Chapter

Building the Regression Model III: Remedial Measures

When the diagnostics indicate that a regression model is not appropriate or that one or several cases are very influential, remedial measures may need to be taken. In earlier chapters, we discussed some remedial measures, such as transformations to linearize the regression relation, to make the error distributions more nearly normal, or to make the variances of the error terms more nearly equal. In this chapter, we take up some additional remedial measures to deal with unequal error variances, a high degree of multicollinearity, and influential observations. We next consider two methods for nonparametric regression in detail, lowess and regression trees. Since these remedial measures and alternative approaches often involve relatively complex estimation procedures, we consider next a general approach, called bootstrapping, for evaluating the precision of these complex estimators. We conclude the chapter by presenting a case that illustrates some of the issues that arise in model building.

11.1 Unequal Error Variances Remedial Measures—Weighted Least Squares

We explained in Chapters 3 and 6 how transformations of Y may be helpful in reducing or eliminating unequal variances of the error terms. A difficulty with transformations of Y is that they may create an inappropriate regression relationship. When an appropriate regression relationship has been found but the variances of the error terms are unequal, an alternative to transformations is weighted least squares, a procedure based on a generalization of multiple regression model (6.7). We shall now denote the variance of the error term ε_i by σ_i^2 to recognize that different error terms may have different variances. The generalized multiple regression model can then be expressed as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad (11.1)$$

where:

$\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters
$X_{i1}, \ldots, X_{i,p-1}$ are known constants
ε_i are independent $N(0, \sigma_i^2)$
$i = 1, \ldots, n$

The variance-covariance matrix of the error terms for the generalized multiple regression model (11.1) is more complex than before:

$$\underset{n\times n}{\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \tag{11.2}$$

The estimation of the regression coefficients in generalized model (11.1) could be done by using the estimators in (6.25) for regression model (6.7) with equal error variances. These estimators are still unbiased and consistent for generalized regression model (11.1), but they no longer have minimum variance. To obtain unbiased estimators with minimum variance, we must take into account that the different Y observations for the n cases no longer have the same reliability. Observations with small variances provide more reliable information about the regression function than those with large variances. We shall first consider the estimation of the regression coefficients when the error variances σ_i^2 are known. This case is usually unrealistic, but it provides guidance as to how to proceed when the error variances are not known.

Error Variances Known

When the error variances σ_i^2 are known, we can use the method of maximum likelihood to obtain estimators of the regression coefficients in generalized regression model (11.1). The likelihood function in (6.26) for the case of equal error variances σ^2 is modified by replacing the σ^2 terms with the respective variances σ_i^2 and expressing the likelihood function in the first form of (1.26):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma_i^2)^{1/2}} \exp\left[-\frac{1}{2\sigma_i^2}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1}X_{i,p-1})^2\right] \tag{11.3}$$

where $\boldsymbol{\beta}$ as usual denotes the vector of the regression coefficients. We define the reciprocal of the variance σ_i^2 as the *weight* w_i:

$$w_i = \frac{1}{\sigma_i^2} \tag{11.4}$$

We can then express the likelihood function (11.3) as follows, after making some simplifications:

$$L(\boldsymbol{\beta}) = \left[\prod_{i=1}^{n}\left(\frac{w_i}{2\pi}\right)^{1/2}\right] \exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1}X_{i,p-1})^2\right] \tag{11.5}$$

We find the maximum likelihood estimators of the regression coefficients by *maximizing* $L(\boldsymbol{\beta})$ in (11.5) with respect to $\beta_0, \beta_1, \ldots, \beta_{p-1}$. Since the error variances σ_i^2 and hence the weights w_i are assumed to be known, maximizing $L(\boldsymbol{\beta})$ with respect to the regression coefficients is equivalent to *minimizing* the exponential term:

$$Q_w = \sum_{i=1}^{n} w_i(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2 \tag{11.6}$$

This term to be minimized for obtaining the maximum likelihood estimators is also the *weighted least squares criterion,* denoted by Q_w. Thus, the methods of maximum likelihood and weighted least squares lead to the same estimators for the generalized multiple regression model (11.1), as is also the case for the ordinary multiple regression model (6.7).

Note how the weighted least squares criterion (11.6) generalizes the ordinary least squares criterion in (6.22) by replacing equal weights of 1 by w_i. Since the weight w_i is inversely related to the variance σ_i^2, it reflects the amount of information contained in the observation Y_i. Thus, an observation Y_i that has a large variance receives less weight than another observation that has a smaller variance. Intuitively, this is reasonable. The more precise is Y_i (i.e., the smaller is σ_i^2), the more information Y_i provides about $E\{Y_i\}$ and therefore the more weight it should receive in fitting the regression function.

It is easiest to express the maximum likelihood and weighted least squares estimators of the regression coefficients for model (11.1) in matrix terms. Let the matrix $\mathbf{W}$ be a diagonal matrix containing the weights w_i:

$$\underset{n\times n}{\mathbf{W}} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} \tag{11.7}$$

The normal equations can then be expressed as follows:

$$(\mathbf{X}'\mathbf{W}\mathbf{X})\mathbf{b}_w = \mathbf{X}'\mathbf{W}\mathbf{Y} \tag{11.8}$$

and the weighted least squares and maximum likelihood estimators of the regression coefficients are:

$$\underset{p\times 1}{\mathbf{b}_w} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} \tag{11.9}$$

where $\mathbf{b}_w$ is the vector of the estimated regression coefficients obtained by weighted least squares. The variance-covariance matrix of the weighted least squares estimated regression coefficients is:

$$\underset{p\times p}{\sigma^2\{\mathbf{b}_w\}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \tag{11.10}$$

Note that this variance-covariance matrix is known since the variances σ_i^2 are assumed to be known.

The weighted least squares and maximum likelihood estimators of the regression coefficients in (11.9) are unbiased, consistent, and have minimum variance among unbiased linear estimators. Thus, when the weights are known, $\mathbf{b}_w$ generally exhibits less variability than the ordinary least squares estimator $\mathbf{b}$.

Many computer regression packages will provide the weighted least squares estimated regression coefficients. The user simply needs to provide the weights w_i.

Error Variances Known up to Proportionality Constant

We now relax the requirement that the variances σ_i^2 are known by considering the case where only the relative magnitudes of the variances are known. For instance, if we know that σ_2^2 is twice as large as σ_1^2, we might use the weights $w_1 = 1, w_2 = 1/2$. In that case, the relative weights w_i are a constant multiple of the unknown true weights $1/\sigma_i^2$:

$$w_i = k\left(\frac{1}{\sigma_i^2}\right) \tag{11.11}$$

where k is the proportionality constant. It can be shown that the weighted least squares and maximum likelihood estimators are unaffected by the unknown proportionality constant k and are still given by (11.9). The reason is that the proportionality constant k appears on both sides of the normal equations (11.8) and cancels out. The variance-covariance matrix of the weighted least squares regression coefficients is now as follows:

$$\underset{p\times p}{\boldsymbol{\sigma}^2\{\mathbf{b}_w\}} = k(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \tag{11.12}$$

This matrix is unknown because the proportionality constant k is not known. It can be estimated, however. The estimated variance-covariance matrix of the regression coefficients $\mathbf{b}_w$ is:

$$\underset{p\times p}{\mathbf{s}^2\{\mathbf{b}_w\}} = MSE_w(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \tag{11.13}$$

where MSE_w is based on the weighted squared residuals:

$$MSE_w = \frac{\sum w_i(Y_i - \hat{Y}_i)^2}{n-p} = \frac{\sum w_i e_i^2}{n-p} \tag{11.13a}$$

Thus, MSE_w here is an estimator of the proportionality constant k.

Error Variances Unknown

If the variances σ_i^2 were known, or even known up to a proportionality constant, the use of weighted least squares with weights w_i would be straightforward. Unfortunately, one rarely has knowledge of the variances σ_i^2. We are then forced to use estimates of the variances. These can be obtained in a variety of ways. We discuss two methods of obtaining estimates of the variances σ_i^2.

Estimation of Variance Function or Standard Deviation Function. The first method of obtaining estimates of the error term variances σ_i^2 is based on empirical findings that the magnitudes of σ_i^2 and σ_i often vary in a regular fashion with one or several predictor variables X_k or with the mean response $E\{Y_i\}$. Figure 3.4c, for example, shows a typical "megaphone" prototype residual plot where σ_i^2 increases as the predictor variable X becomes larger. Such a relationship between σ_i^2 and one or several predictor variables can be estimated because the squared residual e_i^2 obtained from an ordinary least squares regression fit is an estimate of σ_i^2, provided that the regression function is appropriate. We know from (A.15a) that

the variance of the error term ε_i, denoted by σ_i^2, can be expressed as follows:

$$\sigma_i^2 = E\{\varepsilon_i^2\} - (E\{\varepsilon_i\})^2 \tag{11.14}$$

Since $E\{\varepsilon_i\} = 0$ according to the regression model, we obtain:

$$\sigma_i^2 = E\{\varepsilon_i^2\} \tag{11.15}$$

Hence, the squared residual e_i^2 is an estimator of σ_i^2. Furthermore, the absolute residual $|e_i|$ is an estimator of the standard deviation σ_i, since $\sigma_i = |\sqrt{\sigma_i^2}|$.

We can therefore estimate the variance function describing the relation of σ_i^2 to relevant predictor variables by first fitting the regression model using unweighted least squares and then regressing the squared residuals e_i^2 against the appropriate predictor variables. Alternatively, we can estimate the standard deviation function describing the relation of σ_i to relevant predictor variables by regressing the absolute residuals $|e_i|$ obtained from fitting the regression model using unweighted least squares against the appropriate predictor variables. If there are any outliers in the data, it is generally advisable to estimate the standard deviation function rather than the variance function, because regressing absolute residuals is less affected by outliers than regressing squared residuals. Reference 11.1 provides a detailed discussion of the issues encountered in estimating variance and standard deviation functions.

We illustrate the use of some possible variance and standard deviation functions:

1. A residual plot against X_1 exhibits a megaphone shape. Regress the absolute residuals against X_1.
2. A residual plot against $\hat{Y}$ exhibits a megaphone shape. Regress the absolute residuals against $\hat{Y}$.
3. A plot of the squared residuals against X_3 exhibits an upward tendency. Regress the squared residuals against X_3.
4. A plot of the residuals against X_2 suggests that the variance increases rapidly with increases in X_2 up to a point and then increases more slowly. Regress the absolute residuals against X_2 and X_2^2.

After the variance function or the standard deviation function is estimated, the fitted values from this function are used to obtain the estimated weights:

$$w_i = \frac{1}{(\hat{s}_i)^2} \quad \text{where } \hat{s}_i \text{ is fitted value from standard deviation function} \tag{11.16a}$$

$$w_i = \frac{1}{\hat{v}_i} \quad \text{where } \hat{v}_i \text{ is fitted value from variance function} \tag{11.16b}$$

The estimated weights are then placed in the weight matrix $\mathbf{W}$ in (11.7) and the estimated regression coefficients are obtained by (11.9), as follows:

$$\mathbf{b}_w = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} \tag{11.17}$$

The weighted error mean square MSE_w may be viewed here as an estimator of the proportionality constant k in (11.11). If the modeling of the variance or standard deviation function is done well, the proportionality constant will be near 1 and MSE_w should then be near 1.

We summarize the estimation process:

1. Fit the regression model by unweighted least squares and analyze the residuals.
2. Estimate the variance function or the standard deviation function by regressing either the squared residuals or the absolute residuals on the appropriate predictor(s).
3. Use the fitted values from the estimated variance or standard deviation function to obtain the weights w_i.
4. Estimate the regression coefficients using these weights.

If the estimated coefficients differ substantially from the estimated regression coefficients obtained by ordinary least squares, it is usually advisable to iterate the weighted least squares process by using the residuals from the weighted least squares fit to reestimate the variance or standard deviation function and then obtain revised weights: Often one or two iterations are sufficient to stabilize the estimated regression coefficients. This iteration process is often called *iteratively reweighted least squares*.

Use of Replicates or Near Replicates. A second method of obtaining estimates of the error term variances σ_i^2 can be utilized in designed experiments where replicate observations are made at each combination of levels of the predictor variables. If the number of replications is large, the weights w_i may be obtained directly from the sample variances of the Y observations at each combination of levels of the X variables. Otherwise, the sample variances or sample standard deviations should first be regressed against appropriate predictor variables to estimate the variance or standard deviation function, from which the weights can then be obtained. Note that each case in a replicate group receives the same weight with this method.

In observational studies, replicate observations often are not present. Near replicates may then be used. For example, if the residual plot against X_1 shows a megaphone appearance, cases with similar X_1 values can be grouped together and the variance of the residuals in each group calculated. The reciprocals of these variances are then used as the weights w_i if the number of replications is large. Otherwise, a variance or standard deviation function may be estimated to obtain the weights. Again, all cases in a near-replicate group receive the same weight. If the estimated regression coefficients differ substantially from those obtained with ordinary least squares, the procedure may be iterated, as when an estimated variance or standard deviation function is used.

Inference Procedures when Weights Are Estimated. When the error variances σ_i^2 are unknown so that the weights w_i need to be estimated, which almost always is the case, the variance-covariance matrix of the estimated regression coefficients is usually estimated by means of (11.13), using the estimated weights, provided the sample size is not very small. Confidence intervals for regression coefficients are then obtained by means of (6.50), with the estimated standard deviation $s\{b_{w k}\}$ obtained from the matrix (11.13). Confidence intervals for mean responses are obtained by means of (6.59), using $\mathbf{s}^2\{\mathbf{b}_w\}$ from (11.13) in (6.58). These inference procedures are now only approximate, however, because the estimation of the variances σ_i^2 introduces another source of variability. The approximation is often quite good when the sample size is not too small. One means of determining whether the approximation is good is to use bootstrapping, a statistical procedure that will be explained in Section 11.5.

Use of Ordinary Least Squares with Unequal Error Variances. If one uses $\mathbf{b}$ (not $\mathbf{b}_w$) with unequal error variances, the ordinary least squares estimators of the regression coefficients are still unbiased and consistent, but they are no longer minimum variance estimators. Also, $\sigma^2\{\mathbf{b}\}$ is no longer given by $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The correct variance-covariance matrix is:

$$\sigma^2\{\mathbf{b}\} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\sigma^2\{\boldsymbol{\varepsilon}\}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

If error variances are unequal and unknown, an appropriate estimator of $\sigma^2\{\mathbf{b}\}$ can still be obtained using ordinary least squares. The *White estimator* (Ref. 11.2) is:

$$\mathbf{S}^2\{\mathbf{b}\} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{S}_0\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

where:

$$\underset{n\times n}{\mathbf{S}_0} = \begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{bmatrix}$$

and where $e_1, \ldots, e_n$ are the ordinary least squares estimators of the residuals. White's estimator is sometimes referred to as a robust covariance matrix, because it can be used to make appropriate inferences about the regression parameters based on ordinary least squares, without having to specify the form of the nonconstant error variance.

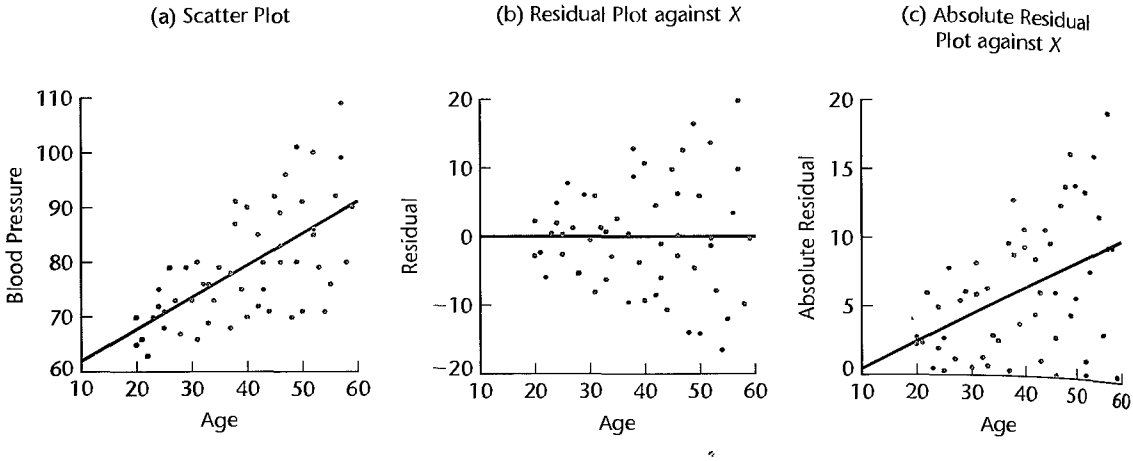
Example

A health researcher, interested in studying the relationship between diastolic blood pressure and age among healthy adult women 20 to 60 years old, collected data on 54 subjects. A portion of the data is presented in Table 11.1, columns 1 and 2. The scatter plot of the data in Figure 11.1a strongly suggests a linear relationship between diastolic blood pressure and age but also indicates that the error term variance increases with age. The researcher fitted a linear regression function by unweighted least squares to conduct some preliminary analyses of the residuals. The fitted regression function and the estimated standard deviations of b_0

TABLE 11.1 Weighted Least Squares—Blood Pressure Example.

Subject i	(1) Age X_i	(2) Diastolic Blood Pressure Y_i	(3) e_i	(4) $\lvert e_i \rvert$	(5) $\hat{s}_i$	(6) w_i
1	27	73	1.18	1.18	3.801	.06921
2	21	66	−2.34	2.34	2.612	.14656
3	22	63	−5.92	5.92	2.810	.12662
...	...	...	...	...	...	...
52	52	100	13.68	13.68	8.756	.01304
53	58	80	−9.80	9.80	9.944	.01011
54	57	109	19.78	19.78	9.746	.01053

FIGURE 11.1 Diagnostic Plots Detecting Unequal Error Variances—Blood Pressure Example.



and b_1 are:

$$\hat{Y} = 56.157 + .58003X \qquad (11.18)$$
$$(3.994) \quad (.09695)$$

The residuals are shown in Table 11.1, column 3, and the absolute residuals are presented in column 4. Figure 11.1a presents this estimated regression function. Figure 11.1b presents a plot of the residuals against X, which confirms the nonconstant error variance. A plot of the absolute residuals against X in Figure 11.1c suggests that a linear relation between the error standard deviation and X may be reasonable. The analyst therefore regressed the absolute residuals against X and obtained:

$$\hat{s} = -1.54946 + .198172X \qquad (11.19)$$

Here, $\hat{s}$ denotes the estimated expected standard deviation. The estimated standard deviation function in (11.19) is shown in Figure 11.1c.

To obtain the weights w_i, the analyst obtained the fitted values from the standard deviation function in (11.19). For example, for case 1, for which $X_1 = 27$, the fitted value is:

$$\hat{s}_1 = -1.54946 + .198172(27) = 3.801$$

The fitted values are shown in Table 11.1, column 5. The weights are then obtained by using (11.16a). For case 1, we obtain:

$$w_1 = \frac{1}{(\hat{s}_1)^2} = \frac{1}{(3.801)^2} = .0692$$

The weights w_i are shown in Table 11.1, column 6.

Using these weights in a regression program that has weighted least squares capability, the analyst obtained the following estimated regression function:

$$\hat{Y} = 55.566 + .59634X \qquad (11.20)$$

Note that the estimated regression coefficients are not much different from those in (11.18) obtained with unweighted least squares. Since the regression coefficients changed only a little, the analyst concluded that there was no need to reestimate the standard deviation function and the weights based on the residuals for the weighted regression in (11.20).

The analyst next obtained the estimated variance-covariance matrix of the estimated regression coefficients by means of (11.13) to find the approximate estimated standard deviation $s\{b_{w1}\} = .07924$. It is interesting to note that this standard deviation is somewhat smaller than the standard deviation of the estimate obtained by ordinary least squares in (11.18), .09695. The reduction of about 18 percent is the result of the recognition of unequal error variances when using weighted least squares.

To obtain an approximate 95 percent confidence interval for β_1, the analyst employed (6.50) and required $t(.975; 52) = 2.007$. The confidence limits then are $.59634 \pm 2.007(.07924)$ and the approximate 95 percent confidence interval is:

$$.437 \leq \beta_1 \leq .755$$

We shall consider the appropriateness of this inference approximation in Section 11.5.

Comments

1. The condition of the error variance not being constant over all cases is called *heteroscedasticity,* in contrast to the condition of equal error variances, called *homoscedasticity.*

2. Heteroscedasticity is inherent when the response in regression analysis follows a distribution in which the variance is functionally related to the mean. (Significant nonnormality in Y is encountered as well in most such cases.) Consider, in this connection, a regression analysis where X is the speed of a machine which puts a plastic coating on cable and Y is the number of blemishes in the coating per thousand feet of cable. If Y is Poisson distributed with a mean which increases as X increases, the distributions of Y cannot have constant variance at all levels of X since the variance of a Poisson variable equals the mean, which is increasing with X.

3. Estimation of the weights by means of an estimated variance or standard deviation function or by means of groups of replicates or near replicates can be very helpful when there are major differences in the variances of the error terms. When the differences are only small or modest, however, weighted least squares with these approximate methods will not be particularly helpful.

4. The weighted least squares output of some multiple regression software packages includes R^2, the coefficient of multiple determination. Users of these packages need to treat this measure with caution, because R^2 does not have a clear-cut meaning for weighted least squares.

5. The weighted least squares estimators of the regression coefficients in (11.9) for the case of known error variances σ_i^2 can be derived readily. The derivation also shows that weighted least squares may be viewed as ordinary least squares of transformed variables. The generalized multiple regression model in (11.1) may be expressed as follows in matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{11.21}$$

where:

$$\mathbf{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0}$$
$$\sigma^2\{\boldsymbol{\varepsilon}\} = \mathbf{W}^{-1}$$

Note that the variance-covariance matrix of the error terms in (11.2) is the inverse of the weight matrix defined in (11.7).

We now define a diagonal matrix containing the square roots of the weights w_i and denote it by $\mathbf{W}^{1/2}$:

$$\underset{n\times n}{\mathbf{W}^{1/2}} = \begin{bmatrix} \sqrt{w_1} & 0 & \cdots & 0 \\ 0 & \sqrt{w_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{w_n} \end{bmatrix} \tag{11.22}$$

Note that $\mathbf{W}^{1/2}$ is symmetric and that $\mathbf{W}^{1/2}\mathbf{W}^{1/2} = \mathbf{W}$. The latter relation also holds for the corresponding inverse matrices: $\mathbf{W}^{-1/2}\mathbf{W}^{-1/2} = \mathbf{W}^{-1}$.

We premultiply the terms on both sides of regression model (11.21) by $\mathbf{W}^{1/2}$ and obtain:

$$\mathbf{W}^{1/2}\mathbf{Y} = \mathbf{W}^{1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{W}^{1/2}\boldsymbol{\varepsilon} \tag{11.23}$$

which can be expressed as:

$$\mathbf{Y}_w = \mathbf{X}_w\boldsymbol{\beta} + \boldsymbol{\varepsilon}_w \tag{11.23a}$$

where:

$$\begin{aligned} \mathbf{Y}_w &= \mathbf{W}^{1/2}\mathbf{Y} \\ \mathbf{X}_w &= \mathbf{W}^{1/2}\mathbf{X} \\ \boldsymbol{\varepsilon}_w &= \mathbf{W}^{1/2}\boldsymbol{\varepsilon} \end{aligned} \tag{11.23b}$$

By (5.45) and (5.46), we obtain:

$$E\{\boldsymbol{\varepsilon}_w\} = \mathbf{W}^{1/2}E\{\boldsymbol{\varepsilon}\} = \mathbf{W}^{1/2}\mathbf{0} = \mathbf{0} \tag{11.24a}$$

$$\begin{aligned} \sigma^2\{\boldsymbol{\varepsilon}_w\} &= \mathbf{W}^{1/2}\sigma^2\{\boldsymbol{\varepsilon}\}\mathbf{W}^{1/2} = \mathbf{W}^{1/2}\mathbf{W}^{-1}\mathbf{W}^{1/2} \\ &= \mathbf{W}^{1/2}\mathbf{W}^{-1/2}\mathbf{W}^{-1/2}\mathbf{W}^{1/2} = \mathbf{I} \end{aligned} \tag{11.24b}$$

Thus, regression model (11.23a) involves independent error terms with mean zero and constant variance $\sigma_i^2 \equiv 1$. We can therefore apply standard regression procedures to this transformed regression model.

For example. the ordinary least squares estimators of the regression coefficients in (6.25) here become:

$$\mathbf{b}_w = (\mathbf{X}_w'\mathbf{X}_w)^{-1}\mathbf{X}_w'\mathbf{Y}_w$$

Using the definitions in (11.23b), we obtain the result for weighted least squares given in (11.9):

$$\begin{aligned} \mathbf{b}_w &= [(\mathbf{W}^{1/2}\mathbf{X})'\mathbf{W}^{1/2}\mathbf{X}]^{-1}(\mathbf{W}^{1/2}\mathbf{X})'\mathbf{W}^{1/2}\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{W}^{1/2}\mathbf{W}^{1/2}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}\mathbf{W}^{1/2}\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} \end{aligned}$$

6. Weighted least squares is a special case of *generalized least squares* where the error terms not only may have different variances but pairs of error terms may also be correlated.

7. For simple linear regression, the weighted least squares normal equations in (11.8) become:

$$\begin{aligned} \sum w_i Y_i &= b_{w0}\sum w_i + b_{w1}\sum w_i X_i \\ \sum w_i X_i Y_i &= b_{w0}\sum w_i X_i + b_{w1}\sum w_i X_i^2 \end{aligned} \tag{11.25}$$

and the weighted least squares estimators b_{w0} and b_{w1} in (11.9) are:

$$b_{w1} = \frac{\sum w_i X_i Y_i - \frac{\sum w_i X_i \sum w_i Y_i}{\sum w_i}}{\sum w_i X_i^2 - \frac{\left(\sum w_i X_i\right)^2}{\sum w_i}} \tag{11.26a}$$

$$b_{w0} = \frac{\sum w_i Y_i - b_1 \sum w_i X_i}{\sum w_i} \tag{11.26b}$$

Note that if all weights are equal so w_i is identically equal to a constant, the normal equations (11.25) for weighted least squares reduce to the ones for unweighted least squares in (1.9) and the weighted least squares estimators (11.26) reduce to the ones for unweighted least squares in (1.10). ■

11.2 Multicollinearity Remedial Measures—Ridge Regression

We consider first some remedial measures for serious multicollinearity that can be implemented with ordinary least squares, and then take up ridge regression, a method of overcoming serious multicollinearity problems by modifying the method of least squares.

Some Remedial Measures

1. As we saw in Chapter 7, the presence of serious multicollinearity often does not affect the usefulness of the fitted model for estimating mean responses or making predictions, provided that the values of the predictor variables for which inferences are to be made follow the same multicollinearity pattern as the data on which the regression model is based. Hence, one remedial measure is to restrict the use of the fitted regression model to inferences for values of the predictor variables that follow the same pattern of multicollinearity.

2. In polynomial regression models, as we noted in Chapter 7, use of centered data for the predictor variable(s) serves to reduce the multicollinearity among the first-order, second-order, and higher-order terms for any given predictor variable.

3. One or several predictor variables may be dropped from the model in order to lessen the multicollinearity and thereby reduce the standard errors of the estimated regression coefficients of the predictor variables remaining in the model. This remedial measure has two important limitations. First, no direct information is obtained about the dropped predictor variables. Second, the magnitudes of the regression coefficients for the predictor variables remaining in the model are affected by the correlated predictor variables not included in the model.

4. Sometimes it is possible to add some cases that break the pattern of multicollinearity. Often, however, this option is not available. In business and economics, for instance, many predictor variables cannot be controlled, so that new cases will tend to show the same intercorrelation patterns as the earlier ones.

5. In some economic studies, it is possible to estimate the regression coefficients for different predictor variables from different sets of data and thereby avoid the problems of multicollinearity. Demand studies, for instance, may use both cross-section and time series data to this end. Suppose the predictor variables in a demand study are price and income,

and the relation to be estimated is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{11.27}$$

where Y is demand, X_1 is income, and X_2 is price. The income coefficient β_1 may then be estimated from cross-section data. The demand variable Y is thereupon adjusted:

$$Y_i' = Y_i - b_1 X_{i1} \tag{11.28}$$

Finally, the price coefficient β_2 is estimated by regressing the adjusted demand variable Y' on X_2.

6. Another remedial measure for multicollinearity that can be used with ordinary least squares is to form one or several composite indexes based on the highly correlated predictor variables, an index being a linear combination of the correlated predictor variables. The methodology of *principal components* provides composite indexes that are uncorrelated. Often, a few of these composite indexes capture much of the information contained in the predictor variables. These few uncorrelated composite indexes are then used in the regression analysis as predictor variables instead of the original highly correlated predictor variables. A limitation of principal components regression, also called latent root regression, is that it may be difficult to attach concrete meanings to the indexes.

More information about these remedial approaches as well as about Bayesian regression, where prior information about the regression coefficients is incorporated into the estimation procedure, may be obtained from specialized works such as Reference 11.3.

Ridge Regression

Biased Estimation. Ridge regression is one of several methods that have been proposed to remedy multicollinearity problems by modifying the method of least squares to allow biased estimators of the regression coefficients. When an estimator has only a small bias and is substantially more precise than an unbiased estimator, it may well be the preferred estimator since it will have a larger probability of being close to the true parameter value. Figure 11.2 illustrates this situation. Estimator b is unbiased but imprecise, whereas estimator b^R is much more precise but has a small bias. The probability that b^R falls near the true value β is much greater than that for the unbiased estimator b.

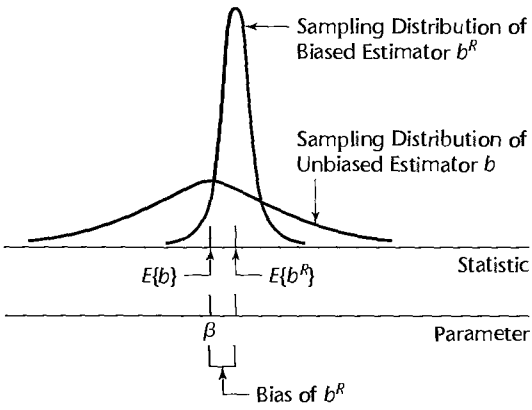


FIGURE 11.2 Biased Estimator with Small Variance May Be Preferable to Unbiased Estimator with Large Variance.

A measure of the combined effect of bias and sampling variation is the mean squared error, a concept that we encountered in Chapter 9 in connection with the C_p criterion. Here, the mean squared error is the expected value of the squared deviation of the biased estimator b^R from the true parameter β. As before, this expected value is the sum of the variance of the estimator and the squared bias:

$$E\{b^R - \beta\}^2 = \sigma^2\{b^R\} + (E\{b^R\} - \beta)^2 \tag{11.29}$$

Note that if the estimator is unbiased, the mean squared error is identical to the variance of the estimator.

Ridge Estimators. For ordinary least squares, the normal equations are given by (6.24):

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y} \tag{11.30}$$

When all variables are transformed by the correlation transformation (7.44), the transformed regression model is given by (7.45):

$$Y_i^* = \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^* \tag{11.31}$$

and the least squares normal equations are given by (7.52a):

$$\mathbf{r}_{XX}\mathbf{b} = \mathbf{r}_{YX} \tag{11.32}$$

where $\mathbf{r}_{XX}$ is the correlation matrix of the X variables defined in (7.47) and $\mathbf{r}_{YX}$ is the vector of coefficients of simple correlation between Y and each X variable defined in (7.48).

The ridge standardized regression estimators are obtained by introducing into the least squares normal equations (11.32) a biasing constant $c \geq 0$, in the following form:

$$(\mathbf{r}_{XX} + c\mathbf{I})\mathbf{b}^R = \mathbf{r}_{YX} \tag{11.33}$$

where $\mathbf{b}^R$ is the vector of the standardized ridge regression coefficients b_k^R:

$$\underset{(p-1)\times 1}{\mathbf{b}^R} = \begin{bmatrix} b_1^R \\ b_2^R \\ \vdots \\ b_{p-1}^R \end{bmatrix} \tag{11.33a}$$

and $\mathbf{I}$ is the $(p-1) \times (p-1)$ identity matrix. Solution of the normal equations (11.33) yields the ridge standardized regression coefficients:

$$\mathbf{b}^R = (\mathbf{r}_{XX} + c\mathbf{I})^{-1}\mathbf{r}_{YX} \tag{11.34}$$

The constant c reflects the amount of bias in the estimators. When $c = 0$, (11.34) reduces to the ordinary least squares regression coefficients in standardized form, as given in (7.52b). When $c > 0$, the ridge regression coefficients are biased but tend to be more stable (i.e., less variable) than ordinary least squares estimators.

Choice of Biasing Constant c. It can be shown that the bias component of the total mean squared error of the ridge regression estimator $\mathbf{b}^R$ increases as c gets larger (with all b_k^R tending toward zero) while the variance component becomes smaller. It can further be shown that there always exists some value c for which the ridge regression estimator $\mathbf{b}^R$ has

a smaller total mean squared error than the ordinary least squares estimator **b**. The difficulty is that the optimum value of c varies from one application to another and is unknown.

A commonly used method of determining the biasing constant c is based on the *ridge trace* and the variance inflation factors $(VIF)_k$ in (10.41). The ridge trace is a simultaneous plot of the values of the $p-1$ estimated ridge standardized regression coefficients for different values of c, usually between 0 and 1. Extensive experience has indicated that the estimated regression coefficients b_k^R may fluctuate widely as c is changed slightly from 0, and some may even change signs. Gradually, however, these wide fluctuations cease and the magnitudes of the regression coefficients tend to move slowly toward zero as c is increased further. At the same time, the values of $(VIF)_k$ tend to fall rapidly as c is changed from 0, and gradually the $(VIF)_k$ values also tend to change only moderately as c is increased further. One therefore examines the ridge trace and the *VIF* values and chooses the smallest value of c where it is deemed that the regression coefficients first become stable in the ridge trace and the *VIF* values have become sufficiently small. The choice is thus a judgmental one.

Example

In the body fat example with three predictor variables in Table 7.1, we noted previously several informal indications of severe multicollinearity in the data. Indeed, in the fitted model with three predictor variables (Table 7.2d), the estimated regression coefficient b_2 is negative even though it was expected that amount of body fat is positively related to thigh circumference. Ridge regression calculations were made for the body fat example data in Table 7.1 (calculations not shown). The ridge standardized regression coefficients for selected values of c are presented in Table 11.2, and the variance inflation factors are given in Table 11.3. The coefficients of multiple determination R^2 are also shown in the latter table. Figure 11.3 presents the ridge trace of the estimated standardized regression coefficients based on calculations for many more values of c than those shown in Table 11.2. To facilitate the analysis, the horizontal c scale in Figure 11.3 is logarithmic.

TABLE 11.2 Ridge Estimated Standardized Regression Coefficients for Different Biasing Constants c—Body Fat Example with Three Predictor Variables.

c	b_1^R	b_2^R	b_3^R
.000	4.264	−2.929	−1.561
.002	1.441	−.4113	−.4813
.004	1.006	−.0248	−.3149
.006	.8300	.1314	−.2472
.008	.7343	.2158	−.2103
.010	.6742	.2684	−.1870
.020	.5463	.3774	−.1369
.030	.5004	.4134	−.1181
.040	.4760	.4302	−.1076
.050	.4605	.4392	−.1005
.100	.4234	.4490	−.0812
.500	.3377	.3791	−.0295
1.000	.2798	.3101	−.0059

TABLE 11.3 *VIF* Values for Regression Coefficients and R^2 for Different Biasing Constants c—Body Fat Example with Three Predictor Variables.

c	$(VIF)_1$	$(VIF)_2$	$(VIF)_3$	R^2
.000	708.84	564.34	104.61	.8014
.002	50.56	40.45	8.28	.7901
.004	16.98	13.73	3.36	.7864
.006	8.50	6.98	2.19	.7847
.008	5.15	4.30	1.62	.7838
.010	3.49	2.98	1.38	.7832
.020	1.10	1.08	1.01	.7818
.030	.63	.70	.92	.7812
.040	.45	.56	.88	.7808
.050	.37	.49	.85	.7804
.100	.25	.37	.76	.7784
.500	.15	.21	.40	.7427
1.000	.11	.14	.23	.6818

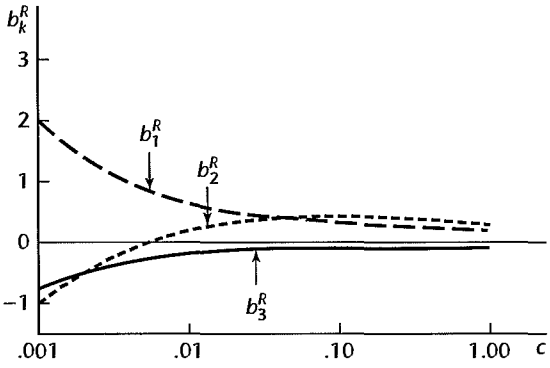


FIGURE 11.3 Ridge Trace of Estimated Standardized Regression Coefficients—Body Fat Example with Three Predictor Variables.

Note the instability in Figure 11.3 of the regression coefficients for very small values of c. The estimated regression coefficient b_2^R, in fact, changes signs. Also note the rapid decrease in the *VIF* values in Table 11.3. It was decided to employ $c = .02$ here because for this value of the biasing constant the ridge regression coefficients have *VIF* values near 1 and the estimated regression coefficients appear to have become reasonably stable. The resulting fitted model for $c = .02$ is:

$$\hat{Y}^* = .5463X_1^* + .3774X_2^* - .1369X_3^*$$

Transforming back to the original variables by (7.53), we obtain:

$$\hat{Y} = -7.3978 + .5553X_1 + .3681X_2 - .1917X_3$$

where $\bar{Y} = 20.195$, $\bar{X}_1 = 25.305$, $\bar{X}_2 = 51.170$, $\bar{X}_3 = 27.620$, $s_Y = 5.106$, $s_1 = 5.023$, $s_2 = 5.235$, and $s_3 = 3.647$.

The improper sign on the estimate for β_2 has now been eliminated, and the estimated regression coefficients are more in line with prior expectations. The sum of the squared residuals for the transformed variables, which increases with c, has only increased from .1986 at $c = 0$ to .2182 at $c = .02$ while R^2 decreased from .8014 to .7818. These changes are relatively modest. The estimated mean body fat when $X_{h1} = 25.0$, $X_{h2} = 50.0$, and $X_{h3} = 29.0$ is 19.33 for the ridge regression at $c = .02$ compared to 19.19 utilizing the ordinary least squares solution. Thus, the ridge solution at $c = .02$ appears to be quite satisfactory here and a reasonable alternative to the ordinary least squares solution.

Comments

1. The normal equations (11.33) for the ridge estimators are as follows:

$$\begin{aligned}
(1+c)b_1^R + r_{12}b_2^R + \cdots + r_{1,p-1}b_{p-1}^R &= r_{Y1} \\
r_{21}b_1^R + (1+c)b_2^R + \cdots + r_{2,p-1}b_{p-1}^R &= r_{Y2} \\
&\vdots \\
r_{p-1,1}b_1^R + r_{p-1,2}b_2^R + \cdots + (1+c)b_{p-1}^R &= r_{Y,p-1}
\end{aligned} \tag{11.35}$$

where r_{ij} is the coefficient of simple correlation between the ith and jth X variables and r_{Yj} is the coefficient of simple correlation between the response variable Y and the jth X variable.

2. *VIF* values for ridge regression coefficients b_k^R are defined analogously to those for ordinary least squares regression coefficients. Namely, the *VIF* value for b_k^R measures how large is the variance of b_k^R relative to what the variance would be if the predictor variables were uncorrelated. It can be shown that the *VIF* values for the ridge regression coefficients b_k^R are the diagonal elements of the following $(p-1) \times (p-1)$ matrix:

$$(\mathbf{r}_{XX} + c\mathbf{I})^{-1}\mathbf{r}_{XX}(\mathbf{r}_{XX} + c\mathbf{I})^{-1} \tag{11.36}$$

3. The coefficient of multiple determination R^2, which for ordinary least squares is given in (6.40):

$$R^2 = 1 - \frac{SSE}{SSTO} \tag{11.37}$$

can be defined analogously for ridge regression. A simplification occurs, however, because the total sum of squares for the correlation-transformed dependent variable Y^* in (7.44a) is:

$$SSTO_R = \sum (Y_i^* - \bar{Y}^*)^2 = 1 \tag{11.38}$$

The fitted values with ridge regression are:

$$\hat{Y}_i^* = b_1^R X_{i1}^* + \cdots + b_{p-1}^R X_{i,p-1}^* \tag{11.39}$$

where the X_{ik}^* are the X variables transformed according to the correlation transformation (7.44b). The error sum of squares, as usual, is:

$$SSE_R = \sum (Y_i^* - \hat{Y}_i^*)^2 \tag{11.40}$$

where $\hat{Y}_i^*$ is given in (11.39). R^2 for ridge regression then becomes:

$$R_R^2 = 1 - SSE_R \tag{11.41}$$

4. Ridge regression estimates can be obtained by the method of *penalized least squares*. The penalized least squares criterion combines the usual sum of squared errors with a penalty for large regression coefficients:

$$Q = \sum_{i=1}^{n} [Y_i^* - (\beta_1^* X_{i1}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^*)]^2 + c \left[\sum_{j=1}^{p-1} (\beta_j^*)^2 \right]$$

The penalty is a biasing constant, c, times the sum of squares of the regression coefficients. Large absolute regression parameters lead to a large penalty; thus, it can be seen that for $c > 0$ the "best" coefficients generally will be smaller in absolute magnitude than the ordinary least squares estimates. For this reason, ridge estimators are sometimes referred to as *shrinkage* estimators.

5. Ridge regression estimates tend to be stable in the sense that they are usually little affected by small changes in the data on which the fitted regression is based. In contrast, ordinary least squares estimates may be highly unstable under these conditions when the predictor variables are highly multicollinear. Predictions of new observations made from ridge estimated regression functions tend to be more precise than predictions made from ordinary least squares regression functions when the predictor variables are correlated and the new observations follow the same multicollinearity pattern (see, for instance, Reference 11.4). The prediction precision advantage with ridge regression is especially great when the intercorrelations among the predictor variables are high.

6. Ridge estimated regression functions at times will provide good estimates of mean responses or predictions of new observations for levels of the predictor variables outside the region of the observations on which the regression function is based. In contrast, estimated regression functions based on ordinary least squares may perform quite poorly in such circumstances. Of course, any estimation or prediction well outside the region of the observations should always be made with great caution.

7. A major limitation of ridge regression is that ordinary inference procedures are not applicable and exact distributional properties are not known. Bootstrapping, a computer-intensive procedure to be discussed in Section 11.5, can be employed to evaluate the precision of ridge regression coefficients. Another limitation of ridge regression is that the choice of the biasing constant c is a judgmental one. Although a variety of formal methods have been developed for making this choice, these have their own limitations.

8. The ridge regression procedures have been generalized to allow for differing biasing constants for the different estimated regression coefficients; see, for instance, Reference 11.3.

9. Ridge regression can be used to help in reducing the number of potential predictor variables in exploratory observational studies by analyzing the ridge trace. Variables whose ridge trace is unstable, with the coefficient tending toward the value of zero, are dropped with this approach. Also, variables whose ridge trace is stable but at a very small value are dropped. Finally, variables with unstable ridge traces that do not tend toward zero are considered as candidates for dropping. ■

11.3 Remedial Measures for Influential Cases—Robust Regression

We noted in Chapter 10 that the hat matrix and studentized deleted residuals are valuable tools for identifying cases that are outlying with respect to the X and Y variables. In addition, we considered there how to measure the influence of these outlying cases on the fitted values and estimated regression coefficients by means of the *DFFITS*, Cook's distance, and *DFBETAS* measures. The reason for our concern with outlying cases is that the method of least squares is particularly susceptible to these cases, resulting sometimes in a seriously distorted fitted model for the remaining cases. A crucial question that arises now is how to handle highly influential cases.

A first step is to examine whether an outlying case is the result of a recording error, breakdown of a measurement instrument, or the like. For instance, in a study of the waiting time in a telephone reservation system, one waiting time was recorded as 1,000 rings. This observation was so extreme and unrealistic that it was clearly erroneous. If erroneous data can be corrected, this should be done. Often, however, erroneous data cannot be corrected later on and should be discarded. Many times, unfortunately, it is not possible after the data have been obtained to tell for certain whether the observations for an outlying case are erroneous. Such cases should usually not be discarded.

If an outlying influential case is not clearly erroneous, the next step should be to examine the adequacy of the model. Scientists frequently have primary interest in the outlying cases because they deviate from the currently accepted model. Examination of these outlying cases may provide important clues as to how the model needs to be modified. In a study of the yield of a process, a first-order model was fitted for the two important factors under consideration because previous studies had not found any interaction effects between these factors on the yield. One case in the current study was outlying and highly influential, with extremely high yield; it corresponded to unusually high levels of the two factors. The tentative conclusion drawn was that an interaction effect is present; this was subsequently confirmed in a follow-up study. The improved model, resulting from the outlying case, led to greatly improved process productivity.

Outlying cases may also lead to the finding of other types of model inadequacies, such as the omission of an important variable or the choice of an incorrect functional form (e.g., a quadratic function instead of an exponential function). The analysis of outlying influential

cases can frequently lead to valuable insights for strengthening the model such that the outlying case is no longer an outlier but is accounted for by the model.

Discarding of outlying influential cases that are not clearly erroneous and that cannot be accounted for by model improvements should be done only rarely, such as when the model is not intended to cover the special circumstances related to the outlying cases. For example, a few cases in an industrial study were outlying and highly influential. These cases occurred early in the study, when the plant was in transition from one process to the new one under study. Discarding of these early cases was deemed to be reasonable since the model was intended for use after the new process had stabilized.

An alternative to discarding outlying cases that is less severe is to dampen the influence of these cases. That is the purpose of robust regression.

Robust Regression

Robust regression procedures dampen the influence of outlying cases, as compared to ordinary least squares estimation, in an effort to provide a better fit for the majority of cases. They are useful when a known, smooth regression function is to be fitted to data that are "noisy," with a number of outlying cases, so that the assumption of a normal distribution for the error terms is not appropriate. Robust regression procedures are also useful when automated regression analysis is required. For example, a complex measurement instrument used for internal medical examinations must be calibrated for each use. There is no time for a thorough identification of outlying cases and an analysis of their influence, nor for a careful consideration of remedial measures. Instead, an automated regression calibration must be used. Robust regression procedures will automatically guard against undue influence of outlying cases in this situation.

Numerous robust regression procedures have been developed. They are described in specialized texts, such as References 11.5 and 11.6. We mention briefly a few of these procedures and then describe in more detail one commonly used procedure based on iteratively reweighted least squares.

LAR or LAD Regression. Least absolute residuals (LAR) or least absolute deviations (LAD) regression, also called *minimum L_1-norm regression*, is one of the most widely used robust regression procedures. It is insensitive to both outlying data values and inadequacies of the model employed. The method of least absolute residuals estimates the regression coefficients by minimizing the sum of the absolute deviations of the Y observations from their means. The criterion to be minimized, denoted by L_1, is:

$$L_1 = \sum_{i=1}^{n} |Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1})| \tag{11.42}$$

Since absolute deviations rather than squared ones are involved here, the LAR method places less emphasis on outlying observations than does the method of least squares.

The estimated LAR regression coefficients can be obtained by linear programming techniques. Details about computational aspects may be found in specialized texts, such as Reference 11.7. The LAR fitted regression model differs from the least squares fitted model in that the residuals ordinarily will not sum to zero. Also, the solution for the estimated regression coefficients with the method of least absolute residuals may not be unique.

IRLS Robust Regression. Iteratively reweighted least squares (IRLS) robust regression uses the weighted least squares procedures discussed in Section 11.1 to dampen the influence of outlying observations. Instead of weights based on the error variances, IRLS robust regression uses weights based on how far outlying a case is, as measured by the residual for that case. The weights are revised with each iteration until a robust fit has been obtained. We shall discuss this procedure in more detail shortly.

LMS Regression. Least median of squares (LMS) regression replaces the sum of squared deviations in ordinary least squares by the median of the squared deviations, which is a robust estimator of location. The criterion for this procedure is to minimize the median squared deviation:

$$\text{median}\{[Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1})]^2\} \tag{11.43}$$

with respect to the regression coefficients. Thus, this procedure leads to estimated regression coefficients $b_0, b_1, \ldots, b_{p-1}$ that minimize the median of the squared residuals.

Other Robust Regression Procedures. There are many other robust regression procedures. Some involve trimming one or several of the extreme squared deviations before applying the least squares criterion; others are based on ranks. Many of the robust regression procedures require extensive computing.

IRLS Robust Regression

Iteratively reweighted least squares was encountered in Section 11.1 as a remedial measure for unequal error variances in connection with the obtaining of weights from an estimated variance or standard deviation function. For robust regression, weighted least squares is used to reduce the influence of outlying cases by employing weights that vary inversely with the size of the residual. Outlying cases that have large residuals are thereby given smaller weights. The weights are revised as each iteration yields new residuals until the estimation process stabilizes. A summary of the steps follows:

1. Choose a weight function for weighting the cases.
2. Obtain starting weights for all cases.
3. Use the starting weights in weighted least squares and obtain the residuals from the fitted regression function.
4. Use the residuals in step 3 to obtain revised weights.
5. Continue the iterations until convergence is obtained.

We now discuss each of the steps in IRLS robust regression.

Weight Function. Many weight functions have been proposed for dampening the influence of outlying cases. Two widely used weight functions are the Huber and bisquare weight functions:

$$\text{Huber: } w = \begin{cases} 1 & |u| \le 1.345 \\ \dfrac{1.345}{|u|} & |u| > 1.345 \end{cases} \tag{11.44}$$

$$\text{Bisquare: } w = \begin{cases} \left[1 - \left(\dfrac{u}{4.685}\right)^2\right]^2 & |u| \le 4.685 \\ 0 & |u| > 4.685 \end{cases} \tag{11.45}$$

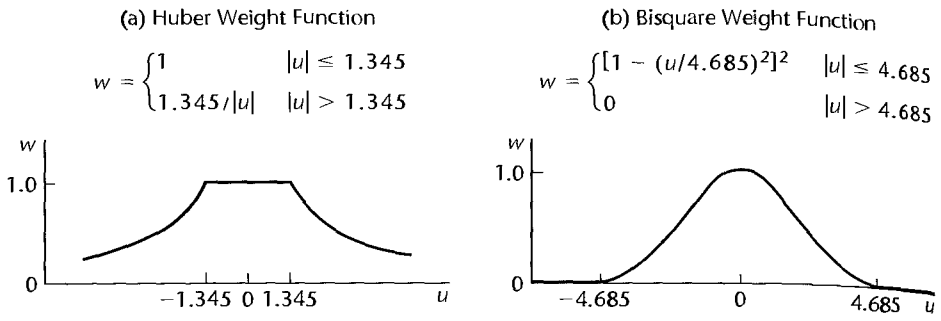


FIGURE 11.4 Two Weight Functions Used in IRLS Robust Regression.

As before, w denotes the weight, and u denotes the scaled residual to be defined shortly. The constant 1.345 in the Huber weight function and the constant 4.685 in the bisquare weight function are called *tuning constants*. They were chosen to make the IRLS robust procedure 95 percent efficient for data generated by the normal error regression model (6.7). Figure 11.4 shows graphs of the two weight functions. Note how the weight w according to each weight function declines as the absolute scaled residual gets larger, and that each weight function is symmetric around $u = 0$. Also note that the Huber weight function does not reduce the weight of a case from 1.0 until the absolute scaled residual exceeds 1.345, and that all cases receive some positive weight, no matter how large the absolute scaled residual. In contrast, the bisquare weight function reduces the weights of all cases from 1.0 (unless the residual is zero). In addition, the bisquare weight function gives weight 0 to all cases whose absolute scaled residual exceeds 4.685, thereby entirely excluding these extreme cases.

Starting Values. Calculations with some of the weight functions are very sensitive to the starting values; with others, this is less of a problem. When the Huber weight function is employed, the initial residuals may be those obtained from an ordinary least squares fit. The bisquare function calculations, on the other hand, are more sensitive to the starting values. To obtain good starting values for the bisquare weight function, the Huber weight function is often used to obtain an initial robust regression fit, and the residuals for this fit are then employed as starting values for several iterations with the bisquare weight function. Alternatively, least absolute residuals regression in (11.42) may be used to obtain starting residuals when the bisquare weight function is used.

Scaled Residuals. The weight functions (11.44) and (11.45) are each designed to be used with scaled residuals. The semistudentized residuals in (3.5) are scaled residuals and could be employed. However, in the presence of outlying observations, $\sqrt{MSE}$ is not a resistant estimator of the error term standard deviation σ; the magnitude of $\sqrt{MSE}$ can be greatly influenced by one or several outlying observations. Also, $\sqrt{MSE}$ is not a robust estimator of σ when the distribution of the error terms is far from normal. Instead, the resistant and robust median absolute deviation (MAD) estimator is often employed:

$$MAD = \frac{1}{.6745} \text{median}\{|e_i - \text{median}\{e_i\}|\} \tag{11.46}$$

The constant .6745 provides an unbiased estimate of σ for independent observations from a normal distribution. Here, it serves to provide an estimate that is approximately unbiased.

The scaled residual u_i based on (11.46) then is:

$$u_i = \frac{e_i}{MAD} \tag{11.47}$$

Number of Iterations. The iterative process of obtaining a new fit, new residuals and thereby new weights, and then refitting with the new weights continues until the process converges. Convergence can be measured by observing whether the weights change relatively little, whether the residuals change relatively little, whether the estimated regression coefficients change relatively little, or whether the fitted values change relatively little.

Example 1: Mathematics Proficiency with One Predictor

The Educational Testing Service Study *America's Smallest School: The Family* (Ref. 11.8) investigated the relation of educational achievement of students to their home environment. Although earlier studies examined the relation of educational achievement to family socioeconomic status (e.g., parents' education, family income, parents' occupation), this study employed more direct measures of the home environment. Specifically, the relation of educational achievement of eighth-grade students in mathematics to the following five explanatory variables was investigated:

PARENTS (X_1)—percentage of eighth-grade students with both parents living at home

HOMELIB (X_2)—percentage of eighth-grade students with three or more types of reading materials at home (books, encyclopedias, magazines, newspapers)

READING (X_3)—percentage of eighth-grade students who read more than 10 pages a day

TVWATCH (X_4)—percentage of eighth-grade students who watch TV for six hours or more per day

ABSENCES (X_5)—percentage of eighth-grade students absent three days or more last month

Data on average mathematics proficiency (MATHPROF) and the home environment variables were obtained from the 1990 National Assessment of Educational Progress for 37 states, the District of Columbia, Guam, and the Virgin Islands. A portion of the data is shown in Table 11.4.

Our first example of robust regression using iteratively reweighted least squares involves only one predictor, HOMELIB (X_2). In this way, simple plots can be used to present the data and the fitted regression function.

Figure 11.5a presents a scatter plot of the data, together with a plot of a first-order (simple linear) regression model fit by ordinary least squares and a lowess smooth. The lowess smooth suggests that the relationship between home reading resources and average mathematics proficiency is curvilinear—possibly second order—for the majority of states, but three points are clear outliers. The District of Columbia and the Virgin Islands are outliers with respect to mathematics proficiency (Y), and Guam appears to be an outlier with respect to both mathematics proficiency and available reading resources (X). Figure 11.5b presents a plot against X of the residuals obtained from the fitted first-order model in Figure 11.5a. This plot shows clearly the three outlying Y cases. Note also from the residual plot that there is a group of six states with low reading resources levels, between 68 and 73, whose average mathematics proficiency scores are all above the fitted regression line. This is another indication that a second-order polynomial model may be appropriate.

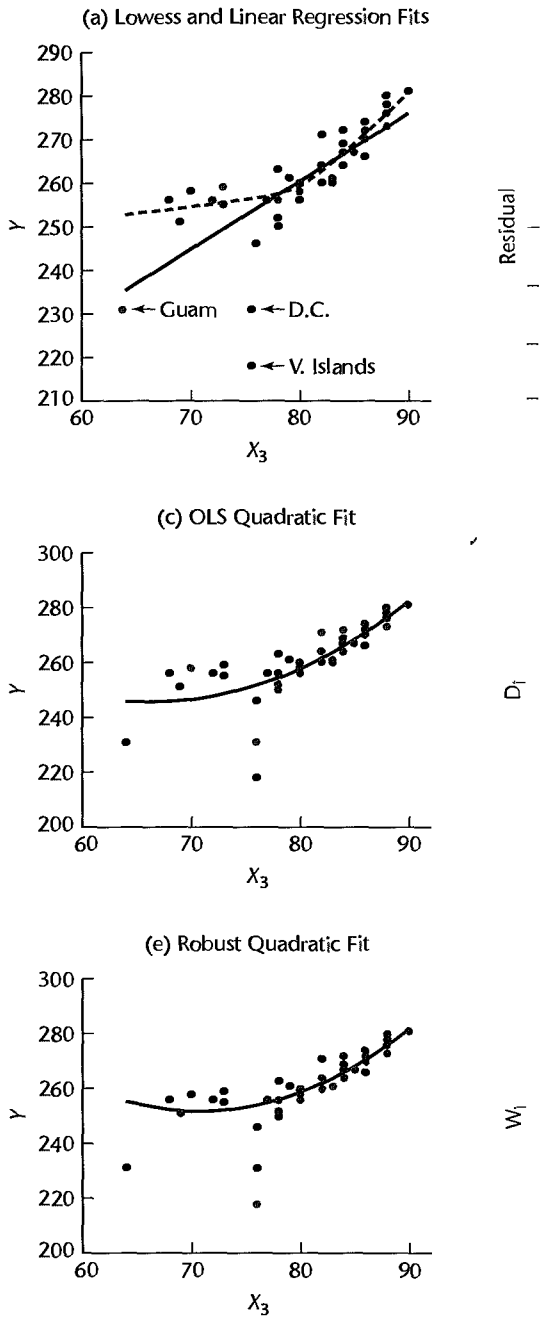


FIGURE 11.5
Comparison of Lowess, Ordinary Least Squares Fits, and Robust Quadratic Fits—Mathematics Proficiency Example.

TABLE 11.4 Data Set—Mathematics Proficiency Example.

	State	MATHPROF Y	PARENTS X_1	HOMELIB X_2	READING X_3	TVWATCH X_4	ABSENCES X_5
1	Alabama	252	75	78	34	18	18
2	Arizona	259	75	73	41	12	26
3	Arkansas	256	77	77	28	20	23
4	California	256	78	68	42	11	28
	...	...	...	...	...	...	...
9	D.C.	231	47	76	24	33	37
	...	...	...	...	...	...	...
	Guam	231	81	64	32	20	28
	...	...	...	...	...	...	...
35	Texas	258	77	70	34	15	18
36	Virgin Islands	218	63	76	23	27	22
37	Virginia	264	78	82	33	16	24
38	West Virginia	256	82	80	36	16	25
39	Wisconsin	274	81	86	38	8	21
40	Wyoming	272	85	86	43	7	23

Source: ETS Policy Information Center, *America's Smallest School: The Family* (Princeton, New Jersey: Educational Testing Service, 1992).

Second-order model (8.2):

$$Y_i = \beta_0 + \beta_2 x_{i2} + \beta_{22} x_{i2}^2 + \varepsilon_i \qquad \textbf{(11.48)}$$

was next fit, again using ordinary least squares. Recall that this model requires calculation of the centered predictor $x_{i2} = X_{i2} - \bar{X}_{i2}$ and its square, x_{i2}^2. A plot of the fit of the second-order model, superimposed on a scatter-plot of the data, is shown in Figure 11.5c. Though improved, the fit is again unsatisfactory: the six points that fell above the first-order fit are still above the fitted second-order model. The regression line is clearly being influenced by the three outliers identified above. The Cook's distance measures for the second-order fit are displayed in an index plot in Figure 11.5d. The plot confirms the influence of Guam and the Virgin Islands.

In an effort to dampen the effect of the three outliers, we shall fit second-order model (8.2) robustly, using iteratively reweighted least squares and the Huber weight function (11.44). We illustrate the calculations for case 1, Alabama. The regression model to be fitted is the first-order model. An ordinary least squares fit of this model yields:

$$\hat{Y} = 258.436 + 1.8327 x_2 + 0.06491 x_2^2 \qquad \textbf{(11.49)}$$

The residual for Alabama is $e_1 = -2.4109$. The residuals are shown in Column 1 of Table 11.5. The median of the 40 residuals is median$\{e_i\} = 0.7063$. Hence, $e_1 - \text{median}\{e_i\} = -2.4109 - 0.7063 = -3.1172$, and the absolute deviation is $|e_1 - \text{median}\{e_i\}| = 3.1172$. The median of the 40 absolute deviations is:

$$\text{median}\{|e_i - \text{median}\{e_i\}|\} = 3.1488$$

TABLE 11.5 Iteratively Huber-Reweighted Least Squares Calculations—Mathematics Proficiency Example.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Iteration 0		Iteration 1		Iteration 2		Iteration 7	
i	e_i	u_i	w_i	e_i	w_i	e_i	w_i	e_i
1	−2.4109	−0.51643	1.00000	−3.7542	1.00000	−4.0354	1.00000	−4.1269
2	10.5724	2.26466	0.59391	8.4297	0.71515	7.4848	0.86011	6.7698
3	3.0454	0.65234	1.00000	1.5411	1.00000	1.1559	1.00000	0.9731
4	10.3104	2.20853	0.60900	7.3822	0.81663	5.4138	1.00000	3.6583
...	...	...	...	...	...	...	...	...
8	−20.6282	−4.41866	0.30439	−22.2929	0.27042	−22.7964	0.25263	−23.0873
...	...		...	...	...	...	...	...
11	−14.8358	−3.17791	0.42323	−18.3824	0.32795	−21.4287	0.24019	−24.3167
...	...	...	...	...	...	...	...	...
36	−33.6282	−7.20333	0.18672	−35.2929	0.17081	−35.7964	0.16161	−36.0873
37	2.4659	0.52821	1.00000	1.7722	1.00000	1.7627	1.00000	1.8699
38	−1.7129	−0.36691	1.00000	−2.7325	1.00000	−2.8490	1.00000	−2.8079
39	3.2658	0.69954	1.00000	3.2305	1.00000	3.2624	1.00000	3.3014
40	1.2658	0.27113	1.00000	1.2305	1.00000	1.2624	1.00000	1.3014

so that the *MAD* estimator (11.46) is:

$$MAD = \frac{3.1488}{.6745} = 4.6683$$

Hence, the scaled residual (11.47) for Alabama is:

$$u_1 = \frac{-2.4109}{4.6683} = -.5164$$

The scaled residuals are shown in Table 11.5, column 2. Since $|u_1| = .5164 \leq 1.345$, the initial Huber weight for Alabama is $w_1 = 1.0$. The initial weights are shown in Table 11.5, column 3. To interpret these weights, remember that ordinary least squares may be viewed as a special case of weighted least squares with the weights for all cases being equal to 1. We note in column 3 that the initial weights for cases 8, 11, and 36 (District of Columbia, Guam, and Virgin Islands) are substantially reduced, and that the weights for some other states are reduced somewhat.

The first iteration of weighted least squares uses the initial weights in column 3, leading to the fitted regression model:

$$\hat{Y} = 259.390 + 1.6701x_2 + 0.06463x_2^2 \qquad \textbf{(11.50)}$$

This fitted regression function differs considerably from the ordinary least squares fit in (11.49). The coefficient of x_2 has decreased from $b_2 = 1.8327$ to $b_2 = 1.6701$, while the curvature term $b_{22} = 0.06463$ changed little from its previous value of $b_{22} = 0.06491$. This has permitted the estimated regression function to increase for smaller values of X_2 and to therefore conform more closely to the six values that previously fell above the fitted line.

Iteration 2 uses the residuals in column 4 of Table 11.5, scales them, and obtains revised Huber weights, which are then used in iteration 2 of weighted least squares. The weights

obtained for the eighth iteration differed relatively little from those for the seventh iteration; hence the iteration process was stopped with the seventh iteration. The final weights are shown in Table 11.5, column 7. Note that only minor changes in the weights occurred between iterations 2 and 7. Use of the weights in column 7 leads to the final fitted model:

$$\hat{Y} = 259.421 + 1.5649x_2 + 0.08016x_2^2 \qquad \textbf{(11.51)}$$

The residuals for the final fit are shown in Table 11.5, column 8. Just as the weights changed only moderately between iterations 2 and 7, so the residuals changed only to a small extent after iteration 2. Note that the coefficient of the curvature term did change a bit more substantially—from $b_{22} = .06463$ to $b_{22} = .08016$.

Figure 11.5e shows the scatter plot and the IRLS fitted second-order regression function, and Figure 11.5f contains an index plot of the weights used in the final iteration. The robust fit now tracks the responses to the 37 states extremely well, and the fit to the six cases that were previously above the regression line is now satisfactory. The plot of the final weights in Figure 11.5f shows clearly the downweighting of the three outliers.

We conclude from the robust fit in Figure 11.5e that there is a clear upward-curving relationship between availability of reading resources in the home and average mathematics proficiency at the state level. This does not necessarily imply a causal relation, of course. The availability of reading resources may be positively correlated with other variables that are causally related to mathematics proficiency.

Example 2: Mathematics Proficiency with Five Predictors

We shall explore from a descriptive perspective the relationship between average mathematics proficiency and the five home environment variables. A MINITAB scatter plot matrix of the data is presented in Figure 11.6a and the correlation matrix is presented in Figure 11.6b. The scatter plot matrix also shows the lowess nonparametric regression fits, where $q = .9$ (the proportion defining a neighborhood) is used in the local fitting.

We see from the first row of the scatter plot matrix that average mathematics proficiency is related to each of the five explanatory variables and that there are three clear outliers. They are District of Columbia, Guam, and Virgin Islands, as noted earlier in this section. The lowess fits show positive relations for PARENTS, HOMELIB, and READING and a negative relation for ABSENCES. The lowess fit for TVWATCH is distorted because of the outliers. If these are ignored, the relation is negative. The correlation matrix shows fairly strong linear association with average mathematics proficiency for all explanatory variables except ABSENCES, where the degree of linear association is moderate.

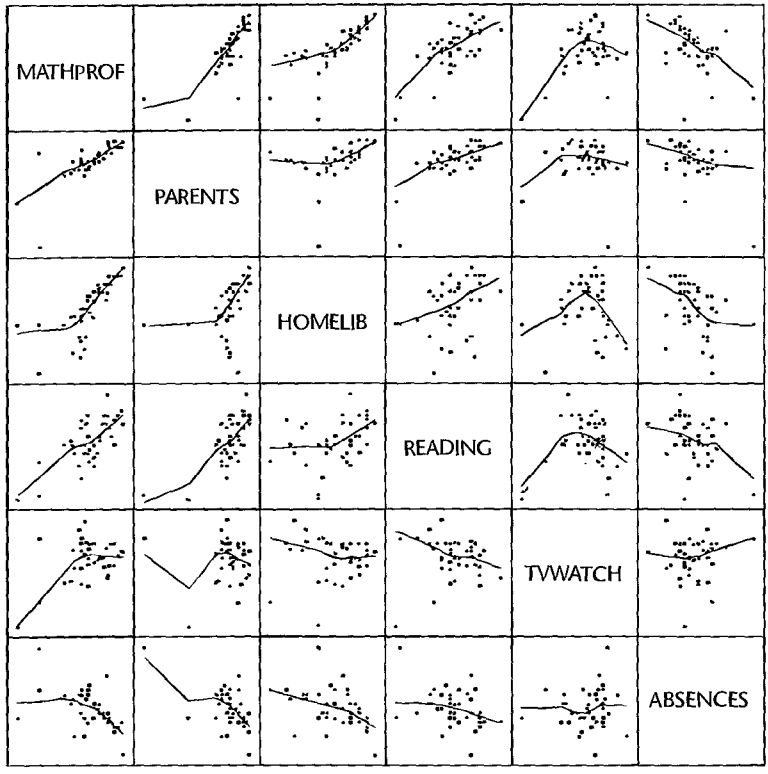
The relationships with mathematics proficiency found in Figure 11.6a must be interpreted with caution. We see from the remainder of the scatter plot matrix and from the correlation matrix in Figure 11.6b that the explanatory variables are correlated with each other, some fairly strongly. Also, some of the explanatory variables are correlated with other important variables not considered in this study. For example, the percentage of students with both parents at home is related to family income.

For simplicity, we consider only first-order terms in this example. An initial fit of the first-order model to the data using ordinary least squares yields the following estimated regression function:

$$\hat{Y} = 155.03 + .3911X_1 + .8639X_2 + .3616X_3 - .8467X_4 + .1923X_5 \qquad \textbf{(11.52)}$$

FIGURE 11.6
Scatter Plot Matrix with Lowess Smooths, and Correlation Matrix—Mathematics Proficiency Example.

(a) SYGRAPH Scatter Plot Matrix



(b) Correlation Matrix

	MATHPROF	PARENTS	HOMELIB	READING	TVWATCH
PARENTS	0.741				
HOMELIB	0.745	0.395			
READING	0.717	0.693	0.377		
TVWATCH	−0.873	−0.831	−0.594	−0.792	
ABSENCES	−0.480	−0.565	−0.443	−0.357	0.512

The signs of the regression coefficients, except for b_5, are in the expected directions. The coefficient of multiple determination for this fitted model is $R^2 = .86$, suggesting that the explanatory variables are strongly related to average mathematics proficiency.

Table 11.6 presents some diagnostics for the fitted model in (11.52): leverage h_{ii}, studentized deleted residual t_i, and Cook's distance D_i. We see that the District of Columbia, Guam, Texas, and Virgin Islands have leverage values equal to or exceeding $2p/n = 12/40 = .30$

TABLE 11.6 Diagnostics for First-Order Model with All Five Explanatory Variables—Mathematics Proficiency Example.

i	State	h_{ii}	t_i	D_i
1	Alabama	.16	−.05	.00
2	Arizona	.19	.40	.01
3	Arkansas	.16	1.41	.06
4	California	.29	.10	.00
...	...	...	...	...
8	D.C.	.69	1.41	.72
...	...	...	...	...
11	Guam	.34	−2.83	.57
...	...	...	...	...
35	Texas	.30	2.25	.33
36	Virgin_Islands	.32	−5.21	1.21
37	Virginia	.06	.90	.01
38	West_Virginia	.13	−.91	.02
39	Wisconsin	.08	.39	.00
40	Wyoming	.08	−.91	.01

We also see that the Virgin Islands is outlying with respect to its Y value; the absolute value of its studentized deleted residual $t_{36} = -5.21$ exceeds the Bonferroni critical value at $\alpha = .05$ of $t(1 - \alpha/2n; n - p - 1) = t(.99938; 33) = 3.53$. Of these outlying cases, the Virgin Islands is clearly influential according to Cook's distance measure, and District of Columbia and Guam are somewhat influential; the 50th percentile of the $F(6, 34)$ distribution is .91, and the 25th percentile is .57.

Residual plots against each of the explanatory variables and against $\hat{Y}$ (not shown here) presented no strong indication of nonconstancy of the error variance for the states aside from the outliers. Since the explanatory variables are correlated among themselves, the question arises whether a simpler model can be obtained with almost as much descriptive ability as the model containing all five explanatory variables. Figure 11.7 presents the MINITAB best subsets regression output, showing the two models with highest R^2 for each number of X variables. We see that the two best models for three variables ($p = 4$ parameters) contain relatively little bias according to the C_p criterion and have R^2 values almost as high as the model with all five variables.

We explore now one of these two models, the one containing HOMELIB, READING, and TVWATCH. In view of the outlying and influential cases, we employ IRLS robust regression with the Huber weight function (11.44). We find that after eight iterations, the weights change very little, so the iteration process is ended with the eighth iteration. The final robust fitted regression function is:

$$\hat{Y} = 207.83 + .7942X_2 + .1637X_3 - 1.1695X_4 \qquad \textbf{(11.53)}$$

The signs of the regression coefficients agree with expectations. For comparison, the regression function fitted by ordinary least squares is:

$$\hat{Y} = 199.61 + .7804X_2 + .4012X_3 - 1.1565X_4 \qquad \textbf{(11.54)}$$

FIGURE 11.7 MINITAB Best Subsets Regression—Mathematics Proficiency Example.

Best Subsets Regression of MATHPROF

Vars	R-sq	Adj. R-sq	C-p	S	PARENTED	HOMEENRG	RACE	TEACHEXP	ASSESSMENT
1	76.3	75.7	22.0	6.5079				X	
1	55.5	54.3	72.8	8.9157		X			
2	84.2	83.4	4.6	5.3810		X		X	
2	79.2	78.1	16.8	6.1743	X	X			
3	85.1	83.9	4.4	5.2939		X	X	X	
3	85.1	83.8	4.5	5.3062	X	X		X	
4	85.9	84.3	4.5	5.2327	X	X	X	X	
4	85.4	83.7	5.8	5.3285	X	X		X	X
5	86.1	84.1	6.0	5.2680	X	X	X	X	X

Notice that the robust regression led to a deemphasis of X_3 (READING), with the other regression coefficients remaining almost the same.

To obtain an indication of how well the robust regression model (11.53) describes the relation between average mathematics proficiency of eighth-grade students and the three home environment variables, we have ranked the 40 states according to their average mathematics proficiency score and according to their corresponding fitted value. The Spearman rank correlation coefficient (2.97), is .945. This indicates a fairly good ability of the three explanatory variables to distinguish between states whose average mathematics proficiency is very high or very low.

The analysis of the mathematics proficiency data set in Table 11.4 presented here is by no means exhaustive. We have not analyzed higher-order effects, nor have we explored other subsets that might be reasonable to use. We have not recognized that the precision of the state data varies because the data are based on samples of different sizes, nor have we considered other explanatory variables that are related to mathematics proficiency, such as parents' education and family income. Furthermore, we have analyzed state averages, which may obscure important insights into relations between the variables at the family level.

Comments

1. Robust regression requires knowledge of the regression function. When the appropriate regression function is not clear, nonparametric regression may be useful. Nonparametric regression is discussed in Section 11.4.

2. Robust regression can be employed to identify outliers in situations where there are multiple outliers whose presence is masked with diagnostic measures that delete one case at a time. Cases whose final weights are relatively small are outlying.

3. As illustrated by the mathematics proficiency example, robust regression is often useful for confirming the reasonableness of ordinary least squares results. When robust regression yields similar results to ordinary least squares (for example, the residuals are similar), one obtains some reassurance that ordinary least squares is not unduly influenced by outlying cases.

4. A limitation of robust regression is that the evaluation of the precision of the estimated regression coefficients is more complex than for ordinary least squares. Some large-sample results have been obtained (see, for example, Reference 11.5), but they may not perform well in the presence of outliers. Bootstrapping (to be discussed in Section 11.5) may also be used for evaluating the precision of robust regression results.

5. When the Huber, bisquare, and other weight functions are based on the scaled residuals in (11.47), they primarily reduce the influence of cases that are outlying with respect to their Y values. To make the robust regression fit more sensitive to cases that are outlying with respect to their X values, studentized residuals in (10.20) or studentized deleted residuals in (10.24) may be used instead of the scaled residuals in (11.47). Again, $\sqrt{MSE}$ may be replaced by MAD in (11.46) for better resistance and robustness when calculating the studentized or studentized deleted residuals.

In addition, the weights w_i obtained from the weight function may be modified to reduce directly the influence of cases with large X leverage. One suggestion is to multiply the weight function weight w_i by $\sqrt{1 - h_{ii}}$, where h_{ii} is the leverage value of the ith case defined in (10.18).

Methods that reduce the influence of cases that are outlying with respect to their X values are called *bounded influence regression* methods. ■

11.4 Nonparametric Regression: Lowess Method and Regression Trees

We considered nonparametric regression in Chapter 3 when there is one predictor variable in the regression model. We noted there that nonparametric regression fits are useful for exploring the nature of the response function, to confirm the nature of a particular response function that has been fitted to the data, and to obtain estimates of mean responses without specifying the nature of the response function.

Nonparametric regression can be extended to multiple regression when there are two or more predictor variables. Additional complexities are encountered, however, when making this extension. With more than two predictor variables, it is not possible to show the fitted response surface graphically, so one cannot see its appearance. Unlike parametric regression, no analytic expression for the response surface is provided by nonparametric regression. Also, as the number of predictor variables increases, there may be fewer and fewer cases in a neighborhood, leading to erratic smoothing. This latter problem is less serious when the predictor variables are highly correlated and interest in the response surface is confined to the region of the X observations.

Numerous procedures have been developed for fitting a response surface when there are two or more predictor variables without specifying the nature of the response function. Reference 11.9 discusses a number of these procedures. These include locally weighted regressions (Ref. 11.10), regression trees (Ref. 11.11), projection pursuit (Ref. 11.12), and smoothing splines (Ref. 11.13). We discuss the lowess method and regression trees in this section. We first extend the lowess method to multiple regression. In doing so, we will be able to describe it in far greater detail because we have established the necessary foundation of weighted least squares in Section 11.1.

Lowess Method

We described the lowess method briefly in Chapter 3 for regression with one predictor variable. The lowess method for multiple regression, developed by Cleveland and Devlin

(Ref. 11.10), assumes that the predictor variables have already been selected, that the response function is smooth, and that appropriate transformations have been made or other remedial steps taken so that the error terms are approximately normally distributed with constant variance. For any combination of X levels, the lowess method fits either a first-order model or a second-order model based on cases in the neighborhood, with more distant cases in the neighborhood receiving smaller weights. We shall explain the lowess method for the case of two predictor variables when we wish to obtain the fitted value at (X_{h1}, X_{h2}).

Distance Measure. We need a distance measure showing how far each case is from (X_{h1}, X_{h2}). Usually, a Euclidean distance measure is employed. For the ith case, this measure is denoted by d_i and is defined:

$$d_i = [(X_{i1} - X_{h1})^2 + (X_{i2} - X_{h2})^2]^{1/2} \tag{11.55}$$

When the predictor variables are measured on different scales, each should be scaled by dividing it by its standard deviation. The median absolute deviation estimator in (11.46) can be used in place of the standard deviation if outliers are present.

Weight Function. The neighborhood about the point (X_{h1}, X_{h2}) is defined in terms of the proportion q of cases that are nearest to the point. Let d_q denote the Euclidean distance of the furthest case in the neighborhood. The weight function used in the lowess method is the tricube weight function, which is defined as follows:

$$w_i = \begin{cases} [1 - (d_i/d_q)^3]^3 & d_i < d_q \\ 0 & d_i \geq d_q \end{cases} \tag{11.56}$$

Thus, cases outside the neighborhood receive weight zero and cases within the neighborhood receive weights between 0 and 1, the weight decreasing with greater distance. In this way, the mean response at (X_{h1}, X_{h2}) is estimated locally.

The choice of the proportion q defining the neighborhood requires a balancing of two opposing tendencies. The larger is q, the smoother will be the fit but at the same time the greater may be the bias in the fitted value. A choice of q between .4 and .6 may often be appropriate.

Local Fitting. Given the weights for the n cases based on (11.55) and (11.56), weighted least squares is then used to fit either the first-order model (6.1) or the second-order model (6.16). The second-order model is helpful when the response surface has substantial curvature; moderate curvilinearities can be detected by using the first-order model. After the regression model is fitted by weighted least squares, the fitted value $\hat{Y}_h$ at (X_{h1}, X_{h2}) then serves as the nonparametric estimate of the mean response at these X levels. By recalculating the weights for different (X_{h1}, X_{h2}) levels, fitting the response function repeatedly, and each time obtaining the fitted value $\hat{Y}_h$, we obtain information about the response surface without making any assumptions about the nature of the response function.

Example

We shall fit a nonparametric regression function for the life insurance example in Chapter 10. A portion of the data for a second group of 18 managers is given in Table 11.7, columns 1–3. The relation between amount of life insurance carried (Y) and income (X_1) and risk aversion (X_2) is to be investigated, the data pertaining to managers in the 30–39 age group.

TABLE 11.7 Lowess Calculations for Nonparametric Regression Fit at $X_{h1} = 30$, $X_{h2} = 3$—Life Insurance Example.

i	(1) X_{i1}	(2) X_{i2}	(3) Y_i	(4) d_i	(5) w_i
1	66.290	7	240	3.013	0
2	40.964	5	73	1.143	.300
3	72.996	10	311	4.212	0
...	...	...	...	...	...
16	79.380	1	316	3.461	0
17	52.766	8	154	2.663	0
18	55.916	6	164	2.188	0

The local fitting will be done using the first-order model in (6.1) because the number of available cases is not too large. For the same reason, the proportion of cases defining the local neighborhoods is set at $q = .5$; in other words, each local neighborhood is to consist of half of the cases.

The exploration of the response surface begins at $X_{h1} = 30$, $X_{h2} = 3$. To obtain a locally fitted value at $X_{h1} = 30$, $X_{h2} = 3$, we need to obtain the Euclidean distances of each case from this point. We shall use the sample standard deviations of the two predictor variables to standardize the variables in obtaining the Euclidean distance since the two variables are measured on different scales. The sample standard deviations are $s_1 = 14.739$ and $s_2 = 2.3044$. For case 1, the Euclidean distance from $X_{h1} = 30$, $X_{h2} = 3$ is obtained as follows:

$$d_1 = \left[\left(\frac{66.290 - 30}{14.739} \right)^2 + \left(\frac{7 - 3}{2.3044} \right)^2 \right]^{1/2} = 3.013$$

The Euclidean distances are shown in Table 11.7, column 4. The Euclidean distance of the furthest case in the neighborhood of $X_{h1} = 30$, $X_{h2} = 3$ for $q = .5$ is for the ninth case when these are ordered according to their Euclidean distance. It is $d_q = 1.653$. Since $d_1 = 3.013 > 1.653$, the weight assigned for case 1 is $w_1 = 0$. For case 2, the Euclidean distance is $d_2 = 1.143$. Since this is less than 1.653, the weight for case 2 is:

$$w_2 = [1 - (1.143/1.653)^3]^3 = .300$$

The weights are shown in Table 11.7, column 5.

The fitted first-order regression function using these weights is:

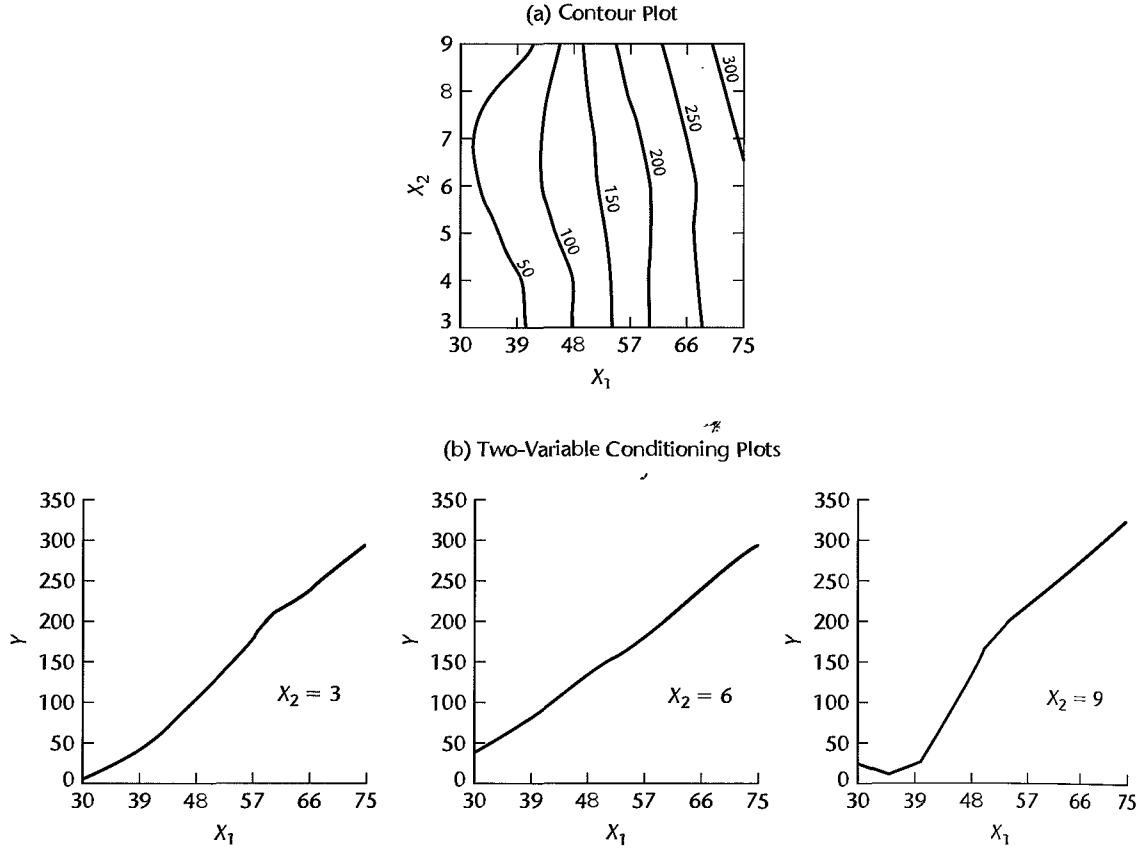
$$\hat{Y} = -134.076 + 3.571X_1 + 10.532X_2$$

The fitted value for $X_{h1} = 30$, $X_{h2} = 3$ therefore is:

$$\hat{Y}_h = -134.076 + 3.571(30) + 10.532(3) = 4.65$$

In the same fashion, locally fitted values at other values of X_{h1} and X_{h2} are calculated. Figure 11.8a contains a contour plot of the fitted response surface. The surface clearly ascends as X_1 increases, but the effect of X_2 is more difficult to see from the contour plot. The effect of X_2 can be seen more easily by the conditional effects plots of Y against X_1 at low, middle, and high levels of X_2 in Figure 11.8b. The conditional effects plots in Figure 11.8b are also called *two-variable conditioning plots*. Note that the expected amount of life insurance carried increases with income (X_1) at all levels of risk aversion (X_2). The

FIGURE 11.8 Contour and Conditioning Plots for Lowess Nonparametric Regression—Life Insurance Example.



response functions for $X_2 = 3$ and $X_2 = 6$ appear to be approximately linear. The dip in the left part of the response function for $X_2 = 9$ may be the result of an interaction or of noisy data and inadequate smoothing. Note also from Figure 11.8b that the expected amount of life insurance carried at the higher income levels increases as the risk aversion becomes very high.

Comments

1. The fitted nonparametric response surface can be used, just as for simple regression, for examining the appropriateness of a fitted parametric regression model. If the fitted nonparametric response surface falls within the confidence band in (6.60) for the parametric regression function, the nonparametric fit supports the appropriateness of the parametric regression function.

2. Reference 11.10 discusses a procedure to assist in choosing the proportion q for defining a local neighborhood. It also describes how the precision of any fitted value $\hat{Y}_h$ obtained with lowess nonparametric multiple regression can be approximated.

3. The assumptions of normality and constant variance of the error terms required by the lowess nonparametric procedure can be checked in the usual fashion. The residuals are obtained by fitting the lowess nonparametric regression function for each case and calculating $e_i = Y_i - \hat{Y}_i$ as usual. These residuals will not have the least squares property of summing to zero, but can be examined for normality and constancy of variance. The residuals can also serve to identify outliers that might not be disclosed by standard diagnostic procedures.

4. A discussion of some of the advantages of the lowess smoothing procedure is presented in Reference 11.14. ■

Regression Trees

Regression trees are a very powerful, yet conceptually simple, method of nonparametric regression. For the case of a single predictor, the range of the predictor is partitioned into segments and within each segment the estimated regression fit is given by the mean of the responses in the segment. For two or more predictors, the X space is partitioned into rectangular regions, and again, the estimated regression surface is given by the mean of the responses in each rectangle. Regression trees have become a popular alternative to multiple regression for exploratory studies, especially for extremely large data sets. Along with neural networks (see Chapter 13), regression trees are one of the standard methods used in the emerging field of data mining. Regression trees are easy to calculate, require virtually no assumptions, and are simple to interpret.

One Predictor Tree: Steroid Level Example. Figure 1.3 on page 5 presents data on age and level of a steroid in plasma for 27 healthy females between 8 and 25 years of age. The data are shown in the first two columns of Table 11.8. A regression tree based on five regions is obtained by partitioning the range of X (age) into five segments or regions, and using the sample average of the Y responses in each region for the fitted regression surface. We will use R_{51} through R_{55} to denote the regions of a 5-region tree, and $\bar{Y}_{R_{51}}$ through $\bar{Y}_{R_{55}}$ to denote the corresponding sample averages. These values are shown for the steroid level example in columns 4–6 of Table 11.8. The fitted regression tree is shown in Figure 11.9a. Note that the regression tree is a step function that steps up rapidly for girls between the ages of 8 and 14, after which point steroid level is roughly constant.

A plot of residuals versus fitted values is shown in Figure 11.9b. Note that the variance of the residuals in each region seems roughly constant, an indication that further splitting may be unnecessary. We discuss the determination of appropriate tree size below.

TABLE 11.8 Data Set and 5-Region Regression Tree Fit—Steroid Level Example.

(1) Case i	(2) Steroid Level Y_i	(3) Age X_i	(4) Region Number k	(5) Region R_{5k}	(6) Fitted Value $\bar{Y}_{R_{5k}}$
1	27.1	23	1	$8 \le X < 9$	3.550
2	22.1	19	2	$9 \le X < 10$	8.133
3	21.9	25	3	$10 \le X < 13$	13.675
...	...	...	4	$13 \le X < 14$	16.950
25	12.8	13	5	$14 \le X < 25$	22.200
26	20.8	14			
27	20.6	18			

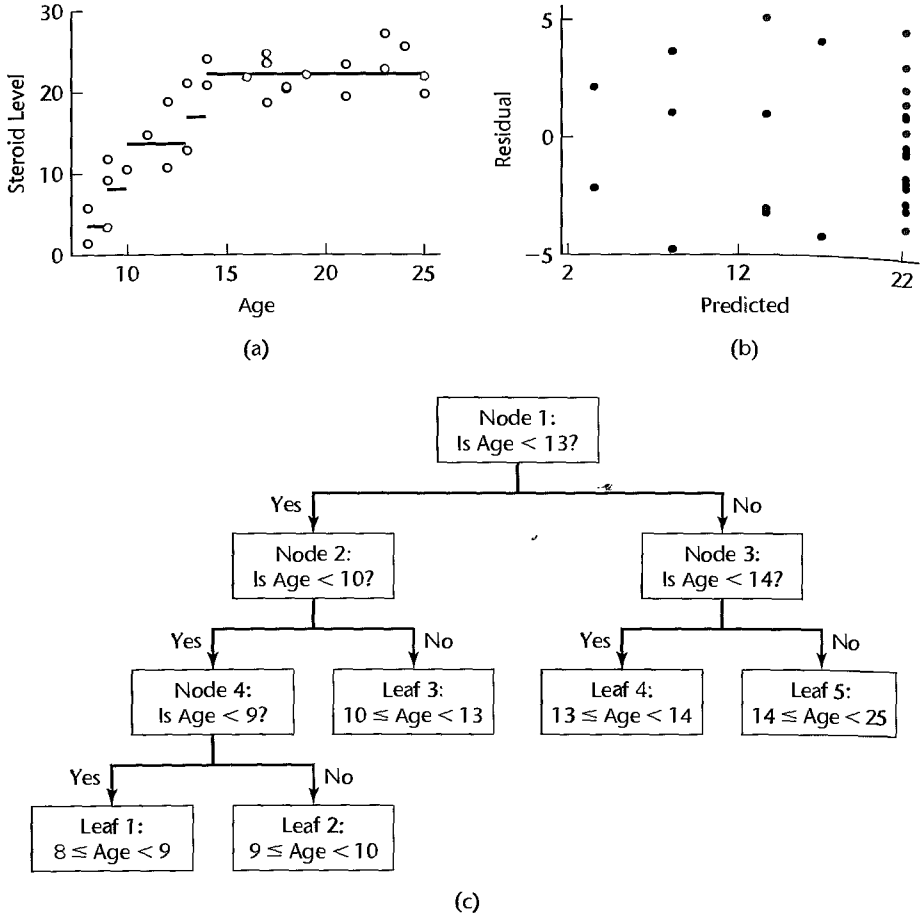


FIGURE 11.9
Fitted Regression Tree, Residual Plot, and Regression Tree Diagram—Steroid Level Example.

Determining the predicted value for a given X_h is accomplished with the help of a tree diagram, such as the one shown in Figure 11.9c. Suppose we wish to determine the predicted value at $X_h = 12.5$. Starting at node 1—the *root node*—we ask, "Is Age < 13?" Since 12.5 < 13, we follow the left branch to node 2 where we ask, "Is Age < 10?" Since Age is not less than 10, we branch right to the terminal node labeled *Leaf* 3, where we find from Table 11.8 that $\bar{Y}_{R_{53}} = 13.675$. Tree diagrams such as that shown in Figure 11.9c are particularly helpful when more than a single predictor is present.

Growing a Regression Tree. To find a "best" regression tree, it is necessary to specify the number of regions, r, and the boundaries, or *split points,* between the regions. The process of determining a best value for r and the associated split points is referred to as *growing the tree.*

First consider the case of a single predictor, and assume that the range of X is to be divided into $r = 2$ regions, R_{21} and R_{22}. We need to find the split point X_s that optimally divides the data into two sets. The best point is chosen to minimize the error sum of squares

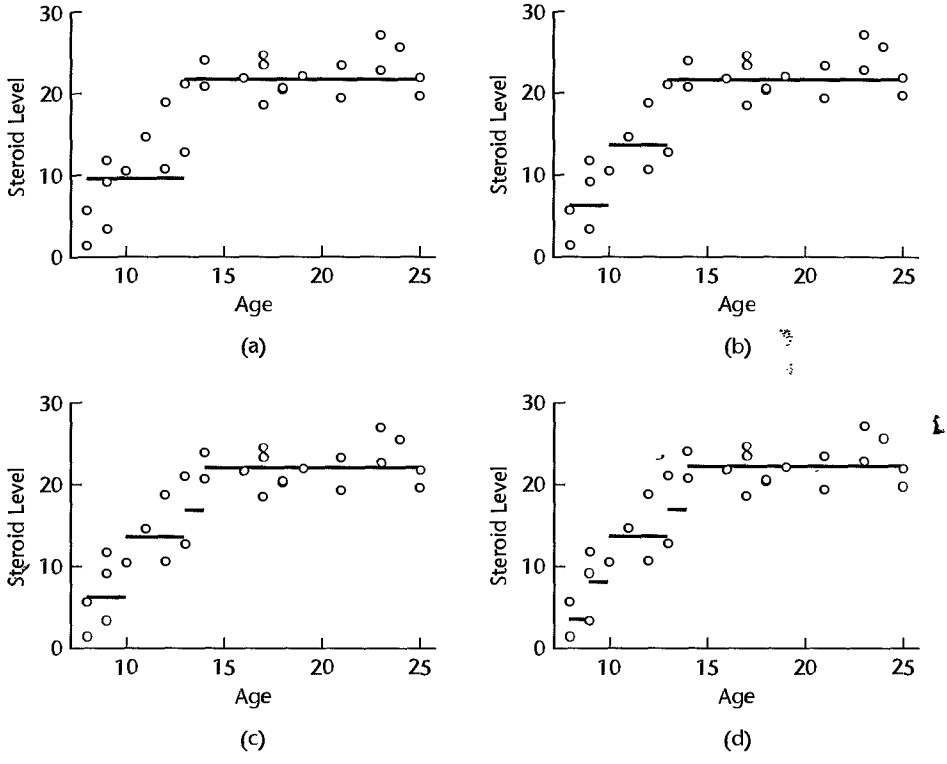


FIGURE 11.10 Growing the Regression Tree—Steroid Level Example.

for the resulting regression tree:

$$SSE = SSE(R_{21}) + SSE(R_{22})$$

where $SSE(R_{rj})$ is the sum of squared residuals in region R_{rj}:

$$SSE(R_{rj}) = \sum (Y_i - \bar{Y}_{R_{jk}})^2$$

For the steroid level data, the best split point is shown in Figure 11.10a to be $X_s = 13.0$. For this tree, we have:

$$R_{21} = \{X | X < 13\}$$
$$R_{22} = \{X | X \geq 13\}$$

for which we obtain:

$$SSE = SSE(R_{21}) + SSE(R_{22}) = 238.55 + 167.79 = 406.35$$

From (2.72), the coefficient of determination for the regression tree is:

$$R^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{406.35}{1284.8} = .684$$

Also, $MSE = SSE/(n - r) = 406.35/(27 - 2) = 16.254$.

At this point, there are two regions, and growing the tree further will require the identification of a third region. We have two choices: (1) we can work sequentially and split one of the two existing regions, or (2) start from scratch and identify simultaneously two entirely new split points that globally minimize the resulting *SSE* criterion. The second approach will always lead to a criterion value that is at least as good as the first; however, as the tree grows, so do the computational demands associated this approach (particularly if there is more than one predictor). For this reason, regression trees are generally grown sequentially, according to the following rule: If the tree currently is based on r regions, we determine the best split point for each of the regions, and then split the region that leads to the greatest decrease in *SSE*.

For the steroid-level example, the next step involves splitting R_{21} at $X_s = 10$, resulting in three regions:

$$R_{21} = \{X | X < 10\}$$
$$R_{32} = \{X | 10 \leq X < 13\}$$
$$R_{33} = \{X | X \geq 13\}$$

A plot of this tree is shown in Figure 11.10b. Continuing this process, we next split R_{33} at $X_s = 14$, and a final split occurs at $X_s = 19$. The 4-region and 5-region regression trees are shown in Figures 11.10c and 11.10d.

For two or more predictors, the procedure is the same, except that in addition to determining the best region and split point, we must also determine the best predictor upon which to base the split. The rule is as follows: assuming the tree is based currently on r rectangular regions, we determine the best split point for each of the r regions for each of the $p - 1$ predictors, and then implement a new split based on the region and predictor that leads to the largest decrease in *SSE*. Note that we are choosing the best predictor-and-split-point combination from $r(p - 1)$ possibilities.

This process is illustrated for two predictors in Figure 11.11. We first consider splitting the rectangular X space either on the basis of X_1 or X_2. We find the best split points X_{1s} and X_{2s} for X_1 and X_2 respectively, and then we base our next partition on the split point that leads to the greatest decrease in *SSE*. According to Figure 11.11a, the first split is based on X_1, resulting in two rectangular regions R_{21} and R_{22}. For each of these two regions, we determine the best predictor upon which to split and the associated split point, and choose the combination that leads to the largest decrease in *SSE*. Figure 11.11b indicates that region R_{22} was partitioned in this step on the basis of X_2. Finally, in the third split, region R_{31} is partitioned on the basis of X_1, resulting in a 4-region tree, as shown in Figure 11.11c.

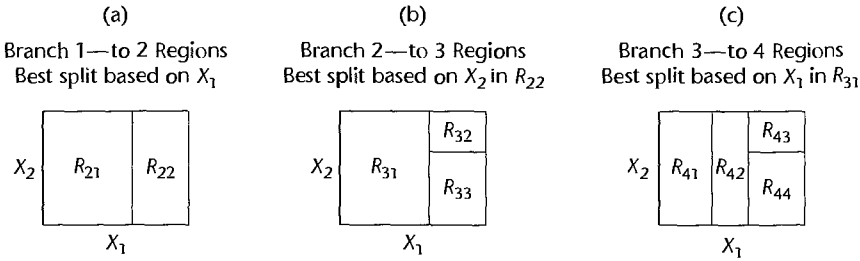


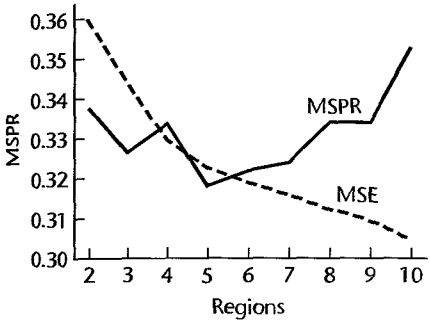
FIGURE 11.11 Regression Tree Growth—Two-Predictor Example.

Determining the Number of Regions, r. If the tree-growing process is allowed to continue indefinitely, there will eventually be n regions, with each region containing a single observation, and further partitioning will be impossible. A "best" number of regions will generally fall between 1 and n, and is usually chosen through validation studies. For example, for each split we determine, in addition to *SSE*, the mean square for prediction error *MSPR* for data in a hold-out or validation sample. We then choose the tree that minimizes *MSPR*.

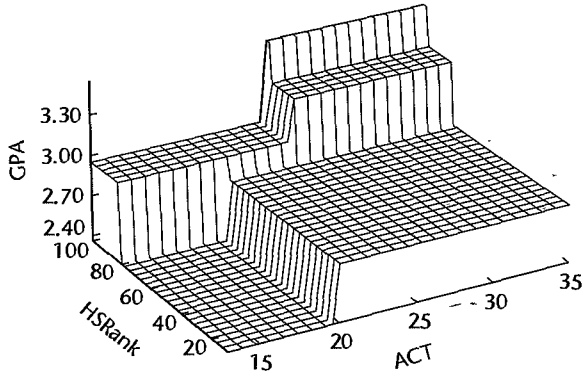
Example

We illustrate the use of regression trees with the University admissions data set in Appendix C.4. We fit GPA at the end of freshman year (Y) as a function of ACT entrance test score (X_1) and high school rank (X_2). The data consist of 705 cases, and a random sample of $n^* = 353$ records was selected for the validation set. Figure 11.12a provides a plot of *MSPR* versus the number of regions, or terminal nodes. The plot shows that the ability to predict improves as nodes are added until $r = 5$, for which $MSPR = .318$ ($\hat{MSE}$ for this

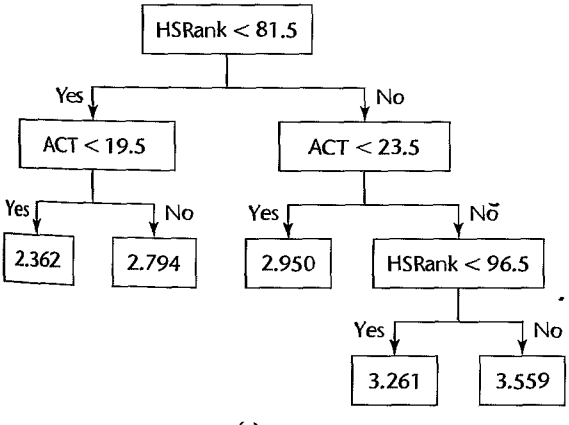
FIGURE 11.12 S-Plus Regression Tree Results—University Admissions Example.



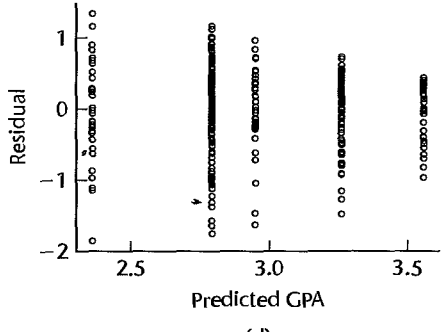
(a)



(b)



(c)



(d)

model is .322). For $r > 5$, the ability to predict responses in the validation set deteriorates as the number of regions increases. A plot of *MSE* is also included, and as expected, *MSE* decreases monotonically with the size of the tree. The fitted regression tree surface is shown in Figure 11.12b and the corresponding tree diagram is shown in Figure 11.12c.

A plot of residuals versus predicted values is shown for this tree in Figure 11.12d. Note that the variance of the residuals appears to be somewhat constant, and indication that further partitions may not be required.

It is instructive to compare qualitatively the fit of the regression tree to the fit obtained using standard regression methods. Using a full second-order model leads to the equation:

$$\hat{Y} = 1.77 - .0223X_1 + .0780X_2 + .000187X_1^2 - .00133X_2^2 + .000342X_1X_2$$

MSPR for the second-order regression model is .296, which is slightly better than the value obtained by the regression tree (.318). Interestingly the *MSE* value obtained by the second-order regression model (.333) is about the same as that obtained by the regression tree (.322).

In summary, the regression tree surface suggests as expected that college GPA increases with both ACT score and high school rank. Overall, high school rank seems to have a slightly more pronounced effect than ACT score. For this tree, R^2 is .256 for the training data set, and .157 for the validation data set. We conclude that GPA following freshman year is related to high school rank and ACT score, but the fraction of variation in GPA explained by these predictors is quite small.

Comments

1. The number of regions r is sometimes chosen by minimizing the *cost complexity criterion:*

$$C_\lambda(r) = \sum_{k=1}^{r} SSE(R_{rk}) + \lambda r$$

The cost complexity criterion has two components: the sum of squared residuals plus a penalty, λr, for the number of regions r employed. The tuning parameter $\lambda \geq 0$ determines the balance between the size of the tree (complexity) and the goodness of fit. Larger values of λ lead to smaller trees. Note that this criterion is a form of *penalized least squares,* which, as we commented in Section 11.2, can be used to obtain ridge regression estimates. Penalized least squares is also used in connection with neural networks as described in Section 13.6. A "best" value for λ is generally chosen through validation studies.

2. Regression trees are often used when the response Y is qualitative. In such cases, predicting a response at X_h is equivalent to determining to which response category X_h belongs. This is a classification problem, and the resulting tree is referred to as a classification tree. Details are provided in References 11.11 and 11.15. ■

11.5 Remedial Measures for Evaluating Precision in Nonstandard Situations—Bootstrapping

For standard fitted regression models, methods described in earlier chapters are available for evaluating the precision of estimated regression coefficients, fitted values, and predictions of new observations. However, in many nonstandard situations, such as when nonconstant error

variances are estimated by iteratively reweighted least squares or when robust regression estimation is used, standard methods for evaluating the precision may not be available or may only be approximately applicable when the sample size is large. Bootstrapping was developed by Efron (Ref. 11.16) to provide estimates of the precision of sample estimates for these complex cases. A number of bootstrap methods have now been developed. The bootstrap method that we shall explain is simple in principle and nonparametric in nature. Like all bootstrap methods, it requires extensive computer calculations.

General Procedure

We shall explain the bootstrap method in terms of evaluating the precision of an estimated regression coefficient. The explanation applies identically to any other estimate, such as a fitted value. Suppose that we have fitted a regression model (simple or multiple) by some procedure and obtained the estimated regression coefficient b_1; we now wish to evaluate the precision of this estimate by the bootstrap method. In essence, the bootstrap method calls for the selection from the observed sample data of a random sample of size n with replacement. Sampling with replacement implies that the bootstrap sample may contain some duplicate data from the original sample and omit some other data in the original sample. Next, the bootstrap method calculates the estimated regression coefficient from the bootstrap sample, using the same fitting procedure as employed for the original fitting. This leads to the first bootstrap estimate b_1^*. This process is repeated a large number of times; each time a bootstrap sample of size n is selected with replacement from the original sample and the estimated regression coefficient is obtained for the bootstrap sample. The estimated standard deviation of all of the bootstrap estimates b_1^*, denoted by $s^*\{b_1^*\}$, is an estimate of the variability of the sampling distribution of b_1 and therefore is a measure of the precision of b_1.

Bootstrap Sampling

Bootstrap sampling for regression can be done in two basic ways. When the regression function being fitted is a good model for the data, the error terms have constant variance, and the predictor variable(s) can be regarded as fixed, *fixed X sampling* is appropriate. Here the residuals e_i from the original fitting are regarded as the sample data to be sampled with replacement. After a bootstrap sample of the residuals of size n has been obtained, denoted by $e_1^*, \ldots, e_n^*$, the bootstrap sample residuals are added to the fitted values from the original fitting to obtain new bootstrap Y values, denoted by $Y_1^*, \ldots, Y_n^*$:

$$Y_i^* = \hat{Y}_i + e_i^* \tag{11.57}$$

These bootstrap Y^* values are then regressed on the original X variable(s) by the same procedure used initially to obtain the bootstrap estimate b_1^*.

When there is some doubt about the adequacy of the regression function being fitted, the error variances are not constant, and/or the predictor variables cannot be regarded as fixed, *random X sampling* is appropriate. For simple regression, the pairs of X and Y data in the original sample are considered to be the data to be sampled with replacement. Thus, this second procedure samples cases with replacement n times, yielding a bootstrap sample of n pairs of (X^*, Y^*) values. This bootstrap sample is then used for obtaining the bootstrap estimate b_1^*, as with fixed X sampling.

The number of bootstrap samples to be selected for evaluating the precision of an estimate depends on the special circumstances of each application. Sometimes, as few

as 50 bootstrap samples are sufficient. Often, 200–500 bootstrap samples are adequate. One can observe the variability of the bootstrap estimates by calculating $s^*\{b_1^*\}$ as the number of bootstrap samples is increased. When $s^*\{b_1^*\}$ stabilizes fairly reasonably, bootstrapping can be terminated.

Bootstrap Confidence Intervals

Bootstrapping can also be used to arrive at approximate confidence intervals. Much research is ongoing on different procedures for obtaining bootstrap confidence intervals (see, for example, References 11.17 and 11.18). A relatively simple procedure for setting up a $1-\alpha$ confidence interval is the *reflection method*. This procedure often produces a reasonable approximation, but not always. The reflection method confidence interval for β_1 is based on the $(\alpha/2)100$ and $(1-\alpha/2)100$ percentiles of the bootstrap distribution of b_1^*. These percentiles are denoted by $b_1^*(\alpha/2)$ and $b_1^*(1-\alpha/2)$, respectively. The distances of these percentiles from b_1, the estimate of β_1 from the original sample, are denoted by d_1 and d_2:

$$d_1 = b_1 - b_1^*(\alpha/2) \tag{11.58a}$$

$$d_2 = b_1^*(1-\alpha/2) - b_1 \tag{11.58b}$$

The approximate $1-\alpha$ confidence interval for β_1 then is:

$$b_1 - d_2 \le \beta_1 \le b_1 + d_1 \tag{11.59}$$

Bootstrap confidence intervals by the reflection method require a larger number of bootstrap samples than do bootstrap estimates of precision because tail percentiles are required. About 500 bootstrap samples may be a reasonable minimum number for reflection bootstrap confidence intervals.

Examples

We illustrate the bootstrap method by two examples. In the first one, standard analytical methods are available and bootstrapping is used simply to show that it produces similar results. In the second example, the estimation procedure is complex, and bootstrapping provides a means for assessing the precision of the estimate.

Example 1—Toluca Company

We use the Toluca Company example of Table 1.1 to illustrate how the bootstrap method approximates standard analytical results. We found in Chapter 2 that the estimate of the slope β_1 is $b_1 = 3.5702$, that the estimated precision of this estimate is $s\{b_1\} = .3470$, and that the 95 percent confidence interval for β_1 is $2.85 \le \beta_1 \le 4.29$.

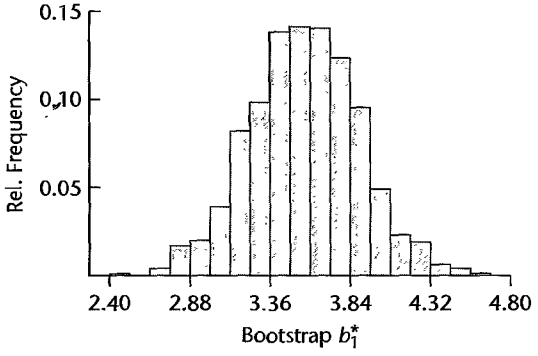
To evaluate the precision of the estimate $b_1 = 3.5702$ by the bootstrap method, we shall use fixed X sampling. Here, the simple linear regression function fits the data well, the error variance appears to be constant, and it is reasonable to consider a repetition of the study with the same lot sizes. A portion of the data on lot size (X) and work hours (Y) is repeated in Table 11.9, columns 1 and 2. The fitted values and residuals obtained from the original sample are repeated from Table 1.2 in columns 3 and 4. Column 5 of Table 11.9 shows the first bootstrap sample of n residuals e_i^*, selected from column 4 with replacement. Finally, column 6 shows the first bootstrap sample Y_i^* observations. For example, by (11.57), we obtain $Y_1^* = \hat{Y}_1 + e_1^* = 347.98 - 19.88 = 328.1$.

When the Y_i^* values in column 6 are regressed against the X values in column 1, based on simple linear regression model (2.1), we obtain $b_1^* = 3.7564$. In the same way, 999 other bootstrap samples were selected and b_1^* obtained for each. Figure 11.13 contains a histogram

TABLE 11.9 Bootstrapping with Fixed X Sampling—Toluca Company Example.

	(1)	(2)	(3)	(4)	(5)	(6)
	Original Sample				Bootstrap Sample 1	
i	X_i	Y_i	$\hat{Y}_i$	e_i	e_i^*	Y_i^*
1	80	399	347.98	51.02	−19.88	328.1
2	30	121	169.47	−48.47	10.72	180.2
3	50	221	240.88	−19.88	−6.68	234.2
...	...	...	...	...	...	...
23	40	244	205.17	38.83	4.02	209.2
24	80	342	347.98	−5.98	−45.17	302.8
25	70	323	312.28	10.72	51.02	363.3

FIGURE 11.13 Histogram of Bootstrap Estimates b_1^*—Toluca Company Example.



$b_1^*(.025) = 2.940$ $\quad s^*\{b_1^*\} = .3251$ $\quad b_1^*(.975) = 4.211$

of the 1,000 bootstrap b_1^* estimates. Note that this bootstrap sampling distribution is fairly symmetrical and appears to be close to a normal distribution. We also see in Figure 11.13 that the standard deviation of the 1,000 b_1^* estimates is $s^*\{b_1^*\} = .3251$, which is quite close to the analytical estimate $s\{b_1\} = .3470$.

To obtain an approximate 95 percent confidence interval for β_1 by the bootstrap reflection method, we note in Figure 11.13 that the 2.5th and 97.5th percentiles of the bootstrap sampling distribution are $b_1^*(.025) = 2.940$ and $b_1^*(.975) = 4.211$, respectively. Using (11.58), we obtain:

$$d_1 = 3.5702 - 2.940 = .630$$
$$d_2 = 4.211 - 3.5702 = .641$$

Finally, we use (11.59) to obtain the confidence limits $3.5702 + .630 = 4.20$ and $3.5702 - .641 = 2.93$ so that the approximate 95 percent confidence interval for β_1 is:

$$2.93 \leq \beta_1 \leq 4.20$$

Note that these limits are quite close to the confidence limits 2.85 and 4.29 obtained by analytical methods.

Example 2—Blood Pressure

For the blood pressure example in Table 11.1, the analyst used weighted least squares in order to recognize the unequal error variances and fitted a standard deviation function to estimate the unknown weights. The standard inference procedures employed by the analyst for estimating the precision of the estimated regression coefficient $b_{w1} = .59634$ and for obtaining a confidence interval for β_1 are therefore only approximate. To examine whether the approximation is good here, we shall evaluate the precision of the estimated regression coefficient in a way that recognizes the impreciseness of the weights by using bootstrapping. The X variable (age) probably should be regarded as random and the error variance varies with the level of X, so we shall use random X sampling. Table 11.10 repeats from Table 11.1 the original data for age (X) and diastolic blood pressure (Y) in columns 1 and 2. Columns 3 and 4 contain the (X_i^*, Y_i^*) observations for the first bootstrap sample selected with replacement from columns 1 and 2. When we now regress Y^* on X^* by ordinary least squares, we obtain the fitted regression function:

$$\hat{Y}^* = 50.384 + .7432X^*$$

The residuals for this fitted function are shown in column 5. When the absolute values of these residuals are regressed on X^*, the fitted standard deviation function obtained is:

$$\hat{s}^* = -5.409 + .32745X^*$$

The fitted values $\hat{s}_i^*$ are shown in column 6. Finally, the weights $w_i^* = 1/(\hat{s}_i^*)^2$ are shown in column 7. For example, $w_1^* = 1/(10.64)^2 = .0088$. Finally, Y^* is regressed on X^* by using the weights in column 7, to yield the bootstrap estimate $b_1^* = .838$.

This process was repeated 1,000 times. The histogram of the 1,000 bootstrap values b_1^* is shown in Figure 11.14 and appears to approximate a normal distribution. The standard deviation of the 1,000 bootstrap values is shown in Figure 11.14; it is $s^*\{b_1^*\} = .0825$. When we compare this precision with that obtained by the approximate use of (11.13), .0825 versus .07924, we see that recognition of the use of estimated weights has led here only to a small increase in the estimated standard deviation. Hence, the variability in b_{w1} associated with the use of estimated variances in the weights is not substantial and the standard inference procedures therefore provide a good approximation here.

A 95 percent bootstrap confidence interval for β_1 can be obtained from (11.59) by using the percentiles $b_1^*(.025) = .4375$ and $b_1^*(.975) = .7583$ shown in Figure 11.14. The

TABLE 11.10 Bootstrapping with Random *X* Sampling—Blood Pressure Example.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Original Sample		Bootstrap Sample 1				
i	X_i	Y_i	X_i^*	Y_i^*	e_i^*	$\hat{s}_i^*$	w_i^*
1	27	73	49	101	14.20	10.64	.0088
2	21	66	34	73	−2.65	5.72	.0305
3	22	63	49	101	14.20	10.64	.0088
...	...	...	...	...	...	...	...
52	52	100	46	89	4.43	9.65	.0107
53	58	80	27	73	2.55	3.43	.0850
54	57	109	40	70	−10.11	7.69	.0169

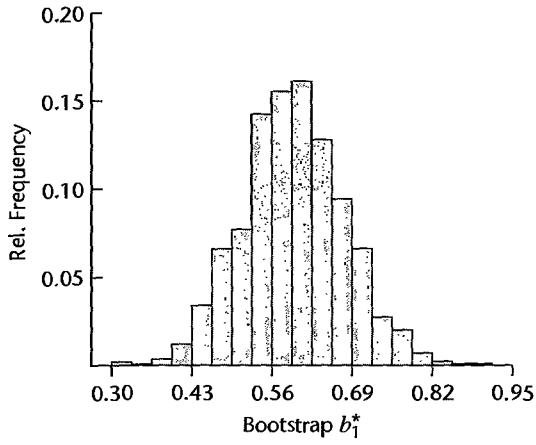


FIGURE 11.14 Histogram of Bootstrap Estimates b_1^*—Blood Pressure Example.

$b_1^*(.025) = .4375$ $\quad s^*\{b_1^*\} = .0825$ $\quad b_1^*(.975) = .7583$

approximate 95 percent confidence limits are [recall from (11.20) that $b_{w1} = .59634$]:

$$b_{w1} - d_2 = .59634 - (.7583 - .59634) = .4344$$
$$b_{w1} + d_1 = .59634 + (.59634 - .4375) = .7552$$

and the confidence interval for β_1 is:

$$.434 \leq \beta_1 \leq .755$$

Note that this confidence interval is almost the same as that obtained earlier by standard inference procedures ($.437 \leq \beta_1 \leq .755$). This again confirms that it is appropriate to use standard inference procedures here even though the weights were estimated.

Comment

The reason why d_1 is associated with the upper confidence limit in (11.59) and d_2 with the lower limit is that the upper $(1 - \alpha/2)100$ percentile in the sampling distribution of b_1 identifies the lower confidence limit for β_1, whereas the lower $(\alpha/2)100$ percentile identifies the upper confidence limit. To see this, consider the sampling distribution for b_1, for which we can state with probability $1 - \alpha$ that b_1 will fall between:

$$b_1(\alpha/2) \leq b_1 \leq b_1(1 - \alpha/2) \tag{11.60}$$

where $b_1(\alpha/2)$ and $b_1(1 - \alpha/2)$ denote the $(\alpha/2)100$ and $(1 - \alpha/2)100$ percentiles of the sampling distribution of b_1. We now express these percentiles in terms of distances from the mean of the sampling distribution, $E\{b_1\} = \beta_1$:

$$\begin{aligned} D_1 &= \beta_1 - b_1(\alpha/2) \\ D_2 &= b_1(1 - \alpha/2) - \beta_1 \end{aligned} \tag{11.61}$$

and obtain:

$$\begin{aligned} b_1(\alpha/2) &= \beta_1 - D_1 \\ b_1(1 - \alpha/2) &= \beta_1 + D_2 \end{aligned} \tag{11.62}$$

Substituting (11.62) into (11.60) and rearranging the inequalities so that β_1 is in the middle leads to the limits:

$$b_1 - D_2 \leq \beta_1 \leq b_1 + D_1$$

The confidence interval in (11.59) is obtained by replacing D_1 and D_2 by d_1 and d_2, which involves using the percentiles of the bootstrap sampling distribution as estimates of the corresponding percentiles of the sampling distribution of b_1 and using b_1 as the estimate of the mean β_1 of the sampling distribution. ■

11.6 Case Example—MNDOT Traffic Estimation

Traffic monitoring involves the collection of many types of data, such as traffic volume, traffic composition, vehicle speeds, and vehicle weights. These data provide information for highway planning, engineering design, and traffic control, as well as for legislative decisions concerning budget allocation, selection of state highway routes, and the setting of speed limits. One of the most important traffic monitoring variables is the average annual daily traffic (AADT) for a section of road or highway. AADT is defined as the average, over a year, of the number of vehicles that pass through a particular section of a road each day. Information on AADT is often collected by means of automatic traffic recorders (ATRs). Since it is not possible to install these recorders on all state road segments because of the expense involved, Cheng (Ref. 11.19) investigated the use of regression analysis for estimating AADT for road sections that are not monitored in the state of Minnesota.

The AADT Database

Seven potential predictors of traffic volume were chosen from the Minnesota Department of Transportation (MNDOT) road-log database, including type of road section, population density in the vicinity of road section, number of lanes in road section, and road section's width. Four of the seven variables were qualitative, requiring 19 indicator variables. Preliminary regression analysis indicated that the large number of levels of two of the qualitative variables was not helpful. Consequently, judgment and statistical information about marginal reductions in the error sum of squares were used to collapse the categories, so only 10 instead of 19 indicator variables remained in the AADT database.

The variables included in the initial analysis were as follows:

CTYPOP (X_1)—population of county in which road section is located (best proxy available for population density in immediate vicinity of road section)

LANES (X_2)—number of lanes in road section

WIDTH (X_3)—width of road section (in feet)

CONTROL (X_4)—two-category qualitative variable indicating whether or not there is control of access to road section (1 = access control; 2 = no access control)

CLASS (X_5, X_6, X_7)—four-category qualitative variable indicating road section function (1 = rural interstate; 2 = rural noninterstate; 3 = urban interstate, 4 = urban noninterstate)

TRUCK (X_8, X_9, X_{10}, X_{11})—five-category qualitative variable indicating availability status of road section to trucks (e.g., tonnage and time-of-year restrictions)

TABLE 11.11 Data—MNDOT Traffic Estimation Example.

Road Section i	AADT Y_i	County Population X_{i1}	Lanes X_{i2}	Width X_{i3}	Access Control Category X_{i4}	Function Class Category (X_{i5} to X_{i7})	Truck Route Category (X_{i8} to $X_{i,11}$)	Locale Category ($X_{i,12}$, $X_{i,13}$)
1	1,616	13,404	2	52	2	2	5	1
2	1,329	52,314	2	60	2	2	5	1
3	3,933	30,982	2	57	2	4	5	2
...	...	...	...	...	...	...	...	...
119	14,905	459,784	4	68	2	4	5	2
120	15,408	459,784	2	40	2	4	5	3
121	1,266	43,784	2	44	2	4	5	2

Source: C. Cheng, "Optimal Sampling for Traffic Volume Estimation," unpublished Ph.D. dissertation, University of Minnesota, Carlson School of Management, 1992.

LOCALE (X_{12}, X_{13})—three-category qualitative variable indicating type of locale (1 = rural; 2 = urban, population $\leq$ 50,000; 3 = urban, population > 50,000)

A portion of the data is shown in Table 11.11. Altogether, complete records for 121 ATRs were available. For conciseness, only the category is shown for a qualitative variable and not the coding of the indicator variables.

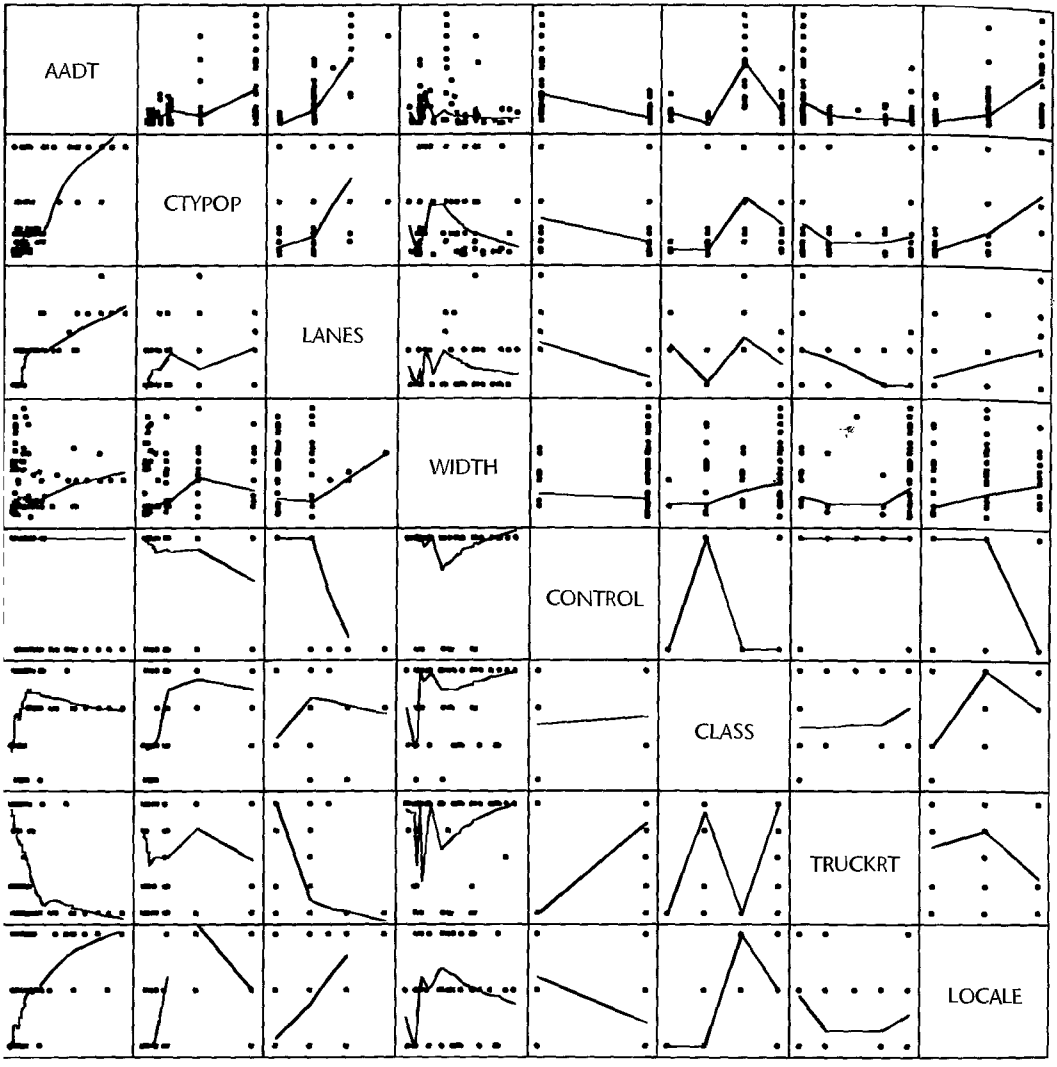
Model Development

A SYSTAT scatter plot matrix of the data set, with lowess fits added, is presented in Figure 11.15. We see from the first row of the matrix that several of the predictor variables are related to AADT. The lowess fits suggest a potentially curvilinear relationship between LANES and AADT. Although the lowess fits of AADT to the qualitative categories designated 1, 2, 3, etc., are meaningless, they do highlight the average traffic volume for each category. For example, the lowess fit of AADT to CLASS shows that average AADT for the third category of CLASS is higher than for the other three categories. The scatter plot matrix also suggests that the variability of AADT may be increasing with some predictor variables, for instance, with CTYPOP.

An initial regression fit of a first-order model with ordinary least squares, using all predictor variables, indicated that CTYPOP and LANES are important variables. Regression diagnostics for this initial fit suggested two potential problems. First, the residual plot against predicted values revealed that the error variance might not be constant. Also, the maximum variance inflation factor (10.41) was 24.55, suggesting a severe degree of multicollinearity. The maximum Cook's distance measure (10.33) was .2076, indicating that none of the individual cases is particularly influential. Since many of the variables appeared to be unimportant, we next considered the use of subset selection procedures to identify promising, initial models.

The SAS all-possible-regressions procedure, PROC RSQUARE, was used for subset selection. To reduce the volume of computation, CTYPOP and LANES were forced to be included. The SAS output is given in Figure 11.16. The left column indicates the number of X variables in the model, i.e., $p - 1$. The names of the qualitative variables identify the

FIGURE 11.15 SYSTAT Scatter Plot Matrix—MNDOT Traffic Estimation Example.



predictor variable and the category for which the indicator variable is coded 1. For example, CLASS1 refers to the first indicator variable for the predictor variable CLASS; i.e., it refers to X_5, which is coded 1 for category 1 (rural interstate). Two simple models look particularly promising. The three-variable model consisting of X_1 (CTYPOP), X_2 (LANES), and X_7 (CLASS = 3) stands out as the best three-variable model, with $R_p^2 = .805$ and $C_p = 5.23$. Since $p = 4$ for this model, the C_p statistic suggests that this model contains little bias. The best four-variable model includes X_1 (CTYPOP), X_2 (LANES), X_4 (CONTROL = 1), and X_5 (CLASS = 1). With this model, some improvements in the selection criteria are realized: $R_p^2 = .812$ and $C_p = 2.65$. On the basis of these results, it was decided to investigate

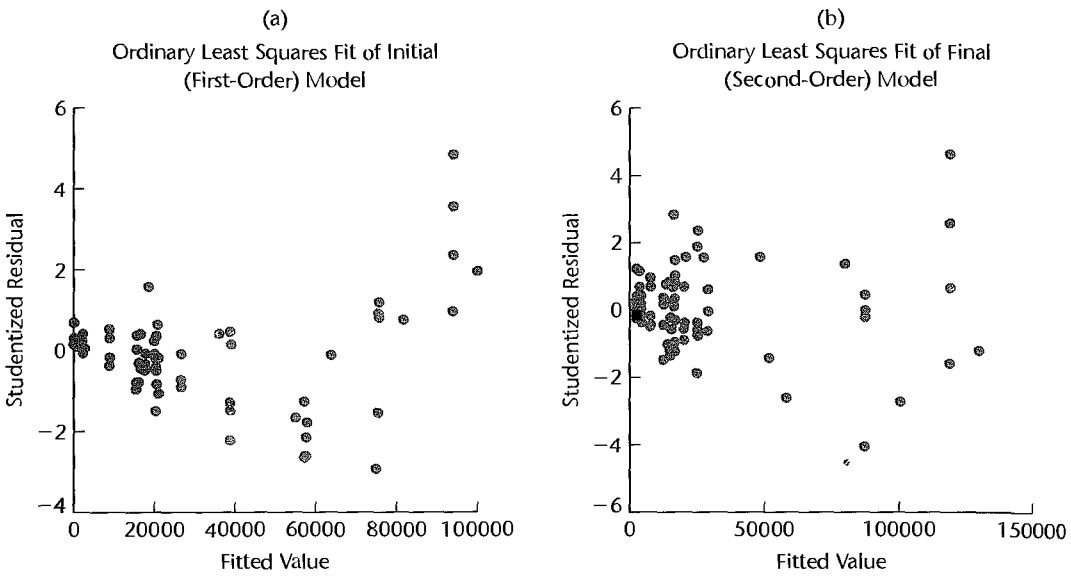
FIGURE 11.16
SAS All-Possible-Regressions Output—MNDOT Traffic Estimation Example.

```
N = 121      Regression Models for Dependent Variable: AADT

          R-square     C(p)  Variables in Model
    In

     2    0.694589  69.7231  CTYPOP LANES

NOTE: The above variables are included in all models to follow

-----------------------------------------
     3    0.804522   5.2315  CLASS3
     3    0.751353  37.3903  CONTROL1
     3    0.725755  52.8725  TRUCK1
     3    0.704495  65.7318  LOCALE2
     3    0.704250  65.8798  CLASS1
-------------------------------------
     4    0.812099   2.6490  CONTROL1 CLASS1
     4    0.810364   3.6986  CLASS3 LOCALE2
     4    0.808001   5.1275  CLASS3 LOCALE1
     4    0.807122   5.6590  CLASS2 CLASS3
     4    0.806300   6.1562  CLASS3 TRUCK4
----------------------------------------------
     5    0.816245   2.1414  CONTROL1 CLASS1 LOCALE2
     5    0.815842   2.3848  CONTROL1 CLASS1 LOCALE1
     5    0.814362   3.2803  CONTROL1 CLASS1 CLASS2
     5    0.813901   3.5589  CONTROL1 CLASS1 TRUCK4
     5    0.812788   4.2321  CONTROL1 CLASS1 TRUCK2
-----------------------------------------------------
     6    0.818304   2.8958  WIDTH CONTROL1 CLASS1 LOCALE1
     6    0.817992   3.0845  CONTROL1 CLASS1 TRUCK4 LOCALE2
     6    0.817915   3.1309  CONTROL1 CLASS1 TRUCK2 LOCALE2
     6    0.817741   3.2367  CONTROL1 CLASS1 TRUCK2 LOCALE1
     6    0.817738   3.2383  WIDTH CONTROL1 CLASS1 LOCALE2
----------------------------------------------------------
     7    0.820443   3.6023  WIDTH CONTROL1 CLASS1 TRUCK4 LOCALE1
     7    0.819942   3.9050  WIDTH CONTROL1 CLASS1 TRUCK4 LOCALE2
     7    0.819473   4.1891  WIDTH CONTROL1 CLASS1 TRUCK2 LOCALE1
     7    0.819180   4.3663  CONTROL1 CLASS1 TRUCK2 TRUCK4 LOCALE2
     7    0.819007   4.4705  WIDTH CONTROL1 CLASS1 CLASS2 LOCALE1
---------------------------------------------------------------------
```

a model based on the five predictor variables included in these two models: X_1 (CTYPOP), X_2 (LANES), X_4 (CONTROL = 1), X_5 (CLASS = 1), and X_7 (CLASS = 3). Note that because X_6 (CLASS = 2) has been dropped from further consideration, the rural noninterstate (CLASS = 2) and urban noninterstate (CLASS = 4) categories of the CLASS variable have been collapsed into one category.

Figure 11.17a contains a plot of the studentized residuals against the fitted values for the five-variable model. The plot reveals two potential problems: (1) The residuals tend to be positive for small and large values of $\hat{Y}$ and negative for intermediate values, suggesting a curvilinearity in the response function. (2) The variability of the residuals tends to increase with increasing $\hat{Y}$, indicating nonconstancy of the error variance.

FIGURE 11.17 Plots of Studentized Residuals versus Fitted Values—MNDOT Traffic Estimation Example.



Curvilinearity was investigated next, together with possible interaction effects. A squared term for each of the two quantitative variables (CTYPOP and LANES) was added to the pool of potential X variables. To reduce potential multicollinearity problems, each of these variables was first centered. In addition, nine cross-product terms were added to the pool of potential X variables, consisting of the cross products of the X variables for the four predictor variables.

The SAS all-possible-regressions procedure was run again for this enlarged pool of potential X variables (output not shown). Analysis of the results suggested a model with five X variables: CTYPOP, LANES, LANES^2, CONTROL1, and CTYPOP $\times$ CONTROL1. For this model, R_p^2 is .925, and all P-values for the regression coefficients are 0+. Although this model does not have the largest R_p^2 value among five-term models, it is desirable because it is easy to interpret and does not differ substantially from other models favorably identified by the C_p or R_p^2 criteria. A plot of the studentized residuals against $\hat{Y}$, shown in Figure 11.17b, indicates that curvilinearity is no longer present. Also, neither Cook's distance measure (maximum = .47) nor the variance inflation factors (maximum = 2.5) revealed serious problems at this stage. Nonconstancy of the error term variance has persisted, however, as confirmed by the Breusch-Pagan test.

Weighted Least Squares Estimation

To remedy the problem with nonconstancy of the error term variance, weighted least squares was implemented by developing a standard deviation function. Residual plots indicated that the absolute residuals vary with CTYPOP and LANES. A fit of a first-order model where the absolute residuals are regressed on CTYPOP and LANES yielded an estimated standard deviation function for which $R^2 = .386$ and the P-values for the regression coefficients for CTYPOP and LANES are .001 and 0+. Note that, as is often the case, the R^2 value for

FIGURE 11.18 MINITAB Weighted Least Squares Regression Results—MNDOT Traffic Estimation Example.

```
The regression equation is
AADT = 9602 + 0.0146 CTYPOP + 6162 LANES + 16556 CONTROL1 + 2250 LANES2
           + 0.0637 POPXCTL1

Predictor        Coef      Stdev   t-ratio       p
Constant         9602       1432      6.71   0.000
CTYPOP       0.014567   0.003047      4.78   0.000
LANES          6161.8      933.9      6.60   0.000
CONTROL1        16556       2966      5.58   0.000
LANES2         2249.7      755.8      2.98   0.004
POPXCTL1     0.063696   0.008421      7.56   0.000

Analysis of Variance

SOURCE          DF         SS         MS        F        P
Regression       5     919.55     183.91    93.13    0.000
Error          115     227.10       1.97
Total          120    1146.65
```

the estimated standard deviation function (.386) is substantially smaller than that for the estimated response function (.925).

Using the weights obtained from the standard deviation function, weighted least squares estimates of the regression coefficients were obtained. Since some of the estimated regression coefficients differed substantially from those obtained with unweighted least squares, the residuals from the weighted least squares fit were used to reestimate the standard deviation function, and revised weights were obtained. Two more iterations of this iteratively reweighted least squares process led to stable estimated coefficients.

MINITAB regression results for the weighted least squares fit based on the final weights are shown in Figure 11.18. Note that the signs of the regression coefficients are all positive, as might be expected:

CTYPOP: Traffic increases with local population density

LANES: Traffic increases with number of lanes

CONTROL1: Traffic is highest for road sections under access control

$LANES^2$: An upward-curving parabola is consistent with the shape of the lowess fit of AADT to LANES in Figure 11.15

CTYPOP × CONTROL1: Traffic increase with access control is more pronounced for higher population density

Figure 11.19a contains a plot of the studentized residuals against the fitted values, and Figure 11.19b contains a normal probability plot of the studentized residuals. Notice that the variability of the studentized residuals is now approximately constant. While the normal probability plot in Figure 11.19b indicates some departure from normality (this was confirmed by the correlation test for normality), the departure does not appear to be serious, particularly in view of the large sample size.

To assess the usefulness of the model for estimating AADT, approximate 95 percent confidence intervals for mean traffic for typical rural, suburban, and urban road sections

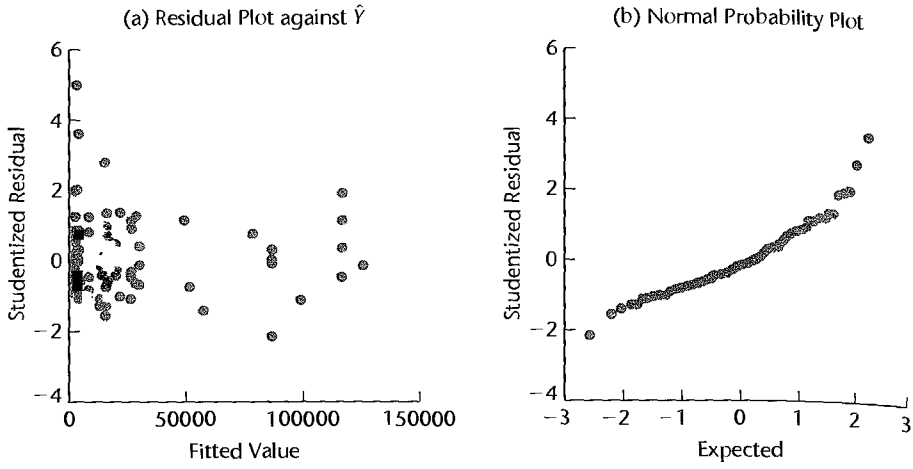


FIGURE 11.19 **Residual Plots for Final Weighted Least Squares Regression Fit—MNDOT Traffic Estimation Example.**

TABLE 11.12 **95 Percent Approximate Confidence Limits for Mean Responses—MNDOT Traffic Estimation Example.**

	(1)	(2)	(3)	(4)	(5)	(6) Confidence Limits	(7)
Road Section	CTYPOP	LANES	CONTROL1	$\hat{Y}_h$	$s\{\hat{Y}_h\}$	Lower	Upper
Rural	113,571	2	0	3,365	354	2,663	4,066
Suburban	222,229	4	0	16,379	1,827	12,758	19,999
Urban	941,411	6	1	116,024	6,597	102,953	129,095

were constructed. The levels of the predictor variables for these road sections are given in Table 11.12, columns 1–3. The estimated mean traffic is given in column 4. The approximate estimated standard deviations of the estimated mean responses for each of these road sections, shown in column 5, were obtained by using $\mathbf{s}^2\{\mathbf{b}_w\}$ from (11.13) in (6.58):

$$s^2\{\hat{Y}_h\} = \mathbf{X}_h'\mathbf{s}^2\{\mathbf{b}_w\}\mathbf{X}_h = MSE_w\mathbf{X}_h'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_h \qquad (11.63)$$

where the vector $\mathbf{X}_h$ is defined in (6.53). Since the estimated standard deviations in column 5 are only approximations because the least squares weights were estimated by means of a standard deviation function, bootstrapping with random X sampling was employed to assess the precision of the fitted values. The standard deviations of the bootstrap sampling distributions were close to the estimated standard deviations in column 5. The consistency of the results shows that the iterative estimation of the weights by means of the standard deviation function did not have any substantial effect here on the precision of the fitted values.

The approximate 95 percent confidence limits for $E\{Y_h\}$, computed using (6.59), are presented in columns 6 and 7 of Table 11.12. The precision of these estimates was considered to be sufficient for planning purposes. However, because the suburban and rural road estimates

have the poorest relative precision, it was recommended that better records be developed for population density in the immediate vicinity of a road section, since county population does not always reflect local population density. The improved information could lead to a better regression model, with more precise estimates for road sections in rural and suburban settings.

The approach for developing the regression model described here is not, of course, the only approach that can lead to a useful regression model, nor is the analysis complete as described. For example, the residual plot in Figure 11.19a suggests the presence of at least one outlier ($r_{92} = 5.02$). Possible remedial measures for this case should be considered. In addition, the departure from normality might be remedied by a transformation of the response variable. This transformation might also stabilize the variance of the error terms sufficiently so that weighted least squares would not be needed. In fact, subsequent analysis using the Box-Cox transformation approach found that a cube root transformation of the response is very effective in this instance. A final choice between the model fit obtained by weighted least squares and a model fit developed by an alternative approach can be made on the basis of model validation studies.

Cited References

11.1. Davidian, M., and R. J. Carroll. "Variance Function Estimation," *Journal of the American Statistical Association* 82 (1987), pp. 1079–91.

11.2. Greene, W. H. *Econometric Analysis,* 5th ed. Upper Saddle River, New Jersey: Prentice Hall, 2003.

11.3. Belsley, D. A. *Conditioning Diagnostics: Collinearity and Weak Data in Regression.* New York: John Wiley & Sons, 1991.

11.4. Frank, I. E., and J. H. Friedman. "A Statistical View of Some Chemometrics Regression Tools," *Technometrics* 35 (1993), pp. 109–35.

11.5. Hoaglin, D. C.; F. Mosteller; and J. W. Tukey. *Exploring Data Tables, Trends, and Shapes.* New York: John Wiley & Sons, 1985.

11.6. Rousseeuw, P. J., and A. M. Leroy. *Robust Regression and Outlier Detection.* New York: John Wiley & Sons, 1987.

11.7. Kennedy, W. J., Jr., and J. E. Gentle. *Statistical Computing.* New York: Marcel Dekker, 1980.

11.8. ETS Policy Information Center. *America's Smallest School: The Family.* Princeton, N.J.: Educational Testing Service, 1992.

11.9. Haerdle, W. *Applied Nonparametric Regression.* Cambridge: Cambridge University Press, 1992.

11.10. Cleveland, W. S., and S. J. Devlin. "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association* 83 (1988), pp. 596–610.

11.11. Breiman, L.; J. H. Friedman; R. A. Olshen; and C. J. Stone. *Classification and Regression Trees.* Belmont, Calif.: Wadsworth, 1984.

11.12. Friedman, J. H., and W. Stuetzle. "Projection Pursuit Regression," *Journal of the American Statistical Association* 76 (1981), pp. 817–23.

11.13. Eubank, R. L. *Spline Smoothing and Nonparametric Regression,* 2nd ed. New York: Marcel Dekker, 1999.

11.14. Hastie, T., and C. Loader. "Local Regression: Automatic Kernel Carpentry" (with discussion), *Statistical Science* 8 (1993), pp. 120–43.

11.15. Hastie, T., Tibshirani, R., and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer-Verlag, 2001.

11.16. Efron, B. *The Jackknife, The Bootstrap, and Other Resampling Plans*. Philadelphia, Penn.: Society for Industrial and Applied Mathematics, 1982.

11.17. Efron, B., and R. Tibshirani. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science* 1 (1986), pp. 54–77.

11.18. Efron, B. "Better Bootstrap Confidence Intervals" (with discussion), *Journal of the American Statistical Association* 82 (1987), pp. 171–200.

11.19. Cheng, C. "Optimal Sampling for Traffic Volume Estimation," unpublished Ph.D. dissertation, University of Minnesota, Carlson School of Management, 1992.

Problems

11.1. One student remarked to another: "Your residuals show that nonconstancy of error variance is clearly present. Therefore, your regression results are completely invalid." Comment.

11.2. An analyst suggested: "One nice thing about robust regression is that you need not worry about outliers and influential observations." Comment.

11.3. Lowess smoothing becomes difficult when there are many predictors and the sample size is small. This is sometimes referred to as the "curse of dimensionality." Discuss the nature of this problem.

11.4. Regression trees become difficult to utilize when there are many predictors and the sample size is small. Discuss the nature of this problem.

11.5. Describe how bootstrapping might be used to obtain confidence intervals for regression coefficients when ridge regression is employed.

11.6. **Computer-assisted learning.** Data from a study of computer-assisted learning by 12 students, showing the total number of responses in completing a lesson (X) and the cost of computer time (Y, in cents), follow.

i:	1	2	3	4	5	6	7	8	9	10	11	12
X_i:	16	14	22	10	14	17	10	13	19	12	18	11
Y_i:	77	70	85	50	62	70	55	63	88	57	81	51

a. Fit a linear regression function by ordinary least squares, obtain the residuals, and plot the residuals against X. What does the residual plot suggest?

b. Divide the cases into two groups, placing the six cases with the smallest fitted values $\hat{Y}_i$ into group 1 and the other six cases into group 2. Conduct the Brown-Forsythe test for constancy of the error variance, using $\alpha = .05$. State the decision rule and conclusion.

c. Plot the absolute values of the residuals against X. What does this plot suggest about the relation between the standard deviation of the error term and X?

d. Estimate the standard deviation function by regressing the absolute values of the residuals against X, and then calculate the estimated weight for each case using (11.16a). Which case receives the largest weight? Which case receives the smallest weight?

e. Using the estimated weights, obtain the weighted least squares estimates of β_0 and β_1. Are these estimates similar to the ones obtained with ordinary least squares in part (a)?

f. Compare the estimated standard deviations of the weighted least squares estimates b_{w0} and b_{w1} in part (e) with those for the ordinary least squares estimates in part (a). What do you find?

g. Iterate the steps in parts (d) and (e) one more time. Is there a substantial change in the estimated regression coefficients? If so, what should you do?

*11.7. **Machine speed.** The number of defective items produced by a machine (Y) is known to be linearly related to the speed setting of the machine (X). The data below were collected from recent quality control records.

i:	1	2	3	4	5	6	7	8	9	10	11	12
X_i:	200	400	300	400	200	300	300	400	200	400	200	300
Y_i:	28	75	37	53	22	58	40	96	46	52	30	69

a. Fit a linear regression function by ordinary least squares, obtain the residuals, and plot the residuals against X. What does the residual plot suggest?

b. Conduct the Breusch-Pagan test for constancy of the error variance, assuming $\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i$; use $\alpha = .10$. State the alternatives, decision rule, and conclusion.

c. Plot the squared residuals against X. What does the plot suggest about the relation between the variance of the error term and X?

d. Estimate the variance function by regressing the squared residuals against X, and then calculate the estimated weight for each case using (11.16b).

e. Using the estimated weights, obtain the weighted least squares estimates of β_0 and β_1. Are the weighted least squares estimates similar to the ones obtained with ordinary least squares in part (a)?

f. Compare the estimated standard deviations of the weighted least squares estimates b_{w0} and b_{w1} in part (e) with those for the ordinary least squares estimates in part (a). What do you find?

g. Iterate the steps in parts (d) and (e) one more time. Is there a substantial change in the estimated regression coefficients? If so, what should you do?

11.8. **Employee salaries.** A group of high-technology companies agreed to share employee salary information in an effort to establish salary ranges for technical positions in research and development. Data obtained for each employee included current salary (Y), a coded variable indicating highest academic degree obtained (1 = bachelor's degree, 2 = master's degree, 3 = doctoral degree), years of experience since last degree (X_3), and the number of persons currently supervised (X_4). The data follow.

Employee i	Y_{i1}	Degree	X_{i3}	X_{i4}
1	58.8	3	4.49	0
2	34.8	1	2.92	0
3	163.7	3	29.54	42
...	...	...	...	...
63	40.0	2	.44	0
64	60.5	3	2.10	0
65	104.8	3	19.81	24

a. Create two indicator variables for highest degree attained:

Degree	X_1	X_2
Bachelor's	0	0
Master's	1	0
Doctoral	0	1

b. Regress Y on X_1, X_2, X_3, and X_4, using a first-order model and ordinary least squares, obtain the residuals, and plot them against $\hat{Y}$. What does the residual plot suggest?

c. Divide the cases into two groups, placing the 33 cases with the smallest fitted values $\hat{Y}_i$ into group 1 and the other 32 cases into group 2. Conduct the Brown-Forsythe test for constancy of the error variance, using $\alpha = .01$. State the decision rule and conclusion.

d. Plot the absolute residuals against X_3 and against X_4. What do these plots suggest about the relation between the standard deviation of the error term and X_3 and X_4?

e. Estimate the standard deviation function by regressing the absolute residuals against X_3 and X_4 in first-order form, and then calculate the estimated weight for each case using (11.16a).

f. Using the estimated weights, obtain the weighted least squares fit of the regression model. Are the weighted least squares estimates of the regression coefficients similar to the ones obtained with ordinary least squares in part (b)?

g. Compare the estimated standard deviations of the weighted least squares coefficient estimates in part (f) with those for the ordinary least squares estimates in part (b). What do you find?

h. Iterate the steps in parts (e) and (f) one more time. Is there a substantial change in the estimated regression coefficients? If so, what should you do?

11.9. Refer to **Cosmetics sales** Problem 10.13. Given below are the estimated ridge standardized regression coefficients, the variance inflation factors, and R^2 for selected biasing constants c.

c:	.00	.01	.02	.04	.06	.08	.09	.10
b_1^R:	.490	.461	.443	.463	.410	.401	.398	.394
b_2^R:	.296	.322	.336	.349	.354	.356	.356	.356
b_3^R:	.169	.167	.167	.166	.165	.164	.164	.164
$(VIF)_1$:	20.07	10.36	6.37	3.20	1.98	1.38	1.20	1.05
$(VIF)_2$:	20.72	10.67	6.55	3.27	2.07	1.40	1.21	1.06
$(VIF)_3$:	1.22	1.17	1.14	1.08	1.02	.98	.95	.93
R^2:	.7417	.7416	.7145	.7412	.7409	.7045	.7402	.7399

a. Make a ridge trace plot for the given c values. Do the ridge regression coefficients exhibit substantial changes near $c = 0$?

b. Suggest a reasonable value for the biasing constant c based on the ridge trace, the *VIF* values, and R^2.

c. Transform the estimated standardized regression coefficients selected in part (b) back to the original variables and obtain the fitted values for the 44 cases. How similar are these fitted values to those obtained with the ordinary least squares fit in Problem 10.13a?

*11.10. **Chemical shipment.** The data to follow, taken on 20 incoming shipments of chemicals in drums arriving at a warehouse, show number of drums in shipment (X_1), total weight of shipment (X_2, in hundred pounds), and number of minutes required to handle shipment (Y).

i:	1	2	3	...	18	19	20
X_{i1}:	7	18	5	...	21	6	11
X_{i2}:	5.11	16.72	3.20	...	15.21	3.64	9.57
Y_i:	58	152	41	...	155	39	90

Given below are the estimated ridge standardized regression coefficients, the variance inflation factors, and R^2 for selected biasing constants c.

c:	.000	.005	.01	.05	.07	.09	.10	.20
b_1^R:	.451	.453	.455	.460	.460	.459	.458	.444
b_2^R:	.561	.556	.552	.526	.517	.508	.504	.473
$(VIF)_1 = (VIF)_2$:	7.03	6.20	5.51	2.65	2.03	1.61	1.46	.71
R^2:	.9869	.9869	.9869	.9862	.9856	.9852	.9844	.9780

a. Fit regression model (6.1) to the data and find the fitted values.

b. Make a ridge trace plot for the given c values. Do the ridge regression coefficients exhibit substantial changes near $c = 0$?

c. Why are the $(VIF)_1$ values the same as the $(VIF)_2$ values here?

d. Suggest a reasonable value for the biasing constant c based on the ridge trace, the VIF values, and R^2.

e. Transform the estimated standardized regression coefficients selected in part (c) back to the original variables and obtain the fitted values for the 20 cases. How similar are these fitted values to those obtained with the ordinary least squares fit in part (a)?

*11.11. Refer to **Copier maintenance** Problem 1.20. Two cases had been held out of the original data set because special circumstances led to unusually long service times:

Case i	X_i	Y_i
46	6	132
47	5	166

a. Using the enlarged (47-case) data set, fit a simple linear regression model using ordinary least squares and plot the data together with the fitted regression function. What is the effect of adding cases 46 and 47 on the fitted response function?

b. Obtain the scaled residuals in (11.47) and use the Huber weight function (11.44) to obtain the case weights for a first iteration of IRLS robust regression. Which cases receive the smallest Huber weights? Why?

c. Using the weights calculated in part (b), obtain the weighted least squares estimates of the regression coefficients. How do these estimates compare to those found in part (a) using ordinary least squares?

d. Continue the IRLS procedure for two more iterations. Which cases receive the smallest weights in the final iteration? How do the final IRLS robust regression estimates compare to the ordinary least squares estimates obtained in part (a)?

e. Plot the final IRLS estimated regression function, obtained in part (d), on the graph constructed in part (a). Does the robust fit differ substantially from the ordinary least squares fit? If so, which fit is preferred here?

11.12. **Weight and height.** The weights and heights of twenty male students in a freshman class are recorded in order to see how well weight (Y, in pounds) can be predicted from height (X, in inches). The data are given below. Assume that first-order regression (1.1) is appropriate.

i:	1	2	3	...	18	19	20
X_i:	74	65	72	...	69	68	67
Y_i:	185	195	216	...	177	145	137

a. Fit a simple linear regression model using ordinary least squares, and plot the data together with the fitted regression function. Also, obtain an index plot of Cook's distance (10.33). What do these plots suggest?

b. Obtain the scaled residuals in (11.47) and use the Huber weight function (11.44) to obtain case weights for a first iteration of IRLS robust regression. Which cases receive the smallest Huber weights? Why?

c. Using the weights calculated in part (b), obtain the weighted least squares estimates of the regression coefficients. How do these estimates compare to those found in part (a) using ordinary least squares?

d. Continue the IRLS procedure for two more iterations. Which cases receive the smallest weights in the final iteration? How do the final IRLS robust regression estimates compare to the ordinary least squares estimates obtained in part (a)?

Exercises

11.13. (Calculus needed.) Derive the weighted least squares normal equations for fitting a simple linear regression function when $\sigma_i^2 = kX_i$, where k is a proportionality constant.

11.14. Express the weighted least squares estimator b_{w1} in (11.26a) in terms of the centered variables $Y_i - \bar{Y}_w$ and $X_i - \bar{X}_w$, where $\bar{Y}_w$ and $\bar{X}_w$ are the weighted means.

11.15. Refer to **Computer-assisted learning** Problem 11.6. Demonstrate numerically that the weighted least squares estimates obtained in part (e) are identical to those obtained using transformation (11.23) and ordinary least squares.

11.16. Refer to **Machine speed** Problem 11.7. Demonstrate numerically that the weighted least squares estimates obtained in part (e) are identical to those obtained when using transformation (11.23) and ordinary least squares.

11.17. Consider the weighted least squares criterion (11.6) with weights given by $w_i = .3/X_i$. Set up the variance-covariance matrix for the error terms when $i = 1, \ldots, 4$. Assume $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$ for $i \neq j$.

11.18. Derive the variance-covariance matrix $\boldsymbol{\sigma}^2\{\mathbf{b}_w\}$ in (11.10) for the weighted least squares estimators when the variance-covariance matrix of the observations Y_i is $k\mathbf{W}^{-1}$, where $\mathbf{W}$ is given in (11.7) and k is a proportionality constant.

11.19. Derive the mean squared error in (11.29).

11.20. Refer to the body fat example of Table 7.1. Employing least absolute residuals regression, the LAR estimates of the regression coefficients are $b_0 = -17.027$, $b_1 = .4173$, and $b_2 = .5203$.

a. Find the sum of the absolute residuals based on the LAR fit.

b. For the least squares estimated regression coefficients $b_0 = -19.174$, $b_1 = .2224$, and $b_2 = .6594$, find the sum of the absolute residuals. Is this sum larger than the sum obtained in part (a)? Is this to be expected?

Projects

11.21. Observations on Y are to be taken when $X = 10, 20, 30, 40$, and 50, respectively. The true regression function is $E\{Y\} = 20 + 10X$. The error terms are independent and normally distributed, with $E\{\varepsilon_i\} = 0$ and $\sigma^2\{\varepsilon_i\} = .8X_i$.

a. Generate a random Y observation for each X level and calculate both the ordinary and weighted least squares estimates of the regression coefficient β_1 in the simple linear regression function.

b. Repeat part (a) 200 times, generating new random numbers each time.

c. Calculate the mean and variance of the 200 ordinary least squares estimates of β_1 and do the same for the 200 weighted least squares estimates.

d. Do both the ordinary least squares and weighted least squares estimators appear to be unbiased? Explain. Which estimator appears to be more precise here? Comment.

11.22. Refer to **Patient satisfaction** Problem 6.15.

a. Obtain the estimated ridge standardized regression coefficients, variance inflation factors, and R^2 for the following biasing constants: $c = .000, .005, .01, .02, .03, .04, .05$.

b. Make a ridge trace plot for the given c values. Do the ridge regression coefficients exhibit substantial changes near $c = 0$?

c. Suggest a reasonable value for the biasing constant c based on the ridge trace, the *VIF* values, and R^2.

d. Transform the estimated standardized regression coefficients selected in part (c) back to the original variables and obtain the fitted values for the 46 cases. How similar are these fitted values to those obtained with the ordinary least squares fit in Problem 6.15c?

11.23. **Cement composition.** Data on the effect of composition of cement on heat evolved during hardening are given below. The variables collected were the amount of tricalcium aluminate (X_1), the amount of tricalcium silicate (X_2), the amount of tetracalcium alumino ferrite (X_3), the amount of dicalcium silicate (X_4), and the heat evolved in calories per gram of cement (Y).

i:	1	2	3	...	11	12	13
X_{i1}:	7	1	11	...	1	11	10
X_{i2}:	26	29	56	...	40	66	68
X_{i3}:	6	15	8	...	23	9	8
X_{i4}:	60	52	20	...	34	12	12
Y_i:	78.5	74.3	104.3	...	83.8	113.3	109.4

Adapted from H. Woods, H. H. Steinour, and H. R. Starke, "Effect of Composition of Portland Cement on Heat Evolved During Hardening," *Industrial and Engineering Chemistry*, 24, 1932, 1207–1214.

a. Fit regression model (6.5) for four predictor variables to the data. State the estimated regression function.

b. Obtain the estimated ridge standardized regression coefficients, variance inflation factors, and R^2 for the following biasing constants: $c = .000, .002, .004, .006, .008, .02, .04, .06, .08, .10$.

c. Make a ridge trace plot for the biasing constants listed in part (b). Do the ridge regression coefficients exhibit substantial changes near $c = 0$?

d. Suggest a reasonable value for the biasing constant c based on the ridge trace, *VIF* values, and R^2 values.

e. Transform the estimated standardized ridge regression coefficients selected in part (d) to the original variables and obtain the fitted values for the 13 cases. How similar are these fitted values to those obtained with the ordinary least squares fit in part (a)?

11.24. Refer to **Commercial properties** Problem 6.18.

a. Use least absolute residuals regression to obtain estimates of the parameters $\beta_0, \beta_1, \beta_2, \beta_3$, and β_4.

b. Find the sum of the absolute residuals based on the LAR fit in part (a).

c. For the least squares estimated regression function in Problem 6.18c, find the sum of the absolute residuals. Is this sum larger than the sum obtained in part (b)? Is this to be expected?

11.25. **Crop yield.** An agronomist studied the effects of moisture (X_1, in inches) and temperature (X_2, in C) on the yield of a new hybrid tomato (Y). The experimental data follow.

i:	1	2	3	...	23	24	25
X_{i1}:	6	6	6	...	14	14	14
X_{i2}:	20	21	22	...	22	23	24
Y_i:	49.2	48.1	48.0	...	42.1	43.9	40.5

The agronomist expects that second-order polynomial regression model (8.7) with independent normal error terms is appropriate here.

a. Fit a second-order polynomial regression model omitting the interaction term and the quadratic effect term for temperature.

b. Construct a contour plot of the fitted surface obtained in part (a).

c. Use the lowess method to obtain a nonparametric estimate of the yield response surface as a function of moisture and temperature. Employ weight function (11.53), $q = 9/25$, and a Euclidean distance measure with unscaled variables. Obtain fitted values $\hat{Y}_h$ for the 9×9 rectangular grid of (X_{h1}, X_{h2}) values where $X_{h1} = 6.7, \ldots, 13.14$ and $X_{h2} = 20, 20.5, \ldots, 23.5, 24$, using a local first-order model.

d. Construct a contour plot of the resulting lowess surface. Are the lowess contours consistent with the contours in part (b) for the polynomial model? Discuss.

11.26. Refer to **Computer-assisted learning** Problem 11.6.

a. Based on the weighted least squares fit in Problem 11.6e, construct an approximate 95 percent confidence interval for β_1 by means of (6.50), using the estimated standard deviation $s\{b_{w1}\}$.

b. Using random X sampling, obtain 750 bootstrap samples of size 12. For each bootstrap sample, (1) use ordinary least squares to regress Y on X and obtain the residuals, (2) estimate the standard deviation function by regressing the absolute residuals on X and then use the fitted standard deviation function and (11.16a) to obtain weights, and (3) use weighted least squares to regress Y on X and obtain the bootstrap estimated regression coefficient b_1^*. (Note that for each bootstrap sample, only one iteration of the iteratively reweighted least squares procedure is to be used.)

c. Construct a histogram of the 750 bootstrap estimates b_1^*. Does the bootstrap sampling distribution of b_1^* appear to approximate a normal distribution?

d. Calculate the sample standard deviation of the 750 bootstrap estimates b_1^*. How does this value compare to the estimated standard deviation $s\{b_{w1}\}$ used in part (a)?

e. Construct a 95 percent bootstrap confidence interval for β_1 using reflection method (11.59). How does this confidence interval compare with that obtained in part (a)? Does the approximate interval in part (a) appear to be useful for this data set?

11.27. Refer to **Machine speed** Problem 11.7.

a. On the basis of the weighted least squares fit in Problem 11.7e, construct an approximate 90 percent confidence interval for β_1 by means of (6.50), using the estimated standard deviation $s\{b_{w1}\}$.

b. Using random X sampling, obtain 800 bootstrap samples of size 12. For each bootstrap sample, (1) use ordinary least squares to regress Y on X and obtain the residuals, (2) estimate

the standard deviation function by regressing the absolute residuals on X and then use the fitted standard deviation function and (11.16a) to obtain weights, and (3) use weighted least squares to regress Y on X and obtain the bootstrap estimated regression coefficient b_1^*. (Note that for each bootstrap sample, only one iteration of the iteratively reweighted least squares procedure is to be used.)

c. Construct a histogram of the 800 bootstrap estimates b_1^*. Does the bootstrap sampling distribution of b_1^* appear to approximate a normal distribution?

d. Calculate the sample standard deviation of the 800 bootstrap estimates b_1^*. How does this value compare to the estimated standard deviation $s\{b_{w1}\}$ used in part (a)?

e. Construct a 90 percent bootstrap confidence interval for β_1 using reflection method (11.59). How does this confidence interval compare with that obtained in part (a)? Does the approximate interval in part (a) appear to be useful for this data set?

11.28. **Mileage study.** The effectiveness of a new experimental overdrive gear in reducing gasoline consumption was studied in 12 trials with a light truck equipped with this gear. In the data that follow, X_i denotes the constant speed (in miles per hour) on the test track in the ith trial and Y_i denotes miles per gallon obtained.

i:	1	2	3	4	5	6	7	8	9	10	11	12
X_i:	35	35	40	40	45	45	50	50	55	55	60	60
Y_i:	22	20	28	31	37	38	41	39	34	37	27	30

Second-order regression model (8.2) with independent normal error terms is expected to be appropriate.

a. Fit regression model (8.2). Plot the fitted regression function and the data. Does the quadratic regression function appear to be a good fit here?

b. Automotive engineers would like to estimate the speed $X_{\max}$ at which the average mileage $E\{Y\}$ is maximized. It can be shown for second-order model (8.2) that $X_{\max} = \bar{X} - (.5\beta_1/\beta_{11})$, provided that β_{11} is negative. Estimate the speed $X_{\max}$ at which the average mileage is maximized, using $\hat{X}_{\max} = \bar{X} - (.5b_1/b_{11})$. What is the estimated mean mileage at the estimated optimum speed?

c. Using fixed X sampling, obtain 1,000 bootstrap samples of size 12. For each bootstrap sample, fit regression model (8.2) and obtain the bootstrap estimate $\hat{X}_{\max}^*$.

d. Construct a histogram of the 1,000 bootstrap estimates $\hat{X}_{\max}^*$. Does the bootstrap sampling distribution of $\hat{X}_{\max}^*$ appear to approximate a normal distribution?

e. Construct a 90 percent bootstrap confidence interval for $X_{\max}$ using reflection method (11.56). How precisely has $X_{\max}$ been estimated?

11.29. Refer to **Muscle mass** Problem 1.27.

a. Fit a two-region regression tree. What is the first split point based on age? What is *SSE* for this two-region tree?

b. Find the second split point given the two-region tree in part (a). What is *SSE* for the resulting three-region tree?

c. Find the third split point given the three-region tree in part (b). What is *SSE* for the resulting four-region tree?

d. Prepare a scatter plot of the data with the four-region tree in part (c) superimposed. How well does the tree fit the data? What does the tree suggest about the change in muscle mass with age?

e. Prepare a residual plot of e_i versus $\hat{Y}_i$ for the four-region tree in part (d). State your findings.

11.30. Refer to **Patient satisfaction** Problem 6.15. Consider only the first two predictors (patient's age, X_1, and severity of illness, X_2).

a. Fit a two-region regression tree. What is the first split point, and on which predictor is it based? What is *SSE* for the resulting two-region tree?

b. Find the second split point given the two-region tree in part (a). Is it based on X_1 or X_2? What is *SSE* for the resulting three-region tree?

c. Find the third split point given the three-region tree in part (b). Is it based on X_1 or X_2? What is *SSE* for the resulting four-region tree?

d. Find the fourth split point given the four-region tree in part (c). Is it based on X_1 or X_2? What is *SSE* for the resulting five-region tree?

e. Prepare a three-dimensional surface plot of the five-region tree obtained in part (d). What does this tree suggest about the relative importance of the two predictors?

f. Prepare a residual plot of e_i versus $\hat{Y}_i$ for the five-region tree in part (d). State your findings.

Case Studies

11.31. Refer to the **Prostate cancer** data set in Appendix C.5 and Case Study 9.30. Select a random sample of 65 observations to use as the model-building data set.

a. Develop a regression tree for predicting PSA. Justify your choice of number of regions (tree size), and interpret your regression tree.

b. Assess your model's ability to predict and discuss its usefulness to the oncologists.

c. Compare the performance of your regression tree model with that of the best regression model obtained in Case Study 9.30. Which model is more easily interpreted and why?

11.32. Refer to the **Real estate sales** data set in Appendix C.7 and Case Study 9.31. Select a random sample of 300 observations to use as the model-building data set.

a. Develop a regression tree for predicting sales price. Justify your choice of number of regions (tree size), and interpret your model.

b. Assess your model's ability to predict and discuss its usefulness as a tool for predicting sales prices.

c. Compare the performance of your regression tree model with that of the best regression model obtained in Case Study 9.31. Which model is more easily interpreted and why?