Linear Regression

13

Chapter Outline
13.1 Simple Linear Regression
13.2 Inference about the Coefficients
13.3 Interval Estimation for $\mu_{Y|X}$ and $Y|X$
13.4 Multiple Linear Regression

In this chapter we present methods for examining the relation between a response or dependent variable and one or more predictor or independent variables. The methods are based on the linear model introduced in Chapter 12. In linear regression, we examine the relation between a normally distributed response or dependent variable and one or more continuous predictor or independent variables. In a sense, linear regression is an extension of the correlation coefficient. Although linear regression was created for the examination of the relation between continuous variables, in practice, people often use the term linear regression even when continuous and discrete independent variables are used in the analysis.

Linear regression is one of the more frequently used techniques in statistics today. These methods are often used because problems, particularly those concerning humans, usually involve several independent variables. For example, in the creation of norms for lung functioning, age, race, and sex are taken into account. Linear regression is one approach that allows multiple independent variables to be used in the analysis. In the linear regression model, the dependent variable is the observed pulmonary function test value and age, race, and sex are the independent variables. When the dependent variable is a discrete variable as in the disease status (presence or absence), logistic regression (the topic of the next chapter) is used to consider many possible risk factors related to the disease.

13.1 Simple Linear Regression

Simple linear regression is used to examine the relation between a normally distributed dependent variable and a continuous independent variable. An example of a situation where simple linear regression is useful is the following.

Some physicians believe that there should be a standard — a value that only a small percentage of the population exceeds — for blood pressure in children (NHLBI Task Force 1987). When a standard is used, it is desirable that it be easy for the physician to quickly and accurately determine how the patient relates to the standard. Therefore, the standards should be based on a small number of variables that are easy to measure. Since it is known that blood pressure is related to maturation, the variables used in the development of the standard should, therefore, reflect maturation. Two variables that are

related to maturation and are easy to measure are age and height. Of these two variables, height appears to be the more appropriate variable for the development of standards (Forthofer 1991; Gillum, Prineas, and Horibe 1982; Voors et al. 1977). Because of physiological differences, the standards are developed separately for females and males. In the following, we shall focus on systolic blood pressure (SBP).

In developing the standards, we are going to assume that the mean SBP for girls increases by a constant amount for each one unit increase in height. The use of the mean instead of the individual SBP values reflects the fact that there is variation in the SBP of girls of the same height. Not all the girls who are 50 inches tall have the same SBP value; their SBPs vary about the mean SBP of girls who are 50 inches tall. The assumption of a constant increase in the mean SBP for each one unit increase in height is characteristic of a linear relation. Thus, in symbols, the relation between Y, the SBP variable, and X, the height variable, can be expressed as

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

where $\mu_{Y|X}$ is the mean SBP for girls who are X units tall, β_0 is a constant term, and β_1 is the coefficient of the height variable — that is, β_1 is the increase in the mean SBP for each one unit change in height. The β_0 coefficient is the Y intercept and β_1 is the slope of the straight line.

In general, the X variable shown in the preceding expression may represent the square, the reciprocal, the logarithm, or some other nonlinear transformation of a variable. This is acceptable in linear regression because the expression is really a linear combination of the β_i's, not of the independent variables.

The preceding equation is similar to the linear growth model in Chapter 3 and the linear model representation of ANOVA. In the ANOVA model, values of the X variables, 1 or 0, indicate which effect should be added in the model. In the regression model, the values of the X variable are the individual observations of the continuous independent variable. The parameters in the ANOVA model are the effects of the different levels of the independent variable. In the regression model, the parameters are the Y-intercept and the slope of the line.

Figure 13.1 shows the graph of this simple linear regression equation. The ⊗ symbols show the values of the mean SBP for the different values of height that we are considering. As we can see, a straight line does indeed have a rate of increase in the mean SBP that is constant for each one unit increase in height. The ■ symbols show the projected values of the mean SBP, assuming that the relationship holds for very small height values as well. It is usually inappropriate to estimate the values of $\mu_{Y|X}$ for values of X outside the range of observation. The point at which the projected line intersects the $\mu_{Y|X}$ axis is β_0. Since β_1 is the amount of increase in $\mu_{Y|X}$ for each one unit increase in X, the bracketed change in $\mu_{Y|X}$ is 8 β_1, since X has increased 8 units from x_1 to x_2. Note that if the regression line is flat — that is, parallel to the X axis — this means that there is no change in $\mu_{Y|X}$ regardless of how much X changes. Thus, if the regression line is flat, then β_1 is zero and there is no linear relation between $\mu_{Y|X}$ and X.

If we wish to express this relationship in terms of individual observations, we must take the variation in SBP for each height into account. The model that does this is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

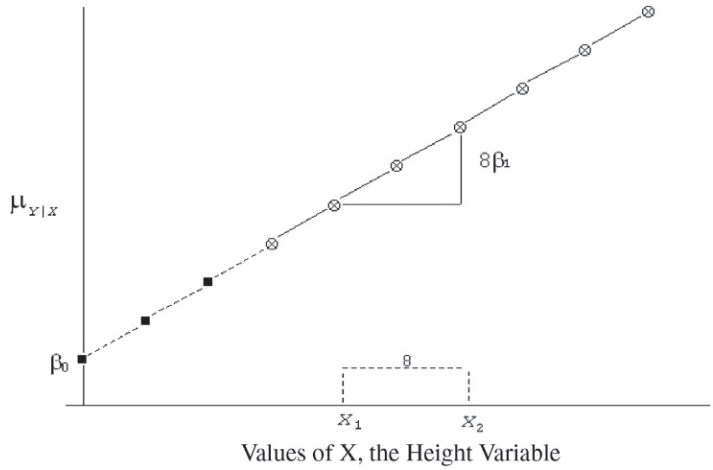

Figure 13.1 Line showing the regression of $\mu_{Y|X}$ on X.

where ε_i represents the difference between the mean SBP value at height x_i and the SBP of the ith girl who is also x_i units tall. The ε term is also referred to as the residual or error term. Knowledge of β_0 and β_1 is necessary in developing the standards for SBP. However, we do not know them and we have to collect data to estimate these values.

13.1.1 Estimation of the Coefficients

There are a variety of ways of estimating β_0 and β_1. We must decide on what criterion we will use to find the "best" estimators of these two coefficients. Possible criteria include minimization of the following:

1. The sum of the differences of y_i and $\hat{y}_i$, where y_i is the observed value of the SBP and $\hat{y}_i$ is the estimated value of the SBP for the ith girl. The value of $\hat{y}_i$ is found by substituting the estimates of β_0 and β_1 in the simple linear regression equation — that is, $\hat{y}_i = \hat{\beta}_0 + x_i\hat{\beta}_1$, where x_i is the observed value of height for the ith girl.
2. The sum of the absolute differences of y_i and $\hat{y}_i$.
3. The sum of the squared differences of y_i and $\hat{y}_i$.

The first criterion can be made to equal zero by setting $\hat{\beta}_1$ to zero and letting $\hat{\beta}_0$ equal to the sample mean. The use of the absolute value yields interesting estimators, but the testing of hypotheses is more difficult with these estimators. Based on considerations similar to those discussed in Chapter 3 in the presentation of the variance, we are going to use the third criterion to determine our "best" estimators.

Thus our estimators of the coefficients will be derived based on the minimization of the sum of squares of the differences of the observed and estimated values of SBP. In symbols, this is the minimization of

$$\sum_i (y_i - \hat{y}_i)^2.$$

The use of this criterion provides estimators that are called *least squares estimators* because they minimize the sum of squares of the differences.

The least squares estimators of the coefficients are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

The second formula for $\hat{\beta}_1$ is provided because it is easier to calculate. Let's use these formulas to calculate the least squares estimates for the data in Table 13.1. The hypothetical values of the SBP and height variables for the 50 girls are based on data from the NHANES II (Forthofer 1991).

The value of $\hat{\beta}_1$ is found from

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{269902 - 50(52.5)(101.5)}{142319 - 50(52.5)^2} = 0.7688.$$

The calculation of $\hat{\beta}_0$ is easier to perform, and its value is found from

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 101.5 - 0.7688(52.5) = 61.138.$$

Table 13.1 Hypothetical data — SBP and predicted SBP[a] (mmHg) and height (inches) for 50 girls.

SBP	Predicted SBP	Height	SBP	Predicted SBP	Height	SBP	Predicted SBP	Height
105	88.8	36	120	98.0	48	94	106.5	59
90	89.6	37	114	98.8	49	88	107.3	60
82	90.4	38	78	98.8	49	110	107.3	60
96	90.4	38	116	99.6	50	124	107.3	60
82	91.1	39	74	99.6	50	86	108.0	61
74	91.1	39	80	100.3	51	120	108.0	61
104	91.9	40	98	101.1	52	112	108.8	62
100	91.9	40	90	101.9	53	100	109.6	63
80	92.7	41	92	102.7	54	122	110.3	64
98	93.4	42	80	102.7	54	122	110.3	64
96	94.2	43	88	102.7	54	110	111.1	65
86	95.0	44	104	103.4	55	124	111.1	65
88	95.0	44	100	104.2	56	122	111.9	66
128	95.0	44	126	105.0	57	94	112.6	67
118	95.7	45	108	105.7	58	110	112.6	67
90	96.5	46	106	106.5	59	140	114.2	69
108	98.0	48	98	106.5	59			

[a]Predicted using the least squares estimates of the regression coefficients.

The estimated coefficient of the height variable is about 0.8, which means that there is an increase of 0.8 mmHg in SBP for an increase of 1 inch in height for girls between the heights of 36 and 69 inches. The estimate of the β_0 coefficient is about 60 mmHg and that is the *Y* intercept. Based on projecting the regression line beyond the data values observed, the *Y* intercept gives the value of SBP for a girl 0 inches tall. However, it does not make sense to talk about the SBP for a girl 0 inches tall, and this shows one of the dangers of extrapolating the regression line beyond the observed data.

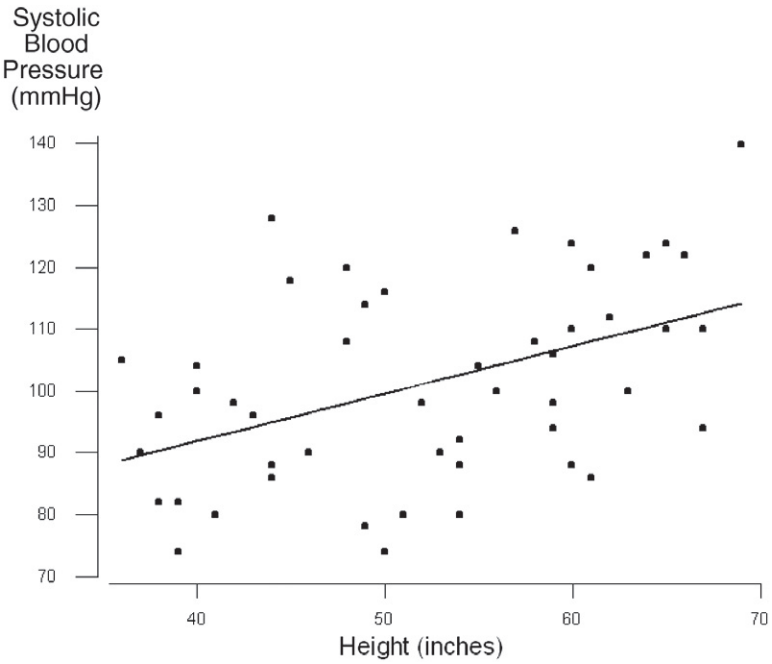

Figure 13.2 Plot of systolic blood pressure versus height for 50 girls shown in Table 13.1.

Figure 13.2 is a plot of SBP versus height for the data shown in Table 13.1. From this plot, we can see that there is a slight tendency for the larger values of SBP to be associated with the larger values of height, but the relationship is not particularly strong. The path of the regression line is shown within the range of observations.

We can use the preceding estimates of the population coefficients in predicting SBP values for the hypothetical data shown in Table 13.1. For example, the predicted value of SBP for the first observation in Table 13.1, a girl 36 inches tall, is

$$61.138 + 0.7688(36) = 88.82 \text{ mmHg}.$$

The other predicted SBP values are found in the same way, and they are also shown in Table 13.1.

13.1.2 The Variance of *Y|X*

Before going forward with the use of the regression line in the development of the standards, we should examine whether or not the estimated regression line is an improvement over simply using the sample mean as an estimate of the observed values. One way of obtaining a feel for this is to examine the sum of squares of deviations of Y from $\hat{Y}$ — that is,

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

If we subtract and add $\bar{y}$ in this expression, we can rewrite this sum of squares as

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

and we have not changed the value of the sum of squares. However, this sum of squares can be rewritten as

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

because the crossproduct terms, $(y_i-\overline{y})(\hat{y}_i-\overline{y})$, sum to zero. In regression terminology, the first sum of squares is called the *sum of squares about regression* or the *residual or error sum of squares*. The second sum of squares, about the sample mean, is called the *total sum of squares* (corrected for the mean) and the third sum of squares is called the *sum of squares due to regression*. If we rewrite this equation, putting the total sum of squares (corrected for the mean) on the left side of the equal sign, we have

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

This equation shows the partition of the total sum of squares into two components, the sum of squares about regression, and the sum of squares due to regression.

Figure 13.3 is a graph which shows the differences, $(y_i-\overline{y})$, $(y_i-\hat{y}_i)$ and $(\hat{y}_i-\overline{y})$, for one y_i. In Figure 13.3, the regression line is shown as well as a horizontal line that shows the value of the sample mean. We have focused on the last point, the girl who is 69 inches tall and who has an SBP of 140 mmHg. For this point, the deviation of the observed SBP of 140 from the sample mean of 101.5 can be partitioned into two components. The first component is the difference between the observed value and 114.2, the value predicted from the regression line. The second component is the difference between this predicted value and the sample mean. This partitioning cannot be done for many of the points, since, for example, the sample mean may be closer to the observed point than the regression line is.

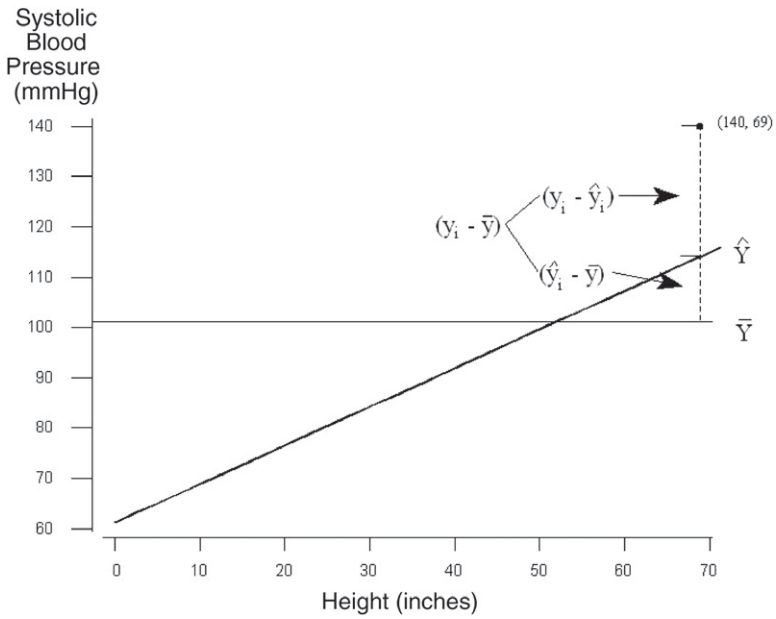

Figure 13.3 An observed value in relation to the regression line and the sample mean.

Ideally, we would like the sum of squares about the regression line to be close to zero. From the last preceding equation, we see that the sum of the square deviations

from the regression line must be less than or equal to the sum of the square deviations from the sample mean. However, the direct comparison of the sum of squares is not fair, since they are based on different degrees of freedom. The sum of squares about the sample mean has $n - 1$ degrees of freedom, as we discussed in the material about the variance. Since we estimated two coefficients in obtaining the least squares estimator of Y, there are thus $n - 2$ degrees of freedom associated with sum of squares about $\hat{Y}$. Thus, let us compare s_Y^2 with $s_{Y|X}^2$ — that is,

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

If $s_{Y|X}^2$ is much less than s_Y^2, then the regression was worthwhile; if not, then we should use the sample mean as there appears to be little linear relation between Y and X.

Let us calculate the sample variance of Y and the sample variance of Y, taking X into account. The sample variance of Y is

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{12780}{50-1} = 260.827$$

and the sample variance of Y given X is

$$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{10117}{50-2} = 210.772.$$

Thus, $s_{Y|X}^2$ is less than s_Y^2. The use of the height variable has reduced the sample variance from 260.827 to 210.772, about a 20 percent reduction. It appears that the inclusion of the height variable has allowed for somewhat better estimation of the SBP values.

13.1.3 The Coefficient of Determination (R^2)

An additional way of examining whether or not the regression was helpful is to divide the sum of squares due to regression by the sum of squares about the mean — that is,

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

If the regression line provides estimates of the SBP values that closely match the observed SBP values, this ratio will be close to one. If the regression line is close to the mean line, then this ratio will be close to zero. Hence, the ratio provides a measure that varies from 0 to 1, with 0 indicating no linear relation between Y and X, and 1 indicating a perfect linear relation between Y and X. This ratio is denoted by R^2, and is called the *coefficient of determination*. It is a measure of how much of the variation in Y is accounted for by X. R^2 is also the square of the sample Pearson correlation coefficient between Y and X.

For the SBP example, the value of R^2 is

$$\frac{12780-10117}{12780}=0.2084.$$

Approximately 21 percent of the variation in SBP is accounted for by height for girls between 36 to 69 inches tall. This is not an impressive amount. Almost 80 percent of the variation in SBP remains to be explained. Even though this measure of the relation between SBP and height is only 21 percent, it is larger than its corresponding value for the relation between SBP and age.

The derivation of the R^2 term is based on a linear model that has both a β_0 and a β_1 term. If the model does not include β_0, then a different expression must be used to calculate R^2.

The sample Pearson correlation coefficient, r, is defined as

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

and its numerical value is

$$r=\frac{3464.5}{\sqrt{4506.6(12780)}}=0.4565.$$

If we square r, r^2 is 0.2084, which agrees with R^2, as it must.

Although, symbolically, R^2 is the square of the sample Pearson correlation coefficient, R^2 does not necessarily measure the strength of the linear association between Y and X. In correlation analysis, the observed pairs of values of Y and X are obtained by simple random sampling from a population. In correlation analysis, we don't necessarily consider one of the variables to be the dependent variable and the other the independent variable. The sample r simply measures the strength of the linear association between the two variables. In contrast, linear regression provides a formula that describes the linear relation between a dependent variable and an independent variable(s). To discover that relationship, we often use stratified random sampling — that is, we select simple random samples of Y for specified values of X; however, as Ranney and Thigpen (1981) show, the value of R^2 depends on the range of the values of X used in the analysis, the number of repeated observations at given values of X, and the location of the X values. Hence, although symbolically R^2 is the square of the correlation coefficient between two variables, it does not necessarily measure the strength of the linear association between the variables. It does reflect how much of the variation in Y is accounted for by knowledge of X. Korn and Simon provide more on the interpretation of R^2 (Korn and Simon 1991).

There is also a relation between the sample correlation coefficient and the estimator of β_1. From Chapter 3, we had another form for r than the defining formula given above and it was

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})/(n-1)}{s_x s_y}.$$

The estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})/(n-1)}{s_x^2}.$$

If we multiply r by s_y and divide r by s_x, we have

$$\left(\frac{s_y}{s_x}\right) r = \left(\frac{s_y}{s_x}\right) \frac{\sum (x_i - \bar{x})(y_i - \bar{y})/(n-1)}{s_x s_y}$$

or

$$\left(\frac{s_y}{s_x}\right) r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})/(n-1)}{s_x^2} = \hat{\beta}_1.$$

As the preceding relation shows, if the correlation coefficient is zero, the slope coefficient is also zero and vice versa.

13.2 Inference about the Coefficients

The parametric approach to testing hypotheses about a parameter requires that we know the probability distribution of the sample estimator of the parameter. The standard approach to finding the probability distributions of the sample estimators of β_0 and β_1 is based on the following assumptions.

13.2.1 Assumptions for Inference in Linear Regression

We assume that the y_i's are independent, normally distributed for each value of X, and that the normal distributions at the different values of X all have the same variance, σ^2. Figure 13.4 graphically shows these assumptions. The regression line, showing the relation between $\mu_{Y|X}$ and X, is graphed as well as the distributions of Y at the selected values of

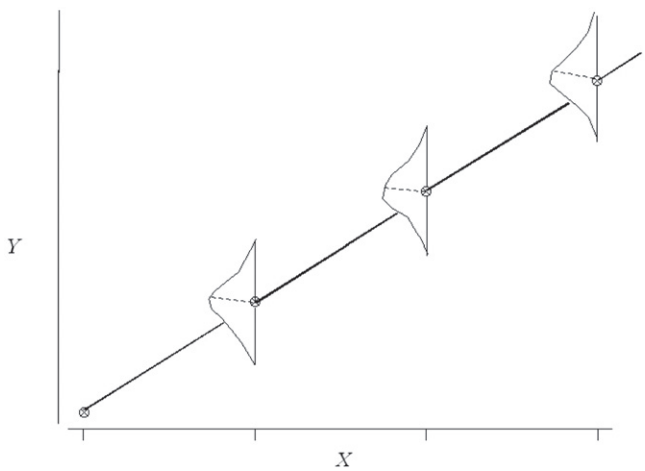

Figure 13.4
Distribution of Y at selected values of X.

X. Note that Y is normally distributed at each of the selected X values and that the normal distributions have the same shapes — that is, the same variance, σ^2. The mean of the normal distribution, $\mu_{Y|X}$, is obtained from the regression equation and is $\beta_0 + \beta_1 X$.

In the following, we shall consider the values of the X variable to be fixed. There are two ways that X can be viewed as being fixed. First, we may have used a stratified sample, stratified on height, to select girls with the heights shown in Table 13.1. Since we have chosen the values of the height variable, they are viewed as being fixed. In a second way, we consider our results to be conditional on the observed values of X. The conditional approach is usually used with simple random samples in which both Y and X otherwise would be considered to be random variables. This is the conventional approach, and it means that the error or residual term, ε, also follows a normal distribution with mean 0 and variance σ^2. Note that the least squares estimation of the regression coefficients did not require this specification of the probability distribution of Y.

Before testing hypotheses about the regression coefficients, we should attempt to determine whether or not the assumptions just stated are true. We should also examine whether or not any single data point is exercising a large influence on the estimates of the regression coefficients. These two issues are discussed in the next section.

13.2.2 Regression Diagnostics

In our brief introduction to regression diagnostics — methods for examining the regression equation — we consider only two of the many methods that exist. More detail on other methods is given in Kleinbaum et al. (1998). The first method we shall present involves plotting of the residuals. Plots are used in an attempt to determine whether or not the residuals or errors are normally distributed or to see if there are any patterns in the residuals. The second method tries to discover the existence of data points that play a major role in the estimation of the regression coefficients.

Residuals and Standardized Residuals: The sample estimator of ε_i is the residual e_i, defined as the difference between y_i and $\hat{y}_i$, and the e_i can be used to examine the regression assumptions. Since we are used to dealing with standardized variables, people often consider a standardized residual, $e_i/s_{Y|X}$, instead of e_i itself. The standardized residuals should approximately follow a standard normal distribution if the regression assumptions are met. Thus, values of the standardized residuals larger than 2.5 or less than −2.5 are unusual. Table 13.2 shows these residuals and a quantity called leverage (described in the next section) for the data in Table 13.1.

We use the standardized residuals in our examination of the normality assumption. Other residuals could also be used for this examination (Kleinbaum 1998). The normal scores of the standardized residuals are plotted in Figure 13.5. The normal scores plot looks reasonably straight; thus the assumption that the error term is normally distributed does not appear to be violated.

If this plot deviates sufficiently from a straight line to cause us to question the assumption of normality, then it may be necessary to consider a transformation of the dependent variable. There are a number of mathematical functions which can be used to transform nonnormally distributed data to normality (Kleinbaum 1998; Lin and Vonesh 1989; Miller 1984).

Table 13.2 Residuals and leverage for the data in Table 13.1.

Y	Residual	Standardized Residual	Leverage h_i	Y	Residual	Standardized Residual	Leverage h_i
105	16.1848	1.16253	0.08041	92	−10.6532	−0.74143	0.02049
90	0.4161	0.02977	0.07331	80	−22.6532	−1.57659	0.02049
82	−8.3527	−0.59552	0.06665	88	−14.6532	−1.01982	0.02049
96	5.6473	0.40264	0.06665	104	0.5781	0.04025	0.02138
82	−9.1215	−0.64818	0.06044	100	−4.1907	−0.29199	0.02271
74	−17.1215	−1.21667	0.06044	126	21.0405	1.46735	0.02449
104	12.1097	0.85790	0.05467	108	2.2717	0.15861	0.02671
100	8.1097	0.57452	0.05467	106	−0.4971	−0.03475	0.02937
80	−12.6590	−0.89430	0.04934	98	−8.4971	−0.59407	0.02937
98	4.5722	0.32218	0.04446	94	−12.4971	−0.87373	0.02937
96	1.8034	0.12678	0.04002	88	−19.2658	−1.34912	0.03248
86	−8.9654	−0.62897	0.03603	110	2.7342	0.19146	0.03248
88	−6.9654	−0.48866	0.03603	124	16.7342	1.17184	0.03248
128	33.0346	2.31756	0.03603	86	−22.0346	−1.54585	0.03603
118	22.2658	1.55920	0.03248	120	11.9654	0.83944	0.03603
90	−6.5029	−0.45465	0.02937	112	3.1966	0.22473	0.04002
108	9.9595	0.69457	0.02449	100	−9.5722	−0.67450	0.04446
120	21.9595	1.53144	0.02449	122	11.6590	0.82366	0.04934
114	15.1907	1.05843	0.02271	122	11.6590	0.82366	0.04934
78	−20.8093	−1.44991	0.02271	110	−1.1097	−0.07862	0.05467
116	16.4219	1.14344	0.02138	124	12.8903	0.91320	0.05467
74	−25.5781	−1.78096	0.02138	122	10.1215	0.71924	0.06044
80	−20.3468	−1.41608	0.02049	94	−18.6473	−1.32950	0.06665
98	−3.1156	−0.21679	0.02005	110	−2.6473	−0.18874	0.06665
90	−11.8844	−0.82693	0.02005	140	25.8152	1.85426	0.08041

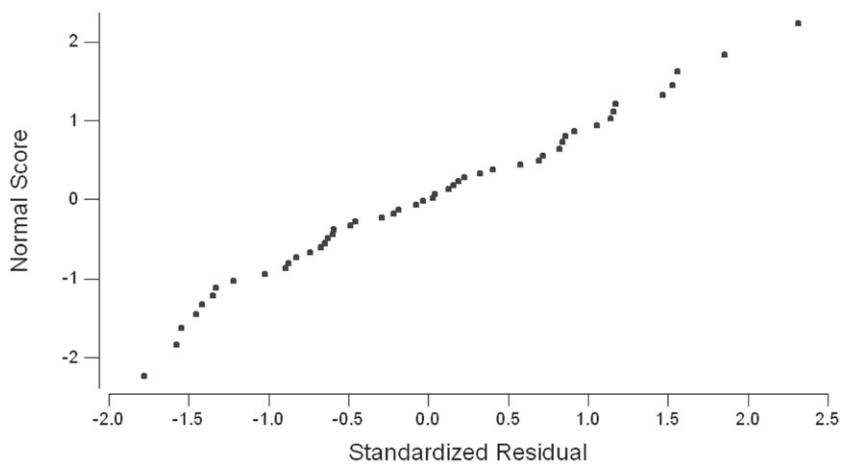

Figure 13.5 Normal scores plot of the standardized residuals from the linear regression of systolic blood pressure on height.

It is also of interest to plot the standardized residuals against the values of the X variable(s). If any pattern is observed in this plot, it suggests that another term involving the X variable — for example, X^2, might be needed in the model. Figure 13.6 shows the plot of the standardized residuals versus the height variable. No pattern is immediately obvious from an examination of this plot. Again, there is no evidence to cause us to reject this model. If the data have been collected in time sequence, it is also useful to examine a plot of the residuals against time.

***Leverage*:** The predicted values of Y are found from

$$\hat{\beta}_0 + \hat{\beta}_1 X$$

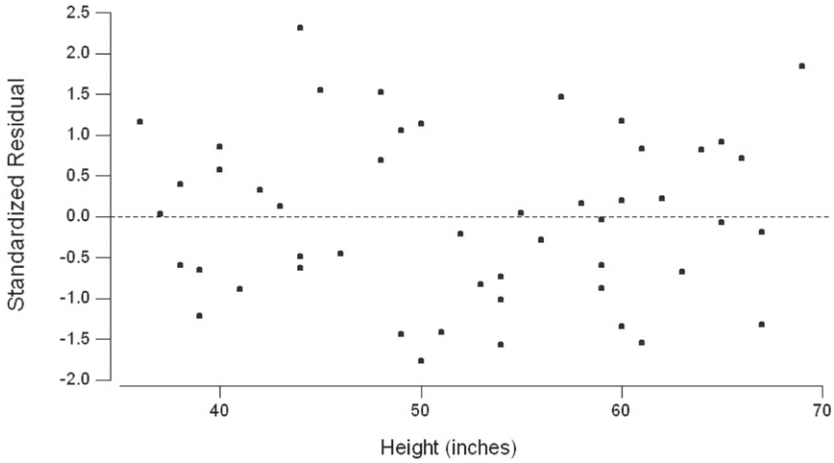

Figure 13.6 Plot of standardized residuals versus height.

where the estimators of β_0 and β_1 are linear combinations of the observed values of Y. Thus, the predicted values of Y are also linear combinations of the observed values of Y. An expression for the predicted value of y_i reflecting this relation is

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n$$

where h_{ij} is the coefficient of y_j in the expression for $\hat{y}_i$. For simplicity, h_{ii} is denoted by h_i. The effect of y_i on its predicted value is denoted by h_i and this effect is called *leverage*. Leverage shows how much change there is in the predicted value of y_i per unit change in y_i. The possible values of the h_i are greater than or equal to zero and less than or equal to one. The average value of the leverages is the number of estimated coefficients in the regression equation divided by the sample size. In our problem, we estimated two coefficients and there were 50 observations. Thus the average value of the leverages is 0.04 (= 2/50). If any of the leverages are large — some statisticians consider large to be greater than twice the average leverage and others say greater than three times the average — the points with these large leverages should be examined. Perhaps there was a mistake in recording the values or there is something unique about the points that should be examined. If there is nothing wrong or unusual with the points, it is useful to perform the regression again excluding these points. A comparison of the two regression equations can be made, and the effect of the excluded points can be observed.

In our problem, we can see from Table 13.2 that there are two points, the first and the last, with the larger leverages. Both of these points had leverages slightly larger than twice the average leverage value. The first girl had a large SBP value relative to her height, and the last girl had the highest SBP value. At this stage, we will assume that there was no error in recording or entering the data. We could perform the regression again and see if there is much difference in the results. However, since the leverages are only slightly larger than twice the average leverage, we shall not perform any additional regressions.

Based on these looks at the data, we have no reason to doubt the appropriateness of the regression assumptions and there do not appear to be any really unusual data points that would cause us concern. Therefore, it is appropriate to move into the inferential part of the analysis, that is, to test hypotheses and to form confidence and prediction intervals. We begin the inferential stage with consideration of the slope coefficient.

13.2.3 The Slope Coefficient

Even though there is an indication of a linear relation between SBP and height — that is, it appears that β_1 is not zero — we do not know if β_1 is statistically significantly different from zero. To determine this, we must estimate the standard error of $\hat{\beta}_1$, which is used in both confidence intervals and tests of hypotheses about β_1. To form the confidence interval about β_1 or to test a hypothesis about it, we also must know the probability distribution of $\hat{\beta}_1$.

Since we are assuming that Y is normally distributed, this means that $\hat{\beta}_1$, a linear combination of the observed Y values, is also normally distributed. Therefore, to form a confidence interval or to test a hypothesis about β_1, we now need to know the standard error of its estimator. The standard error (*s.e.*) of $\hat{\beta}_1$ is

$$s.e.(\beta_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and, because σ is usually unknown, the standard error is estimated by substituting $s_{Y|X}$ for σ. From the above equation, we can see that the magnitude of the standard error depends on the variability in the X variable. Larger variability decreases the standard error of $\hat{\beta}_1$. Thus, we should be sure to include some values of X at the extremes of X over the range of interest.

To test the hypothesis that β_1 is equal to β_{10} — that is,

$$H_0\colon \beta_1 = \beta_{10},$$

we use the statistic

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{est.s.e.(\hat{\beta}_1)} = \frac{(\hat{\beta}_1 - \beta_{10})\sqrt{\sum(x_i - \bar{x})^2}}{s_{Y|X}}.$$

If σ were known, the test statistic, using σ instead of $s_{Y|X}$, would follow the standard normal distribution; however, σ is usually unknown, and the test statistic using $s_{Y|X}$ follows the t distribution with $n - 2$ degrees of freedom. The degrees of freedom parameter has the value of $n - 2$, since we have estimated two coefficients, β_0 and β_1.

If the alternative hypothesis is

$$H_a\colon \beta_1 \neq \beta_{10}$$

the rejection region consists of values of t less than or equal to $t_{n-2,\alpha/2}$ or greater than or equal to $t_{n-2,1-\alpha/2}$.

The hypothesis usually of interest is that β_{10} is zero — that is, there is no linear relation between Y and X. If, however, our study is one attempting to replicate previous findings, we may wish to determine if our slope coefficient is the same as that reported in the original work. Then β_{10} will be set equal to the previously reported value. Let us test the hypothesis that there is no linear relation between SBP and height versus the alternative hypothesis that there is some linear relation at the 0.05 significance level.

The test statistic, t, is

$$t = \frac{\left(\hat{\beta}_1 - \beta_{10}\right)\sqrt{\sum (x_i - \bar{x})^2}}{s_{X|Y}}$$

which is

$$t = \frac{(0.7688 - 0)\sqrt{4506.5}}{14.518} = 3.555.$$

This value is compared with –2.01 (= $t_{48,0.025}$) and 2.01 (= $t_{48,0.975}$). Since 3.555 is greater than 2.01, we reject the hypothesis of no linear relation between SBP and height. The *p*-value of this test is approximately 0.001.

The $(1 - \alpha)*100$ percent confidence interval for β_1 is formed by

$$\hat{\beta}_1 \pm t_{\mathrm{n}-2,1-\alpha/2} * est.\ s.e.\ (\hat{\beta}_1)$$

which is

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} * \frac{s_{Y|X}}{\sqrt{\sum (x_i - \bar{x})^2}}.$$

The 95 percent confidence interval for β_1 is found using

$$0.7688 \pm 2.01\frac{14.518}{\sqrt{4506.5}} = 0.7688 \pm 0.4347$$

and this gives a confidence interval from 0.3341 to 1.2035. The confidence interval is consistent with the test given above. Since zero is not contained in the confidence interval for β_1, there appears to be a linear relation between SBP and height. Since there is evidence to suggest that β_1 is not zero, this also means that the correlation coefficient between *Y* and *X* is not zero.

13.2.4 The *Y*-intercept

It is also possible to form confidence intervals and to test hypotheses about β_0, although these are usually of less interest than those for β_1. The location of the *Y* intercept is relatively unimportant compared to determining whether or not there is a relation between the dependent and independent variables. However, sometimes we wish to compare whether or not both our coefficients — slope and *Y* intercept — agree with those presented in the literature. In this case, we are interested in examining β_0 as well as β_1.

Since the estimator of β_0 is also a linear combination of the observed values of the normally distributed dependent variable, $\hat{\beta}_0$ also follows a normal distribution. The standard error of $\hat{\beta}_0$ is estimated by

$$est.s.e.\left(\beta_0\right) = s_{Y|X}\sqrt{\frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}}.$$

The hypothesis of interest is

$$H_0\colon \beta_0 = \beta_{00}$$

versus either a one- or two-sided alternative hypothesis. The test statistic for this hypothesis is

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{s_{Y|X}\sqrt{\sum x_i^2 \Big/ \left[n\sum (x_i - \overline{x})^2\right]}}$$

and this is compared to $\pm t_{n-2,1-\alpha/2}$ for the two-sided alternative hypothesis. If the alternative hypothesis is that β_0 is greater than β_{00}, we reject the null hypothesis in favor of the alternative when t is greater than $t_{n-2,1-\alpha}$. If the alternative hypothesis is that β_0 is less than β_{00}, we reject the null hypothesis in favor of the alternative when t is less than $-t_{n-2,1-\alpha}$.

The $(1 - \alpha/2)*100$ percent confidence interval for β_0 is given by

$$\hat{\beta}_0 \pm t_{n-2,1-\alpha/2} s_{Y|X} \sqrt{\frac{\sum x_i^2}{n\sum (x_i - \overline{x})^2}}.$$

Let us form the 99 percent confidence interval for β_0 for these SBP data. The 0.995 value of the t distribution with 48 degrees of freedom is approximately 2.68. Therefore, the confidence interval is found from the following calculations

$$61.14 \pm 2.68(14.52)\sqrt{\frac{142319}{50(4506.5)}} = 61.14 \pm 30.93$$

which gives an interval from 30.21 to 92.07, a wide interval.

13.2.5 An ANOVA Table Summary

Table 13.3 shows the information required to test the hypothesis of no relation between the dependent and independent variables in an ANOVA table similar to that used in Chapter 12. The test statistic for the hypothesis of no linear relation between the dependent and independent variables is the F ratio, which is distributed as an F variable with 1 and $n - 2$ degrees of freedom. Large values of the F ratio cause us to reject the null hypothesis of no linear relation in favor of the alternative hypothesis of a linear relation. The F statistic is the ratio of the mean square due to regression to the mean square about regression (mean square error or residual mean square). The degrees of freedom parameters for the F ratio come from the two mean squares involved in the ratio. The degrees of freedom due to regression is the number of parameters estimated minus one. The degrees of freedom associated with the about regression source of variation is the sample size minus the number of coefficients estimated in the regression model.

Table 13.3 An ANOVA table for the simple linear regression model.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	*F* Ratio[a]
Due to Regression	1	$\Sigma(\hat{y}_i - \overline{y})^2$	$\Sigma(\hat{y}_i - \overline{y})^2/1$	MSR/MSE
About Regression or Error	$n - 2$	$\Sigma(y_i - \hat{y}_i)^2$	$\Sigma(y_i - \hat{y}_i)^2/(n-2)$	
Corrected Total	$n - 1$	$\Sigma(y_i - \overline{y})^2$		

[a]MSR is the mean square due to regression, and MSE is the mean square error term.

Table 13.4 ANOVA table for the regression of SBP on height.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	*F* Ratio
Due to Regression	1	2,663	2,663	12.63
About Regression or Error	48	10,117	210.77	
Corrected Total	49	12,780		

The ANOVA table for the SBP and height data is shown in Table 13.4. If we perform this test at the 0.05 significance level, we will compare the calculated F ratio to $F_{1,48,0.95}$, which is approximately 4.04. Since the calculated value, 12.63, is greater than the tabulated value, 4.04, we reject the null hypothesis in favor of the alternative hypothesis. There appears to be a linear relation between SBP and height at the 0.05 significance level.

Note that if we take the square root of 12.63, we obtain 3.554. With allowance for rounding, we have obtained the value of the t statistic calculated in the section for testing the hypothesis that β_1 is zero. This equality is additional verification of the relation, pointed out in Chapter 12, between the t and F statistics. An F statistic with 1 and $n - p$ degrees of freedom is the square of the t statistic with $n - p$ degrees of freedom. Examination of the t and F tables shows that $t^2_{n-p,1-\alpha/2}$ equals $F_{1,n-p,1-\alpha}$. Hence, we have two equivalent ways of testing whether or not the dependent and independent variables are linearly related at a given significance level. As we shall see in the multiple regression material, the F statistic directly extends to simultaneously testing several variables, whereas the t can be used with only one variable at a time.

These calculations associated with regression analysis require much time, care, and effort. However, they can be quickly and accurately performed with computer packages (see **Program Note 13.1** on the website).

13.3 Interval Estimation for $\mu_{Y|X}$ and $Y|X$

Even though the relation between SBP and height is not impressive, we will continue with the idea of developing a height-based standard for SBP for children. We would be much more comfortable doing this if the relation between height and SBP were stronger. The height-based standards that we shall create are the SBP levels such that 95 percent of the girls of a given height have lower SBP and 5 percent have a higher SBP. This standard is not based on the occurrence of any disease or other undesirable property. When using a standard created in this manner, approximately 5 percent of the girls will be said to have undesirably high SBP, regardless of whether or not that is really a problem.

The standard will be based on a one-sided prediction interval for the SBP variable. Also of interest is the confidence interval for the SBP variable and we shall consider the confidence interval first.

13.3.1 Confidence Interval for $\mu_{Y|X}$

The regression line provides estimates of the mean of the dependent variable for different values of the independent variable. How confident are we about these estimates or

predicted values? The confidence interval provides one way of answering this question. To create the confidence interval, we require knowledge of the distribution of $\hat{Y}$ and also an estimate of its standard error.

Since the predicted value of $\mu_{Y|X}$ at a given value of x, say x_k, is also a linear combination of normal values, it is normally distributed. Its standard error is estimated by

$$est.s.e.(\mu_{Y|X_k}) = s_{Y|X}\sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x_i - \bar{x})^2}}.$$

The estimated standard error increases with increases in the distance between x_k and $\bar{x}$, and there is a unique estimate of the standard error for each x_k.

Because we are using $s_{Y|X}$ to estimate σ, we must use the t distribution in place of the normal in the formation of the confidence interval. The confidence interval for $\mu_{Y|X}$ has the form

$$\hat{\mu}_{Y|X} \pm t_{n-2,1-\alpha/2} est.\ s.e.\ (\hat{\mu}_{Y|X}).$$

Figure 13.7 shows the 95 percent confidence interval for SBP as a function of height. As we can see from the graph, the confidence interval widens as the values of height move away from the mean of the height variable. This is in accord with the expression for the confidence interval, which has the term $(x_k - \bar{x})^2$ in the numerator. We are thus less sure of our prediction for the extreme values of the independent variable. The confidence interval is about 17 mmHg wide for girls 35 or 70 inches tall and narrows to about 8 mmHg for girls about 50 to 55 inches tall.

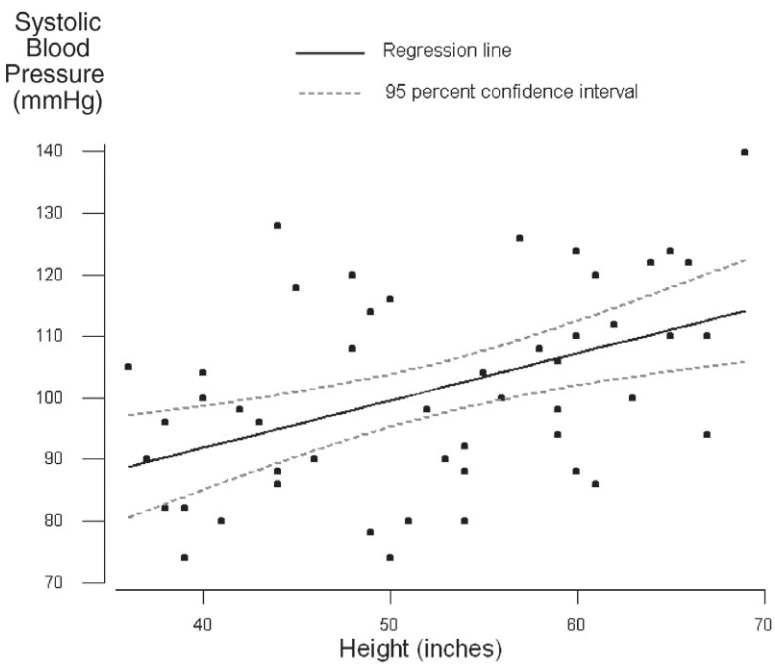

Figure 13.7 Ninety-five percent confidence interval for $\mu_{Y|X}$.

13.3.2 Prediction Interval for *Y|X*

In the preceding section, we saw how to form the confidence interval for the mean of SBP for a height value. In this section, we shall form the prediction interval — the interval for a single observation. The prediction interval is of interest to a physician because the physician is examining a single person, not an entire community. How does the person's SBP value relate to the standard?

As we saw in Chapter 7 in the material on intervals based on the normal distribution, the prediction interval is wider than the corresponding confidence interval because we must add the individual variation about the mean to the mean's variation. Similarly, the formula for the prediction interval based on the regression equation adds the individual variation to the mean's variation. Thus, the estimated standard error for a single observation is

$$est.s.e.(\hat{y}_k) = s_{Y|X}\sqrt{1+\frac{1}{n}+\frac{(x_k-\bar{x})^2}{\sum(x_i-\bar{x})^2}}.$$

The corresponding two-sided $(1-\alpha)*100$ percent prediction interval is

$$\hat{y}_k \pm t_{n-2,1-\alpha/2}\ est.\ s.e.\ (\hat{y}_k).$$

Figure 13.8 shows the 95 percent prediction interval for the data in Table 13.1. The prediction interval is much wider than the corresponding confidence interval because of the addition of the individual variation in the standard error term. The prediction interval here is about 60 mmHg wide. Note that most of the data points are within the prediction interval band. Inclusion of the individual variation term has greatly reduced the effect of the $(x_k - \bar{x})^2$ term in the estimated standard error in this example. The upper

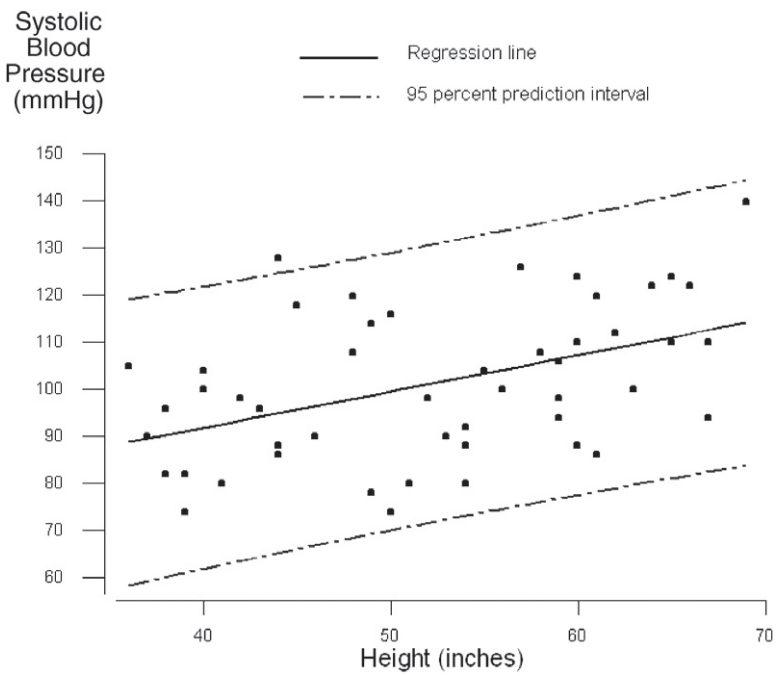

Figure 13.8 Ninety-five percent prediction interval for y_k.

and lower limits are essentially straight lines, in contrast to the shape of the upper and lower limits of the confidence interval.

Software packages can be used to perform the calculations necessary to create the 95 percent confidence and prediction intervals (see **Program Note 13.2** on the website).

Example 13.1

We apply the prediction interval to develop the standard for systolic blood pressure**.** Since we are only concerned about systolic blood pressures that may be too high, we shall use a one-sided prediction interval in the creation of the height-based standard for SBP for girls. The upper $(1 - \alpha) * 100$ percent prediction interval for SBP is found from

$$\hat{y}_k \pm t_{n-2,1-\alpha}\, est.\ s.e.(\hat{y}_k).$$

Because the standard is the value such that 95 percent of the SBP values fall below it and 5 percent of the values are greater than it, we shall use the upper 95 percent prediction interval to obtain the standard.

The data shown in Figure 13.8 can be used to help create the height-based standards for SBP. The difference between the one- and two-sided interval is the use of $t_{n-2,1-\alpha}$ in place of $t_{n-2,1-\alpha/2}$. Thus, the amount to be added to $\hat{y}_k$ for the upper one-sided interval is simply 0.834 $(= t_{48,0.95}/t_{48,0.975})$ times the amount added for the two-sided interval. To find the amount added for the two-sided interval, we subtract the predicted SBP value shown from the upper limit of the 95 percent prediction interval. For example, for a girl 35 inches tall, the amount added, using the two-sided interval, is found by subtracting 88.05 (predicted value) from 118.50 (upper limit of the two-sided prediction interval). This yields a difference of 30.45 mmHg. If we multiply this difference by 0.834, we have the amount to add to the 88.05 value. Thus, the standard for a girl 35 inches tall is

$$0.834\ (118.50 - 88.05) + 88.05 = 113.45\ \text{mmHg}.$$

Table 13.5 shows these calculations and the height-based standards for SBP for girls. As just shown, the calculations in Table 13.5 consist of taking column 2 minus

Table 13.5 Creation of height-based standards for SBP (mmHg) for girls.

x_k (Inches) (1)	Upper Limit of Prediction Interval (2)	$\hat{y}_k$ (3)	Difference (4)	Difference Times 0.834 (5)	Standard (6)
35	118.50	88.05	30.45	25.40	113.45
40	121.87	91.89	29.98	25.00	116.89
45	125.40	95.93	29.67	24.74	120.67
50	129.09	99.58	29.51	24.61	124.19
55	132.93	103.42	29.51	24.61	128.03
60	136.93	107.27	29.66	24.74	132.01
65	141.09	111.11	29.98	25.00	136.11
70	145.41	114.95	30.46	25.40	140.35

column 3. This is stored in column 4. Column 5 contains 0.834 times column 4. The standard, column 6, is the sum of column 3 with column 5.

The upper one-sided prediction interval is one way of creating height-based standards for SBP. It has the advantage over simply using the observed 95th percentiles of the SBP at the different heights in that it does not require such a large sample size to achieve the same precision. If SBP is really linearly related to height, standards based on the prediction interval also smooth out random fluctuations that may be found in considering each height separately.

The standards developed here are illustrative of the procedure. If one were going to develop standards, a larger sample size would be required. We would also prefer to use additional variables or another variable to increase the amount of variation in the SBP that is accounted for by the independent variable(s). In addition, as we just stated, the rationale for having standards for blood pressure in children is much weaker than that for having standards in adults. In adults, there is a direct linkage between high blood pressure and disease, whereas in children no such linkage exists. Additionally, the evidence that relatively high blood pressure in children carries over into adulthood is inconclusive. Use of the 95th percentile or other percentiles as the basis of a standard implies that some children will be identified as having a problem when none may exist.

So far we have focused on a single independent variable. In the next section, we consider multiple independent variables.

13.4 Multiple Linear Regression

For many chronic diseases, there is no one single cause associated with the occurrence of the disease. There are many factors, called risk factors, that play a role in the development of the disease. In the study of the occurrence of air pollution, there are many factors — for example, wind, temperature, and time of day — that must be considered. In comparing mortality rates for hospitals, factors such as the mean age of the patients, severity of the diseases seen, and the percentage of patients admitted from the emergency room must be taken into account in the analysis. As these examples suggest, it is uncommon for an analysis to include only one independent variable. Therefore, in this section we introduce multiple linear regression, a method for examining the relation between one normally distributed dependent variable and more than one continuous independent variable. We also extend the mode to include categorical independent variables.

13.4.1 The Multiple Linear Regression Model

The equation showing the hypothesized relation between the dependent and $(p - 1)$ independent variables is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{p-1} x_{p-1,i} + \varepsilon_i.$$

The coefficient β_i describes how much change there is in the dependent variable when the ith independent variable changes by one unit and the other independent variables

are held constant. Again, the key hypothesis is whether or not β_i is equal to zero. If β_i is equal to zero, we probably would drop the corresponding X_i from the equation because there is no linear relation between X_i and the dependent variable once the other independent variables are taken into account.

The regression coefficients of $(p - 1)$ independent variables and the intercept can be estimated by the least squares method, the same approach we used in the simple model presented above. We are also making the same assumptions — independence, normality, and constant variance — about the dependent variable and the error term in this model as we did in the simple linear regression model. We can also partition the sums of squares for the multiple regression model similarly to the partition used in the simple linear regression situation. The corresponding ANOVA table is

Source	DF	Sum of Squares	Mean Square	*F*-ratio
Regression	$p - 1$	$\sum(\hat{y}_i - \bar{y})^2 = SSR$	$SSR/(p - 1) = MSR$	MSR/MSE
Residual	$n - p$	$\sum(y_i - \hat{y}_i)^2 = SSE$	$SSE/(n - p) = MSE$	
Total	$n - 1$	$\sum(y_i - \bar{y})^2$		

and the overall F ratio now tests the hypothesis that the $p - 1$ regression coefficients (excluding the intercept) are equal to zero.

A goal of multiple regression is to obtain a small set of independent variables that makes sense substantively and that does a reasonable job in accounting for the variation in the dependent variable. Often we have a large number of variables as candidates for the independent variables, and our job is to reduce that larger set to a parsimonious set of variables. As we just saw, we do not want to retain a variable in the equation if it is not making a contribution. Inclusion of redundant or noncontributing variables increases the standard errors of the other variables and may also make it more difficult to discern the true relationship among the variables. A number of approaches have been developed to aid in the selection of the independent variables, and we show a few of these approaches.

The calculations and the details of multiple linear regression are much more than we can cover in this text. For more information on this topic, see books by Kleinbaum, Kupper, and Muller and by Draper and Smith, both excellent texts that focus on linear regression methods. We consider examples for the use of multiple linear regression based on NHANES III sample data that are shown in Table 13.6.

13.4.2 Specification of a Multiple Regression Model

There are no firm sample size requirements for performing a multiple regression analysis. However, a reasonable guideline is that the sample size should be at least 10 times as large as the number of independent variables to be used in the final multiple linear regression equation. In our example, there are 50 observations, and we will probably use no more than three independent variables in the final regression equation. Hence, our sample size meets the guideline, assuming that we do not add interaction terms or higher-order terms of the three independent variables.

Before beginning any formal analysis, it is highly recommend that we look at our data to see if we detect any possible problems or questionable data points. The

Table 13.6 Adult (≥18 years of age) sample data from NHANES III, Phase II (1991–1994).

Row	Race[a]	Sex[b]	Age[c]	Education[d]	Height[e]	Weight[f]	Smoke[g]	SBP[h]	BMI[i]
1	1	1	28	16	68	160	7	111	24.33
2	1	1	26	12	68	165	1	101	25.09
3	2	2	31	15	68	175	1	120	26.61
4	2	1	18	12	76	265	7	158	32.26
5	1	1	50	17	67	145	1	125	22.71
6	2	1	42	12	69	247	1	166	36.48
7	1	2	20	12	66	156	7	114	25.18
8	1	1	29	12	76	180	1	143	21.91
9	1	2	35	12	63	166	2	111	29.41
10	1	1	47	16	66	169	1	133	27.28
11	1	2	20	14	69	120	7	95	17.72
12	1	2	33	16	68	133	7	113	20.22
13	4	1	24	13	71	185	7	128	25.80
14	1	1	28	14	72	150	1	110	20.34
15	1	2	32	8	61	126	1	117	23.81
16	2	1	21	10	68	190	1	112	28.89
17	1	1	28	17	71	150	7	110	20.92
18	1	2	60	12	61	130	7	117	24.56
19	1	1	55	12	66	215	2	142	34.70
20	1	2	74	12	65	130	7	105	21.63
21	1	2	38	16	68	126	7	94	19.16
22	1	1	26	14	66	160	2	131	25.82
23	1	1	52	9	74	328	2	128	42.11
24	1	2	25	16	69	125	7	93	18.46
25	1	2	24	12	67	133	1	103	20.83
26	1	2	26	16	59	105	1	114	21.21
27	1	2	51	13	64	119	7	130	20.43
28	2	2	29	16	62	98	7	105	17.92
29	4	1	26	0	64	150	7	117	25.75
30	1	2	60	12	64	175	1	124	30.04
31	1	1	22	9	70	190	1	122	27.26
32	1	2	19	12	65	125	7	112	20.80
33	3	1	39	12	73	210	1	135	27.71
34	3	2	77	4	62	138	7	150	25.24
35	1	1	39	12	73	230	2	125	30.34
36	1	1	40	11	69	170	1	126	25.10
37	1	2	44	13	62	115	7	99	21.03
38	3	2	27	9	61	140	7	114	26.45
39	1	1	29	14	73	220	7	139	29.03
40	1	2	78	11	63	110	7	150	19.49
41	1	1	62	13	65	208	7	112	34.61
42	1	1	22	10	71	125	1	127	17.43
43	1	2	37	11	64	176	7	125	30.21
44	1	1	38	17	72	195	7	136	26.45
45	3	1	22	12	65	140	7	108	23.30
46	3	1	79	0	61	125	2	156	23.62
47	1	2	24	12	62	146	7	108	26.70
48	1	2	32	13	67	141	2	105	22.08
49	1	1	42	16	70	192	7	121	27.55
50	1	1	42	14	68	185	7	126	28.13

[a](1 = white, 2 = black, 3 = Hispanic, 4 = other); [b](1 = male; 2 = female); [c]Age in years; [d]Number of years of education; [e]Height (inches); [f]Weight (pounds); [g](1 = current smoker, 2 = never, 7 = previous); [h]Systolic blood pressure (mmHg); [i]Body mass index

descriptive statistics, such as the minimum and maximum, along with different graphical procedures, such as the box plot, are certainly very useful. A simple examination of the data in Table 13.6 finds that there are two people with zero years of education. One of these people is 26 years old and the other is 79 years old. Is it possible that someone 26 years old didn't go to school at all? It is possible but highly unlikely. Before

using the education variable in any analysis, we should try to determine more about these values.

We consider building a model for SBP based on weight, age, and height. Before starting with the multiple regression analysis, it may be helpful to examine the relationship among these variables using a *scatterplot matrix* shown in Figure 13.9. It is essentially a grid of scatterplots for each pair of variables. Such a display is often useful in assessing the general relationships between the variables and in identifying possible outliers. The individual relationships of SBP to each of the explanatory variables shown in the first column of the scatterplot matrix do not appear to be particularly impressive, apart perhaps from the weight variable.

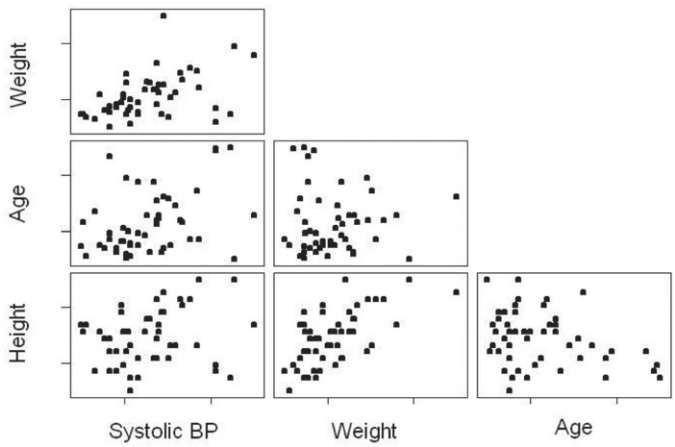

Figure 13.9 Scatterplot matrix for systolic blood pressure, weight, age, and height.

It may also be helpful to examine the correlation among the variables under consideration. The simple correlation coefficients among these variables can be represented in the format shown in Table 13.7. The correlation between SBP and weight is 0.465, the largest of the correlations between SBP and any of the variables. The correlation between height and weight is 0.636, the largest correlation in this table. It is clear from these estimates of the correlations among these three independent variables that they are not really independent of one another. We prefer the use of the term *predictor* variables, but the term *independent* variables is so widely accepted that it is unlikely to be changed.

Table 13.7 Correlations among systolic blood pressure, weight, age, and height for 50 adults in Table 13.6.

	Systolic Blood Pressure	Weight	Age
Weight	0.465		
Age	0.393	−0.004	
Height	0.214	0.636	−0.327

In this multiple regression situation, we have three variables that are candidates for inclusion in the multiple linear regression equation to help account for the variation in SBP. As just mentioned, we wish to obtain a parsimonious set of independent variables that account for much of the variation in SBP. We shall use a stepwise regression procedure and an all possible regressions procedure to demonstrate two approaches to selecting the independent variables to be included in the final regression model.

There are many varieties of stepwise regression, and we shall consider forward stepwise regression. In forward stepwise regression, independent variables are added to the equation in steps, one per each step. The first variable to be added to the equation is the independent variable with the highest correlation with the dependent variable, provided that the correlation is high enough. The analyst provides the level that is used to determine whether or not the correlation is high enough. Instead of actually using the value of the correlation coefficient, the criterion for inclusion into the model is expressed in terms of the significance levels of the F ratio for the test that the regression coefficient is zero.

After the first variable is entered, the next variable to enter the model is the one that has the highest correlation with the residuals from the earlier model. This variable must also satisfy the significance level of the F ratio requirement for inclusion. This process continues in this stepwise fashion, and an independent variable may be added or deleted at each step. An independent variable that had been added previously may be deleted from the model if, after the inclusion of other variables, it no longer meets the required F ratio.

Table 13.8 shows the results of applying the forward stepwise regression procedure to our example. In the stepwise output, we see that the weight variable is the independent variable that entered the model first. It is highly significant with a t-value of 3.64, and the R^2 for the model is 21.61 percent. In the second step the age variable is added to the model. The default significance level of the F ratio for adding or deleting a variable is 0.15. The age variable is also highly significant with a t-value of 3.42 and as a result the R^2 value increased to 37.23 percent. Thus, this is the model selected by the forward stepwise process.

Table 13.8 Forward stepwise regression: Systolic blood pressure regressed on weight, age, and height.

Predictor	Step 1	Step 2
Constant	92.50	77.18
Weight	0.177	0.177
(t-value)	(3.64)	(4.04)
(p-value)	(0.001)	(<0.001)
Age		0.41
(t-value)		(3.42)
(p-value)		(0.001)
$S_{Y\|X}$	15.1	13.7
R^2	21.61	37.23
Adjusted R^2	19.98	34.55
Cp	11.8	2.3

In Table 13.8 there are four different statistics shown: R^2, adjusted R^2, C_p, and $s_{Y|X}$. Adjusted R^2 is similar to R^2, but it takes the number of variables in the equation into account. If a variable is added to the equation, but its associated F ratio is less than one, the adjusted R^2 will decrease. In this sense, the adjusted R^2 is a better measure than R^2. One minor problem with adjusted R^2 is that it can be slightly less than zero. The formula for calculating the adjusted R^2 is

$$Adjusted\ R_p^2 = 1 - (1 - R_p^2)\left(\frac{n}{n-p}\right)$$

where R_p^2 is the coefficient of determination for a model with p coefficients.

The statistic C_p was suggested by Mallows (1973) as a possible alternative criterion for selecting variables. It is defined as

$$C_p = \frac{SSE_p}{s^2} - (n - 2p)$$

where s^2 is the mean square error from the regression including all the independent variables under consideration and SSE_p is the residual sum of squares for a model that includes a given subset of $p - 1$ independent variables. It is generally recommended that we choose the model where C_p first approaches p.

The all possible regression procedure in effect considers all possible regressions with one independent variable, with two independent variables, with three independent variables, and so on, and it provides a summary report of the results for the "best" models. "Best" here is defined in statistical terms, but the actual determination of what is best must use substantive knowledge as well as statistical measures. Table 13.9 shows the results of applying the all possible regression procedure to our example.

Table 13.9 All possible (best subsets) regression: Systolic blood pressure regressed on weight, age, and height.

Number of Variables Entered	R^2	Adjusted R^2	Cp	$S_{Y\|X}$	Variables Entered		
					Weight	Age	Height
1	21.6	20.0	11.8	15.110	X		
1	15.5	13.7	16.4	15.692		X	
2	37.2	34.6	2.3	13.665	X	X	
2	28.6	25.6	8.7	14.573	X		X
3	37.7	33.6	4.0	13.764	X	X	X

From the all possible regressions output, we see that the model including weight was the best model with one independent variable. The second best model, with only one independent variable, used the age variable. The best two-independent-variable model used weight and age. The second best model, with two independent variables, used weight and height. The only three-independent-variable model has the highest R^2 value, but its adjusted R^2 is less than that for the best two independent variable model. Thus, on statistical grounds, we should select the model with weight and age as independent variables. It has the highest adjusted R^2 and the lowest value of $s_{Y|X}$. It also has C_p value closest to 2.

Again, these automatic selection procedures should be used with caution. We cannot treat the selected subset as containing the only variables that have an effect on the dependent variable. The excluded variables may still be important when different variables are in the model. Often it is necessary to force certain variables to be included in the model based on substantive considerations.

We also must realize that, since we are performing numerous tests, the p-values now only reflect the relative importance of the variables instead of the actual significance level associated with a variable.

13.4.3 Parameter Estimates, ANOVA, and Diagnostics

Let us now proceed to the multiple regression analysis with the full three-independent-variable model and compare it with the selected model that uses weight and age. Table 13.10 shows the regression with the three independent variables. The main features of interest are the tests of hypotheses and the parameter estimates. In the ANOVA table the F ratio of 9.27 is the value of the test statistic for the hypothesis that all the coefficients are simultaneously zero. Since its associated p-value is <0.001, we reject the hypothesis in favor of the alternative hypothesis that at least one of the coefficients is not zero. In general, however, this overall test is of little real interest because it is unlikely that none of the independent variables are related to the response variable. Of greater interest is the examination of the regression coefficients to see which independent variables are related to the response variable. In this model with the three independent variables, weight and age are statistically significant, but height is not, as shown by the t-values and the associated p-values. We should remove the statistically unimportant variables from the model unless there is a substantive reason to retain them. In fitting the model with the statistically unimportant variables eliminated, the estimated coefficients and standard errors will likely change in value due to the lack of independence of the predictor variables.

Table 13.10 also shows the sequential sum of squares. These sums of squares show the added contribution of the variables when they are entered in the order specified in the model. The contribution of height is very small after weight and age are already in the model. The table also shows *VIF* (variance inflation factor), and this term is discussed in the next section.

Table 13.10 Multiple regression analysis I: Systolic blood pressure regressed on weight, age, and height.

Predictor	Coef	SE Coef	T	*p*	VIF
Constant	53.96	41.54	1.30	0.200	
Weight	0.15435	0.05969	2.59	0.013	1.8
Age	0.4381	0.1319	3.32	0.002	1.2
Height	0.3845	0.6725	0.57	0.570	2.0

S = 13.76 R – Sq = 37.7% R – Sq(adj) = 33.6%

Analysis of Variance:

Source	DF	SS	MS	*F*	*p*
Regression	3	5,266.5	1,755.5	9.27	<0.001
Residual Error	46	8,714.4	189.4		
Total	49	13,980.9			

Source	DF	Seq SS
Weight	1	3,021.9
Age	1	2,182.6
Height	1	61.9

Table 13.11 shows the multiple regression analysis of the model selected by the variable selection procedures. In this model, the coefficients for the weight and age variables are highly significant with the t-values of 4.04 and 3.42, respectively, and an F ratio for the overall test of the model is 13.94. The estimated coefficient for the weight variable (0.177) increased slightly from its value in the three-independent-variable model (0.154), and its standard error has decreased to 0.044 from 0.060 in the previous model. Inclu-

Table 13.11 Multiple regression analysis II: Systolic blood pressure regressed on weight and age.

Predictor	Coef	SE Coef	T	*p*	VIF
Constant	77.185	8.668	8.91	<0.001	
Weight	0.17727	0.04391	4.04	<0.001	1.0
Age	0.4064	0.1189	3.42	0.001	1.0

S = 13.66 R – Sq = 37.2% R – Sq(adj) = 34.6%

ANOVA table:

Source	DF	SS	MS	*F*	*p*
Regression	2	5,204.5	2,602.3	13.94	<0.001
Residual Error	47	8,776.3	186.7		
Total	49	13,980.9			

Source	DF	Seq SS
Weight	1	3,021.9
Age	1	2,182.6

sion of an unnecessary term in the three-independent-variable model has caused the increase in the estimated standard errors and thus makes it harder to discern the significance of any of the independent variables.

The R^2 statistic indicates that the selected model is not able to account for the great majority of the variation in SBP. Much work needs to be done to discover these additional sources of variation before standards are created. It is likely that the effects of weight and age would be altered if we include other variables that have not been considered in the current models. A key message is that conclusions drawn about the importance of independent variables depend on the model that is being considered.

Having arrived at a final multiple regression model for the data set, it is important to go further and check the assumptions we made in selecting the important variables. Most useful at this stage is an examination of residuals from the fitted model. Among many regression diagnostics now available in computer packages, the following graphic plots are often used.

(1) A normal probability plot of the residuals: In creating the regression model, we assume that the errors (ε_i) are distributed normally. After the systematic variation associated with the independent variables in the model has been removed from the data, the residuals should therefore resemble a sample from a normal distribution. The normal probability plot of standardized residuals is shown in Figure 13.10. The points appear to lie along a line with the exception of the one large residual value, giving support to the normality assumption. If the normality assumption does not appear to be valid, then we may need to transform the response variable. However, transformations are not innocuous and must be done with care (Kleinbaum et al. 1998).

(2) A plot of the residuals against the fitted values: Figure 13.11 shows the standardized residuals plotted against the estimated values of the dependent variable with a useful reference line at zero. There is no strong pattern shown in the plot although the larger residuals in absolute value show a tendency to occur with estimated systolic blood pressure values over 130. If the trend were stronger, the equal variance assumption might be invalid and a transformation of the response variable might be required. If any clear patterns are shown in this plot, it raises concerns.

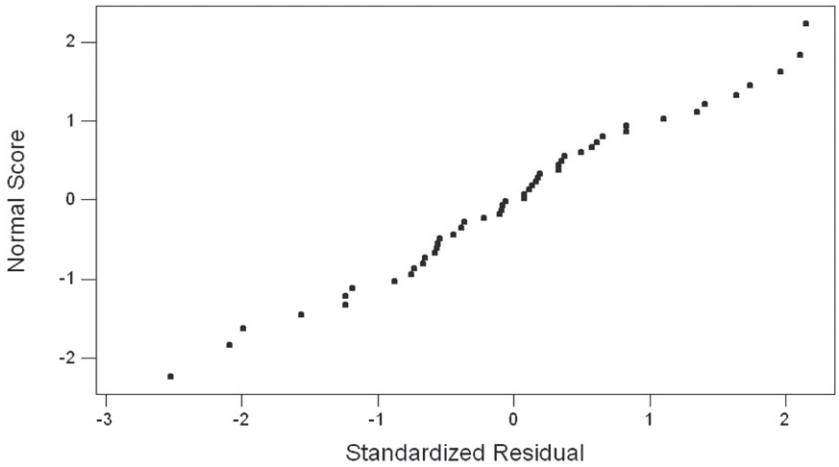

Figure 13.10 Normal probability plot of the standardized residuals.

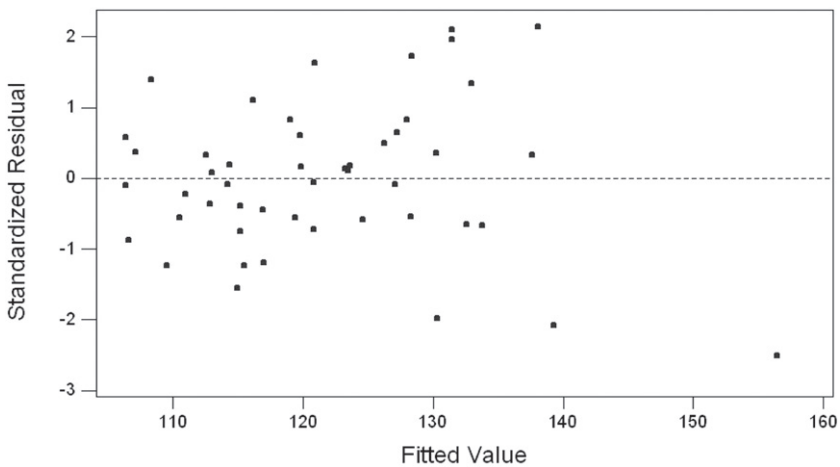

Figure 13.11 Plot of standardized residual versus the fitted value.

(3) A plot of residuals against each independent variable in the model: This plot helps in determining whether or not there may be a nonlinear relationship between the response variable and the independent variable used in the plot. Figure 13.12 shows the plot of residuals with the weight variable, and Figure 13.13 is a plot of the standardized residuals with the age variable. Neither plot shows the existence of any pattern. The presence of a curvilinear relationship, for example, would suggest that a higher-order term such as a quadratic term in the independent variable may be needed.

13.4.4 Multicollinearity Problems

In a multiple regression situation, it is not uncommon to have independent variables that are interrelated to a certain extent especially when survey data are used. *Multicollinearity* occurs when an explanatory variable is strongly related to a linear combination of the other independent variables. Multicollinearity does not violate the assumptions of the model, but it does increase the variance of the regression coefficients. This increase means that the parameter estimates are less reliable. Severe multicollinearity also makes determining the importance of a given explanatory variable difficult because the effects of explanatory variables are confounded.

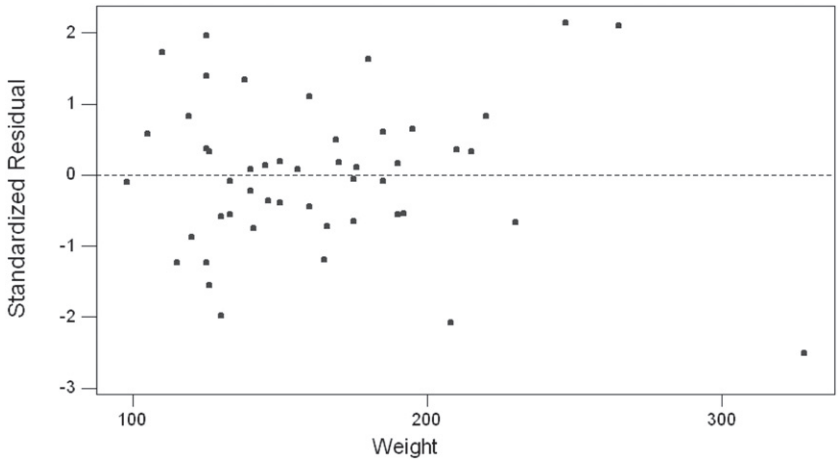

Figure 13.12 Plot of the standardized residual versus the weight variable.

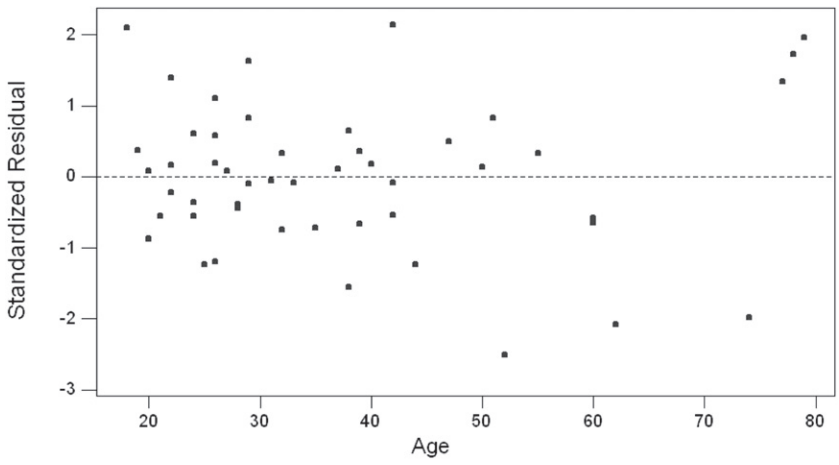

Figure 13.13 Plot of the standardized residual versus the age variable.

Recognizing multicollinearity among a set of explanatory variables is not necessarily easy. Obviously, we can simply examine the scatterplot matrix or the correlations between these variables, but we may miss more subtle forms of multicollinearity. An alternative and more useful approach is to examine what are known as the *variance inflation factors* (*VIF*) of the explanatory variables. The *VIF* for the *j*th independent variable is given by

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 from the regression of the *j*th explanatory variable on the remaining explanatory variables. The *VIF* of an explanatory variable indicates the strength of the linear relationship between the variable and the remaining explanatory variables. A rough rule of thumb is that the *VIF*s greater than 10 give some cause for concern.

Now let us review the multiple regression results shown in Tables 13.10 and 13.11. The *VIF*s shown in these tables are all less than 10, indicating that the multicollinearity does not pose a serious problem for those models. As a demonstration for a severe multicollinearity, we added to the model shown in Table 13.10 another independent variable

that is closed associated with weight and height. Table 13.12 shows the multiple regression analysis of SBP on weight, age, height, and the body mass index (BMI) defined as your weight in kilograms divided by the square of your height in meters. The *VIF*s for weight, height, and BMI are all greater than 10 in Table 13.12. More important, the variances of the regression coefficients for weight and height increased, and these variables are no longer statistically significant. The effect of weight on SBP shown in the earlier model cannot be demonstrated if we add BMI. A solution to a severe multicollinearity is to delete one of correlated variables. If we drop the BMI variable, we would eliminate the extreme multicollinearity.

Table 13.12 Multiple regression analysis III: Systolic blood pressure versus weight, age, height, body mass index.

Predictor	Coef	SE Coef	T	*p*	VIF
Constant	105.2	154.4	0.68	0.499	
Weight	0.3052	0.4413	0.69	0.493	97.6
Age	0.4364	0.1333	3.27	0.002	1.2
Height	−0.354	2.246	−0.16	0.875	22.3
BMI	−1.040	3.016	−0.34	0.732	60.9

S = 13.90 R − Sq = 37.8% R − Sq(adj) = 32.3%

Analysis of Variance:

Source	DF	SS	MS	*F*	*p*
Regression	4	5,289.4	1,322.4	6.85	<0.001
Residual Error	45	8,691.4	193.1		
Total	49	13,980.9			

Source	DF	Seq SS
Weight	1	3,021.9
Age	1	2,182.6
Height	1	61.9
BMI	1	23.0

13.4.5 Extending the Regression Model: Dummy Variables

So far we limited our analysis to continuous independent variables. As we discussed briefly in the previous chapter in conjunction with unbalanced ANOVA models, the independent variables can be categorical as well as continuous. It is easy to incorporate categorical explanatory variables into a multiple regression equation, provided we code the categorical variables with care. Let us consider the smoking status variable shown in Table 13.6. It has three levels: current smoker, never smoked, and previous smoker. Let us consider the never smoked category the baseline level and measure the effects of being a current smoker or a previous smoker from the never smoked level. We will then create two indicator variables to represent smoking status. The first indicator variable will have the value of 1 if the person is a current smoker and a value of 0 otherwise. The second indicator will have the value of 1 if the person is a former smoker and 0 otherwise. If the person has never smoked, both the indicator variables are 0.

	Indicator Variables		
Category	x_1	x_2	
Never Smoked	0	0	(reference)
Current Smoker	1	0	
Previous Smoker	0	1	

Table 13.13 Multiple regression analysis IV: Systolic blood pressure versus weight, age, sex (dummy variable).

Predictor	Coef	SE Coef	T	*p*	VIF
Constant	81.218	8.689	9.35	<0.001	
Weight	0.11749	0.05287	2.22	0.031	1.5
Age	0.4295	0.1162	3.69	0.001	1.0
Sex	8.990	4.685	1.92	0.061	1.5

S = 13.29 R – Sq = 41.9% R – Sq(adj) = 38.1%

Analysis of Variance:

Source	DF	SS	MS	*F*	*p*
Regression	3	5,854.8	1,951.6	11.05	<0.001
Residual Error	46	8,126.1	176.7		
Total	49	13,980.9			

Source	DF	Seq SS
Weight	1	3,021.9
Age	1	2,182.6
Sex	1	650.3

The number of indicator variables we need to represent a categorical variable is one less than the number of categories, corresponding to the degrees of freedom for the variable.

To demonstrate the use of an indicator variable into a regression analysis, we added the gender variable (female = 0; male = 1) to the multiple regression model shown in Table 13.11. The regression analysis of systolic blood pressure on weight, age, and gender is shown in Table 13.13. The gender variable accounted for some variation in SBP, although it did not quite reach statistical significance at the 0.05 level. The estimated regression equation is

$$\text{SBP} = 81.218 + 0.117\ (\text{weight}) + 0.430\ (\text{age}) + 8.990\ (\text{sex}).$$

The predicted SBP for females with weight of 100 lbs and age of 50 is

$$81.218 + 0.117(100) + 0.430(50) + 8.990(0) = 114.418.$$

The predicted SBP for males with the same weight and age is

$$81.218 + 0.117(100) + 0.430(50) + 8.990(1) = 123.408.$$

The predicted value for males is 8.990 mmHg higher than the predicted value for females. In other words, the regression coefficient for sex represents the difference in the mean SBP between the indicated category (coded as 1, males in this case) and the reference category (coded as 0, females in this case), holding the other independent variables constant.

Multiple regression analysis is a very useful technique. It becomes even more useful through its ability to incorporate categorical predictor variables along with continuous predictor variables. If only categorical explanatory variables are used, we have the analysis of variance situation. All of these situations — linear regression, ANOVA, and multiple linear regression with a mixture of continuous and discrete predictor variables — fit under the rubric of the *General Linear Model* (GLM).

See **Program Note 13.3** on the website for conducting multiple regression analysis including the use of variable selection procedures and residual plots.

Conclusion

In this chapter, we showed how to examine the relation between a normally distributed dependent variable and a continuous independent variable via linear regression analysis. We also demonstrated how this method could be extended to include many independent variables. We further expanded the linear regression model to include discrete predictor variables. These discrete predictor variables are incorporated through binary coding. Often we wish to use the linear regression or ANOVA idea, but the dependent variable is a binary variable — for example, the occurrence of a disease. In this case, the logistic regression method, discussed in the next chapter, can be used.

EXERCISES

13.1 Restenosis — narrowing of the blood vessels — frequently occurs after coronary angioplasty, but accurate prediction of which individuals will have this problem is problematic. In a study by Simons et al. (1993), the authors hypothesized that restenosis is more likely to occur if activated smooth-muscle cells in coronary lesions at the time of surgery are present. They used the number of reactive nuclei in the coronary lesions as an indicator of the presence of the activated smooth-muscle cells. The number of reactive nuclei in the lesions and the degree of stenosis at follow-up for 16 patients who underwent a second angiography are shown here.

Patient	Degree of Stenosis (%) at Follow-up	Number of Reactive Nuclei at Initial Surgery
1	28	5
2	15	3
3	22	2
4	93	10
5	60	12
6	90	25
7	42	8
8	53	3
9	72	15
10	0	13
11	79	17
12	28	0
13	82	13
14	28	14
15	100	17
16	21	1

Are you suspicious of any of these data points? If so, why? Does there appear to be a linear relation between the degree of stenosis and the number of reactive nuclei? If there is, describe the relation. Are there any points that have a large influence on the estimated regression line? If there are, eliminate the point with the greatest leverage and refit the equation. Is there much difference between the two regression equations? Are there any points that have a large standardized residual? Explain why the residuals are large for these points. Do you think that Simons et al. have a promising lead for predicting which patients will undergo restenosis?

13.2 Use the following data (NCHS 2005) to determine whether or not there is a linear relation between the U.S. national health expenditures as a percent of gross domestic product (GDP) and time.

Year	National Health Expenditures as Percentage of GDP	Year	National Health Expenditures as Percentage of GDP
1960	5.1	1999	13.2
1970	7.0	2000	13.3
1980	8.8	2001	14.1
1990	12.0	2002	14.9
1995	13.4		
1997	13.1		
1998	13.2		

What is your predicted value for national health expenditures as a percent of GDP for 2010? What is the 95 percent confidence interval for your estimate? What data have you used as the basis of your predictions? What assumptions have you made?

13.3 The estimated age-adjusted percent of persons 18 years of age and over who smoke cigarettes are shown below for females and males for selected years (NCHS 2005).

	Estimated Age-Adjusted Percent Smoking Cigarettes	
Year	Female	Male
1965	33.7	51.2
1974	32.2	42.8
1979	30.1	37.0
1985	27.9	32.2
1990	22.9	28.0
1995	22.7	26.5
1998	22.1	25.9
1999	21.6	25.2
2000	21.1	25.2
2001	20.7	24.6
2002	20.0	24.6
2003	19.4	23.7

Describe the linear relation between the estimated age-adjusted percent smoking and time for females and males separately. How much of the variation in the percents is accounted for by time for females and for males? Do females and males appear to have the same rate of decrease in the estimated age-adjusted percent smoking? Provide an estimate when the age-adjusted percent of males who smoke will equal the corresponding percent for females. What assumption(s) have you made in coming up with the estimate of this time point? Do you think this assumption is reasonable? Explain your answer.

13.4 Use the data in Table 13.1 to construct height-based standards for systolic blood pressure for girls. In constructing these standards, you should be concerned about values that may be too low as well as too high.

13.5 Anderson et al. (Anderson 1990) provide serum cholesterol and body mass index (BMI) values for subjects who participated in a study to examine the effects of oat-bran cereal on serum cholesterol. The values of serum cholesterol and BMI for the 12 subjects included in the analysis are shown next.

Subject	Serum cholesterol (mmol/L)	BMI
1	7.29	29.0
2	8.04	26.3
3	8.43	21.6
4	7.96	21.8
5	5.43	27.2
6	5.77	24.8
7	6.96	25.2
8	6.23	24.5
9	6.65	25.1
10	6.26	23.5
11	8.20	27.9
12	6.21	24.8

Plot serum cholesterol versus BMI. Calculate the correlation coefficient between serum cholesterol and BMI. Regress serum cholesterol on BMI. Does there appear to be any linear relation between these two variables? Form a new variable that is BMI minus its mean. Square this new variable. Include this new independent variable in the regression equation along with the BMI variable. Does there appear to be any linear relation between these two independent variables and serum cholesterol? Why do you think that we suggested that this new variable be added to the regression equation?

13.6 The following data are a sample of observations from the NHANES II. We wish to determine whether or not diastolic blood pressure (DBP) of adults can be predicted based on knowledge of the person's body mass index (BMI — weight in kilograms divided by the square of height in meters), age, sex (females coded as 0 and males coded as 1), smoking status (not currently a smoker is coded as 0 and currently a smoker is coded as 1), race (0 represents nonblack and 1 represents black), years of education, poverty status (household income expressed as a multiple of the poverty level for households of the same size), and vitamin status (0 indicates not taking supplements and 1 indicates taking supplements).

Select an appropriate multiple regression model that shows the relation between DBP and the set or a subset of the independent variables shown here. Note that the independent variables include both continuous and discrete variables. Provide an interpretation of the estimated regression coefficients for each discrete independent variable used in the model. From these independent variables, are we able to do a good job of predicting DBP? What other independent variables, if any, should be included to improve the prediction of DBP?

Vitamin Status	BMI	Sex	Race	Education	Age	Poverty Index	DBP	Smoking Status
1	18.46	0	0	13	24	1.93	50	0
0	32.98	1	0	14	24	3.97	98	0
1	29.48	1	0	12	39	1.71	80	1
1	19.20	0	0	12	29	1.62	62	1
0	24.76	0	0	12	45	5.49	90	0
1	20.60	0	0	14	24	4.78	70	0
0	24.80	1	0	8	65	3.63	80	0
1	24.24	0	0	12	25	4.55	56	1
0	29.95	1	0	16	24	2.77	90	0

(continued)

Vitamin Status	BMI	Sex	Race	Education	Age	Poverty Index	DBP	Smoking Status
0	21.80	1	0	17	29	2.15	78	0
0	23.19	1	0	13	29	1.09	56	0
0	28.34	0	0	12	18	1.71	78	0
0	22.00	1	0	12	28	5.49	70	1
0	24.60	1	0	8	65	3.35	70	1
1	21.83	0	0	16	26	0.77	74	0
0	30.50	0	0	3	73	1.10	70	0
1	19.63	0	0	13	33	5.48	62	1
0	27.92	0	0	12	65	3.83	78	0
1	26.77	1	0	12	59	3.57	90	0
1	21.02	1	0	15	21	1.25	64	0
1	19.40	0	0	16	26	3.25	70	0
0	31.12	0	0	12	58	1.91	100	0
0	20.68	0	0	7	57	4.63	74	0
0	22.48	0	0	12	28	1.75	75	0
0	24.89	0	0	14	23	3.25	74	0
1	21.08	0	0	12	56	5.04	68	0
1	23.67	1	0	14	23	4.47	86	1
1	28.19	1	0	12	24	3.38	82	1
0	22.09	0	1	7	58	1.73	80	0
0	23.46	1	0	14	66	5.12	70	0
1	21.11	1	0	13	18	0.64	70	1
1	21.35	0	1	12	20	0.26	60	1
0	20.36	0	1	14	23	2.85	78	0
1	25.00	0	1	4	36	0.72	80	0
1	20.47	0	0	17	37	3.97	88	1
0	24.73	0	1	8	44	1.36	82	0
0	27.87	0	0	12	50	3.31	70	1
0	28.22	1	0	15	50	3.41	112	0
0	26.05	1	0	13	33	5.85	80	0
0	24.51	0	0	12	42	3.17	92	0
1	28.09	0	1	16	46	2.39	92	0
1	18.85	0	1	11	36	1.62	56	1
0	25.99	0	1	12	74	1.40	80	0
1	23.47	1	0	16	35	1.97	96	1
0	26.57	0	0	12	55	6.11	86	0
0	25.09	1	0	12	33	2.15	104	1
0	30.78	0	0	12	38	1.37	74	0
0	28.89	1	0	14	49	1.82	90	1
1	23.82	1	0	17	35	2.85	70	0
0	28.29	1	0	12	62	6.89	60	0

13.7 Find an article from a health-related journal that used a multiple regression analysis and review it thoroughly. Is the multiple regression model an appropriate choice of analysis? Would you conduct the analysis or interpret the result differently? Did your article report all the necessary analytical results that would convince you to accept the author's conclusions?

13.8 The following data set consists of infant mortality rates (IMR) for 50 states in 1997–1998, along with the following eight potential explanatory variables (NCHS 2004).

1. Low birthweight: Percent of live births with weight less than 2500 grams, 1997–1999
2. Vaccination: Percent of children 19–35 months of age vaccinated against selected diseases, 1998
3. Medicaid expenditures as percent of total personal health care expenditures, 1998

4. Prenatal care: Percent of live births with prenatal care started in the first trimester, 1998
5. Uninsured: Percent of people under 65 years of age without health insurance, 1998
6. Hospital care: Per capita expenditure in dollars for hospital care, 1998
7. Personal care: Per capita expenditure for personal health care, 1998
8. Personal care: Per capita expenditure for personal health care, 1996

States are grouped in regions and the region can be another potential explanatory variable(s). Build an appropriate multiple regression model that would show the relationship between infant mortality rate and a subset of the potential explanatory variables (five or fewer considering the total number of observations). Apply different criteria for selecting a subset and see whether different criteria give different results. Check whether various assumptions are met in your final model. Interpret your analytical results, taking into account that these variables are measurements made at the state level. Do you think all relevant explanatory variables are represented in your model?

State	IMR	Low Birth Weight	Vaccination	Medicaid	Prenatal Care	Uninsured	Hospital Care	Personal Care	
								1998	1996
East (New England & Mideast)									
CT	6.8	7.56	90	17.5	88.8	14.3	1,478	4,656	4,250
ME	5.3	5.93	86	21.1	89.0	14.6	1,501	4,025	3,512
MA	5.1	6.99	87	19.3	89.3	11.6	1,807	4,810	4,347
NH	4.5	5.91	82	15.6	90.0	12.5	1,234	3,840	3,441
RI	6.5	7.43	86	21.6	90.1	7.6	1,626	4,497	3,978
VT	6.7	6.15	86	18.0	87.8	11.0	1,328	3,654	3,273
DE	6.5	8.01	82	12.5	81.4	17.1	1,581	5,258	3,847
MD	6.6	7.83	85	12.7	80.9	18.9	1,486	3,848	3,573
NJ	7.5	7.69	83	14.0	84.6	18.0	1,481	4,197	4,009
NY	8.5	7.96	78	31.5	82.6	19.7	1,769	4,706	4,346
PA	8.1	7.84	78	16.3	80.2	12.1	1,599	4,168	3,791
Midwest (Great Lakes & Plains)									
IL	8.2	7.84	78	14.8	84.1	16.6	1,558	3,801	3,535
ID	7.8	7.78	78	12.0	85.7	16.1	1,413	3,566	3,196
MI	7.0	6.53	78	14.9	84.3	14.9	1,489	3,676	3,457
OH	6.5	6.31	82	15.6	87.5	11.7	1,437	3,747	3,542
WI	7.6	7.01	82	13.4	85.7	13.2	1,377	3,845	3,476
IA	5.9	5.92	82	15.4	84.4	10.9	1,520	3,765	3,368
KS	7.6	7.75	85	10.8	86.4	12.2	1,428	3,707	3,412
MN	7.8	6.75	76	15.4	84.1	10.3	1,254	3,986	3,614
MO	6.8	6.31	79	14.4	85.6	12.1	1,566	3,754	3,390
NE	7.4	5.75	74	14.4	82.7	10.2	1,507	3,627	3,287
ND	8.4	8.57	79	13.8	83.2	16.5	1,741	3,881	3,540
SD	7.3	8.09	79	13.4	83.8	16.3	1,534	3,650	3,253
South (Southeast & Southwest)									
AL	8.7	8.68	80	13.0	86.5	19.5	1,432	3,630	3,422
AR	8.6	8.82	77	15.5	87.8	21.7	1,430	3,540	3,177
FL	9.2	8.84	83	10.4	84.5	21.1	1,371	4,046	3,774
GA	9.2	9.52	88	12.2	80.9	19.4	1,329	3,505	3,291
KY	7.7	7.80	80	16.9	85.2	16.0	1,479	3,711	3,300
LA	8.3	8.12	82	19.1	83.6	21.3	1,601	3,742	4,396
MS	10.0	9.28	82	15.8	82.6	22.9	1,551	3,474	3,145
NC	7.3	8.06	82	16.9	86.3	17.0	1,373	3,535	3,232
SC	10.5	10.18	84	16.6	80.7	17.4	1,480	3,529	3,131
TN	8.4	9.01	82	17.4	84.0	14.3	1,375	3,808	3,569
VA	9.0	8.62	73	9.9	77.5	15.8	1,286	3,284	3,009

(continued)

State	IMR	Low Birth Weight	Vaccination	Medicaid	Prenatal Care	Uninsured	Hospital Care	Personal Care 1998	Personal Care 1996
WV	9.2	10.09	78	17.3	82.1	20.8	1,693	4,044	3,649
AZ	8.1	7.28	75	12.0	79.2	26.9	1,085	3,100	2,862
NM	6.3	7.35	74	17.7	79.0	24.0	1,389	3,209	2,942
OK	7.4	6.86	76	11.8	75.5	21.2	1,307	3,397	3,188
TX	6.8	8.60	76	12.5	82.2	26.9	1,274	3,397	3,117
West (Rocky Mountains & Far West)									
CO	7.0	6.15	76	11.4	79.3	16.4	1,147	3,331	3,071
ID	7.1	6.71	82	12.1	82.9	19.7	1,163	3,035	2,765
MT	6.6	7.59	76	13.8	75.3	21.9	1,440	3,314	2,917
UT	6.6	7.68	71	11.8	68.2	15.1	1,016	2,731	2,506
WY	5.9	6.72	76	12.3	82.1	18.8	1,436	3,881	3,046
AK	6.6	8.75	80	16.9	82.3	17.9	1,496	3,442	3,227
CA	6.9	5.90	81	12.7	80.4	24.4	1,145	3,429	3,200
HI	5.9	6.17	76	14.2	82.6	11.3	1,391	3,770	3,656
NV	6.5	7.44	79	9.1	84.8	23.7	1,033	3,147	2,949
OR	5.5	5.41	76	15.3	80.7	16.0	1,112	3,334	3,019
WA	5.7	5.72	81	16.2	83.1	13.4	1,116	3,382	3,142

REFERENCES

Anderson, J. W., D. B. Spencer, C. C. Hamilton, S. F. Smith, J. Tietyen, C. A. Bryant, and P. Oeltgen. "Oat-Bran Cereal Lowers Serum Total and LDL Cholesterol in Hypercholesterolemic Men." *American Journal of Clinical Nutrition* 52:495–499, 1990.

Draper, N. R., and H. Smith. *Applied Regression Analysis*. New York: John Wiley & Sons, 1981.

Forthofer, R. N. "Blood Pressure Standards in Children." Paper presented at the American Statistical Association Meeting, August 1991.

Gillum, R., R. Prineas, and H. Horibe. "Maturation vs Age: Assessing Blood Pressure by Height." *Journal of the National Medical Association* 74:43–46, 1982.

Kleinbaum, D. G., L. L. Kupper, K. E. Muller, and A. Nizam. *Applied Regression Analysis and Other Multivariable Methods*, 3rd ed. Pacific Grove, CA: Brooks/Cole, 1998.

Korn, E. L., and R. Simon. "Explained Residual Variation, Explained Risk, and Goodness of Fit." *The American Statistician* 45:201–206, 1991.

Lin, L. I., and E. F. Vonesh. "An Empirical Nonlinear Data-Fitting Approach for Transforming Data to Normality." *The American Statistician* 43:237–243, 1989.

Mallows, C. L. "Some Comments on *C*p." *Technometrics* 15:661–675, 1973.

Miller, D. M. "Reducing Transformation Bias in Curve Fitting." *The American Statistician* 38:124–126, 1984.

National Center for Health Statistics. *Health, United States, 2004 with Chartbook on Trends in the Health of Americans*. Hyattsville, MD: Public Health Service. DHHS Pub. No. 2004-1232, September 2004, Tables 7, 14, 23, 72, 119, 143, 148, and 153.

National Center for Health Statistics. *Health, United States, 2005 with Chartbook on Trends in the Health of Americans*. Hyattsville, MD: Public Health Service. DHHS Pub. No. 2005-1232, November 2005, Tables 63 and 118.

The NHLBI Task Force on Blood Pressure Control in Children. "The Report of the Second Task Force on Blood Pressure Control in Children, 1987." *Pediatrics* 79:1–25, 1987.

Ranney, G. B., and C. C. Thigpen. "The Sample Coefficient of Determination in Simple Linear Regression." *The American Statistician* 35:152–153, 1981.

Simons, M., G. Leclerc, R. D. Safian, J. M. Isner, L. Weir, and D. S. Baim. "Relation between Activated Smooth-Muscle Cells in Coronary-Artery Lesions and Restenosis after Atherectomy." *The New England Journal of Medicine* 328:608–613, 1993.

Voors, A., L. Webber, R. Frerichs, and G. S. Berrenson. "Body Height and Body Mass as Determinants of Basal Blood Pressure in Children — the Bogalusa Heart Study." *American Journal of Epidemiology* 106:101–108, 1977.