PART ONE

The Multiple Linear Regression Model

CHAPTER ONE

Multiple Linear Regression

1.1 Introduction

This is a book about regression modeling, but when we refer to regression models, what do we mean? The regression framework can be characterized in the following way:

1. We have one particular variable that we are interested in understanding or modeling, such as sales of a particular product, sale price of a home, or voting preference of a particular voter. This variable is called the **target**, **response**, or **dependent** variable, and is usually represented by y.

Handbook of Regression Analysis. By Samprit Chatterjee and Jeffrey S. Simonoff

2. We have a set of p other variables that we think might be useful in predicting or modeling the target variable (the price of the product, the competitor's price, and so on; or the lot size, number of bedrooms, number of bathrooms of the home, and so on; or the gender, age, income, party membership of the voter, and so on). These are called the **predicting**, or **independent** variables, and are usually represented by x_1, x_2, etc.

Typically, a regression analysis is used for one (or more) of three purposes:

1. modeling the relationship between x and y;
2. prediction of the target variable (forecasting);
3. and testing of hypotheses.

In this chapter we introduce the basic multiple linear regression model, and discuss how this model can be used for these three purposes. Specifically, we discuss the interpretations of the estimates of different regression parameters, the assumptions underlying the model, measures of the strength of the relationship between the target and predictor variables, the construction of tests of hypotheses and intervals related to regression parameters, and the checking of assumptions using diagnostic plots.

1.2 Concepts and Background Material

1.2.1 THE LINEAR REGRESSION MODEL

The data consist of n sets of observations $\{x_{1i}, x_{2i}, \ldots, x_{pi}, y_i\}$, which represent a random sample from a larger population. It is assumed that these observations satisfy a linear relationship,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \tag{1.1}$$

where the β coefficients are unknown parameters, and the ε_i are random error terms. By a *linear* model, it is meant that the model is linear in the *parameters*; a quadratic model,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

paradoxically enough, is a linear model, since x and x^2 are just versions of x_1 and x_2.

It is important to recognize that this, or any statistical model, is not viewed as a *true* representation of reality; rather, the goal is that the model be a *useful* representation of reality. A model can be used to explore the relationships between variables and make accurate forecasts based on those relationships even if it is not the "truth." Further, any statistical model is only temporary, representing a provisional version of views about the random process being studied. Models can, and should, change, based on analysis using the current model, selection among several candidate models, the acquisition

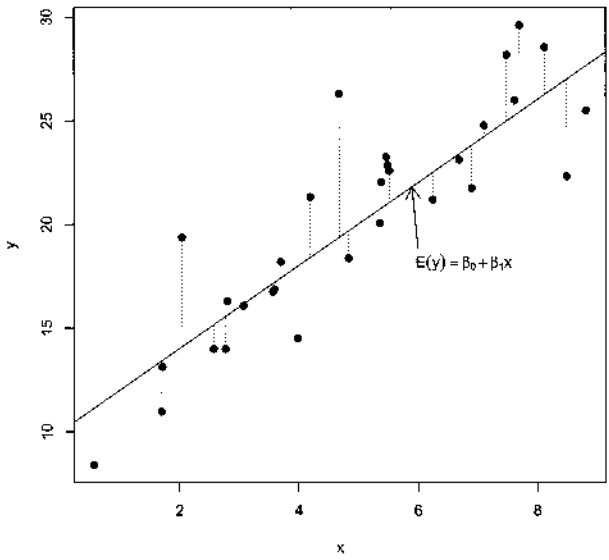

FIGURE 1.1 **The simple linear regression model. The solid line corresponds to the true regression line, and the dotted lines correspond to the random errors ε_i.**

of new data, and so on. Further, it is often the case that there are several different models that are reasonable representations of reality. Having said this, we will sometimes refer to the "true" model, but this should be understood as referring to the underlying form of the currently hypothesized representation of the regression relationship.

The special case of (1.1) with $p = 1$ corresponds to the **simple regression model**, and is consistent with the representation in Figure 1.1. The solid line is the true regression line, the expected value of y given the value of x. The dotted lines are the random errors ε_i that account for the lack of a perfect association between the predictor and the target variables.

1.2.2 ESTIMATION USING LEAST SQUARES

The true regression function represents the expected relationship between the target and the predictor variables, which is unknown. A primary goal of a regression analysis is to estimate this relationship, or equivalently, to estimate the unknown parameters β. This requires a data-based rule, or criterion, that will give a reasonable estimate. The standard approach is **least squares**

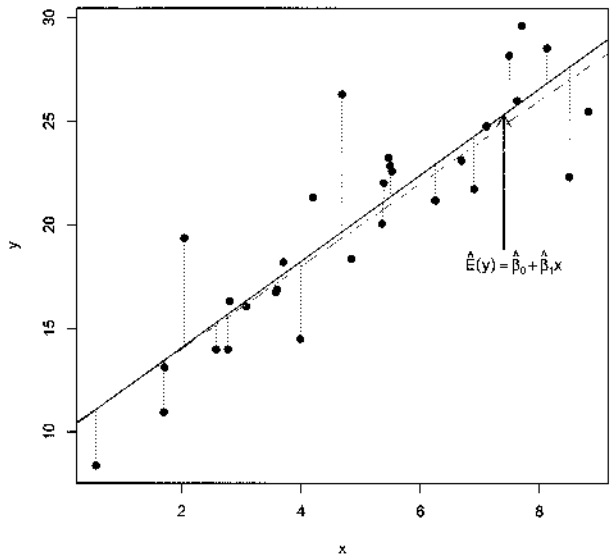

FIGURE 1.2 **Least squares estimation for the simple linear regression model, using the same data as in Figure 1.1. The gray line corresponds to the true regression line, the solid black line corresponds to the fitted least squares line (designed to estimate the gray line), and the lengths of the dotted lines correspond to the residuals. The sum of squared values of the lengths of the dotted lines is minimized by the solid black line.**

regression, where the estimates are chosen to minimize

$$\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})]^2. \tag{1.2}$$

Figure 1.2 gives a graphical representation of least squares that is based on Figure 1.1. Now the true regression line is represented by the gray line, and the solid black line is the estimated regression line, designed to estimate the (unknown) gray line as closely as possible. For any choice of estimated parameters $\hat{\boldsymbol{\beta}}$, the estimated expected response value given the observed predictor values equals

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi},$$

and is called the **fitted value**. The difference between the observed value y_i and the fitted value $\hat{y}_i$ is called the **residual**, the set of which are represented by the lengths of the dotted lines in Figure 1.2. The least squares regression line minimizes the sum of squares of the lengths of the dotted lines; that is, the ordinary least squares (OLS) estimates minimize the sum of squares of the residuals.

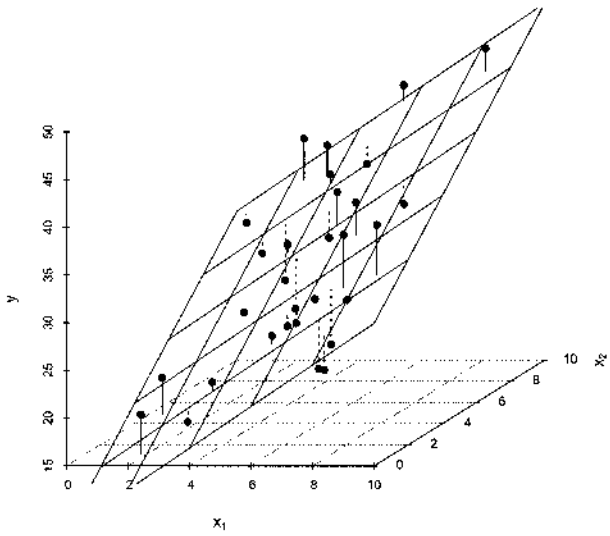

FIGURE 1.3 **Least squares estimation for the multiple linear regression model with two predictors. The plane corresponds to the fitted least squares relationship, and the lengths of the vertical lines correspond to the residuals. The sum of squared values of the lengths of the vertical lines is minimized by the plane.**

In higher dimensions ($p > 1$) the true and estimated regression relationships correspond to planes ($p = 2$) or hyperplanes ($p \geq 3$), but otherwise the principles are the same. Figure 1.3 illustrates the case with two predictors. The length of each vertical line corresponds to a residual (solid lines refer to positive residuals while dashed lines refer to negative residuals), and the (least squares) plane that goes through the observations is chosen to minimize the sum of squares of the residuals.

The linear regression model can be written compactly using matrix notation. Define the following matrix and vectors as follows:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The regression model (1.1) is then

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The normal equations [which determine the minimizer of (1.2)] can be shown (using multivariate calculus) to be

$$(X'X)\hat{\boldsymbol{\beta}} = X'\mathbf{y},$$

which implies that the least squares estimates satisfy

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}.$$

The fitted values are then

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X(X'X)^{-1}X'\mathbf{y} \equiv H\mathbf{y}, \tag{1.3}$$

where $H = X(X'X)^{-1}X'$ is the so-called "hat" matrix (since it takes $\mathbf{y}$ to $\hat{\mathbf{y}}$). The residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ thus satisfy

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X(X'X)^{-1}X'\mathbf{y} = (I - X(X'X)^{-1}X')\mathbf{y}, \tag{1.4}$$

or

$$\mathbf{e} = (I - H)\mathbf{y}.$$

1.2.3 ASSUMPTIONS

The least squares criterion will not necessarily yield sensible results unless certain assumptions hold. One is given in (1.1) — the linear model should be appropriate. In addition, the following assumptions are needed to justify using least squares regression.

1. The expected value of the errors is zero ($E(\varepsilon_i) = 0$ for all i). That is, it cannot be true that for certain observations the model is systematically too low, while for others it is systematically too high. A violation of this assumption will lead to difficulties in estimating β_0. More importantly, this reflects that the model does not include a necessary systematic component, which has instead been absorbed into the error terms.
2. The variance of the errors is constant ($V(\varepsilon_i) = \sigma^2$ for all i). That is, it cannot be true that the strength of the model is more for some parts of the population (smaller σ) and less for other parts (larger σ). This assumption of constant variance is called **homoscedasticity**, and its violation (nonconstant variance) is called **heteroscedasticity**. A violation of this assumption means that the least squares estimates are not as efficient as they could be in estimating the true parameters, and better estimates are available. More importantly, it also results in poorly calibrated confidence and (especially) prediction intervals.
3. The errors are uncorrelated with each other. That is, it cannot be true that knowing that the model underpredicts y (for example) for one particular observation says anything at all about what it does for any other

observation. This violation most often occurs in data that are ordered in time (time series data), where errors that are near each other in time are often similar to each other (such time-related correlation is called **autocorrelation**). Violation of this assumption can lead to very misleading assessments of the strength of the regression.

4. The errors are normally distributed. This is needed if we want to construct any confidence or prediction intervals, or hypothesis tests, which we usually do. If this assumption is violated, hypothesis tests and confidence and prediction intervals can be very misleading.

Since violation of these assumptions can potentially lead to completely misleading results, a fundamental part of any regression analysis is to check them using various plots, tests, and diagnostics.

1.3 Methodology

1.3.1 INTERPRETING REGRESSION COEFFICIENTS

The least squares regression coefficients have very specific meanings. They are often misinterpreted, so it is important to be clear on what they mean (and do not mean). Consider first the intercept, $\hat{\beta}_0$.

$\hat{\beta}_0$: The estimated expected value of the target variable when the predictors all equal zero.

Note that this might not have any physical interpretation, since a zero value for the predictor(s) might be impossible, or might never come close to occurring in the observed data. In that situation, it is pointless to try to interpret this value. If all of the predictors are centered to have mean zero, then $\hat{\beta}_0$ necessarily equals $\overline{Y}$, the sample mean of the target values. Note that if there is any particular value for each predictor that is meaningful in some sense, if each variable is centered around its particular value, then the intercept is an estimate of $E(y)$ when the predictors all have those meaningful values.

The estimated coefficient for the jth predictor ($j = 1, \ldots, p$) is interpreted in the following way.

$\hat{\beta}_j$: The estimated expected change in the target variable associated with a one unit change in the jth predicting variable, holding all else in the model fixed.

There are several noteworthy aspects to this interpretation. First, note the word *associated* — we cannot say that a change in the target variable is *caused* by a change in the predictor, only that they are associated with each other. That is, correlation does not imply causation.

Another key point is the phrase "holding all else in the model fixed," the implications of which are often ignored. Consider the following hypothetical

example. A random sample of college students at a particular university is taken in order to understand the relationship between college grade point average (GPA) and other variables. A model is built with college GPA as a function of high school GPA and the standardized Scholastic Aptitude Test (SAT), with resultant least squares fit

$$\texttt{College GPA} = 1.3 + .7 \times \texttt{High School GPA} - .0001 \times \texttt{SAT}.$$

It is tempting to say (and many people would say) that the coefficient for SAT score has the "wrong sign," because it says that higher values of SAT are associated with lower values of college GPA. This is *not* correct. The problem is that it is likely in this context that what an analyst would find intuitive is the *marginal* relationship between college GPA and SAT score alone (ignoring all else), one that we would indeed expect to be a direct (positive) one. The regression coefficient does not say anything about that marginal relationship. Rather, it refers to the conditional (sometimes called partial) relationship that takes the high school GPA as fixed, which is apparently that higher values of SAT are associated with lower values of college GPA, holding high school GPA fixed. High school GPA and SAT are no doubt related to each other, and it is quite likely that this relationship between the predictors would complicate any understanding of, or intuition about, the conditional relationship between college GPA and SAT score. Multiple regression coefficients should not be interpreted marginally; if you really are interested in the relationship between the target and a single predictor alone, you should simply do a regression of the target on only that variable. This does not mean that multiple regression coefficients are uninterpretable, only that care is necessary when interpreting them.

Another common use of multiple regression that depends on this conditional interpretation of the coefficients is to explicitly include "control" variables in a model in order to try to account for their effect statistically. This is particularly important in observational data (data that are not the result of a designed experiment), since in that case the effects of other variables cannot be ignored as a result of random assignment in the experiment. For observational data it is not possible to physically intervene in the experiment to "hold other variables fixed," but the multiple regression framework effectively allows this to be done statistically.

1.3.2 MEASURING THE STRENGTH OF THE REGRESSION RELATIONSHIP

The least squares estimates possess an important property:

$$\sum_{i=1}^{n}(y_i - \overline{Y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2.$$

This formula says that the variability in the target variable (the left side of the equation, termed the corrected total sum of squares) can be split into two mu-

tually exclusive parts — the variability left over after doing the regression (the first term on the right side, the residual sum of squares), and the variability accounted for by doing the regression (the second term, the regression sum of squares). This immediately suggests the usefulness of R^2 as a measure of the strength of the regression relationship, where

$$R^2 = \frac{\sum_i (\hat{y}_i - \overline{Y})^2}{\sum_i (y_i - \overline{Y})^2} \equiv \frac{\text{Regression SS}}{\text{Corrected total SS}} = 1 - \frac{\text{Residual SS}}{\text{Corrected total SS}}.$$

The R^2 value (also called the **coefficient of determination**) estimates the population proportion of variability in y accounted for by the best linear combination of the predictors. Values closer to 1 indicate a good deal of predictive power of the predictors for the target variable, while values closer to 0 indicate little predictive power. An equivalent representation of R^2 is

$$R^2 = \text{corr}(y_i, \hat{y}_i)^2,$$

where

$$\text{corr}(y_i, \hat{y}_i) = \frac{\sum_i (y_i - \overline{Y})(\hat{y}_i - \overline{\hat{Y}})}{\sqrt{\sum_i (y_i - \overline{Y})^2 \sum_i (\hat{y}_i - \overline{\hat{Y}})^2}}$$

is the sample correlation coefficient between $\mathbf{y}$ and $\hat{\mathbf{y}}$ (this correlation is called the multiple correlation coefficient). That is, R^2 is a direct measure of how similar the observed and fitted target values are.

It can be shown that R^2 is biased upwards as an estimate of the population proportion of variability accounted for by the regression. The **adjusted** R^2 corrects this bias, and equals

$$R_a^2 = R^2 - \frac{p}{n - p - 1}\left(1 - R^2\right). \tag{1.5}$$

It is apparent from (1.5) that unless p is large relative to $n - p - 1$ (that is, unless the number of predictors is large relative to the sample size), R^2 and R_a^2 will be close to each other, and the choice of which to use is a minor concern. What is perhaps more interesting is the nature of R_a^2 as providing an explicit tradeoff between the strength of the fit (the first term, with larger R^2 corresponding to stronger fit and larger R_a^2) and the complexity of the model (the second term, with larger p corresponding to more complexity and smaller R_a^2). This tradeoff of fidelity to the data versus simplicity will be important in the discussion of model selection in Section 2.3.1.

The only parameter left unaccounted for in the estimation scheme is the variance of the errors σ^2. An unbiased estimate is provided by the **residual mean square**,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p - 1}. \tag{1.6}$$

This estimate has a direct, but often underappreciated, use in assessing the practical importance of the model. Does knowing $x_1, \ldots, x_p$ really say anything of value about y? This isn't a question that can be answered completely

statistically; it requires knowledge and understanding of the data (that is, it requires context). Recall that the model assumes that the errors are normally distributed with standard deviation σ. This means that, roughly speaking, 95% of the time an observed y value falls within $\pm 2\sigma$ of the expected response

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

$E(y)$ can be estimated for any given set of **x** values using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p,$$

while the square root of the residual mean square (1.6), termed the **standard error of the estimate**, provides an estimate of σ that can be used in constructing this rough prediction interval $\pm 2\hat{\sigma}$.

1.3.3 HYPOTHESIS TESTS AND CONFIDENCE INTERVALS FOR β

There are two types of hypothesis tests related to the regression coefficients of immediate interest.

1. Do *any* of the predictors provide predictive power for the target variable? This is a test of the overall significance of the regression,

   $$H_0 : \beta_1 = \cdots = \beta_p = 0$$

   versus

   $$H_a : \text{some } \beta_j \neq 0, \qquad j = 1, \ldots, p.$$

   The test of these hypotheses is the F-**test**,

   $$F = \frac{\text{Regression MS}}{\text{Residual MS}} \equiv \frac{\text{Regression SS}/p}{\text{Residual SS}/(n-p-1)}.$$

   This is referenced against a null F-distribution on $(p, n-p-1)$ degrees of freedom.

2. Given the other variables in the model, does a particular predictor provide additional predictive power? This corresponds to a test of the significance of an individual coefficient,

   $$H_0 : \beta_j = 0, \qquad j = 1, \ldots, p$$

   versus

   $$H_a : \beta_j \neq 0.$$

   This is tested using a t-**test**,

   $$t_j = \frac{\hat{\beta}_j}{\widehat{\text{s.e.}}(\hat{\beta}_j)},$$

which is compared to a t-distribution on $n - p - 1$ degrees of freedom. Other values of β_j can be specified in the null hypothesis (say β_{j0}), with the t-statistic becoming

$$t_j = \frac{\hat{\beta}_j - \beta_{j0}}{\widehat{\text{s.e.}}(\hat{\beta}_j)}. \tag{1.7}$$

The values of $\widehat{\text{s.e.}}(\hat{\beta}_j)$ are obtained as the square roots of the diagonal elements of $\hat{V}(\hat{\beta}) = (X'X)^{-1}\hat{\sigma}^2$, where $\hat{\sigma}^2$ is the residual mean square (1.6). Note that for simple regression ($p = 1$) the hypotheses corresponding to the overall significance of the model and the significance of the predictor are identical,

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0.$$

Given the equivalence of the sets of hypotheses, it is not surprising that the associated tests are also equivalent; in fact, $F = t_1^2$, and the associated tail probabilities of the two tests are identical.

A t-test for the intercept also can be constructed as in (1.7), although this does not refer to a hypothesis about a predictor, but rather about whether the expected target is equal to a specified value β_{00} if all of the predictors equal zero. As was noted in Section 1.3.1, this is often not physically meaningful (and therefore of little interest), because the condition that all predictors equal zero cannot occur, or does not come close to occurring in the observed data.

As is always the case, a confidence interval provides an alternative way of summarizing the degree of precision in the estimate of a regression parameter. That is, a $100 \times (1 - \alpha)\%$ confidence interval for β_j has the form

$$\hat{\beta}_j \pm t_{\alpha/2}^{n-p-1} \widehat{\text{s.e.}}(\hat{\beta}_j),$$

where $t_{\alpha/2}^{n-p-1}$ is the appropriate critical value at two-sided level α for a t-distribution on $n - p - 1$ degrees of freedom.

1.3.4 FITTED VALUES AND PREDICTIONS

The rough prediction interval $\hat{y} \pm 2\hat{\sigma}$ discussed in Section 1.3.2 is an approximate 95% interval because it ignores the variability caused by the need to estimate σ, and uses only an approximate normal-based critical value. A more accurate assessment of this is provided by a **prediction interval** given a particular value of x. This interval provides guidance as to how precise $\hat{y}_0$ is as a prediction of y for some particular specified value $\mathbf{x}_0$, where $\hat{y}_0$ is determined by substituting the values $\mathbf{x}_0$ into the estimated regression equation; its width

depends on both $\hat{\sigma}$ and the position of $\mathbf{x}_0$ relative to the centroid of the predictors (the point located at the means of all predictors), since values farther from the centroid are harder to predict as precisely. Specifically, for a simple regression, the estimated standard error of a predicted value based on a value x_0 of the predicting variable is

$$\widehat{\text{s.e.}}(\hat{y}_0^P) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{X})^2}{\sum(x_i - \overline{X})^2}}.$$

More generally, the variance of a predicted value is

$$\hat{V}(\hat{y}_0^P) = [1 + \mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0]\hat{\sigma}^2. \tag{1.8}$$

Here $\mathbf{x}_0$ is taken to include a 1 in the first entry (corresponding to the intercept in the regression model). The prediction interval is then

$$\hat{y}_0 \pm t_{\alpha/2}^{n-p-1}\widehat{\text{s.e.}}(\hat{y}_0^P),$$

where $\widehat{\text{s.e.}}(\hat{y}_0^P) = \sqrt{\hat{V}(\hat{y}_0^P)}$.

This prediction interval should not be confused with a **confidence interval for a fitted value**. The prediction interval is used to provide an interval estimate for a *prediction of y for one member of the population* with a particular value of $\mathbf{x}_0$; the confidence interval is used to provide an interval estimate for the *true expected value of y for all members of the population* with a particular value of $\mathbf{x}_0$. The corresponding standard error, termed the standard error for a fitted value, is the square root of

$$\hat{V}(\hat{y}_0^F) = \mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0\hat{\sigma}^2, \tag{1.9}$$

with corresponding confidence interval

$$\hat{y}_0 \pm t_{\alpha/2}^{n-p-1}\widehat{\text{s.e.}}(\hat{y}_0^F).$$

A comparison of the two estimated variances (1.8) and (1.9) shows that the variance of the predicted value has an extra σ^2 term, which corresponds to the inherent variability in the population. Thus, the confidence interval for a fitted value will always be narrower than the prediction interval, and is often much narrower (especially for large samples), since increasing the sample size will always improve estimation of the expected response value, but cannot lessen the inherent variability in the population associated with the prediction of the target for a single observation.

1.3.5 CHECKING ASSUMPTIONS USING RESIDUAL PLOTS

As was noted earlier, all of these tests, intervals, predictions, and so on, are based on the belief that the assumptions of the regression model hold. Thus, it is crucially important that these assumptions be checked. Remarkably enough, a few very simple plots can provide much of the evidence needed to check the assumptions.

1. A plot of the residuals versus the fitted values. This plot should have no pattern to it; that is, no structure should be apparent. Certain kinds of structure indicate potential problems:
   (a) A point (or a few points) isolated at the top or bottom, or left or right. In addition, often the rest of the points have a noticeable "tilt" to them. These isolated points are unusual points, and can have a strong effect on the regression. They need to be examined carefully, and possibly removed from the data set.
   (b) An impression of different heights of the point cloud as the plot is examined from left to right. This indicates potential heteroscedasticity (nonconstant variance).
2. Plots of the residuals versus each of the predictors. Again, a plot with no apparent structure is desired.
3. If the data set has a time structure to it, residuals should be plotted versus time. Again, there should be no apparent pattern. If there is a cyclical structure, this indicates that the errors are not uncorrelated, as they are supposed to be (that is, there is potentially autocorrelation in the errors).
4. A normal plot of the residuals. This plot assesses the apparent normality of the residuals, by plotting the observed ordered residuals on one axis and the expected positions (under normality) of those ordered residuals on the other. The plot should look like a straight line (roughly). Isolated points once again represent unusual observations, while a curved line indicates that the errors are probably not normally distributed, and tests and intervals might not be trustworthy.

Note that all of these plots should be routinely examined in any regression analysis, although in order to save space not all will necessarily be presented in all of the analyses in the book.

An implicit assumption in any model that is being used for prediction is that the future "looks like" the past; that is, it is not sufficient that these assumptions appear to hold for the available data, as they also must continue to hold for new data on which the estimated model is applied. Indeed, the assumption is stronger than that, since it must be the case that the future is exactly the same as the past, in the sense that all of the properties of the model, including the precise values of all of the regression parameters, are the same. This is unlikely to be exactly true, so a more realistic point of view is that the future should be similar enough to the past so that predictions based on the past are useful. A related point is that predictions should not be based on extrapolation, where the predictor values are far from the values used to build the model. Similarly, if the observations form a time series, predictions far into the future are unlikely to be very useful.

In general, the more complex a model is, the less likely it is that all of its characteristics will remain stable going forward, which implies that a reasonable goal is to try to find a model that is as simple as it can be while still

accounting for the important effects in the data. This leads to questions of **model building**, which is the subject of the next chapter.

1.4 Example — Estimating Home Prices

Determining the appropriate sale price for a home is clearly of great interest to both buyers and sellers. While this can be done in principle by examining the prices at which other similar homes have recently sold, the well-known existence of strong effects related to location means that there are likely to be relatively few homes with the same important characteristics to make the comparison. A solution to this problem is the use of hedonic regression models, where the sale prices of a set of homes in a particular area are regressed on important characteristics of the home such as the number of bedrooms, the living area, the lot size, and so on. Academic research on this topic is plentiful, going back to at least Wabe (1971).

This analysis is based on a sample from public data on sales of one-family homes in the Levittown, NY area from June 2010 through May 2011. Levittown is famous as the first planned suburban community built using mass production methods, being aimed at former members of the military after World War II. Most of the homes in this community were built in the late 1940s to early 1950s, without basements and designed to make expansion on the second floor relatively easy.

For each of the 85 houses in the sample, the number of bedrooms, number of bathrooms, living area (in square feet), lot size (in square feet), the year the house was built, and the property taxes are used as potential predictors of the sale price. In any analysis the first step is to look at the data, and Figure 1.4 gives scatter plots of sale price versus each predictor. It is apparent that there is a positive association between sale price and each variable, other than number of bedrooms and lot size. We also note that there are two houses with unusually large living areas for this sample, two with unusually large property taxes (these are not the same two houses), and three that were built 6 or 7 years later than all of the other houses in the sample.

The output below summarizes the results of a multiple regression fit.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.149e+06  3.820e+06  -1.871 0.065043 .
Bedrooms    -1.229e+04  9.347e+03  -1.315 0.192361
Bathrooms    5.170e+04  1.309e+04   3.948 0.000171 ***
Living.area  6.590e+01  1.598e+01   4.124 9.22e-05 ***
Lot.size    -8.971e-01  4.194e+00  -0.214 0.831197
Year.built   3.761e+03  1.963e+03   1.916 0.058981 .
Property.tax 1.476e+00  2.832e+00   0.521 0.603734
---
```

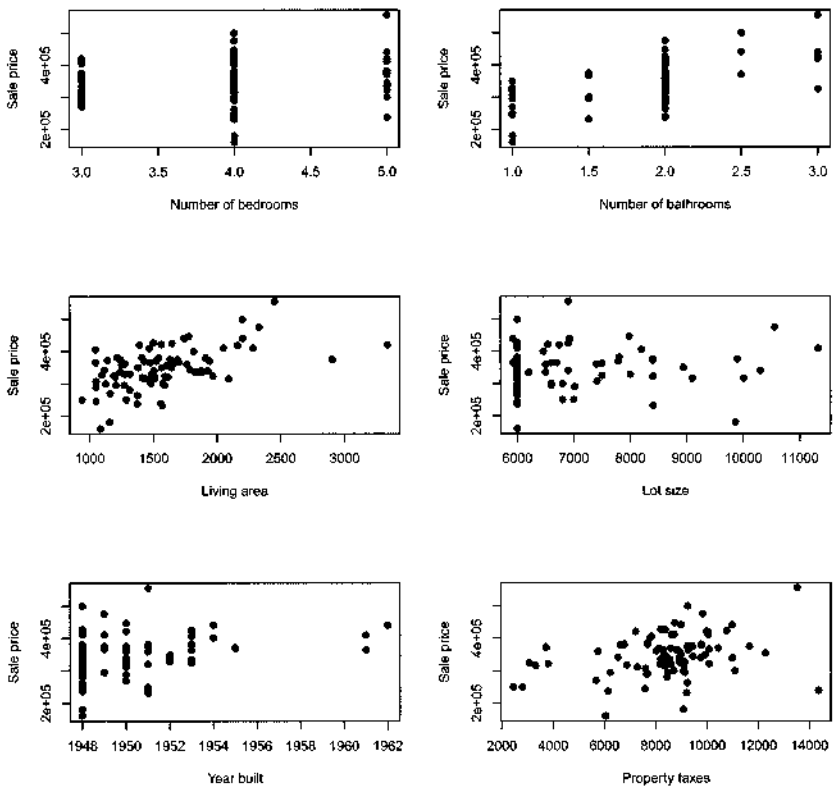

FIGURE 1.4 **Scatter plots of sale price versus each predictor for the home price data.**

```
Signif. codes:
   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47380 on 78 degrees of freedom
Multiple R-squared: 0.5065,     Adjusted R-squared: 0.4685
F-statistic: 13.34 on 6 and 78 DF,  p-value: 2.416e-10
```

The overall regression is strongly statistically significant, with the tail probability of the F-test roughly 10^{-10}. The predictors account for roughly 50% of the variability in sale prices ($R^2 \approx 0.5$). Two of the predictors (number of bathrooms and living area) are highly statistically significant, with tail probabilities less than .0002, and the coefficient of the year built variable is marginally statistically significant. The coefficients imply that given all else in the model is held fixed, one additional bathroom in a house is associated with an estimated expected price that is \$51,700 higher; one additional square

foot of living area is associated with an estimated expected price that is \$65.90 higher (given the typical value of the living area variable, a more meaningful statement would probably be that an additional 100 square feet of living area is associated with an estimated expected price that is \$659 higher); and a house being built one year later is associated with an estimated expected price that is \$3761 higher.

This is a situation where the distinction between a confidence interval for a fitted value and a prediction interval (and which is of more interest to a particular person) is clear. Consider a house with 3 bedrooms, 1 bathroom, 1050 square feet of living area, 6000 square foot lot size, built in 1948, with \$6306 in property taxes. Substituting those values into the above equation gives an estimated expected sale price of a house with these characteristics equal to \$265,360. A buyer or a seller is interested in the sale price of one particular house, so a prediction interval for the sale price would provide a range for what the buyer can expect to pay and the seller expect to get. The standard error of the estimate $\hat{\sigma} = \$47{,}380$ can be used to construct a rough prediction interval, in that roughly 95% of the time a house with these characteristics can be expected to sell for within $\pm(2)(47380) = \pm\$94{,}360$ of that estimated sale price, but a more exact interval might be required. On the other hand, a home appraiser or tax assessor is more interested in the typical (average) sale price for all homes of that type in the area, so they can give a justifiable interval estimate giving the precision of the estimate of the true expected value of the house, so a confidence interval for the fitted value is desired.

Exact 95% intervals for a house with these characteristics can be obtained from statistical software, and turn out to be $(\$167277, \$363444)$ for the prediction interval and $(\$238482, \$292239)$ for the confidence interval. As expected, the prediction interval is much wider than the confidence interval, since it reflects the inherent variability in sale prices in the population of houses; indeed, it is probably too wide to be of any practical value in this case, but an interval with smaller coverage (that is expected to include the actual price only 50% of the time, say) might be useful (a 50% interval in this case would be $(\$231974, \$298746)$, so a seller could be told that there is a 50/50 chance that their house will sell for a value in this range).

The validity of all of these results depends on whether the assumptions hold. Figure 1.5 gives a scatter plot of the residuals versus the fitted values and a normal plot of the residuals for this model fit. There is no apparent pattern in the plot of residuals versus fitted values, and the ordered residuals form a roughly straight line in the normal plot, so there are no apparent violations of assumptions here. The plot of residuals versus each of the predictors (Figure 1.6) also does not show any apparent patterns, other than the houses with unusual living area and year being built, respectively. It would be reasonable to omit these observations to see if they have had an effect on the regression, but we will postpone discussion of that to Chapter 3, where diagnostics for unusual observations are discussed in greater detail.

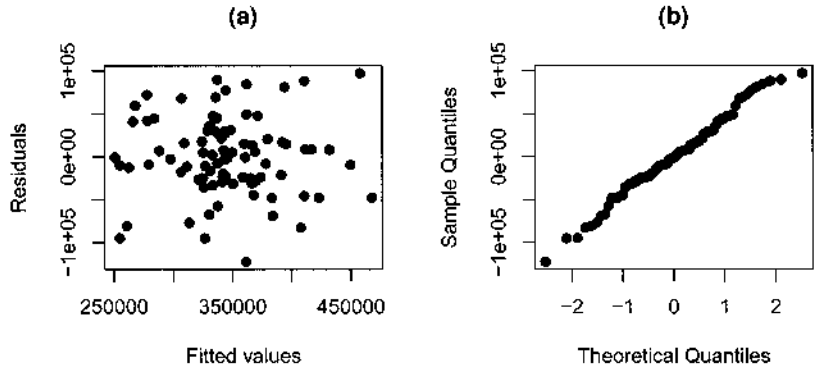

FIGURE 1.5 **Residual plots for the home price data. (a) Plot of residuals versus fitted values. (b) Normal plot of the residuals.**

An obvious consideration at this point is that the models discussed here appear to be overspecified; that is, they include variables that do not apparently add to the predictive power of the model. As was noted earlier, this suggests the consideration of model building, where a more appropriate (simplified) model can be chosen, which will be discussed in the next chapter.

1.5 Summary

In this chapter we have laid out the basic structure of the linear regression model, including the assumptions that justify the use of least squares estimation. The three main goals of regression noted at the beginning of the chapter provide a framework for an organization of the topics covered.

1. Modeling the relationship between $\mathbf{x}$ and y:
   - the least squares estimates $\hat{\beta}$ summarize the expected change in y for a given change in an x, accounting for all of the variables in the model;
   - the standard error of the estimate $\hat{\sigma}$ estimates the standard deviation of the errors;
   - R^2 and R_a^2 estimate the proportion of variability in y accounted for by $\mathbf{x}$;
   - and the confidence interval for a fitted value provides a measure of the precision in estimating the expected target for a given set of predictor values.
2. Prediction of the target variable:

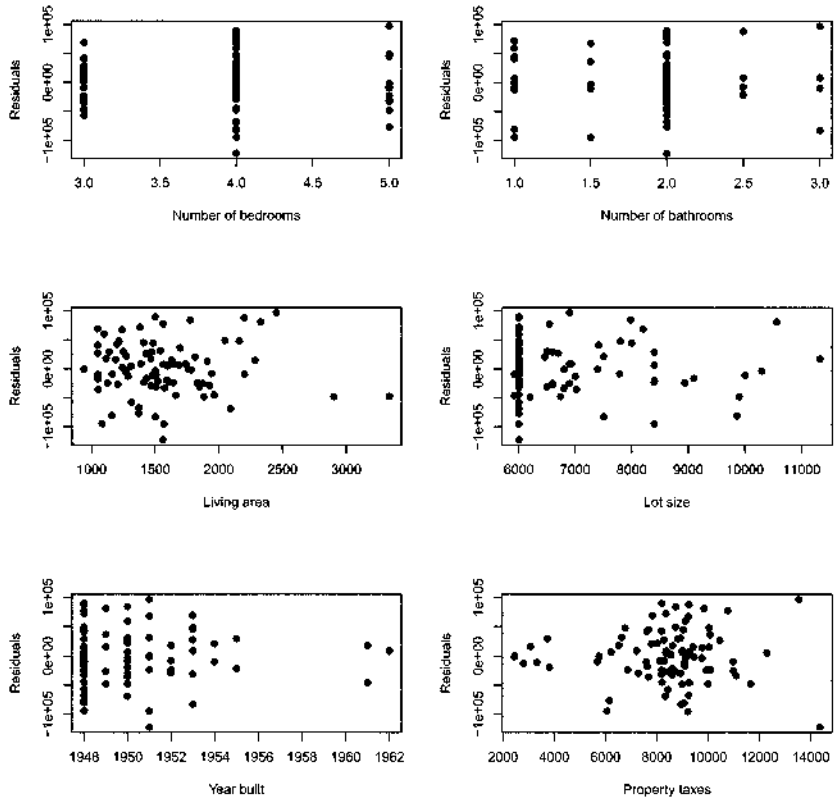

FIGURE 1.6 **Scatter plots of residuals versus each predictor for the home price data.**

- substituting specified values of x into the fitted regression model gives an estimate of the value of the target for a new observation;
- the rough prediction interval $\pm 2\hat{\sigma}$ provides a quick measure of the limits of the ability to predict a new observation;
- and the exact prediction interval provides a more precise measure of those limits.

3. Testing of hypotheses:
   - the F-test provides a test of the statistical significance of the overall relationship;
   - the t-test for each slope coefficient testing whether the true value is zero provides a test of whether the variable provides additional predictive power given the other variables;
   - and the t-tests can be generalized to test other hypotheses of interest about the coefficients as well.

Since all of these methods depend on the assumptions holding, a fundamental part of any regression analysis is to check those assumptions. The residual plots discussed in this chapter are a key part of that process, and other diagnostics and tests will be discussed in future chapters that provide additional support for that task.

KEY TERMS

Autocorrelation: Correlation between adjacent observations in a (time) series. In the regression context it is autocorrelation of the errors that is a violation of assumptions.

Coefficient of determination (R^2): The square of the multiple correlation coefficient, estimates the proportion of variability in the target variable that is explained by the fitted least squares model.

Confidence interval for a fitted value: A measure of precision of the estimate of the expected target value for a given x.

Dependent variable: Characteristic of each member of the sample that is being modeled. This is also known as the **target** or **response** variable.

Fitted value: The least square estimate of the expected target value for a particular observation obtained from the fitted regression model.

Heteroscedasticity: Unequal variance; this can refer to observed unequal variance of the residuals or theoretical unequal variance of the errors.

Homoscedasticity: Equal variance; this can refer to observed equal variance of the residuals or the assumed equal variance of the errors.

Independent variable(s): Characteristic(s) of each member of the sample that could be used to model the dependent variable. These are also known as the **predicting** variables.

Least squares: A method of estimation that minimizes the sum of squared deviations of the observed target values from their estimated expected values.

Prediction interval: The interval estimate for the value of the target variable for an individual member of the population using the fitted regression model.

Residual: The difference between the observed target value and the corresponding fitted value.

Residual mean square: An unbiased estimate of the variance of the errors. It is obtained by dividing the sum of squares of the residuals by $(n - p - 1)$, where n is the number of observations and p is the number of predicting variables.

Standard error of the estimate ($\hat{\sigma}$): An estimate of σ, the standard deviation of the errors, equaling the square root of the residual mean square.