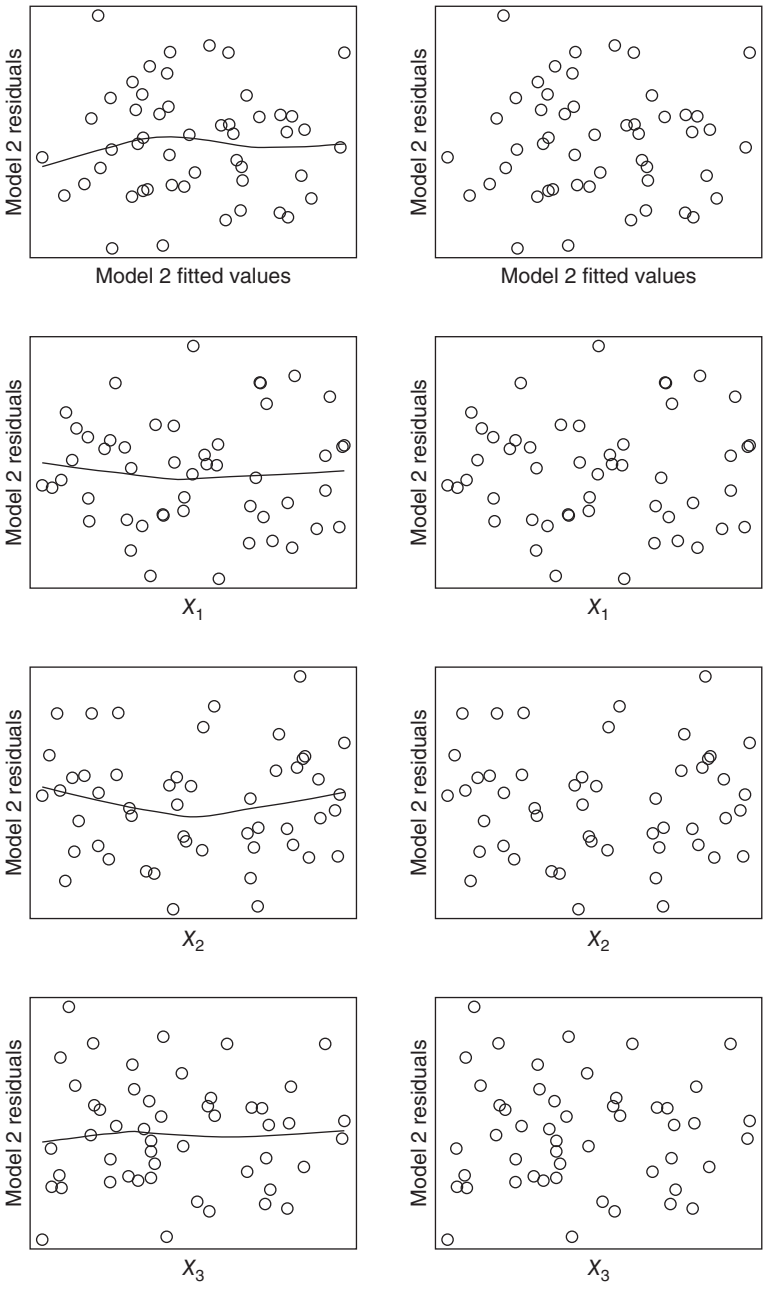

Figure 3.9 Model 2 residual plots for the **MLRA** example. Moving across each plot from left to right, the residuals appear to average close to zero and remain equally variable, providing support for the zero mean and constant variance assumptions. The lack of clear nonrandom patterns supports the independence assumption.

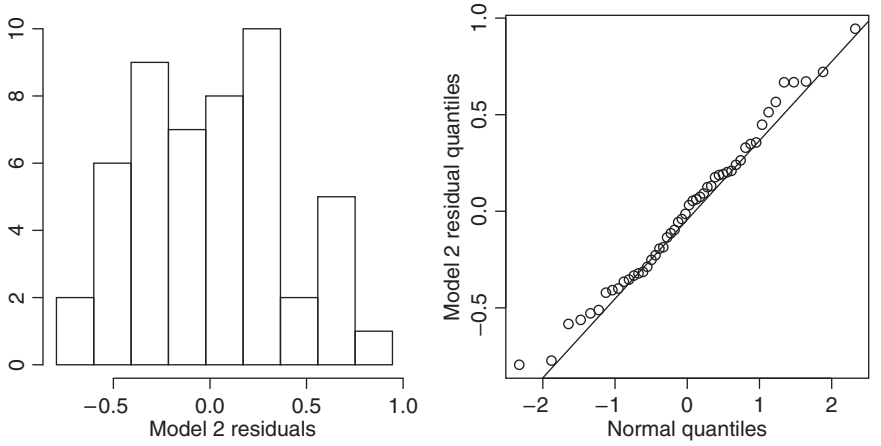

Figure 3.10 Histogram and QQ-plot of the model 2 residuals for the **MLRA** example. The approximately bell-shaped and symmetric histogram and QQ-plot points lying close to the line support the normality assumption.

the categories of a qualitative predictor variable on the horizontal axis can be useful if we "jitter" the points horizontally. This simply adds a small amount of random noise to the points to spread them out so that we can see them more easily. Otherwise, the points are likely to overlap one another, which makes it harder to see how many of them there are. Once we have done this, we can check the zero mean assumption by assessing whether the means of the residuals within each category are approximately zero. We can check the constant variance assumption by assessing whether the residuals are approximately equally variable within each category. To check the normality assumption, we can see whether the residuals are approximately normal within each category using histograms or QQ-plots (although this can be difficult to assess if there are small numbers of observations within the categories). Finally, it is difficult to check the independence assumption in residual plots with a qualitative predictor variable on the horizontal axis since the pattern of the discrete categories in these plots is so dominant.

For checking the zero mean and constant variance assumptions, an alternative to the jittered scatterplot with a qualitative predictor variable on the horizontal axis is to use boxplots of the residuals for the different categories of the qualitative predictor variable. Keep in mind, however, that boxplots work with medians and interquartile ranges rather than means and variances.

As discussed in Section 2.4.1 in relation to simple linear regression, a note of caution regarding residual plots relates to the sample size, n. If n is either very small (say, less than 20) or very large (say, in the thousands or more), then residual plots can become very difficult to use. One possible approach for the latter case is to take an appropriate random sample of, say, 100 residuals and then to use the methods of this section as normal. More comprehensive still would be to take a few such random samples and assess the assumptions for each sample.

In Section 5.2.1, we discuss the constant variance assumption in more detail and outline some remedies for situations where this assumption is in doubt. In Section 5.2.2, we cover the most common way in which the independence assumption can be violated, which is when there is autocorrelation (or serial correlation).

3.4.2 Testing the model assumptions

In many cases, visually assessing residual plots, histograms, and QQ-plots to check the four multiple linear regression assumptions is sufficient. However, in ambiguous cases or when we wish to complement our visual conclusions with quantitative evidence, there are a number of tests available that address each of the assumptions. A full discussion of these tests lies beyond the scope of this book, but brief details and references follow.

- One simple way to test the **zero mean** assumption is to see whether there is a significant linear trend in a residual plot in which the horizontal axis quantity is a potential predictor variable that has not been included in the model. In particular, we can conduct a t-test for that predictor variable in a simple linear regression model applied to the residual plot. For example, for the **MLRA** dataset, if we fit a simple linear regression model with the residuals from model 1 as the response variable and X_3 as the predictor variable, we obtain a significant p-value for X_3 of 0.000. This result confirms the visual impression of Figure 3.8a. If we try to apply this test to a residual plot in which the horizontal axis quantity is a predictor in the model, we will always obtain a nonsignificant p-value of essentially 1. We can illustrate this by fitting a simple linear regression model with the residuals from model 2 as the response variable and X_3 as the predictor variable, and observing the p-value for X_3 to be essentially 1.
- For residual plots in which the horizontal axis quantity is one of the predictors in the model, we can extend the previous test of the **zero mean** assumption by fitting a *quadratic model* (see Section 4.1.2) to the residual plot and applying a t-test to the squared term. For example, if we fit a multiple linear regression model with the residuals from model 2 as the response variable and two predictor variables, X_3 and X_3^2, we obtain a nonsignificant p-value for X_3^2 of 0.984. In other words, there is no evidence from this test that there is a quadratic trend in this residual plot. In practice, we probably would not have applied this test since that there is no suggestion of a quadratic trend in the plot. The test is most useful in cases where we suspect a significant nonlinear trend in a residual plot, such as those in the upper row of residual plots in Figure 2.15. When applying the test for a residual plot in which the horizontal axis quantity is the fitted values from the model, we use the standard normal distribution to conduct the test rather than a t-distribution and the test is called "Tukey's test for nonadditivity."
- For datasets with several observations having the same set of predictor values, called *replicates*, we can test the **zero mean** assumption using a "lack of fit" test of NH: the form of the specified regression model is correct (i.e., there is no lack of fit) versus AH: the form of the specified regression model is incorrect (i.e., there is lack of fit). Each subset of data replicates represents a subpopulation of observations with the same predictor values but potentially different Y-values. Suppose we have g such subpopulations in a sample of n observations. The sum of squares of the differences between the Y-values and their subpopulation means is called the *pure error sum of squares*, PSS. Subtracting this from the residual sum of squares (RSS) gives the *lack of fit sum of squares*. The *pure error degrees of freedom* is $n - g$. The *lack of fit degrees of freedom* is $g - k - 1$, where k is the number of predictor variables. The lack of fit test statistic is $\frac{(\text{RSS}-\text{PSS})/(g-k-1)}{\text{PSS}/(n-g)}$, which has an F-distribution with

$g-k-1$ numerator degrees of freedom and $n-g$ denominator degrees of freedom under NH. The lack of fit test is typically summarized in an expanded ANOVA table:

ANOVA[a]

Model		Sum of squares	df	Mean square	F-statistic	Pr(>F)
1	Regression	$\mathrm{TSS}-\mathrm{RSS}$	k	$(\mathrm{TSS}-\mathrm{RSS})/k$	$\frac{(\mathrm{TSS}-\mathrm{RSS})/k}{\mathrm{RSS}/(n-k-1)}$	p-value
	Residual	RSS	$n-k-1$	$\mathrm{RSS}/(n-k-1)$		
	Lack of fit	$\mathrm{RSS}-\mathrm{PSS}$	$g-k-1$	$(\mathrm{RSS}-\mathrm{PSS})/(g-k-1)$	$\frac{(\mathrm{RSS}-\mathrm{PSS})/(g-k-1)}{\mathrm{PSS}/(n-g)}$	p-value
	Pure error	PSS	$n-g$	$\mathrm{PSS}/(n-g)$		
	Total	TSS	$n-1$			

The lack of fit test statistic and *p*-value are in the row labeled "Lack of fit." We reject NH in favor of AH if the lack of fit test statistic is "large" or the corresponding p-value is less than the selected significance level (generally 5%); in such a case we should question the zero mean assumption. Otherwise, if the p-value is greater than or equal to the selected significance level, we fail to reject AH and conclude that the form of the specified regression model is reasonable and the zero mean assumption is likely satisfied. The lack of fit test is most useful in experimental applications in which replicates are part of the experimental design. The test can sometimes be used in observational applications when replicates arise by chance, particularly in simple linear regression settings. However, the chance of having sufficient replicates to conduct a lack of fit test lessens as more predictors are added to the model in multiple linear regression settings.

- The most common test for the **constant variance** assumption is the test independently developed by Breusch and Pagan (1979) and Cook and Weisberg (1983)—see Section 5.2.1 for further details.
- There are a variety of statistics available to test the **normality** assumption, including the Shapiro and Wilk (1965) W statistic (see also Royston, 1982a, 1982b, and 1995); the Anderson–Darling test (see Stephens, 1986; Thode, 2002, Sec. 5.1.4); the Cramer-von Mises test (see Stephens, 1986; Thode, 2002, Sec. 5.1.3); the Lilliefors (Kolmogorov–Smirnov) test (see Stephens, 1974; Dallal and Wilkinson, 1986; Thode, 2002, Sec. 5.1.1); the Pearson chi-square test (see Moore, 1986; Thode, 2002, Sec. 5.2); and the Shapiro–Francia test (see Royston, 1993; Thode, 2002, Sec. 2.3.2). Each test works in a similar way: reject NH (the regression errors have a normal distribution) in favor of AH (the regression errors do not have a normal distribution) if the test statistic is "large" or the p-value is less than the selected significance level (generally 5%). Otherwise (if the p-value is greater than or equal to the selected significance level), assume that the normality assumption is reasonable. Some statistical software can output normality test results when constructing a QQ-plot (normal probability plot) of the model residuals.
- A common way to test the **independence** assumption is to test for autocorrelation (or serial correlation) using such tests as the Wald and Wolfowitz (1940) runs test, the Durbin and Watson (1950, 1951, 1971) test, and the Breusch (1978) and Godfrey (1978) test—see Section 5.2.2 for further details.

Testing the multiple linear regression model assumptions numerically:

- Test for a significant linear trend in a residual plot with a potential predictor variable that has not been included in the model on the horizontal axis.
- Test for a significant quadratic trend in a residual plot with an included predictor variable or the model fitted values on the horizontal axis.
- If there are several observations with the same set of predictor values, conduct a lack of fit test.
- Assess the constant variance assumption by applying the test of Breusch and Pagan (1979) and Cook and Weisberg (1983).
- Apply a normality test such as the Anderson–Darling test to the residuals.
- Test the independence assumption using an autocorrelation test such as the Breusch and Godfrey test.

3.5 MODEL INTERPRETATION

Once we are satisfied that the four multiple linear regression assumptions seem plausible, we can interpret the model results. This requires relating the numerical information from the statistical software output back to the subject matter. For example, in the shipping dataset, we saw in Section 3.3.4 that the two-predictor model, $\mathrm{E}(Lab) = b_0 + b_1 Tws + b_3 Asw$, was preferable to the four-predictor model, $\mathrm{E}(Lab) = b_0 + b_1 Tws + b_2 Pst + b_3 Asw + b_4 Num$ (see computer help #34 in the software information files available from the book website):

Model summary

Model	R^2	Adjusted R^2	Regression standard error	Change Statistics			
				F-statistic	df1	df2	Pr(>F)
2[b]	0.8082	0.7857	8.815				
1[c]	0.8196	0.7715	9.103	0.472	2	15	0.633

[b]Predictors: (Intercept), Tws, Asw.
[c]Predictors: (Intercept), Tws, Pst, Asw, Num.

Checking regression assumptions with a sample size of 20 is challenging, but there are no clear violations in any residual plots (not shown here) for the two-predictor model. Statistical software output for this model (see computer help #31) is

	Parameters[a]				**95% Confidence interval**	
Model	Estimate	Standard error	t-Statistic	Pr(> |t|)	Lower bound	Upper bound
2 (Intercept)	110.431	24.856	4.443	0.000		
Tws	5.001	2.261	2.212	0.041	0.231	9.770
Asw	−2.012	0.668	−3.014	0.008	−3.420	−0.604

[a]Response variable: Lab.

Model summary

Model	Sample size	Multiple R^2	Adjusted R^2	Regression standard error
2[a]	20	0.8082	0.7857	8.815

[a]Predictors: (Intercept), Tws, Asw.

The corresponding practical interpretations of the results are as follows:

- There is no evidence at the 5% significance level that Pst (proportion shipped by truck) or Num (week number) provide useful information about the response, Lab (weekly labor hours), beyond the information provided by Tws (total weight shipped in thousands of pounds) and Asw (average shipment weight in pounds). (*Nested model test for the regression parameters*, b_2 *and* b_4.)
- There is a linear association between Lab and Tws, holding Asw constant, that is statistically significant at the 5% significance level. (*Hypothesis test for the regression parameter*, b_1.)
- There is a linear association between Lab and Asw, holding Tws constant, that is statistically significant at the 5% significance level. (*Hypothesis test for the regression parameter*, b_3.)
- We expect weekly labor hours to increase by 5.00 for each 1,000-pound increase in total weight shipped when average shipment weight remains constant (for total shipment weights of 2,000–10,000 pounds and average shipment weights of 10–30 pounds). To express our uncertainty due to sampling variation, we could say that we are 95% confident that labor hours increase by between 0.23 and 9.77 for each 1,000-pound increase in total weight shipped when average shipment weight remains constant. (*Point estimate and confidence interval for the regression parameter*, b_1.)
- We expect weekly labor hours to decrease by 2.01 for each 1-pound increase in average shipment weight when total weight shipped remains constant (for total shipment weights of 2,000–10,000 pounds and average shipment weights of 10–30 pounds). To express our uncertainty due to sampling variation, we could say that we are 95% confident that labor hours decrease by between 0.60 and 3.42 for each 1-pound increase in average shipment weight when total weight shipped remains constant. (*Point estimate and confidence interval for the regression parameter*, b_3.)
- If we use a multiple linear regression model to predict weekly labor hours from potential total weight shipped and average shipment weight values, we can expect to be accurate to within approximately ± 17.6 (at a 95% confidence level). (*Regression standard error*, s.)
- 80.8% of the variation in weekly labor hours (about its mean) can be explained by a multiple linear regression association between labor hours and (total weight shipped, average shipment weight). (*Coefficient of determination*, R^2.)

3.6 ESTIMATION AND PREDICTION

As with simple linear regression, there is a distinction between a confidence interval for the population mean, $\mathrm{E}(Y)$, at particular values of the predictor variables, $(X_1, X_2, \ldots, X_k)$,

and a prediction interval for an individual Y-value at those same values of the predictor variables, $(X_1, X_2, \ldots, X_k)$.

3.6.1 Confidence interval for the population mean, E(Y)

Consider estimating the mean (or expected) value of Y at particular values of the predictor variables, $(X_1, X_2, \ldots, X_k)$, based on a multiple linear regression association between Y and $(X_1, X_2, \ldots, X_k)$. Since we have estimated the association to be $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_k X_k$, our best point estimate for $\mathrm{E}(Y)$ is $\hat{Y}$. For example, suppose that for the two-predictor model for the shipping dataset we would like to estimate the average level of weekly labor hours corresponding to total weight shipped of 6,000 pounds and average shipment weight of 20 pounds. Our best point estimate for $\mathrm{E}(Lab)$ at $Tws = 6$ and $Asw = 20$ is $\widehat{Lab} = 110.431 + 5.001 \times 6 - 2.012 \times 20 = 100.2$.

How sure are we about this answer? One way to express our uncertainty is with a confidence interval. For example, a 95% confidence interval for $\mathrm{E}(Y)$ results from the following:

$$\Pr\left(-97.5\text{th percentile} < t_{n-k-1} < 97.5\text{th percentile}\right) = 0.95,$$

$$\Pr\left(-97.5\text{th percentile} < \frac{\hat{Y} - \mathrm{E}(Y)}{s_{\hat{Y}}} < 97.5\text{th percentile}\right) = 0.95,$$

$$\Pr\left(\hat{Y} - 97.5\text{th percentile}\ (s_{\hat{Y}}) < \mathrm{E}(Y) < \hat{Y} + 97.5\text{th percentile}\ (s_{\hat{Y}})\right) = 0.95,$$

where $s_{\hat{Y}}$ is the *standard error of estimation* for the multiple linear regression mean, and the 97.5th percentile comes from the t-distribution with $n - k - 1$ degrees of freedom. In other words, the 95% confidence interval for $\mathrm{E}(Y)$ can be written as $\hat{Y} \pm 97.5\text{th percentile}\ (s_{\hat{Y}})$.

We can use statistical software to calculate $s_{\hat{Y}}$ for particular X-values that we might be interested in, or just use the software to calculate the confidence interval for $\mathrm{E}(Y)$ directly. For interested readers, a formula for $s_{\hat{Y}}$ is provided at the end of this section. The formula shows that $s_{\hat{Y}}$ tends to be smaller (and our estimates more accurate) when n is large, when the particular X-values we are interested in are close to their sample means, and when the regression standard error, s, is small. Also, of course, a lower level of confidence leads to a narrower confidence interval for $\mathrm{E}(Y)$—for example, a 90% confidence interval will be narrower than a 95% confidence interval (all else being equal).

Returning to the shipping dataset, $s_{\hat{Y}} = 2.293$ for $Tws = 6$ and $Asw = 20$ (see following statistical software output), so that the 95% confidence interval for $\mathrm{E}(Lab)$ at $Tws = 6$ and $Asw = 20$ is

$$\begin{aligned}\widehat{Lab} \pm 97.5\text{th percentile}\ (s_{\hat{Y}}) &= 100.2 \pm 2.110 \times 2.293 \\ &= 100.2 \pm 4.838 \\ &= (95.4,\ 105.0).\end{aligned}$$

The 97.5th percentile, 2.110, comes from the t-distribution with $n - k - 1 = 17$ degrees of freedom (see computer help #8 in the software information files available from the book website). The relevant statistical software output (see computer help #29) is

Lab	Tws	Asw	$\widehat{Lab}$	$s_{\hat{Y}}$	CI-low	CI-up
—	6	20	100.192	2.293	95.353	105.031

The point estimate for the population mean, $\mathrm{E}(Lab)$, at particular predictor values is denoted as "$\widehat{Lab}$," while the standard error of estimation is denoted as "$s_{\hat{Y}}$," and the confidence interval goes from "CI-low" to "CI-up."

Now that we have calculated a confidence interval, what exactly does it tell us? Well, for this shipping example, *loosely speaking*, we can say "we are 95% confident that expected weekly labor hours are between 95.4 and 105.0 when total weight shipped is 6,000 pounds and average shipment weight is 20 pounds." To provide a more precise interpretation we would have to say something like "if we were to take a large number of random samples of size 20 from our population of shipping numbers and calculate a 95% confidence interval for $\mathrm{E}(Lab)$ at $Tws = 6$ and $Asw = 20$ in each, then 95% of those confidence intervals would contain the true (unknown) population mean."

For a lower or higher level of confidence than 95%, the percentile used in the calculation must be changed as appropriate. For example, for a 90% interval (i.e., with 5% in each tail), the 95th percentile would be needed, whereas for a 99% interval (i.e., with 0.5% in each tail), the 99.5th percentile would be needed. These percentiles can be obtained using computer help #8. As further practice, calculate a 90% confidence interval for $\mathrm{E}(Lab)$ at $Tws = 6$ and $Asw = 20$ for the shipping example (see Problem 3.6)—you should find that it is (96.2, 104.2).

A confidence interval for the population mean, $\mathrm{E}(Y)$, at a particular set of X-values in multiple linear regression is

$$\hat{Y} \pm \text{t-percentile } (s_{\hat{Y}}),$$

where $\hat{Y}$ is the fitted (or predicted) value of Y at the specified X-values, $s_{\hat{Y}}$ is the standard error of estimation at the specified X-values, and the t-percentile comes from a t-distribution with $n - k - 1$ degrees of freedom (n is the sample size, k is the number of predictor variables).

Suppose we have calculated a 95% confidence interval for the population mean, $\mathrm{E}(Y)$, at a particular set of X-values to be (a, b). Then we can say that we are 95% confident that $\mathrm{E}(Y)$ is between a and b at the specified X-values.

3.6.2 Prediction interval for an individual Y-value

Now, by contrast, consider predicting an individual Y-value at particular values of the predictor variables, $(X_1, X_2, \ldots, X_k)$, based on a multiple linear regression association between Y and $(X_1, X_2, \ldots, X_k)$. To distinguish a prediction from an estimated population mean, $\mathrm{E}(Y)$, we will call this Y-value to be predicted Y^*. Just as with estimating $\mathrm{E}(Y)$, our best point estimate for Y^* is $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_k X_k$. For example, suppose that for the shipping dataset we would like to predict the actual level of weekly labor hours corresponding to total weight shipped of 6,000 pounds and average shipment

weight of 20 pounds. Our best point estimate for Lab^* at $Tws = 6$ and $Asw = 20$ is $\widehat{Lab} = 110.431 + 5.001 \times 6 - 2.012 \times 20 = 100.2$.

How sure are we about this answer? One way to express our uncertainty is with a prediction interval (like a confidence interval, but for a prediction rather than an estimated population mean). For example, a 95% prediction interval for Y^* results from the following:

$$\Pr\left(-97.5\text{th percentile} < t_{n-k-1} < 97.5\text{th percentile}\right) = 0.95,$$

$$\Pr\left(-97.5\text{th percentile} < \frac{\hat{Y}^* - Y^*}{s_{\hat{Y}^*}} < 97.5\text{th percentile}\right) = 0.95,$$

$$\Pr\left(\hat{Y}^* - 97.5\text{th percentile}\ (s_{\hat{Y}^*}) < Y^* < \hat{Y}^* + 97.5\text{th percentile}\ (s_{\hat{Y}^*})\right) = 0.95,$$

where $s_{\hat{Y}^*}$ is the *standard error of prediction* for the multiple linear regression response, and the 97.5th percentile comes from the t-distribution with $n - k - 1$ degrees of freedom. In other words, the 95% prediction interval for Y^* can be written $\hat{Y}^* \pm 97.5\text{th percentile}\ (s_{\hat{Y}^*})$.

We can use statistical software to calculate $s_{\hat{Y}^*}$ for particular X-values that we might be interested in, or just use the software to calculate the prediction interval for Y^* directly. For interested readers, a formula for $s_{\hat{Y}^*}$ is provided at the end of this section. The formula shows that $s_{\hat{Y}^*}$ is always larger than $s_{\hat{Y}}$ (from Section 3.6.1) for any particular set of X-values. This makes sense because it is more difficult to predict an individual Y-value at a particular set of X-values than to estimate the mean of the population distribution of Y at those same X-values. Consider the following illustrative example. Suppose that the business for the shipping dataset plans to ship 6,000 pounds with an average shipment weight of 20 pounds each week over the next quarter. Estimating the average weekly labor hours over the quarter is easier than predicting the actual weekly labor hours in any individual week. In other words, our uncertainty about an individual prediction is always larger than our uncertainty about estimating a population mean, and $s_{\hat{Y}^*} > s_{\hat{Y}}$.

For the shipping dataset, $s_{\hat{Y}^*} = 9.109$ for $Tws = 6$ and $Asw = 20$, and the 95% prediction interval for Lab^* is

$$\begin{aligned}\widehat{Lab}^* \pm 97.5\text{th percentile}\ (s_{\hat{Y}^*}) &= 100.2 \pm 2.110 \times 9.109 \\ &= 100.2 \pm 19.220 \\ &= (81.0,\ 119.4).\end{aligned}$$

The 97.5th percentile, 2.110, comes from the t-distribution with $n - k - 1 = 17$ degrees of freedom (see computer help #8 in the software information files available from the book website). The relevant statistical software output (see computer help #30) is

Lab	Tws	Asw	$\widehat{Lab}$	PI-low	PI-up
—	6	20	100.192	80.974	119.410

The point estimate for the prediction, Lab^*, at particular predictor values is denoted as "$\widehat{Lab}$," while the prediction interval goes from "PI-low" to "PI-up."

Now that we have calculated a prediction interval, what does it tell us? For this shipping example, *loosely speaking*, "we are 95% confident that actual labor hours in a week are between 81.0 and 119.4 when total weight shipped is 6,000 pounds and average shipment weight is 20 pounds." A more precise interpretation would have to say something like "if we were to take a large number of random samples of size 20 from our population of shipping numbers and calculate a 95% prediction interval for Lab^* at $Tws = 6$ and $Asw = 20$ in each, then 95% of those prediction intervals would contain the true (unknown) labor hours for an individual week picked at random when $Tws = 6$ and $Asw = 20$."

As with the standard error of estimation, $s_{\hat{Y}}$, the standard error of prediction, $s_{\hat{Y}^*}$, tends to be smaller (and our predictions more accurate) when n is large, when the particular X-values we are interested in are close to their sample means, and when the regression standard error, s, is small. Also, of course, a lower level of confidence leads to a narrower prediction interval for Y^*—for example, a 90% prediction interval will be narrower than a 95% prediction interval (all else being equal).

For a lower or higher level of confidence than 95%, the percentile used in the calculation must be changed as appropriate. For example, for a 90% interval (i.e., with 5% in each tail), the 95th percentile would be needed, whereas for a 99% interval (i.e., with 0.5% in each tail), the 99.5th percentile would be needed. These percentiles can be obtained using computer help #8. As further practice, calculate a 90% prediction interval for Lab^* at $Tws = 6$ and $Asw = 20$ for the shipping example (see Problem 3.6); you should find that it is (84.4, 116.1).

A prediction interval for an individual Y-value at a particular set of X-values in multiple linear regression is

$$\hat{Y} \pm \text{t-percentile } (s_{\hat{Y}^*}),$$

where $\hat{Y}$ is the fitted (or predicted) value of Y at the specified X-values, $s_{\hat{Y}^*}$ is the standard error of prediction at the specified X-values, and the t-percentile comes from a t-distribution with $n - k - 1$ degrees of freedom (n is the sample size, k is the number of predictor variables).

Suppose we have calculated a 95% prediction interval for an individual Y-value at a particular set of X-values to be (a, b). Then we can say that we are 95% confident that the individual Y-value is between a and b at the specified X-values.

One final note on prediction intervals. The "$\pm 2s$" interpretation we discussed for the regression standard error in Section 3.3.1 is based on an approximation of a 95% prediction interval for datasets with a large sample size, n. For sufficiently large n, $s_{\hat{Y}^*}$ is approximately equal to s, while the 97.5th percentile from the t-distribution with $n - k - 1$ degrees of freedom is close to 2. Thus, the 95% prediction interval for Y^* can be written approximately as $\hat{Y}^* \pm 2s$.

As for simple linear regression, we have now covered four different types of standard error in the context of multiple linear regression:

- the regression standard error, s, which is an estimate of the standard deviation of the error term in the model (Section 3.3.1)
- the standard errors for the regression parameter estimates, $s_{\hat{b}_0}, s_{\hat{b}_1}, s_{\hat{b}_2}, \ldots, s_{\hat{b}_k}$ (Section 3.3.5)

- the standard error of estimation for the multiple linear regression mean, $s_{\hat{Y}}$ (Section 3.6.1)
- the standard error of prediction for the multiple linear regression response, $s_{\hat{Y}^*}$ (Section 3.6.2)

Optional—formulas for standard errors of estimation and prediction. Define the model matrix $\mathbf{X}$ as in Section 3.2 and write the particular X-values we are interested in as a vector, $\mathbf{x}$ (including a one as the first entry to represent the intercept term). Then the standard error of estimation for the multiple linear regression mean at $\mathbf{x}$ is $s_{\hat{Y}} = s\sqrt{\mathbf{x}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}}$, where s is the regression standard error. The standard error of prediction for the multiple linear regression response at $\mathbf{x}$ is $s_{\hat{Y}^*} = s\sqrt{1 + \mathbf{x}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}}$. The two standard errors are linked by the formula, $s_{\hat{Y}^*} = \sqrt{s^2 + s_{\hat{Y}}^2}$, so if we know s (the regression standard error) and $s_{\hat{Y}}$ (the standard error of estimation), we can easily calculate $s_{\hat{Y}^*}$ (the standard error of prediction). Further details and examples are provided in Appendix E (www.wiley.com/go/pardoe/AppliedRegressionModeling3e).

3.7 CHAPTER SUMMARY

The major concepts that we covered in this chapter relating to multiple linear regression are as follows:

Multiple linear regression allows us to model the association between a response variable, Y, and predictor variables, $(X_1, X_2, \ldots, X_k)$, as $\mathrm{E}(Y) = b_0 + b_1X_1 + \cdots + b_kX_k$.

The method of least squares provides an estimated regression equation, $\hat{Y} = \hat{b}_0 + \hat{b}_1X_1 + \cdots + \hat{b}_kX_k$, by minimizing the residual sum of squares (residuals are the differences between the observed Y-values and the fitted $\hat{Y}$-values).

The estimated intercept represents the expected Y-value when $X_1 = X_2 = \cdots = X_k = 0$ (if such values are meaningful and fall within the range of X-values in the sample dataset)—denoted $\hat{b}_0$.

The estimated regression parameters represent expected changes in Y for a unit change in X_p, holding the other X's constant (over the range of X-values in the sample dataset)—denoted $\hat{b}_p$ $(p = 1, \ldots, k)$.

The regression standard error, s, is an estimate of the standard deviation of the random errors. One way to interpret s is to calculate $2s$ and say that when using the multiple linear regression model to predict Y from $(X_1, X_2, \ldots, X_k)$, we can expect to be accurate to within approximately $\pm 2s$ (at a 95% confidence level).

The coefficient of determination, R^2, represents the proportion of variation in Y (about its sample mean) explained by a multiple linear regression association between Y and $(X_1, X_2, \ldots, X_k)$. This is also equivalent to the square of multiple R, the correlation between the observed Y-values and the fitted $\hat{Y}$-values. Adjusted R^2 is a variant on R^2, which takes into account the number of predictor variables in a model to facilitate model comparisons.

Hypothesis testing provides a means of making decisions about the likely values of the regression parameters, $b_1, b_2, \ldots, b_k$. The magnitude of the calculated sample

test statistic indicates whether we can reject a null hypothesis in favor of an alternative hypothesis. Roughly speaking, the p-value summarizes the hypothesis test by representing the weight of evidence for the null hypothesis (i.e., small values favor the alternative hypothesis):

A global usefulness test has null hypothesis $b_1 = b_2 = \cdots = b_k = 0$ and tests whether *any* of the predictors have a linear association with Y.

Nested model tests have null hypothesis $b_{r+1} = b_{r+2} = \cdots = b_k = 0$ (i.e., a subset of the parameters set equal to zero) and test whether the corresponding *subset* of predictors have a linear association with Y, once the linear association with the other r predictors has been accounted for.

Individual tests have null hypothesis $b_p = 0$ and test whether an *individual* predictor has a linear association with Y, once the other $k - 1$ predictors have been accounted for.

Confidence intervals are another method for calculating the sample estimate of a population regression parameter and its associated uncertainty:

$$\hat{b}_p \pm \text{t-percentile}(s_{\hat{b}_p}).$$

Model assumptions should be satisfied before we can rely on the multiple linear regression results. These assumptions relate to the random errors in the model: The probability distributions of the random errors at each set of values of $(X_1, X_2, \ldots, X_k)$ should be normal with zero means and constant variances, and each random error should be independent of every other. We can check whether these assumptions seem plausible by calculating residuals (estimated errors) and visually assessing residual plots, histograms, and QQ-plots.

Confidence intervals are used for presenting sample estimates of the population mean, $\text{E}(Y)$, at particular X-values and their associated uncertainty:

$$\hat{Y} \pm \text{t-percentile}(s_{\hat{Y}}).$$

Prediction intervals, while similar in spirit to confidence intervals tackle the different problem of predicting individual Y-values at particular X-values:

$$\hat{Y}^* \pm \text{t-percentile}(s_{\hat{Y}^*}).$$

PROBLEMS

"Computer help" refers to the numbered items in the software information files available from the book website. There are *brief* answers to the even-numbered problems in Appendix F (www.wiley.com/go/pardoe/AppliedRegressionModeling3e).

3.1 The **MOVIES** data file contains data on 25 movies from "The Internet Movie Database" (www.imdb.com). Based on this dataset, we wish to investigate whether all-time US box office receipts (Box, in millions of US dollars unadjusted for inflation) are associated with any of the following variables:

$Rate$ = Internet Movie Database user rating (out of 10)
$User$ = Internet Movie Database users rating the movie (in thousands)
$Meta$ = "metascore" based on 35 critic reviews (out of 100)
Len = runtime (in minutes)
Win = award wins
Nom = award nominations

Theatrical box office receipts (movie ticket sales) may include theatrical re-release receipts, but exclude video rentals, television rights, and other revenues.

(a) Write out a regression equation (like the equation at the end of Section 3.1) for a multiple linear regression model for predicting response Box from just three predictors: $Rate$, $User$, and $Meta$.
Hint: This question is asking you to write an equation for $\mathrm{E}(Box)$.

(b) Use statistical software to fit this model [computer help #31] and write out the estimated multiple linear regression equation [i.e., replace the b's in part (a) with numbers].
Hint: This question is asking you to write an equation for $\widehat{Box}$.

(c) Interpret the estimated regression parameter for $Rate$ in the context of the problem [i.e., put the appropriate number from part (b) into a meaningful sentence, remembering to include the correct units for any variables that you use in your answer].
Hint: See the fourth or fifth bullet point in Section 3.5 for an example of the type of sentence expected.

3.2 Consider the shipping example in the **SHIPDEPT** data file, introduced in Section 3.3.2.

(a) As suggested in Section 3.3.5, calculate a 90% confidence interval for the population regression parameter b_1 in the four-predictor multiple linear regression model of Lab (weekly labor hours) on Tws (total weight shipped in thousands of pounds), Pst (proportion shipped by truck), Asw (average shipment weight in pounds), and Num (week number). Recall that the estimated regression parameter for this model is $\hat{b}_1 = 6.074$, while the standard error of this estimate is $s_{\hat{b}_1} = 2.662$.

(b) Also calculate a 90% confidence interval for the population regression parameter b_1 in the *two*-predictor multiple linear regression model of Lab on Tws and Asw. Recall that the estimated regression parameter for this model is $\hat{b}_1 = 5.001$, while the standard error of the estimate is $s_{\hat{b}_1} = 2.261$. Compare the width of the resulting interval with the width of the interval in part (a). Which is narrower, and why?

3.3 Six students in a statistics course had the following scores for their midterm and final exams:

Midterm1	100	85	95	80	90	70
Midterm2	90	95	75	90	70	80
Final	90	94	78	74	67	62

A multiple linear regression model with response variable $Final$ and predictor variables $Midterm1$ and $Midterm2$ resulted in the least squares estimated regression equation:

$$\widehat{Final} = -65.218 + 0.723 Midterm1 + 0.960 Midterm2.$$

Answer the following questions using hand calculations. You can check your answers using statistical software.

(a) Use the least squares estimated regression equation to calculate the predicted (fitted) values, $\widehat{Final}$, for each student.

(b) Calculate the residuals, $\hat{e} = Final - \widehat{Final}$, for each student.

(c) Calculate the residual sum of squares, $\text{RSS} = \sum_{i=1}^{n} \hat{e}_i^2$.

(d) Calculate the residual standard error, $s = \sqrt{\frac{\text{RSS}}{n-3}}$.

(e) Calculate the sample mean of $Final$, m_{Final}, and use this to calculate the difference, $Final - m_{Final}$, for each student.

(f) Calculate the total sum of squares, $\text{TSS} = \sum_{i=1}^{n} (Final_i - m_{Final})^2$.

(g) Calculate the coefficient of determination, $\text{R}^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$.

3.4 Doctors assessed 50 patients with a particular disease on three risk factors. The variables $Disease$, $Risk1$, $Risk2$, and $Risk3$ measure the prevalence of the disease and each of the three risk factors with a number between 0 and 10. A multiple linear regression model with response variable $Disease$ and predictor variables $Risk1$, $Risk2$, and $Risk3$ resulted in the following statistical software output:

Parameters[a]

Model		Estimate	Standard error	t-Statistic	Pr(> |t|)
1	(Intercept)	−0.679	0.841	−0.808	0.423
	Risk1	0.406	0.145	2.799	0.007
	Risk2	0.379	0.169	2.245	0.030
	Risk3	0.112	0.173	0.647	0.521

[a]Response variable: *Disease*.

Model summary

Model	Sample size	Multiple R^2	Adjusted R^2	Regression standard error
1[a]	50	0.5635	0.5350	1.459

[a]Predictors: (Intercept), *Risk1*, *Risk2*, *Risk3*.

ANOVA[a]

Model		Sum of squares	df	Mean square	Global F-statistic	Pr(>F)
1	Regression	126.459	3	42.153	19.794	0.000[b]
	Residual	97.961	46	2.130		
	Total	224.420	49			

[a]Response variable: *Disease*.
[b]predictors: (Intercept), *Risk1*, *Risk2*, *Risk3*.

Use the output to fill in the missing numbers in the following statements. In some cases, you can write down the relevant numbers directly, while in others, you may need to do a brief calculation.

(a) The sample size, n, for the dataset is ...

(b) The number of predictor variables, k, is ...

(c) The number of regression parameters, $k + 1$, in the multiple linear regression model is ...

(d) The global usefulness F-statistic for testing whether at least one of the risk factors is linearly associated with the disease is calculated as $\ldots / \ldots = \ldots$

(e) The p-value for testing whether at least one of the risk factors is linearly associated with the disease is ...

(f) The individual t-statistic for testing whether the first risk factor is linearly associated with the disease, after taking into account the other two risk factors, is calculated as $\ldots / \ldots = \ldots$

(g) The p-value for testing whether the first risk factor is linearly associated with the disease, after taking into account the other two risk factors, is ...

(h) The mean square error for the fitted model is calculated as $\ldots / \ldots = \ldots$

(i) The regression standard error, s, is calculated as $\sqrt{\ldots} = \ldots$

(j) Loosely speaking, the observed disease scores are, on average, ... units away from the model-based predicted values of $Disease$.

(k) The coefficient of determination, R^2, is calculated as $1 - \ldots / \ldots = \ldots$

(l) ...% of the sample variation in disease scores about their mean can be "explained" by the multiple linear regression model with predictors $Risk1$, $Risk2$, and $Risk3$.

(m) The least squares estimated regression equation is $\widehat{Disease} = \ldots + \ldots Risk1 + \ldots Risk2 + \ldots Risk3$.

(n) When $Risk1$ and $Risk3$ are held constant, we expect $Disease$ to increase by ... unit(s) when $Risk2$ increases by ... unit(s).

(o) The predicted disease score for a patient with risk factors of $Risk1 = 7$, $Risk2 = 5$, and $Risk3 = 6$ is ...

3.5 Consider the **MOVIES** data file from Problem 3.1 again.

(a) Use statistical software to fit the following (complete) model for Box as a function of all six predictor variables [computer help #31]:

$$\mathrm{E}(Box) = b_0 + b_1 Rate + b_2 User + b_3 Meta + b_4 Len + b_5 Win + b_6 Nom.$$

Write down the residual sum of squares for this model.

(b) Use statistical software to fit the following (reduced) model [computer help #31]:

$$\mathrm{E}(Box) = b_0 + b_1 Rate + b_2 User + b_3 Meta.$$

[This is the model from Problem 3.1 part (b).] Write down the residual sum of squares for this model.

(c) Using the results from parts (a) and (b) together with the nested model test F-statistic formula in Section 3.3.4, test the null hypothesis NH: $b_4 = b_5 = b_6 = 0$

in the complete model, using significance level 5%. Write out all the hypothesis test steps and interpret the result in the context of the problem.
Hint: To solve this part you may find the following information useful. The 95th percentile of the F-distribution with 3 numerator degrees of freedom and 18 denominator degrees of freedom is 3.160.

(d) Check your answer for part (c) by using statistical software to do the nested model test directly [computer help #34]. State the values of the F-statistic and the p-value, and draw an appropriate conclusion.

(e) Another way to see whether we should prefer the reduced model for this example is to see whether the regression standard error (s) is smaller for the reduced model than for the complete model and whether adjusted R^2 is higher for the reduced model than for the complete model. Confirm whether these relationships hold in this example (i.e., compare the values of s and adjusted R^2 in the reduced and complete models).

3.6 Consider the shipping example in the **SHIPDEPT** data file from Problem 3.2 again.

(a) As suggested in Section 3.6.1, calculate a 90% confidence interval for $E(Lab)$ at $Tws = 6$ and $Asw = 20$ in the two-predictor multiple linear regression model of Lab on Tws and Asw. Recall that the estimated regression parameters for this model are $\hat{b}_0 = 110.431$, $\hat{b}_1 = 5.001$, and $\hat{b}_3 = -2.012$, and the standard error of estimation at $Tws = 6$ and $Asw = 20$ is $s_{\hat{Y}} = 2.293$.

(b) As suggested in Section 3.6.2, calculate a 90% prediction interval for Lab^* at $Tws = 6$ and $Asw = 20$ in the two-predictor multiple linear regression model. Use the estimated regression parameters from part (a) and recall that the standard error of prediction at $Tws = 6$ and $Asw = 20$ is $s_{\hat{Y}^*} = 9.109$.

3.7 De Rose and Galarza (2000) used multiple linear regression to study Att = average attendance in thousands, from the first few years of Major League Soccer (MLS, the professional soccer league in the United States). The 12 MLS teams at the time ranged in average attendance from 10,000 to 22,000 per game. De Rose and Galarza used the following predictor variables:

Pop	=	total population of metropolitan area within 40 miles (millions)
$Teams$	=	number of (male) professional sports teams in the four major sports
$Temp$	=	average temperature (April–September,°F)

The regression results reported in the study were

Predictor variable	Parameter estimate	Two tail p-value
Intercept	28.721	0.001
Pop	1.350	0.001
$Teams$	−0.972	0.037
$Temp$	−0.238	0.012

(a) Write out the estimated least squares (regression) equation for predicting Att from Pop, $Teams$, and $Temp$.

(b) R^2 was 91.4%, suggesting that this model may be useful for predicting average attendance (for expansion teams, say). Test the global usefulness of the model using a significance level of 5%.
Hint: You will need to use the second formula for the global F-statistic in Section 3.3.3 to solve this part. Also, you may find the following information useful: the 95th percentile of the F-distribution with 3 numerator degrees of freedom and 8 denominator degrees of freedom is 4.07.

(c) Test, at a 5% significance level, whether the regression parameter estimate for $Teams$ suggests that increasing the number of (male) professional sports teams in the four major sports (football, baseball, basketball, and hockey) in a city is associated with a decrease in average MLS attendance in that city (all else being equal).
Hint: You will need to do a lower-tail hypothesis test using the p-value method, but be careful because the p-values given in the table are *two-tailed*.

(d) According to the model results, how much does average attendance differ for two cities with the same population and average temperature when one city has one fewer (male) professional sports teams in the four major sports?
Hint: Write out the equation from part (a) for predicted average attendance in thousands for one city (plug in $Teams = 1$, say) and then do the same for the other city (plug in $Teams = 2$). The difference between the two equations gives you the answer to the problem. You should find that as long as you plug in values for $Teams$ that differ by 1, you will always get the same answer.

(e) One purpose for the study was to predict attendance for future expansion teams. Since the study was published, some of the included cities are no longer represented in MLS and have been replaced by others. In one case, beginning with the 2006 season, the San Jose Earthquakes MLS franchise relocated to Houston, Texas, which was one of the potential cities considered in the study. A 95% prediction interval for average attendance for a potential Houston MLS team based on the model came to (10,980, 15,340). Briefly discuss how studies like this can help to inform decisions about future expansion teams for professional leagues like MLS.

3.8 Researchers at General Motors analyzed data on 56 US Standard Metropolitan Statistical Areas (SMSAs) to study whether air pollution contributes to mortality. These data are available in the **SMSA** data file and were obtained from the Data and Story Library (the original data source is the US Department of Labor Statistics). The response variable for analysis is $Mort =$ age adjusted mortality per 100,000 population (a mortality rate statistically modified to eliminate the effect of different age distributions in different population groups). The dataset includes predictor variables measuring demographic characteristics of the cities, climate characteristics, and concentrations of the air pollutant nitrous oxide (NO_x).

(a) Fit the (complete) model $E(Mort) = b_0 + b_1 Edu + b_2 Nwt + b_3 Jant + b_4 Rain + b_5 Nox + b_6 Hum + b_7 Inc$, where Edu is median years of education, Nwt is percentage nonwhite, $Jant$ is mean January temperature in degrees Fahrenheit, $Rain$ is annual rainfall in inches, Nox is the natural logarithm of nitrous oxide concentration in parts per billion, Hum is relative humidity, and

Inc is median income in thousands of dollars [computer help #31]. Report the least squares estimated regression equation.

(b) Do a nested model F-test (using a significance level of 5%) to see whether Hum and Inc provide significant information about the response, $Mort$, beyond the information provided by the other predictor variables [computer help #34]. Use the fact that the 95th percentile of the F-distribution with 2 numerator degrees of freedom and 48 denominator degrees of freedom is 3.19 [computer help #8].

(c) Do individual t-tests (using a significance level of 5%) for each predictor in the (reduced) model $\mathrm{E}(Mort) = b_0 + b_1 Edu + b_2 Nwt + b_3 Jant + b_4 Rain + b_5 Nox$ [computer help #31]. Use the fact that the 97.5th percentile of the t-distribution with 50 degrees of freedom is 2.01 [computer help #8].

(d) Check the four model assumptions for the model from part (c) [computer help #35, #28, #15, #36, #14, and #22].
Hint: Ideally, you should look at six residual plots (five with each of the five predictors on the horizontal axis in turn and one with the predicted values on the horizontal axis) to check the zero mean, constant variance, and independence assumptions. You should also use a histogram and/or QQ-plot to check the normality assumption. The more comprehensive you can be in checking the assumptions, the more confident you can be about the validity of your model.

(e) Write out the least squares equation for the model from part (c). Do the signs of the estimated regression parameters make sense in this context?

(f) Based on the model from part (c), calculate a 95% confidence interval for $\mathrm{E}(Mort)$ for cities with the following characteristics: $Edu = 10$, $Nwt = 15$, $Jant = 35$, $Rain = 40$, and $Nox = 2$ [computer help #29].

(g) Based on the model from part (c), calculate a 95% prediction interval for $Mort^*$ for a city with the following characteristics: $Edu = 10$, $Nwt = 15$, $Jant = 35$, $Rain = 40$, and $Nox = 2$ [computer help #30].