The Multiple Linear Regression Model
Multiple Linear Regression

1.1 Introduction

1.2 Conceptual and Philosophical Background
   1.2.1 The Linear Regression Model
   1.2.2 Estimation Using Least Squares
   1.2.3 Assumptions

1.3 Methodology
   1.3.1 Interpreting Regression Coefficients
   1.3.2 Measuring the Strength of the Regression Relationship
   1.3.3 Hypothesis Tests and Confidence Intervals for $\beta$
   1.3.4 Fitted Values and Predictions
   1.3.5 Checking Assumptions Using Residual Plots

1.4 Examples

1.5 Summary

1.5.1 Introduction

This is a book about regression modeling, and 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 (Copyright © 2005, John Wiley & Sons, Inc.)
2. We have a set of predictor variables that we think might be useful in predicting or modeling a 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 house, 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 \( x_1, x_2, \ldots \).

Typically, a regression analysis is used for one (or more) of these 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.

1.2.2 Concepts and Background Material

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon, \]  

(1.1)

where the \( \beta \) coefficients are unknown parameters, and the \( \epsilon \) are random error terms. By a linear model, it is meant that the model is linear in the parameters; a quadratic model,

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \epsilon, \]

paradoxically enough, is a linear model, since \( x_1 \) and \( x_2^2 \) are joint variables of \( x_1 \) and \( x_2 \).

It is important to recognize that this, or any statistical model, is not a true representation of reality; rather, the goal is that the model be a useful representation of reality. A model can be used to capture the relationships between the variables and make accurate forecasts based on those relationships even if it is not quite the "truth." Moreover, any statistical model is only an approximation, representing a provisional version of views about the random process being studied. Models can, and should, change, based on the analysis using the current model, selection among several candidate models, the acquisition
of new data, and so on. Moreover, 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 mechanism that generated the data, rather than a specific model.

The special case of (1, 1) with \( y \approx x \) corresponds to a linear 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 \( e \) 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 \( \beta \). 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 set as in Figure 1.1. The grey line corresponds to the true regression line, the solid black line corresponds to the fitted least squares line (designed to minimize the grey line), and the lengths of the dashed lines correspond to the residuals. The sum of squared values of the lengths of the dashed lines is minimized by the solid black line.

regression, where the estimates are chosen to minimize

\[
\sum_{k=1}^{n} \left( y_k - (\beta_0 + \beta_1 x_{1k} + \cdots + \beta_p x_{pk}) \right)^2.
\]

(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 grey line, and the solid black line is the estimated regression line, designed to estimate the (unknown) grey line as closely as possible. For any choice of estimated parameters \( \hat{\beta} \), the estimated expected response value given the observed predictor values equals

\[
\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_{1k} + \cdots + \hat{\beta}_p x_{pk}
\]

and is called the fitted value. The difference between the observed value \( y_k \) and the fitted value \( \hat{y}_k \) is called the residual, the set of which are represented by the lengths of the dashed lines in Figure 1.2. The least squares regression line minimizes the sum of squares of the lengths of the dashed lines; that is, the ordinary least squares (OLS) estimates minimize the sum of squares of the residuals.
In higher dimensions (p > 1) the true and estimated regression relationships correspond to planes (p = 2) or hyperplanes (p > 2), 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:

\[
\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}
\]

The regression model (1.4) is then

\[
\mathbf{y} = \mathbf{X}\mathbf{\beta} + \mathbf{\epsilon}
\]
CHAPTER 8 - Multiple Linear Regression

The normal equations (which determine the minimizer of (1.2)) can be shown (using matrix algebra) to be

\[(X'X)\hat{\beta} = X'y,\]

which implies that the least squares estimator satisfies

\[\hat{\beta} = (X'X)^{-1}X'y.\]

The fitted values are then

\[\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = \hat{X}y,\]

where \[\hat{X} = X(X'X)^{-1}X',\]

is the so-called "hat" matrix (since it involves \( y \) in \( \hat{y} \)).

The residuals \( e = y - \hat{y} \) thus satisfy

\[e = y - \hat{y} = y - X(X'X)^{-1}X'y = (I - \hat{X})y,\]

or

\[e = (I - \hat{X})y.\]

### 8.3.3 ASSUMPTIONS

The least squares estimates will not necessarily yield sensible results unless certain assumptions hold. Once it is given in (1.2), 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:** \( E(\epsilon_i) = 0 \) for all \( i \). That is, it is not the case that for certain observarions 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 \( \beta \).

2. **The variance of the errors is constant:** \( \text{Var}(\epsilon_i) = \sigma^2 \) for all \( i \). That is, it cannot be true that the strength of the model is more for some parts of the population (smaller \( \sigma \)) and less for other parts (larger \( \sigma \)). This assumption of constant variance is called homoscedasticity. 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 what 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 where 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 occur 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 $\bar{Y}$, 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 $\bar{Y}(\bar{z})$ where the predictors all have those meaningful values.

The estimated coefficient for the jth predictor ($\hat{\beta}_j = 1, \ldots, p$) 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, since the usual 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:

$$\text{College GPA} = 1.2 + 0.7 \times \text{High School GPA} + 0.0001 \times \text{SAT}.$$ 

It is tempting to say (and many people would say) that the coefficient for SAT means that the "higher SAT, better GPA," 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 interesting 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 holds the high school GPA, as fixed, which in appearance (or: that higher values of SAT are associated with lower values of college GPA, holding high school GPA fixed). High school GPA and SAT are two inputs 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 effects 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 identified 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.

The least squares estimates possess an important property:

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

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

tually conclusive parts—the variability left unexplained during the regression (the first term on the right side, the residual sum of squares), and the variability accounted for by the regression (the second term, the regression sum of squares). This conventionally suggests the usefulness of $R^2$ as a measure of the strength of the regression relationship, where

$$
R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \quad \text{Regression SS} \quad \sum (y_i - \hat{y})^2 \quad \text{Residual SS}
$$

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 larger variable, while values closer to 0 indicate little predictive power. An equivalent representation of $R^2$ is

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

where:

$$
\text{var}(y_i, \hat{y}_i) = \frac{1}{n} \sum (y_i - \bar{y})(\hat{y}_i - \bar{y})
$$

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

$$
R_{adj}^2 = R^2 - \frac{p}{n - p - 1} \cdot \left(1 - R^2\right)
$$

It is apparent from (1.5) that unless $y$ is large relative to $x$ (i.e., unless the number of predictors is large relative to the sample size), $R^2$ and $R_{adj}^2$ 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 $R^2$ as providing an explicit tradeoff between the strength of the fit (the first term, with larger $R^2$ corresponding to stronger fit and larger $R^2$) and the complexity of the model (the second term, with larger $p$ corresponding to more complexity and smaller $R^2$). This tradeoff in favor of 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 $\sigma^2$. An unbiased estimate is provided by the residual mean square,

$$
\hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - p - 1}
$$

This estimate has a direct, but often underestimated, 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 access).

Recall that the model assumes that the errors are normally distributed with standard deviation \( \sigma \). This means that, roughly speaking, 95% of the time an observed \( y \) value falls within 1.96 \( \sigma \) of the expected response \( \hat{y} \):

\[
\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.
\]

\( \hat{y} \) 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 \( \hat{\sigma} \), termed the standard error of the estimate, provides an estimate of \( \sigma \) that can be used in constructing this rough prediction interval \( \hat{y} \pm 2\hat{\sigma} \).

**HYPOTHESIS TESTING AND COMPUTATIONAL COMPUTATIONAL COMPUTATIONAL**

**There are two types of hypothesis tests related to the regression coefficients of 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,
   \[
   H_0: \beta_1 = \cdots = \beta_p = 0,
   \]
   \( \alpha \)-Test:
   - \( H_0 : \beta_j = 0, j = 1, \ldots, p \).
   - The test of these hypotheses is the \( \mathbf{F} \)-test,
   \[
   \mathbf{F} = \frac{\text{Regression } \text{MS}}{\text{Residual } \text{MS}} = \frac{\text{Regression } \text{MS}/(p)}{\text{Residual } \text{MS}/(n - p - 1)}.
   \]
   - This is referenced against a \( \mathbf{F} \)-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,
   \[
   H_0: \beta_j = 0, j = 1, \ldots, p
   \]
   \( \alpha \)-Test:
   - \( H_0 : \beta_j = 0 \).
   - This is tested using a \( \hat{t} \)-test
   \[
   t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}.
   \]
which is compared to a t-distribution on \( n - p - 1 \) degrees of freedom. Other values of \( \beta_j \) can be specified in the null hypothesis (say \( \beta_j = 0 \)), with the t-statistic becoming

\[
\frac{\hat{\beta}_j - \beta_{j0}}{SE(\hat{\beta}_j)}
\]

The values of \( SE(\hat{\beta}_j) \) are obtained as the square roots of the diagonal elements of \( V(\hat{\beta}) = (X'X)^{-1} \sigma^2 \), where \( \sigma^2 \) is the residual mean square (1.3).

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, \( \hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma^2/n) \), and the associated t-distributions of the two tests are identical.

A t-test for the intercept also can be constructed as in (1.1), although this does not refer to a hypothesis about a predictor, but rather about whether the expected target is equal to a specified value \( \beta_{00} \) 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. Thus, a \( 100(1-\alpha)\% \) confidence interval for \( \beta_j \) has the form

\[
\hat{\beta}_j \pm t_{n-p, \alpha/2} \cdot SE(\hat{\beta}_j)
\]

where \( t_{n-p, \alpha/2} \) is the appropriate critical value at one-sided level \( \alpha \) for a t-distribution on \( n - p - 1 \) degrees of freedom.

### 1.3.4.4 Model Values and Predictions

The rough prediction interval \( \hat{y} \pm 1 \% \) discussed in Section 1.3.2 is an approximate 95% interval because it ignores the variability caused by the need to estimate \( \sigma \), and thus 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 \( x_0 \), where \( \hat{y}_0 \) is determined by substituting the values \( x_0 \) into the estimated regression equation; its width
depends on both \( \hat{\sigma} \) and the position of \( x_0 \) 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:

\[
\hat{\sigma} \hat{E}(\hat{y}_0^*) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\hat{\beta}_0^2}}.
\]

More generally, the variance of a predicted value is:

\[
\hat{\sigma}^2 \hat{V}(\hat{y}_0^*) = [1 + x_0^T (X'X)^{-1} x_0] \hat{\sigma}^2.
\]

(1.4)

Here \( x_0 \) is taken to include a 1 in the first entry (corresponding to the intercept in the regression model). The prediction interval is then:

\[
\hat{y}_0 = \hat{\beta}_0^* + \hat{\beta}_1^* x_0 + \epsilon_0 \hat{\sigma} \sqrt{\frac{x_0'x_0}{n-2}}
\]

where \( \hat{\sigma} \hat{E}(\hat{y}_0^*) = \sqrt{\hat{\sigma}^2 \hat{V}(\hat{y}_0^*)} \).

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 a new member of the population with a particular value of \( x_0 \); 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 \( x_0 \). The corresponding standard error, \( \hat{\sigma} \), is the square root of

\[
\hat{\sigma}^2 = x_0^T (X'X)^{-1} x_0 \hat{\sigma}^2.
\]

(1.5)

with corresponding confidence interval

\[
\hat{y}_0 = \hat{\beta}_0^* + \hat{\beta}_1^* x_0 + \epsilon_0 \hat{\sigma} \sqrt{\frac{x_0'x_0}{n-2}}.
\]

A comparison of the two estimated variances (1.4) and (1.5) shows that the variance of the predicted value has an extra \( \epsilon_0^2 \) 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.

(1.6)
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, other 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 dataset.

(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 placing the observed ordered residuals on one axis and the expected values under normality 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 mean and intervals might not be meaningful.

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 extrapolated model is applied. Indeed, the assumption is stronger than that, since it must be the case that the future is essentially the same as the past, in the sense that all of the properties of the model, including its precise values of all of the regression parameters, are the same. This is unlikely to be the case, 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 fit 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...
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 noisy effects related to location means that these 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 homes such as the number of bedrooms, the living area, the lot size, and so on. Academic research on this topic is plentiful, going back several decades: Walde (1970).

This analysis is based on a sample from public data on sales of one-family homes in the Lassiteres, NY area from June 2000 through May 2001. Lassiteres is famous as the first planned suburban community built using mass production methods, being aimed at lower-middle members of the military after World War II. Most of the homes in this community were built in the late 1940s or 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. To any analysis the first step is to look at the data, and figure 1.4 presents 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 none that were built 5 or 7 years later than all of the other houses in the sample.

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

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>7.148e+06</td>
<td>3.320e+06</td>
<td>2.14</td>
<td>0.03820</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>1.223e+04</td>
<td>9.347e+03</td>
<td>1.31</td>
<td>0.19236</td>
</tr>
<tr>
<td>Bathrooms</td>
<td>3.176e+04</td>
<td>1.309e+04</td>
<td>3.19</td>
<td>0.00171</td>
</tr>
<tr>
<td>Living area</td>
<td>6.396e+03</td>
<td>1.588e+03</td>
<td>2.12</td>
<td>0.03806</td>
</tr>
<tr>
<td>Lot size</td>
<td>5.978e-01</td>
<td>2.194e+00</td>
<td>-0.28</td>
<td>0.77797</td>
</tr>
<tr>
<td>Year built</td>
<td>3.176e+03</td>
<td>1.963e+03</td>
<td>1.51</td>
<td>0.13598</td>
</tr>
<tr>
<td>Property tax</td>
<td>1.476e+00</td>
<td>2.832e+00</td>
<td>0.52</td>
<td>0.60373</td>
</tr>
</tbody>
</table>
FIGURE 1.4 Scatter plots of sale price versus each predictor for the housing price data.

Significance codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17380 on 76 degrees of freedom
Multiple R-squared: 0.3065, Adjusted R-squared: 0.2685
F-statistic: 3.332 on 6 and 76 DF, p-value: 2.416e-10

The overall regression is weakly statistically significant, with the tail probability of the F-test roughly 10^-4. The predictors account for roughly 30% of the variability in sale prices ($R^2 = 0.30$). Two of the predictors (number of bedrooms and living area) are highly statistically significant, with tail probabilities less than 0.001, 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 bedroom in a house is associated with an estimated expected price that is $51,700 higher; one additional square
front of living area is associated with an estimated expected price that is $50.00 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 $300.00 higher, and a house being built one year later is associated with an estimated expected price that is $200.00 higher.

This is a situation where the distinction between a confidence interval for a fixed 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, 1000 square feet of living area, 6000 square feet lot size, built in 1978, with a 25% 25% 25% for property taxes. Substituting these values into the above equation gives an estimated expected sale price of a house with these characteristics equal to $226,380. 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 \( \hat{\sigma} = 547.380 \) can be used to construct a rough prediction interval, so that roughly 95% of the time a house with these characteristics can be expected to sell for within \( \pm 1.96(47.380) = \pm 947.380 \) of that estimated sale price, but a more exact interval might be required. On the other hand, a house appraiser or tax assessor is interested in the typical (average) sale price for all houses 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 fixed value is desired.

Recent 95% intervals for a house with these characteristics can be obtained from statistical software, and open one to be ($167,276, $306,344) for the prediction interval and ($223,382, $293,200) 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 more useful to be of any practical value in this case, but an interval with smaller coverage (that is expected to include the actual price only 90% of the time, say) might be useful (a 90% interval in this case would be ($223,382, $293,200), 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 1.3 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 centered 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 1.4) 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 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.
An obvious consideration at this point is that the models discussed here appear to be overparameterized; 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.

1.3.5 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 modeled at the beginning of the chapter provide a framework for an organization of the topics covered.

1. Modeling the relationship between $x$ and $y$:
   - $\hat{y}$ summarizes the expected change in $y$ for a given change in $x$, 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 $R^2$ estimate the proportion of variability in $y$ accounted for by $x$;
   - and the confidence interval for a forecast value provides a measure of the precision in estimating the expected value for a given set of predictor values.

2. Prediction of the target variable:
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, $1.925 \text{SSE}$, 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 these limits.

4. 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 $t$ test can be generalized to test other hypotheses of interest about the coefficients as well.
Since all of these methods depend on the assumptions held at the time the regression analysis is conducted, a crucial part of any regression analysis is to check these 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 this work.

BIBLIOGRAPHY

Autocorrelations: Correlations 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 fixed value: A measure of precision of the estimate of the expected target value for a given x.

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

Error variance: 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 unexplained 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 variables: Characteristics 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 intervals: The interval estimate for the future target value for an individual member of the population using the fitted regression model.

Residuals: 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 $\sigma$, the standard deviation of the errors, equaling the square root of the residual mean square.