PART III

Linear-Model Diagnostics

11

Unusual and Influential Data

As we have seen, linear statistical models—particularly linear regression analysis—make strong assumptions about the structure of data, assumptions that often do not hold in applications. The method of least squares, which is typically used to fit linear models to data, is very sensitive to the structure of the data, and can be markedly influenced by one or a few unusual observations.

We could abandon linear models and least-squares estimation in favor of nonparametric regression and robust estimation.[1] A less drastic response is also possible, however: We can adapt and extend the methods for examining and transforming data described in Chapters 3 and 4 to diagnose problems with a linear model that has been fit to data, and—often—to suggest solutions.

I shall pursue this strategy in this and the next two chapters. The current chapter deals with unusual and influential data. Chapter 12 takes up a variety of problems, including nonlinearity, nonconstant error variance, and nonnormality. Collinearity is the subject of Chapter 13.

Taken together, the diagnostic and corrective methods described in these chapters substantially extend the practical application of linear models. These methods are often the difference between a crude, mechanical data analysis, and a careful, nuanced analysis that accurately describes the data and therefore supports meaningful interpretation of them.

[1] Methods for nonparametric and robust regression were introduced informally in Chapter 2 and will be described in more detail in Chapter 14.

11.1 Outliers, Leverage, and Influence

Unusual data are problematic in linear models fit by least squares because they can unduly influence the results of the analysis, and because their presence may be a signal that the model fails to capture important characteristics of the data. Some central distinctions are illustrated in Figure 11.1 for the simple regression model $Y = \alpha + \beta X + \varepsilon$.

In simple regression, an *outlier* is an observation whose dependent-variable value is conditionally unusual given the value of the independent variable. In contrast, a univariate outlier is a value of Y or X that is unconditionally unusual; such a value may or may not be a regression outlier.

Regression outliers appear in Figure 11.1(a) and (b). In Figure 11.1(a), the outlying observation has an X-value that is at the center of the X-distribution; as a consequence, deleting the outlier has little impact on the least-squares fit,

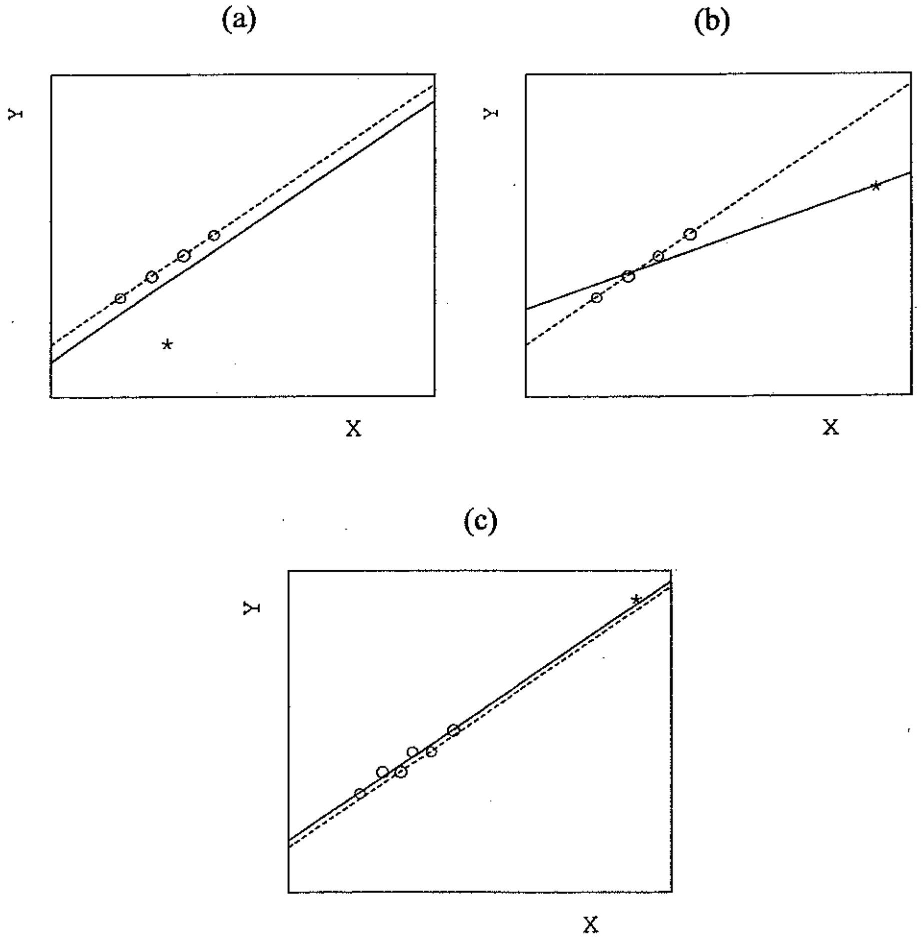

Figure 11.1. Leverage and influence in simple regression. In each graph, the solid line gives the least-squares regression for all of the data, while the broken line gives the least-squares regression with the unusual data point (the asterisk) omitted. (a) An outlier near the mean of X has low leverage and little influence on the regression coefficients. (b) An outlier far from the mean of X has high leverage and substantial influence on the regression coefficients. (c) A high-leverage observation in line with the rest of the data does not influence the regression coefficients. In panel (c), the two regression lines are separated slightly for visual effect, but are, in fact, coincident.

leaving the slope B unchanged, and affecting the intercept A only slightly. In Figure 11.1(b), however, the outlier has an unusual X-value, and thus its deletion markedly affects both the slope and the intercept. Because of its unusual X-value, the outlying last observation in Figure 11.1(b) exerts strong *leverage* on the regression coefficients, while the outlying middle observation in Figure 11.1(a) is at a low-leverage point. The combination of high leverage with a regression outlier therefore produces substantial *influence* on the regression coefficients. In Figure 11.1(c), the last observation has no influence on the regression coefficients even though it is a high-leverage point, because this observation is in line with the rest of the data—it is not a regression outlier.

The following heuristic formula helps to distinguish among the three concepts of influence, leverage, and discrepancy ("outlyingness"):

$$\text{Influence on coefficients} = \text{Leverage} \times \text{Discrepancy}$$

A simple and transparent example, with real data from Davis (1990), appears in Figure 11.2. These data record the measured and reported weight of 183 male and female subjects who engage in programs of regular physical exercise.[2] Davis's data can be treated in two ways:

- We could regress reported weight (RW) on measured weight (MW), a dummy variable for sex (F, coded 1 for women and 0 for men), and an interaction regressor (formed as the product $MW \times F$). This specification follows from the reasonable assumption that measured weight, and possibly sex, can affect reported weight. The results are as follows (with coefficient standard errors in parentheses):

$$\begin{array}{lllll} \widehat{RW} = 1.36 & + 0.990MW & + 40.0F & - 0.725(MW \times F) \\ \quad\;\;\; (3.28) & \;\;(0.043) & \;\;(3.9) & \;\;(0.056) \end{array}$$

$$R^2 = 0.89 \qquad S_E = 4.66$$

Were these results taken seriously, we would conclude that men are unbiased reporters of their weights (because $A = 1.36 \simeq 0$ and $B_1 = 0.990 \simeq 1$), while women tend to overreport their weights if they are relatively light and underreport if they are relatively heavy (the intercept for women is $1.36 + 40.0 = 41.4$ and the slope is $0.990 - 0.725 = 0.265$). Figure 11.2, however, makes it clear that the differential results for women and men are due to one female subject whose reported weight is about average (for women), but whose measured weight is extremely large. Recall that this subject's measured weight in kilograms and height in centimeters were erroneously switched. Correcting the data produces the regression

$$\begin{array}{lllll} \widehat{RW} = 1.36 & + 0.990MW & + 1.98F & - 0.0567(MW \times F) \\ \quad\;\;\; (1.58) & \;\;(0.021) & \;\;(2.45) & \;\;(0.0385) \end{array}$$

$$R^2 = 0.97 \qquad S_E = 2.24$$

which suggests that both women and men are unbiased reporters of their weight.

[2] Davis's data were introduced in Chapter 2.

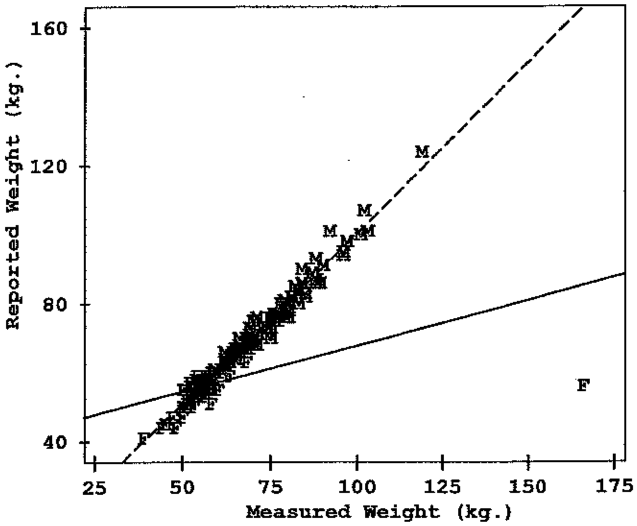

Figure 11.2. Davis's data on reported and measured weight for women (F) and men (M), showing the least-squares linear regression line for each group (the broken line for men, the solid line for women). The outlying observation has a substantial effect on the fitted line for women.

- We could (as in our previous analysis of Davis's data) regress measured weight on reported weight, sex, and their interaction, reflecting a desire to use reported weight as a predictor of measured weight. For the *uncorrected* data:

$$\begin{aligned}\widehat{MW} = \underset{(5.92)}{1.79} + \underset{(0.076)}{0.969RW} + \underset{(9.30)}{2.07F} - \underset{(0.147)}{0.00953(RW \times F)}\end{aligned}$$

$$R^2 = 0.70 \qquad S_E = 8.45$$

The outlier does not have much impact on the coefficients for this regression (both the dummy-variable coefficient and the interaction coefficient are small) precisely because the value of RW for the outlying observation is near $\overline{RW}$ for women. There is, however, a marked effect on the multiple correlation and standard error: For the corrected data, $R^2 = 0.97$ and $S_E = 2.25$.

Unusual data are problematic in linear models fit by least squares because they can substantially influence the results of the analysis, and because they may indicate that the model fails to capture important features of the data.

11.2 Assessing Leverage: Hat Values

The so-called *hat value* h_i is a common measure of leverage in regression.[3] These values are so named because it is possible to express the fitted values $\widehat{Y}_j$ ("Y-hat") in terms of the observed values Y_i:

$$\widehat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{jj}Y_j + \cdots + h_{nj}Y_n = \sum_{i=1}^{n} h_{ij}Y_i$$

Thus, the weight h_{ij} captures the contribution of observation Y_i to the fitted value $\widehat{Y}_j$: If h_{ij} is large, then the *i*th observation can have a substantial impact on the *j*th fitted value. It can be shown that $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, and so the hat value $h_i \equiv h_{ii}$ summarizes the potential influence (the leverage) of Y_i on *all* of the fitted values. The hat values are bounded between $1/n$ and 1 (i.e., $1/n \leq h_i \leq 1$), and the average hat value is $\bar{h} = (k+1)/n$ (where k is the number of regressors in the model, excluding the constant).

In simple-regression analysis,[4] the hat values measure distance from the mean of X:

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}$$

In multiple regression, h_i measures distance from the centroid (point of means) of the X's, taking into account the correlational and variational structure of the X's, as illustrated for $k = 2$ in Figure 11.3. Multivariate outliers in the X-space are thus high-leverage observations. The dependent-variable values are not at all involved in determining leverage.

For Davis's regression of reported weight on measured weight, the largest hat value by far belongs to the 12th subject, whose measured weight was wrongly recorded as 166 kg: $h_{12} = 0.714$. This quantity is many times the average hat value, $\bar{h} = (3+1)/183 = 0.0219$.

Observations with unusual combinations of independent-variable values have high *leverage* in a least-squares regression. The hat values h_i provide a measure of leverage. The average hat value is $\bar{h} = (k+1)/n$.

[3] For derivations of this and other properties of leverage, outlier, and influence diagnostics, see Section 11.8.

[4] See Exercise 11.3. Note that the sum in the denominator is over the subscript j because the subscript i is already in use.

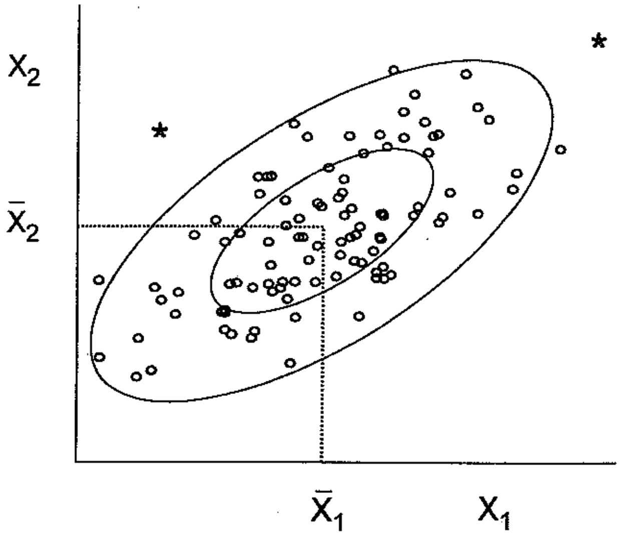

Figure 11.3. Elliptical contours of constant leverage (constant hat values h_i) for $k = 2$ independent variables. Two high-leverage points appear, both represented by asterisks. One point has unusually large values for each of X_1 and X_2, but the other is unusual only in combining a moderately large value of X_2 with a moderately small value of X_1. (These contours of constant leverage are proportional to the standard data ellipse, introduced in Chapter 9.)

11.3 Detecting Outliers: Studentized Residuals

To identify an outlying observation, we need an index of the unusualness of Y given the X's. Discrepant observations usually have large residuals, but it turns out that even if the errors ε_i have equal variances (as assumed in the general linear model), the residuals E_i do not: $V(E_i) = \sigma_\varepsilon^2(1 - h_i)$. High-leverage observations, therefore, tend to have small residuals—an intuitively sensible result, because these observations can coerce the regression surface to be close to them.

Although we can form a *standardized residual* by calculating

$$E_i' = \frac{E_i}{S_E\sqrt{1 - h_i}}$$

this measure is slightly inconvenient because its numerator and denominator are not independent, preventing E_i' from following a t-distribution: When $|E_i|$ is large, $S_E = \sqrt{\sum E_i^2/(n - k - 1)}$, which contains E_i^2, tends to be large as well. Suppose, however, that we refit the model deleting the ith observation, obtaining an estimate $S_{E(-i)}$ of σ_ε that is based on the remaining $n - 1$ observations. Then the *studentized residual*

$$E_i^* = \frac{E_i}{S_{E(-i)}\sqrt{1 - h_i}} \qquad [11.1]$$

has independent numerator and denominator, and follows a t-distribution with $n - k - 2$ degrees of freedom.

An alternative, but equivalent, procedure for finding the studentized residuals employs a “mean-shift” outlier model:

$$Y_j = \alpha + \beta_1 X_{j1} + \cdots + \beta_k X_{jk} + \gamma D_j + \varepsilon_j \qquad [11.2]$$

where D is a dummy regressor set to 1 for observation i and 0 for all other observations:

$$D_j = \begin{cases} 1 & \text{for } j = i \\ 0 & \text{otherwise} \end{cases}$$

Thus,

$$\begin{aligned} E(Y_i) &= \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma \\ E(Y_j) &= \alpha + \beta_1 X_{j1} + \cdots + \beta_k X_{jk} \quad \text{for } j \neq i \end{aligned}$$

It would be natural to specify the model in Equation 11.2 if, before examining the data, we suspected that observation i differed from the others. Then, to test H_0: $\gamma = 0$ (i.e., the null hypothesis that the ith observation is *not* an outlier), we can calculate $t_0 = \hat{\gamma}/\widehat{\mathrm{SE}}(\hat{\gamma})$. This test statistic is distributed as t_{n-k-2} under H_0, and (it turns out) is the studentized residual E_i^* of Equation 11.1.

Hoaglin and Welsch (1978) arrive at the studentized residuals by successively omitting each observation, calculating its residual based on the regression coefficients obtained for the remaining sample, and dividing the resulting residual by its standard error. Finally, Beckman and Trussell (1974) demonstrate the following simple relationship between studentized and standardized residuals:

$$E_i^* = E_i' \sqrt{\frac{n-k-2}{n-k-1-E_i'^2}} \qquad [11.3]$$

If n is large, then the factor under the square root in Equation 11.3 is close to 1, and the distinction between standardized and studentized residuals essentially disappears.[5] Moreover, for large n, the hat values are generally small, and thus it is usually the case that

$$E_i^* \simeq E_i' \simeq \frac{E_i}{S_E}$$

[5] Here, as elsewhere in statistics, terminology is not wholly standard: E_i^* is sometimes called a *deleted studentized residual*, an *externally studentized residual*, or even a standardized residual; likewise, E_i' is sometimes called an *internally studentized residual*, or simply a studentized residual. It is therefore important, especially in small samples, to determine exactly what is being calculated by a computer program before using these quantities.

11.3.1 Testing for Outliers in Linear Models

Because in most applications we do not suspect a particular observation in advance, but rather want to look for *any* outliers that may occur in the data, we can, in effect, refit the mean-shift model n times,[6] once for each observation, producing studentized residuals $E_1^*, E_2^*, \ldots, E_n^*$. Usually, our interest then focuses on the largest absolute E_i^*, denoted E_{max}^*. Because we have picked the biggest of n test statistics, however, it is not legitimate simply to use t_{n-k-2} to find a p-value for E_{max}^*: For example, even if our model is wholly adequate, and disregarding for the moment the dependence among the E_i^*'s, we would expect to obtain about 5% of E_i^*'s beyond $t_{.025} \simeq \pm 2$, about 1% beyond $t_{.005} \simeq \pm 2.6$, and so forth.

One solution[7] to the problem of simultaneous inference is to perform a Bonferroni adjustment to the p-value for the largest absolute E_i^*. The Bonferroni test requires either a special t-table or, even more conveniently, a computer program that returns accurate p-values for values of t far into the tail of the t-distribution. In the latter event, suppose that $p' = \Pr(\ t_{n-k-2} > E_{\text{max}}^*)$. Then the Bonferroni p-value for testing the statistical significance of E_{max}^* is $p = 2np'$. The factor 2 reflects the two-tail character of the test: We want to detect large negative as well as large positive outliers.

Beckman and Cook (1983) have shown that the Bonferroni adjustment is usually exact in testing the largest studentized residual. Note that a much larger E_{max}^* is required for a statistically significant result than would be the case for an ordinary individual t-test.

In Davis's regression of reported weight on measured weight, the largest studentized residual by far belongs to the incorrectly coded 12th observation, with $E_{12}^* = -24.3$. Here, $n - k - 2 = 183 - 3 - 2 = 178$, and $\Pr(t_{178} > 24.3) \simeq 10^{-58}$. The Bonferroni p-value for the outlier test is thus $p \simeq 2 \times 183 \times 10^{-58} \simeq 4 \times 10^{-56}$, an unambiguous result.

Put alternatively, the 5% critical value for E_{max}^* in this regression is the value of t_{178} with probability $.025/183 = 0.0001366$ to the right. That is, $E_{\text{max}}^* = t_{178,\,.0001366} = 3.714$; this critical value contrasts with $t_{178.,\,.025} = 1.973$, which would be appropriate for testing an individual studentized residual identified in advance of inspecting the data.

11.3.2 Anscombe's Insurance Analogy

Thus far, I have treated the identification (and, implicitly, the potential correction, removal, or accommodation) of outliers as a hypothesis-testing problem. Although this is by far the most common procedure in practice, a more reasonable (if subtle) general approach is to assess the potential costs and benefits for estimation of rejecting an unusual observation.

[6] It is not necessary literally to perform n auxiliary regressions. Equation 11.3, for example, permits the computation of studentized residuals with little effort.

[7] A graphical alternative is to construct a quantile-comparison plot for the studentized residuals, comparing the sample distribution of these quantities with the t-distribution for $n-k-2$ degrees of freedom. See the discussion of nonnormality in Section 12.1.

Imagine, for the moment, that the observation with the largest E_i^* is simply an unusual data point, but one generated by the assumed statistical model:

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

with independent errors ε_i that are each distributed as $N(0, \sigma_\varepsilon^2)$. To discard an observation under these circumstances would decrease the efficiency of estimation, because when the model—including the assumption of normality—is correct, the least-squares estimators are maximally efficient among all unbiased estimators of the regression coefficients.

If, however, the observation in question does not belong with the rest (e.g., because the mean-shift model applies), then to eliminate it may make estimation more efficient. Anscombe (1960) developed this insight by drawing an analogy to insurance: To obtain protection against "bad" data, one purchases a policy of outlier rejection, a policy paid for by a small premium in efficiency when the policy inadvertently rejects "good" data.[8]

Let q denote the desired premium, say 0.05—that is, a 5% increase in estimator mean-squared error if the model holds for all of the data. Let z represent the unit-normal deviate corresponding to a tail probability of $q(n-k-1)/n$. Following the procedure derived by Anscombe and Tukey (1963), compute $m = 1.4 + 0.85z$, and then find

$$E'_q = m\left(1 - \frac{m^2 - 2}{4(n-k-1)}\right)\sqrt{\frac{n-k-1}{n}} \qquad [11.4]$$

The largest absolute *standardized* residual can be compared with E'_q to determine whether the corresponding observation should be rejected as an outlier. This cutoff can be translated to the studentized-residual scale using Equation 11.3:

$$E_q^* = E'_q\sqrt{\frac{n-k-2}{n-k-1-E_q'^2}} \qquad [11.5]$$

In a real application, of course, we should inquire about discrepant observations rather than simply throwing them away.[9]

For example, for Davis's regression of reported on measured weight, $n = 183$ and $k = 3$; so, for the premium $q = 0.05$, we have

$$\frac{q(n-k-1)}{n} = \frac{0.05(183-3-1)}{183} = 0.0489$$

From the unit-normal table, $z = 1.66$, from which $m = 1.4 + 0.85 \times 1.66 = 2.81$. Then, using Equation 11.4, $E'_q = 2.76$, and using Equation 11.5, $E_q^* = 2.81$. Because $E_{\max}^* = |E_{12}^*| = 24.3$ is much larger than E_q^*, the 12th observation is identified as an outlier.

[8] An alternative is to employ a robust estimator, which is a bit less efficient than least squares when the model is correct, but much more efficient when outliers are present. See Section 14.3.

[9] See the discussion in Section 11.7.

A regression *outlier* is an observation with an unusual dependent-variable value given its combination of independent-variable values. The studentized residuals E_i^* can be used to identify outliers, through graphical examination, a Bonferroni test for the largest absolute E_i^*, or Anscombe's insurance analogy. If the model is correct (and there are no true outliers), then each studentized residual follows a t-distribution with $n - k - 2$ degrees of freedom.

11.4 Measuring Influence

As noted previously, influence on the regression coefficients combines leverage and discrepancy. The most direct measure of influence simply expresses the impact on each coefficient of deleting each observation in turn:

$$D_{ij} = B_j - B_{j(-i)} \quad \text{for } i = 1, \ldots, n \text{ and } j = 0, 1, \ldots, k$$

where the B_j are the least-squares coefficients calculated for all of the data, and the $B_{j(-i)}$ are the least-squares coefficients calculated with the ith observation omitted. (So as not to complicate the notation here, I denote the least-squares intercept A as B_0.) To assist in interpretation, it is useful to scale the D_{ij} by (deleted) estimates of the coefficient standard errors:

$$D_{ij}^* = \frac{D_{ij}}{\widehat{\text{SE}}_{(-i)}(B_j)}$$

Following Belsley, et al. (1980), the D_{ij} are often termed DFBETA$_{ij}$, and the D_{ij}^* are called DFBETAS$_{ij}$.

One problem associated with using the D_{ij} or the D_{ij}^* is their large number—$n(k+1)$ of each. Of course, these values can be more quickly and effectively examined graphically than in numerical tables. We can, for example, construct an *index plot* of the D_{ij}^*'s for each coefficient, $j = 0, 1, \ldots, k$—a simple scatterplot with D_{ij}^* on the vertical axis versus the observation index i on the horizontal axis. A more informative, if more complex, alternative is to construct a scatterplot matrix of the D_{ij}^* with index plots (or some other univariate display) on the diagonal.[10] Nevertheless, it is useful to have a single summary index of the influence of each observation on the least-squares fit.

Cook (1977) has proposed measuring the "distance" between the B_j and the corresponding $B_{j(-i)}$ by calculating the F-statistic for the "hypothesis" that $\beta_j = B_{j(-i)}$, for $j = 0, 1, \ldots, k$. This statistic is recalculated for each observation

[10] This interesting display was suggested to me by Michael Friendly of the Psychology Department, York University.

$i = 1, \ldots, n$. The resulting values should not literally be interpreted as F-tests—Cook's approach merely exploits an analogy to testing to produce a measure of distance that is independent of the scales of the X variables. Cook's statistic can be written (and simply calculated) as

$$D_i = \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

In effect, the first term in the formula for Cook's D is a measure of discrepancy, and the second is a measure of leverage. We look for values of D_i that are substantially larger than the rest.

> Observations that combine high leverage with a large studentized residual exert substantial *influence* on the regression coefficients. Cook's D-statistic provides a summary index of influence on the coefficients.

Belsley et al. (1980) have suggested the very similar measure[11]

$$\text{DFFITS}_i = E_i^* \sqrt{\frac{h_i}{1-h_i}}$$

Except for unusual data configurations, $D_i \simeq \text{DFFITS}_i^2/(k+1)$.

Because all of the deletion statistics depend on the hat values and residuals, a graphical alternative to either of these general influence measures is to plot the E_i^* against the h_i and to look for observations for which both are big. A slightly more sophisticated (and more informative) version of this plot displays circles of area proportional to Cook's D instead of points (see Figure 11.6 on page 285). We can follow up by examining the D_{ij} or D_{ij}^* for the observations with the largest few D_i, $|\text{DFFITS}_i|$, or combination of large h_i and $|E_i^*|$.

For Davis's regression of reported weight on measured weight, all of the indices of influence point to the obviously discrepant 12th observation:

$$\begin{aligned}
\text{Cook's } D_{12} &= 85.9 \text{ (next largest, } D_{21} = 0.065) \\
\text{DFFITS}_{12} &= -38.4 \text{ (next largest, DFFITS}_{50} = 0.512) \\
\text{DFBETAS}_{0,12} &= \text{DFBETAS}_{1,12} = 0 \\
\text{DFBETAS}_{2,12} &= 20.0, \text{DFBETAS}_{3,12} = -24.8
\end{aligned}$$

Notice that the outlying observation 12, which is for a female subject, has no impact on the male intercept B_0 (i.e., A) and slope B_1.

[11] Other global measures of influence are available (see Chatterjee and Hadi, 1988, Chapter 4, for a comparative treatment).

11.4.1 Influence on Standard Errors

In developing the concept of influence in regression, I have focused on changes in regression coefficients. Other regression outputs are also subject to influence, however. One important regression output is the set of coefficient sampling variances and covariances, which capture the precision of estimation in regression.

Recall, for example, Figure 11.1(c), in which a high-leverage observation exerts no influence on the regression coefficients because it is in line with the rest of the data. Recall, as well, that the estimated standard error of the least-squares slope in simple regression is

$$\widehat{\text{SE}}(B) = \frac{S_E}{\sqrt{\sum(X_i - \overline{X})^2}}$$

By increasing the variance of X, therefore, a high-leverage in-line observation serves to decrease $\widehat{\text{SE}}(B)$ even though it does not influence the regression coefficients A and B. Depending on the context, such an observation may be considered beneficial—because it increases the precision of estimation—or it may cause us to exaggerate our confidence in the estimate B.

In multiple regression, we can examine the impact of deleting each observation in turn on the size of the joint confidence region for the regression coefficients.[12] The size of the joint confidence region is analogous to the length of a confidence interval for an individual regression coefficient, which, in turn, is proportional to the standard error of the coefficient. The squared length of a confidence interval is, therefore, proportional to the sampling variance of the coefficient, and, analogously, the squared size of a joint confidence region is proportional to the "generalized variance" of a set of coefficients.

An influence measure proposed by Belsley et al. (1980) closely approximates the squared ratio of volumes of the deleted and full-data confidence regions for the regression coefficients:[13]

$$\text{COVRATIO}_i = \frac{1}{(1 - h_i)\left(\dfrac{n - k - 2 + E_i^{*2}}{n - k - 1}\right)^{k+1}}$$

Observations that increase the precision of estimation have values of COVRATIO that are larger than 1; those that decrease the precision of estimation have values smaller than 1. Look for values of COVRATIO, therefore, that differ substantially from 1.

As was true of measures of influence on the regression coefficients, both the hat value and the (studentized) residual figure in COVRATIO. A large hat

[12] See Section 9.4.4 for a discussion of joint confidence regions.

[13] Alternative, similar measures have been suggested by several authors. Chatterjee and Hadi (1988, Chapter 4) provide a comparative discussion.

value produces a large COVRATIO, however, even when—indeed, especially when—the studentized residual is small, because a high-leverage in-line observation improves the precision of estimation. In contrast, a discrepant, low-leverage observation might not change the coefficients much, but it decreases the precision of estimation by increasing the estimated error variance; such an observation, with small h_i and large E_i^*, produces a COVRATIO$_i$ substantially below 1.

For Davis's regression of reported weight on measured weight, sex, and their interaction, by far the most extreme value is COVRATIO$_{12} = 0.0103$. The 12th observation, therefore, *decreases* the precision of estimation by a factor of $1/0.0103 \simeq 100$. In this instance, a very large leverage, $h_{12} = 0.714$, is more than offset by a massive residual, $E_{12}^* = -24.3$.

11.4.2 Influence on Collinearity

Other characteristics of a regression analysis can also be influenced by individual observations, including the degree of collinearity among the independent variables.[14] I shall not address this issue in any detail, but the following points may prove helpful:[15]

- Influence on collinearity is one of the factors reflected in influence on coefficient standard errors. Measures such as COVRATIO, however, also reflect influence on the error variance and on the variation of the X's. Moreover, COVRATIO and similar measures examine the sampling variances and covariances of all of the regression coefficients, including the regression constant, while a consideration of collinearity generally excludes the constant. Nevertheless, our concern for collinearity reflects its impact on the precision of estimation, which is precisely what is addressed by COVRATIO.
- Collinearity-influential points are those that either induce or weaken correlations among the X's. Such points usually—but not always—have large hat values. Conversely, points with large hat values often influence collinearity.
- Individual points that induce collinearity are obviously problematic. More subtly, points that substantially weaken collinearity also merit examination, because they may cause us to be overly confident in our results.
- It is frequently possible to detect collinearity-influential points by plotting independent variables against each other, as in a scatterplot matrix or a three-dimensional rotating plot. This approach may fail, however, if the collinear relations in question involve more than two or three independent variables at a time.

11.5 Numerical Cutoffs for Diagnostic Statistics

I have deliberately refrained from suggesting specific numerical criteria for identifying noteworthy observations on the basis of measures of leverage and influence: I believe that it is generally more effective to examine the distributions of these quantities directly to locate unusual values. For studentized residuals, the hypothesis-testing and insurance approaches provide numerical cutoffs, but even these criteria are no substitute for graphical examination of the residuals.

[14] See Chapter 13 for a general treatment of collinearity.

[15] See Chatterjee and Hadi (1988, Chapter 4 and 5) for more information about influence on collinearity.

Nevertheless, numerical cutoffs can be of some use, as long as they are not given too much weight, and especially when they are employed to enhance graphical displays. A line can be drawn on a graph at the value of a numerical cutoff, and observations that exceed the cutoff can be identified individually.[16]

Cutoffs for a diagnostic statistic may be derived from statistical theory, or they may result from examination of the sample distribution of the statistic. Cutoffs may be absolute, or they may be adjusted for sample size.[17] For some diagnostic statistics, such as measures of influence, absolute cutoffs are unlikely to identify noteworthy observations in large samples. In part, this characteristic reflects the ability of large samples to absorb discrepant data without changing the results substantially, but it is still often of interest to identify *relatively* influential points, even if no observation has strong *absolute* influence.

The cutoffs presented below are, as explained briefly here, derived from statistical theory. An alternative, very simple, and universally applicable data-based criterion is to examine the most extreme (e.g., 5% of) values of a diagnostic statistic.

11.5.1 Hat Values

Belsley et al. (1980) suggest that hat values exceeding about twice the average $\bar{h} = (k + 1)/n$ are noteworthy. This size-adjusted cutoff was derived as an approximation identifying the most extreme 5% of cases when the X's are multivariate normal, and the number of regressors k and degrees of freedom for error $n - k - 1$ are relatively large. The cutoff is nevertheless recommended by these authors as a rough general guide even when the regressors are not normally distributed. In small samples, using $2 \times \bar{h}$ tends to nominate too many points for examination, and $3 \times \bar{h}$ can be used instead.[18]

11.5.2 Studentized Residuals

Beyond the issues of "statistical significance" and estimator robustness and efficiency discussed above, it sometimes helps to call attention to residuals that are relatively large. Recall that, under ideal conditions, about 5% of studentized residuals are outside the range $|E_i^*| \leq 2$. It is, therefore, reasonable, for example, to draw lines at ± 2 on a display of studentized residuals to draw attention to observations outside this range.

11.5.3 Measures of Influence

Many cutoffs have been suggested for different measures of influence. A few are presented here:

- *Standardized change in regression coefficients.* The D_{ij}^* are scaled by standard errors, and, consequently, $|D_{ij}^*| > 1$ or 2 suggests itself as an absolute cutoff. As explained above, however, this criterion is unlikely to nominate observations

[16] An example appears in Figure 11.6 on page 285.

[17] See Belsley et al. (1980, Chapter 2) for further discussion of these distinctions.

[18] See Chatterjee and Hadi (1988, Chapter 4) for a discussion of alternative cutoffs for hat values.

in large samples. Belsley et al. (1980) propose the size-adjusted cutoff $2/\sqrt{n}$ for identifying noteworthy D_{ij}^*'s.

- *Cook's D and DFFITS.* Several numerical cutoffs have been recommended for Cook's D and for DFFITS—exploiting the analogy between D and an F-statistic, for example. Chatterjee and Hadi (1988) suggest the size-adjusted cutoff[19]

$$|\text{DFFITS}_i| > 2\sqrt{\frac{k+1}{n-k-1}}$$

Because of the approximate relationship between DFFITS and Cook's D, it is simple to translate this criterion into

$$D_i > \frac{4}{n-k-1}$$

Absolute cutoffs for D, such as $D_i > 1$, risk missing relatively influential data.

- *COVRATIO.* Belsley et al. (1980) suggest the size-adjusted cutoff

$$|\text{COVRATIO}_i - 1| > \frac{3(k+1)}{n}$$

11.6 Joint Influence and Partial-Regression Plots

As illustrated in Figure 11.4, subsets of observations can be *jointly influential* or can offset each other's influence. Influential subsets or multiple outliers can often be identified by applying single-observation diagnostics, such as Cook's D and studentized residuals, sequentially. It can be important, however, to refit the model after deleting each point, because the presence of a single influential value can dramatically affect the fit at other points. Still, the sequential approach is not always successful.

Although it is possible to generalize deletion statistics to subsets of several points, the very large number of subsets usually renders this approach impractical.[20] An attractive alternative is to employ graphical methods, and a particularly useful influence graph is the *partial-regression plot* (also called a *partial-regression leverage plot* or an *added-variable plot*).

Let $Y_i^{(1)}$ represent the residuals from the least-squares regression of Y on all of the X's with the exception of X_1—that is, the residuals from the fitted regression equation

$$Y_i = A^{(1)} + B_2^{(1)}X_{i2} + \cdots + B_k^{(1)}X_{ik} + Y_i^{(1)}$$

[19] Also see Cook (1977), Belsley et al. (1980), and Velleman and Welsch (1981).

[20] Cook and Weisberg (1980), for example, extend the D-statistic to a subset of p observations indexed by the vector subscript $\mathbf{i} = (i_1, i_2, \ldots, i_p)'$:

$$D_{\mathbf{i}} = \frac{\mathbf{d}_{\mathbf{i}}'(\mathbf{X}'\mathbf{X})\mathbf{d}_{\mathbf{i}}}{(k+1)S_E^2}$$

where $\mathbf{d}_{\mathbf{i}} = \mathbf{b} - \mathbf{b}_{(-\mathbf{i})}$ gives the impact on the regression coefficients of deleting the subset $\mathbf{i}$. See Belsley et al. (1980, Chapter 2) and Chatterjee and Hadi (1988) for further discussion of deletion diagnostics based on subsets of observations. Note that there are $n!/[p!(n-p)!]$ subsets of size p—typically a prohibitively large number, even for modest values of p.

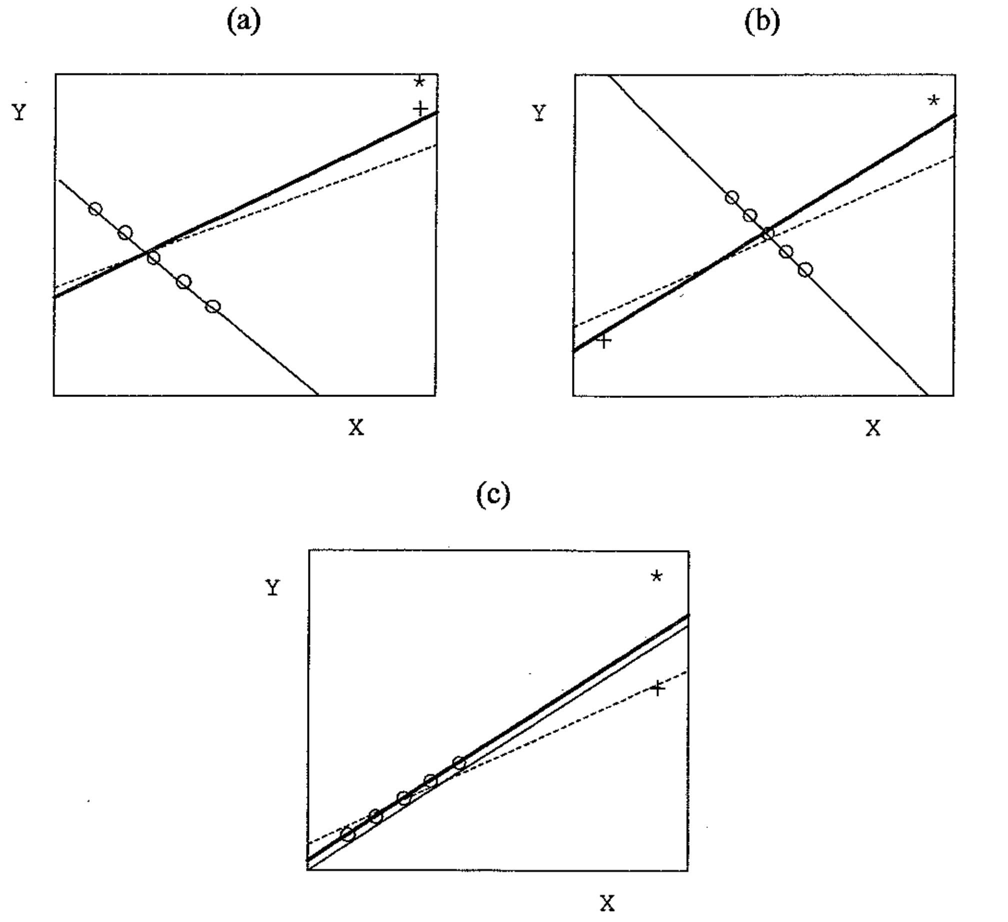

Figure 11.4. Jointly influential data in simple regression. In each graph, the heavy solid line gives the least-squares regression for all of the data; the broken line gives the regression with the asterisk deleted; and the light solid line gives the regression with both the asterisk and the plus deleted. (*a*) Jointly influential observations located close to one another: Deletion of both observations has a much greater impact than deletion of only one. (*b*) Jointly influential observations located on opposite sides of the data. (*c*) Observations that offset one another: The regression with both observations deleted is the same as for the whole dataset (the two lines are separated slightly for visual effect).

The parenthetical superscript (1) indicates the omission of X_1 from the right-hand side of the regression equation. Likewise, $X_i^{(1)}$ is the residual from the least-squares regression of X_1 on all the other X's:

$$X_{i1} = C^{(1)} + D_2^{(1)} X_{i2} + \cdots + D_k^{(1)} X_{ik} + X_i^{(1)}$$

The notation emphasizes the interpretation of the residuals $Y^{(1)}$ and $X^{(1)}$ as the parts of Y and X_1 that remain when the effects of $X_2, \ldots, X_k$ are "removed." The residuals $Y^{(1)}$ and $X^{(1)}$ have the following interesting properties:

1. The slope from the least-squares regression of $Y^{(1)}$ on $X^{(1)}$ is simply the least-squares slope B_1 from the full multiple regression.

2. The residuals from the simple regression of $Y^{(1)}$ on $X^{(1)}$ are the same as those from the full regression; that is,

$$Y_i^{(1)} = B_1 X_i^{(1)} + E_i \qquad [11.6]$$

No constant is required here, because both $Y^{(1)}$ and $X^{(1)}$ are least-squares residuals and therefore have means of 0.

3. The variation of $X^{(1)}$ is the conditional variation of X_1 holding the other X's constant and, as a consequence, the standard error of B_1 in the auxiliary simple regression (Equation 11.6),

$$\widehat{\mathrm{SE}(B_1)} = \frac{S_E}{\sqrt{\sum X_i^{(1)2}}}$$

is the same[21] as the multiple-regression standard error of B_1. Unless X_1 is uncorrelated with the other X's, its conditional variation is smaller than its marginal variation—much smaller, if X_1 is strongly collinear with the other X's.

Plotting $Y^{(1)}$ against $X^{(1)}$ permits us to examine leverage and influence on B_1. Because of properties 1–3, this plot also provides a visual impression of the precision of the estimate B_1. Similar partial-regression plots can be constructed for the other regressors:[22]

$$\text{Plot } Y^{(j)} \text{ versus } X^{(j)} \text{ for each } j = 1, \dots, k$$

Subsets of observations can be jointly influential. Partial-regression plots are useful for detecting joint influence on the regression coefficients. The partial-regression plot for the regressor X_j is formed using the residuals from the least-squares regressions of X_j and Y on all of the other X's.

Illustrative partial-regression plots are shown in Figure 11.5, using data from Duncan's regression of occupational prestige on the income and educational levels of 45 U.S. occupations. Recall (from Chapter 5) that Duncan's regression yields the following least squares fit:

$$\begin{aligned}\widehat{\text{Prestige}} = \underset{(4.27)}{-6.06} + \underset{(0.120)}{0.599} \times \text{ Income} + \underset{(0.098)}{0.546} \times \text{ Education}\end{aligned}$$

$$R^2 = 0.83 \qquad S_E = 13.4$$

The partial-regression plot for income [Figure 11.5(a)] reveals three observations that exert substantial leverage on the income coefficient. Two of these observations serve to decrease the income slope: *ministers*, whose income is un-

[21] There is slight slippage here with respect to the degrees of freedom for error: S_E is from the multiple regression, with $n - k - 1$ degrees of freedom for error. We need not subtract the mean of $X_i^{(1)}$ to calculate the standard error of the slope since the mean is already 0.

[22] We can also construct a partial-regression plot for the intercept A, by regressing the "constant regressor" $X_0 = 1$ and Y on X_1 through X_k, with no constant in these regression equations.

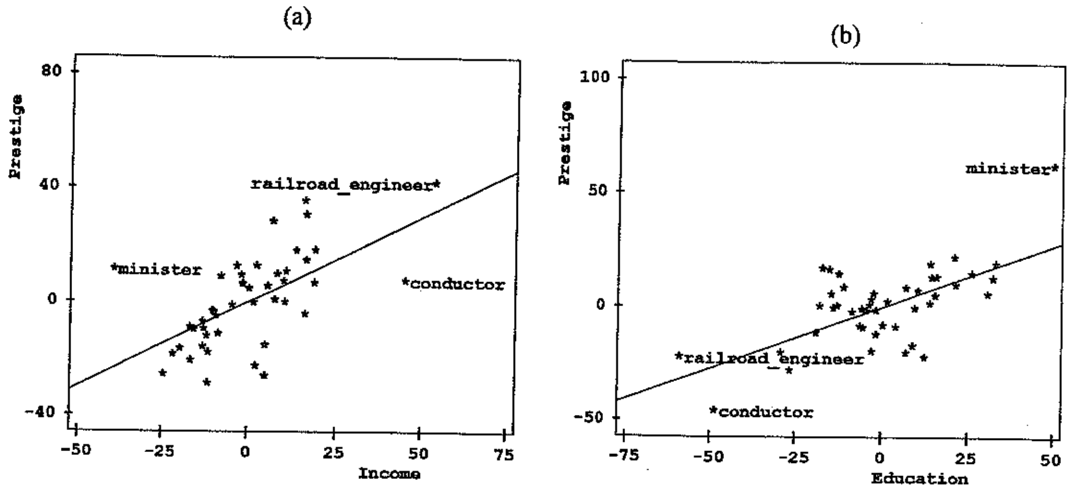

Figure 11.5. Partial-regression plots for Duncan's regression of occupational prestige on the income (*a*) and educational levels (*b*) of 45 U.S. occupations in 1950. Three potentially influential observations (*ministers, railroad conductors*, and *railroad engineers*) are identified on the plots. The partial-regression plot for the intercept *A* is not shown.

usually low given the educational level of the occupation; and *railroad conductors*, whose income is unusually high given education. The third occupation, *railroad engineers*, is above the fitted regression, but is not as discrepant; it, too, has relatively high income given education. Remember that the horizontal variable in this partial-regression plot is the residual from the regression of income on education, and thus values far from 0 in this direction are for occupations with incomes that are unusually high or low given their levels of education.

The partial-regression plot for education [Figure 11.5(*b*)] shows that the same three observations have relatively high leverage on the education coefficient: *Ministers* and *railroad conductors* tend to increase the education slope, while *railroad engineers* appear to be closer in line with the rest of the data.

Examining the single-observation deletion statistics for Duncan's regression reveals that *ministers* have the largest Cook's D ($D_6 = 0.566$) and the largest studentized residual ($E_6^* = 3.14$). This studentized residual is not especially big, however: The Bonferroni p-value for the outlier test is $\Pr(t_{41} > 3.14) \times 2 \times 45 = 0.14$. Figure 11.6 displays a plot of studentized residuals versus hat values, with the areas of the plotted circles proportional to values of Cook's D. The lines on the plot are at $E^* = \pm 2$ (on the vertical axis) and at $h = 2\bar{h}$ and $3\bar{h}$ (on the horizontal axis). Four observations that exceed these cutoffs are identified on the plot. *Reporters* have a relatively large residual but are at a low-leverage point, while *railroad engineers* have high leverage but a small studentized residual.

Deleting *ministers* and *conductors* produces the fitted regression

$$\begin{array}{l}\widehat{\text{Prestige}} = \underset{(3.65)}{-6.41} + \underset{(0.122)}{0.867} \times \text{ Income} + \underset{(0.099)}{0.332} \times \text{ Education} \\ \qquad R^2 = 0.88 \qquad S_E = 11.4\end{array}$$

which, as expected from the partial-regression plots, has a larger income slope and smaller education slope than the original regression. The estimated standard

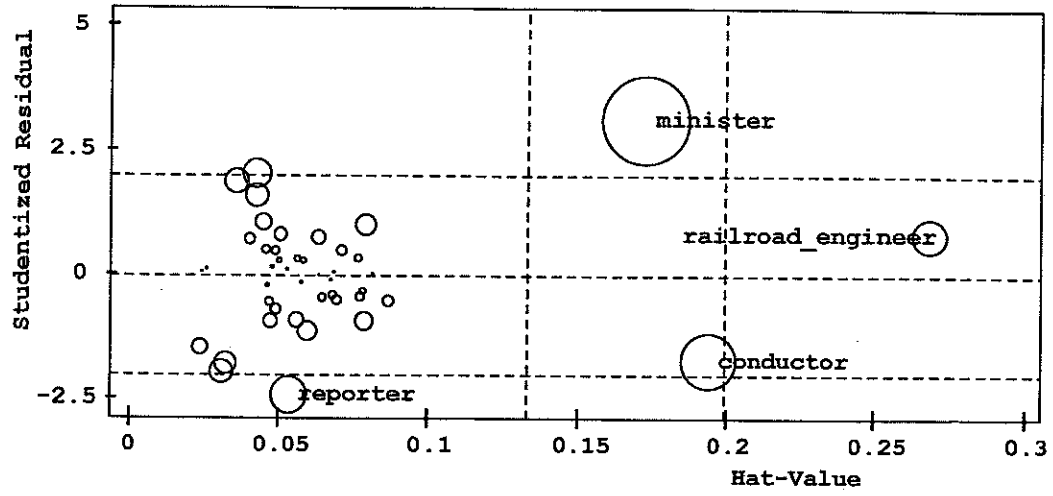

Figure 11.6. "Bubble plot" of Cook's D's, studentized residuals, and hat values, for Duncan's regression of occupational prestige on income and education. Each point is plotted as a circle with area proportional to D. Horizontal reference lines are drawn at studentized residuals of 0 and ±2; vertical reference lines are drawn at hat values of $2\bar{h}$ and $3\bar{h}$. Several observations are identified on the plot: *Ministers* and *conductors* have large hat values and relatively large residuals; *reporters* have a relatively large residual, but a small hat value; *railroad engineers* have a large hat value, but a small residual.

errors are likely optimistic, however, because relative outliers have been trimmed away. Deleting *railroad engineers,* along with *ministers* and *conductors,* further increases the income slope and decreases the education slope, but the change is not dramatic: $B_{\text{Income}} = 0.931$, $B_{\text{Education}} = 0.285$.

Partial-regression plots can be straightforwardly extended to pairs of regressors. We can, for example, regress each of X_1, X_2, and Y on the remaining regressors, $X_3, \ldots, X_k$, obtaining residuals $X_{i1}^{(12)}$, $X_{i2}^{(12)}$, and $Y_i^{(12)}$; $Y^{(12)}$ is then plotted against $X_1^{(12)}$ and $X_2^{(12)}$ to produce a dynamic three-dimensional scatterplot on which the partial-regression plane can be displayed.[23]

11.7 Should Unusual Data Be Discarded?

The discussion thus far in this chapter has implicitly assumed that outlying and influential data are simply discarded. Although problematic data should not be ignored, they also should not be deleted automatically and without reflection:

- It is important to investigate why an observation is unusual. Truly bad data (e.g., an error in data entry as in Davis's regression) can often be corrected or, if correction is not possible, thrown away. When a discrepant data point is correct, we may be able to understand why the observation is unusual. For Duncan's regression, for example, it makes sense that ministers enjoy prestige not accounted for by the income and educational levels of the occupation. In a case like this, we may choose to deal separately with an outlying observation.

[23] See Cook and Weisberg (1989) for a discussion of three-dimensional partial-regression plots. An alternative, two-dimensional extension of partial-regression plots to subsets of coefficients is described in Section 11.8.4.

- Alternatively, outliers or influential data may motivate model respecification. For example, the pattern of outlying data may suggest the introduction of additional independent variables. If, in Duncan's regression, we can identify a variable that produces the unusually high prestige of ministers (net of their income and education), and if we can measure that variable for other observations, then the variable could be added to the regression. In some instances, transformation of the dependent variable or of an independent variable may draw apparent outliers toward the rest of the data, by rendering the error distribution more symmetric or by eliminating nonlinearity. We must, however, be careful to avoid "overfitting" the data—permitting a small portion of the data to determine the form of the model.[24]
- Except in clear-cut cases, we are justifiably reluctant to delete observations or to respecify the model to accommodate unusual data. Some researchers reasonably adopt alternative estimation strategies, such as robust regression, which continuously downweights outlying data rather than simply discarding them. Because these methods assign zero or very small weight to highly discrepant data, however, the result is generally not very different from careful application of least squares, and, indeed, robust-regression weights can be used to identify outliers.[25]

> Outlying and influential data should not be ignored, but they also should not simply be deleted without investigation. "Bad" data can often be corrected. "Good" observations that are unusual may provide insight into the structure of the data, and may motivate respecification of the statistical model used to summarize the data.

EXERCISES

11.1 Employ the methods of this chapter to look for unusual data in each of the following linear-model analyses. In each case, consider the impact of any unusual data that you discover on the results of the analysis, and—within the limits of your knowledge of the datasets—suggest how unusual data should be treated.

(a) Sahlins's regression of acres tended per gardener on consumers per gardener for the households of Mazulu village. (See Exercise 5.6; the data are in Table 2.1 and `sahlins.dat`.)

(b) Angell's dummy-variable regression of moral integration of U.S. cities on ethnic heterogeneity, geographic mobility, and region. (See Exercise 7.3; the data are in Table 2.3 and `angell.dat`.) Are partial-regression plots for dummy regressors interpretable? If so, how?

[24] See Chapter 12.

[25] See Sections 2.3 and 14.3.

(c) Moore and Krupat's two-way analysis of variance of conformity by authoritarianism and partner's status. (See Section 8.2; the data are in Table 8.3 and `moore.dat`.) Repeat your analysis for the analysis-of-covariance model fit to Moore and Krupat's data, treating authoritarianism as a covariate. (See Section 8.4.) Are partial-regression plots for deviation-coded regressors interpretable? If so, how?

(d) Anscombe's regression of state education expenditures on income, proportion under 18, and proportion urban. (See Exercise 5.14; the data are in Table 5.1 and `anscombe.dat`.)

11.2 In order to test two theories of peasant revolt, Chirot and Ragin (1975) gathered data (in Table 11.1 and `chirot.dat`) on a 1907 rebellion in 32 counties of Romania.[26] The dependent variable in their analysis was the intensity of the rebellion (I), an index constructed from the reported level of violence and the degree to which the rebellion spread within a county. According to the "transitional society" theory of peasant rebellion, intensity should be high when *both* the level of commercialization of agriculture (C) and the level of traditionalism (T) are high. Commercialization—the penetration of market forces—was measured by the percentage of land in the county devoted to cultivation of wheat, the major cash crop raised in the region. Traditionalism was measured by the percentage of illiterates. The "structural" theory of peasant revolt implies that the rebellion should be intense where middle peasants (M) are relatively strong and where the inequality of land tenure (G) is high. The strength of the middle peasantry was assessed by the percentage of rural households owning between 7 and 50 hectares of land; inequality of land tenure was measured by a Gini coefficient. Chirot and Ragin tested the two theories by regressing I on C, T, the product of C and T (i.e., $C \times T$), M, and G. The first theory predicts a positive coefficient for $C \times T$, while the second predicts positive coefficients for M and G. Redo Chirot and Ragin's linear-model analysis, using the methods of this chapter to look for unusual data. Consider the impact of any unusual observations that you discover on the results of the study.

11.8 Some Statistical Details*

11.8.1 Hat Values and the Hat Matrix

Recall, from Chapter 9, the matrix form of the general linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Recall, as well, that the fitted model is given by $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, in which the vector of least-squares estimates is $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

The least-squares fitted values are therefore a linear function of the observed dependent-variable values:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

[26] I am grateful to Michael Gillespie of the Sociology Department, University of Alberta, for bringing this dataset to my attention.

TABLE 11.1 Data on the 1907 Romanian Peasant Rebellion: *I*, Intensity of the Rebellion (Corrected from the original); *C*, Commercialization of Agriculture; *T*, Traditionalism; *M*, Market Forces; and *G*, Inequality of Land Tenure.

County	*I*	*C*	*T*	*M*	*G*
1	−1.39	13.8	86.2	6.2	0.60
2	0.65	20.4	86.7	2.9	0.72
3	1.89	27.6	79.3	16.9	0.66
4	−0.15	18.6	90.1	3.4	0.74
5	−0.86	17.2	84.5	9.0	0.70
6	0.11	21.5	81.5	5.2	0.60
7	−0.51	11.6	82.6	5.1	0.52
8	−0.86	20.4	82.4	6.3	0.64
9	−0.24	19.5	87.5	4.8	0.68
10	−0.77	8.9	85.6	9.5	0.58
11	−0.24	25.8	82.2	10.9	0.68
12	−1.57	24.1	83.5	8.4	0.74
13	−0.51	2.0	88.3	6.2	0.70
14	−1.57	24.2	84.9	6.1	0.62
15	−0.51	30.6	76.1	1.3	0.76
16	−1.13	33.9	85.5	5.8	0.70
17	−1.22	28.6	84.2	2.9	0.58
18	−1.22	36.5	78.1	4.3	0.72
19	−0.86	40.9	84.4	2.3	0.64
20	−1.39	6.8	76.3	3.6	0.58
21	2.81	41.9	89.7	6.6	0.66
22	−1.04	25.4	83.2	2.5	0.68
23	1.57	30.5	80.2	4.1	0.76
24	4.32	48.2	91.0	4.2	0.70
25	3.79	46.0	90.5	3.7	0.68
26	3.79	45.1	85.5	5.1	0.64
27	−1.75	12.5	83.8	7.2	0.50
28	0.82	39.3	85.6	4.9	0.60
29	2.59	47.7	87.6	5.2	0.58
30	−0.86	15.2	87.3	10.8	0.42
31	−1.84	11.7	82.3	81.7	0.42
32	−1.84	25.6	80.1	68.4	0.26

Source of Data: Chirot and Ragin (1975).

Here, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the ***hat matrix***, so named because it transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$. The hat matrix is symmetric ($\mathbf{H} = \mathbf{H}'$) and idempotent ($\mathbf{H}^2 = \mathbf{H}$), as can easily be verified.[27] Consequently, the diagonal entries of the hat matrix $h_i \equiv h_{ii}$, which we called the ***hat values***, are

$$h_i \equiv \mathbf{h}_i'\mathbf{h}_i = \sum_{j=1}^{n} h_{ij}^2 = h_i^2 + \sum_{j \neq i} h_{ij}^2 \qquad [11.7]$$

where (because of symmetry) the elements of $\mathbf{h}_i$ comprise both the *i*th row and the *i*th column of $\mathbf{H}$.

Equation 11.7 implies that $0 \leq h_i \leq 1$. If the model matrix $\mathbf{X}$ includes the constant regressor, then $1/n \leq h_i$. Because $\mathbf{H}$ is a projection matrix,[28] projecting

[27] See Exercise 11.4.

[28] See Chapter 10, on the vector geometry of linear models.

$\mathbf{y}$ orthogonally onto the $(k+1)$-dimensional subspace spanned by the columns of $\mathbf{X}$, it follows that $\sum h_i = k+1$, and thus $\bar{h} = (k+1)/n$ (as stated in Section 11.2).

I mentioned as well that when there are several independent variables in the model, the leverage h_i of the ith observation is directly related to the distance of this observation from the center of the independent-variable scatter. To demonstrate this property of the hat-values, it is convenient to rewrite the fitted model with all variables in mean-deviation form: $\mathbf{y}^* = \mathbf{X}^*\mathbf{b}_1 + \mathbf{e}$, where $\mathbf{y}^* \equiv \{Y_i - \overline{Y}\}$ is the "centered" dependent-variable vector; $\mathbf{X}^* \equiv \{X_{ij} - \overline{X}_j\}$ contains the centered independent variables, but no constant regressor, which is no longer required; and $\mathbf{b}_1$ is the vector of least-squares slopes (suppressing the regression intercept). Then the hat value for the ith observation is

$$h_i^* = \mathbf{h}_i^{*\prime}\mathbf{h}_i^* = \mathbf{x}_i^{*\prime}(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{x}_i^* = h_i - \frac{1}{n}$$

where $\mathbf{x}_i^{*\prime} = [X_{i1} - \overline{X}_1, \ldots, X_{ik} - \overline{X}_k]$ is the ith row of $\mathbf{X}^*$ (and $\mathbf{x}_i^*$ is the ith row of $\mathbf{X}^*$ written as a column vector).

As Weisberg (1985, p. 112) has pointed out, $(n-1)h_i^*$ is the *generalized* or *Mahalanobis distance* between $\mathbf{x}_i'$ and $\overline{\mathbf{x}}'$, where $\overline{\mathbf{x}}' = [\overline{X}_1, \ldots, \overline{X}_k]$ is the mean vector or *centroid* of the independent variables. The Mahalanobis distances, and hence the hat values, do not change if the independent variables are rescaled. Indeed, the Mahalanobis distances and hat values are invariant with respect to any nonsingular linear transformation of $\mathbf{X}$.

11.8.2 The Distribution of the Least-Squares Residuals

The least-squares residuals are given by

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}}\\ &= (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\\ &= (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}\end{aligned}$$

Thus,

$$E(\mathbf{e}) = (\mathbf{I} - \mathbf{H})E(\boldsymbol{\varepsilon}) = (\mathbf{I} - \mathbf{H})\mathbf{0} = \mathbf{0}$$

and

$$V(\mathbf{e}) = (\mathbf{I} - \mathbf{H})V(\mathbf{e})(\mathbf{I} - \mathbf{H})' = \sigma_\varepsilon^2(\mathbf{I} - \mathbf{H})$$

because $\mathbf{I} - \mathbf{H}$, like $\mathbf{H}$ itself, is symmetric and idempotent. The matrix $\mathbf{I} - \mathbf{H}$ is not diagonal, and therefore the residuals are generally correlated, even when the errors are (as assumed here) independent. The diagonal entries of $\mathbf{I} - \mathbf{H}$ generally differ from one another, and so the residuals generally have different variances (as stated in Section 11.3):[29] $V(E_i) = \sigma_\varepsilon^2(1 - h_i)$.

[29] Balanced ANOVA models are an exception: Here, all the hat values are equal. (Why?)

11.8.3 Deletion Diagnostics

Let $\mathbf{b}_{(-i)}$ denote the vector of least-squares regression coefficients calculated with the ith observation omitted. Then $\mathbf{d}_i \equiv \mathbf{b} - \mathbf{b}_{(-i)}$ represents the influence of observation i on the regression coefficients. The influence vector $\mathbf{d}_i$ can be calculated efficiently as[30]

$$\mathbf{d}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \frac{E_i}{1-h_i} \qquad [11.8]$$

where $\mathbf{x}_i'$ is the ith row of the model matrix $\mathbf{X}$ (and $\mathbf{x}_i$ is the ith row written as a column vector).

Cook's D_i is the F-statistic for testing the "hypothesis" that $\boldsymbol{\beta} = \mathbf{b}_{(-i)}$:

$$\begin{aligned} D_i &= \frac{(\mathbf{b} - \mathbf{b}_{(-i)})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}_{(-i)})}{(k+1)S_E^2} \\ &= \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})}{(k+1)S_E^2} \end{aligned}$$

An alternative interpretation of D_i, therefore, is that it measures the aggregate influence of observation i on the fitted values $\hat{\mathbf{y}}$. This is why Belsley et al. (1980) call their similar statistic "DFFITS." Using Equation 11.8,

$$\begin{aligned} D_i &= \frac{E_i^2}{S_E^2(k+1)} \times \frac{h_i}{(1-h_i)^2} \\ &= \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i} \end{aligned}$$

which is the formula for Cook's D given in Section 11.4.

11.8.4 Partial-Regression Plots

In vector form, the fitted multiple-regression model is

$$\begin{aligned} \mathbf{y} &= A\mathbf{1}_n + B_1\mathbf{x}_1 + B_2\mathbf{x}_2 + \cdots + B_k\mathbf{x}_k + \mathbf{e} \qquad [11.9] \\ &= \hat{\mathbf{y}} + \mathbf{e} \end{aligned}$$

where the fitted-value vector $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the subspace spanned by the regressors[31] $\mathbf{1}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. Let $\mathbf{y}^{(1)}$ and $\mathbf{x}^{(1)}$ be the projections of $\mathbf{y}$ and $\mathbf{x}_1$, respectively, onto the orthogonal complement of the subspace spanned by $\mathbf{1}_n$ and $\mathbf{x}_2, \ldots, \mathbf{x}_k$ (i.e., the residual vectors from the least-squares regressions of Y and X_1 on the other X's). Then, by the geometry of projections, the orthogonal projection of $\mathbf{y}^{(1)}$ onto $\mathbf{x}^{(1)}$ is $B_1\mathbf{x}^{(1)}$, and

[30] See Exercise 11.5.

[31] See Chapter 10.

$\mathbf{y}^{(1)} - B_1\mathbf{x}^{(1)} = \mathbf{e}$, the residual vector from the overall least-squares regression, given in Equation 11.9.[32]

Sall (1990) suggests the following generalization of partial-regression plots, which he terms *leverage plots*: Consider the general linear hypothesis[33]

$$H_0\colon \underset{(q\times k+1)}{\mathbf{L}}\underset{(k+1\times 1)}{\boldsymbol{\beta}} = \underset{(q\times 1)}{\mathbf{0}} \qquad [11.10]$$

For example, in the regression of occupational prestige (Y) on education (X_1), income (X_2), and type of occupation (represented by the dummy regressors D_1 and D_2),

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 D_1 + \gamma_2 D_2 + \varepsilon$$

the hypothesis matrix

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

is used to test the hypothesis H_0: $\gamma_1 = \gamma_2 = 0$ that there is no effect of type of occupation.[34]

The residuals for the full model, unconstrained by the hypothesis (Equation 11.10), are the usual least-squares residuals, $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$. The estimated regression coefficients under the hypothesis are[35]

$$\mathbf{b}_0 = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\mathbf{u}$$

and the residuals constrained by the hypothesis are given by

$$\mathbf{e}_0 = \mathbf{e} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\mathbf{u}$$

where

$$\mathbf{u} \equiv [\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}\mathbf{Lb}$$

Thus, the incremental sum of squares for H_0 is[36]

$$||\mathbf{e}_0 - \mathbf{e}||^2 = \mathbf{b}'\mathbf{L}'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}\mathbf{Lb}$$

The leverage plot is a scatterplot with

$$\mathbf{v}_x \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\mathbf{u}$$

on the horizontal axis, and

$$\mathbf{v}_y \equiv \mathbf{v}_x + \mathbf{e}$$

[32] See Exercises 11.6 and 11.7.

[33] See Section 9.4.3.

[34] See Exercise 11.9.

[35] For this and other results pertaining to leverage plots, see Sall (1990).

[36] See Exercise 11.8.

on the vertical axis. The leverage plot, so defined, has the following properties:

- The residuals around the horizontal line at $V_y = 0$ are the constrained least-squares residuals E_{0i} under the hypothesis H_0.
- The least-squares line fit to the leverage plot has an intercept of 0 and a slope of 1; the residuals about this line are the unconstrained least-squares residuals, E_i. The incremental sum of squares for H_0 is thus the regression sum of squares for the line.
- When the hypothesis matrix $\mathbf{L}$ is formulated with a single row to test the coefficient of an individual regressor, the leverage plot specializes to the usual partial-regression plot, with the horizontal axis rescaled so that the least-squares intercept is 0 and the slope 1.

EXERCISES

11.3 Show that, in simple-regression analysis, the hat value is

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}$$

[*Hint*: Evaluate $\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ for $\mathbf{x}_i' = (1, X_i)$.]

11.4 Show that the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is symmetric ($\mathbf{H} = \mathbf{H}'$) and idempotent ($\mathbf{H}^2 = \mathbf{H}$).

11.5 Using Duncan's regression of occupational prestige on the educational and income levels of occupations (the data are in Table 3.2 and `duncan.dat`), verify that the influence vector for the deletion of *ministers* on the regression coefficients, $\mathbf{d}_i = \mathbf{b} - \mathbf{b}_{(-i)}$, can be written as

$$\mathbf{d}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \frac{E_i}{1 - h_i}$$

where $\mathbf{x}_i$ is the ith row of the model matrix $\mathbf{X}$ (i.e., the row for *ministers*) written as a column. [A much more difficult problem is to show that this formula works in general; see, e.g., Belsley, et al. (1980, pp. 69–83) or Velleman and Welsch (1981).]

11.6 *Consider the two-independent-variable linear-regression model, with variables written as vectors in mean-deviation form (as in Section 10.2): $\mathbf{y}^* = B_1\mathbf{x}_1^* + B_2\mathbf{x}_2^* + \mathbf{e}$. Let $\mathbf{x}^{(1)}$ and $\mathbf{y}^{(1)}$ represent the residual vectors from the regression (i.e., orthogonal projection) of $\mathbf{x}_1^*$ and $\mathbf{y}^*$, respectively, on $\mathbf{x}_2^*$. Drawing the three-dimensional diagram of the subspace spanned by $\mathbf{x}_1^*$, $\mathbf{x}_2^*$, and $\mathbf{y}^*$, prove geometrically that the coefficient for the orthogonal projection of $\mathbf{y}^{(1)}$ onto $\mathbf{x}^{(1)}$ is B_1.

11.7 *Now consider the more general model $\mathbf{y}^* = B_1\mathbf{x}_1^* + B_2\mathbf{x}_2^* + \cdots + B_k\mathbf{x}_k^* + \mathbf{e}$. Let $\mathbf{x}^{(1)}$ and $\mathbf{y}^{(1)}$ represent the residual vectors from the projections of $\mathbf{x}_1^*$ and $\mathbf{y}^*$, respectively, onto the subspace spanned by $\mathbf{x}_2^*, \ldots, \mathbf{x}_k^*$. Prove that the coefficient for the orthogonal projection of $\mathbf{y}^{(1)}$ onto $\mathbf{x}^{(1)}$ is B_1.

11.8 *Show that the incremental sum of squares for the general linear hypothesis H_0: $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ can be written as

$$||\mathbf{e}_0 - \mathbf{e}||^2 = \mathbf{b}'\mathbf{L}'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}\mathbf{L}\mathbf{b}$$

[*Hint*: $||\mathbf{e}_0 - \mathbf{e}||^2 = (\mathbf{e}_0 - \mathbf{e})'(\mathbf{e}_0 - \mathbf{e})$.]

11.9 Using Duncan's data on the prestige of 45 U. S. occupations, regress prestige on education, income, and two dummy variables to represent the effects of three occupational types. (See Section 7.2; Duncan's data are in Table 3.2 and `duncan.dat`.)

(a) Construct partial-regression plots for education, income, and the two dummy regressors for occupational type.

(b) Construct leverage plots for education, income, and occupational type. Confirm that the leverage plots for education and income are identical to the partial-regression plots in part (a), except for the scaling of the horizontal axis. Compare the information obtained from the leverage plot for occupational type with the two partial-regression plots for the occupational type coefficients in part (a).

11.9 Summary

- Unusual data are problematic in linear models fit by least squares because they can substantially influence the results of the analysis, and because they may indicate that the model fails to capture important features of the data.
- Observations with unusual combinations of independent-variable values have high *leverage* in a least-squares regression. The hat values h_i provide a measure of leverage. A rough cutoff for noteworthy hat values is $h_i > 2\bar{h} = 2(k+1)/n$.
- A regression *outlier* is an observation with an unusual dependent-variable value given its combination of independent-variable values. The studentized residuals E_i^* can be used to identify outliers, through graphical examination, a Bonferroni test for the largest absolute E_i^*, or Anscombe's insurance analogy. If the model is correct (and there are no true outliers), then each studentized residual follows a t-distribution with $n - k - 2$ degrees of freedom.
- Observations that combine high leverage with a large studentized residual exert substantial *influence* on the regression coefficients. Cook's D-statistic provides a summary index of influence on the coefficients. A rough cutoff for noteworthy values of D is $D_i > 4/(n - k - 1)$.
- It is also possible to investigate the influence of individual observations on other regression "outputs," such as coefficient standard errors and collinearity.

- Subsets of observations can be jointly influential. Partial-regression plots are useful for detecting joint influence on the regression coefficients. The partial-regression plot for the regressor X_j is formed using the residuals from the least-squares regressions of X_j and Y on all of the other X's.
- Outlying and influential data should not be ignored, but they also should not simply be deleted without investigation. "Bad" data can often be corrected. "Good" observations that are unusual may provide insight into the structure of the data, and may motivate respecification of the statistical model used to summarize the data.

11.10 Recommended Reading

There is a large journal literature on methods for identifying unusual and influential data. Fortunately, there are several texts that present this literature in a more digestible form:[37]

- Although it is now more than a decade old, Cook and Weisberg (1982) is, in my opinion, still the best book-length presentation of methods for assessing leverage, outliers, and influence. There are also good discussions of others problems, such as nonlinearity and transformations of the dependent and independent variables.
- Chatterjee and Hadi (1988) is a thorough and reasonably up-to-date text dealing primarily with influential data and collinearity; other problems—such as nonlinearity and nonconstant error variance—are treated briefly.
- Belsley, et al. (1980) is a seminal text that discusses influential data and the detection of collinearity.[38]
- Barnett and Lewis (1994) present an encyclopedic survey of methods for outlier detection, including methods for detecting outliers in linear models.

[37] Also see the recommended readings given at the end of the following chapter.

[38] I believe that Belsley et al.'s (1980) approach to diagnosing collinearity is fundamentally flawed—see the discussion of collinearity in Chapter 13.

12

Diagnosing Nonlinearity, Nonconstant Error Variance, and Nonnormality

Chapters 11, 12, and 13 show how to detect and correct problems with linear models that have been fit to data. The previous chapter focused on problems with specific observations. The current chapter and the next deal with more general problems with the specification of the model.

The first three sections of this chapter take up the problems of nonnormally distributed errors, nonconstant error variance, and nonlinearity. The treatment here stresses simple graphical methods for detecting these problems, along with transformations of the data to correct problems that are detected.

Subsequent sections describe tests of nonconstant error variance and nonlinearity for discrete independent variables; diagnostic methods based on embedding the usual linear model in a more general nonlinear model that incorporates transformations as parameters; and diagnostics that seek to detect the underlying dimensionality of the regression.

12.1 Nonnormally Distributed Errors

The assumption of normally distributed errors is almost always arbitrary. Nevertheless, the central-limit theorem assures that, under very broad conditions, inference based on the least-squares estimator is approximately valid in all but small samples. Why, then, should we be concerned about nonnormal errors?

- Although the *validity* of least-squares estimation is robust—the levels of tests and confidence intervals are approximately correct in large samples even when the assumption of normality is violated—the *efficiency* of least squares is not robust: Statistical theory assures us that the least-squares estimator is the most efficient unbiased estimator only when the errors are normal. For some types of error distributions, however, particularly those with heavy tails, the efficiency of least-squares estimation decreases markedly. In these cases, the least-squares estimator becomes much less efficient than robust estimators (or least-squares augmented by diagnostics).[1] To a substantial extent, heavy-tailed error distributions are problematic because they give rise to outliers, a problem that I addressed in the previous chapter.

A commonly quoted justification of least-squares estimation—called the Gauss-Markov theorem—states that the least-squares coefficients are the most efficient unbiased estimators that are *linear* functions of the observations Y_i. This result depends on the assumptions of linearity, constant error variance, and independence, but does not require the assumption of normality.[2] Although the restriction to linear estimators produces simple formulas for coefficient standard errors, it is not compelling in the light of the vulnerability of least squares to heavy-tailed error distributions.

- Highly skewed error distributions, aside from their propensity to generate outliers in the direction of the skew, compromise the interpretation of the least-squares fit. This fit is a conditional mean (of Y given the X's), and the mean is not a good measure of the center of a highly skewed distribution. Consequently, we may prefer to transform the data to produce a symmetric error distribution.
- A multimodal error distribution suggests the omission of one or more discrete independent variables that divide the data naturally into groups. An examination of the distribution of the residuals may, therefore, motivate respecification of the model.

Although there are tests for nonnormal errors, I shall instead describe graphical methods for examining the distribution of the residuals, employing univariate displays introduced in Chapter 3.[3] These methods are more useful for pinpointing the character of a problem and for suggesting solutions.

One such graphical display is the quantile comparison plot. We typically compare the sample distribution of the studentized residuals, E_i^*, with the quantiles of the unit-normal distribution, $N(0, 1)$, or with those of the t-distribution for $n - k - 2$ degrees of freedom. Unless n is small, of course, the normal and t-distributions are nearly identical. We choose to plot studentized residuals because they have equal variances and are t-distributed, but, in larger samples, standardized or raw residuals will convey much the same impression.

Even if the model is correct, however, the studentized residuals are not an independent random sample from t_{n-k-2}: Different residuals are correlated with one another.[4] These correlations depend on the configuration of the X-values, but they are generally negligible unless the sample size is small. Furthermore, at the cost of some computation, it is possible to adjust for the dependencies among the residuals in interpreting a quantile comparison plot.[5]

[1] Robust estimation is discussed in Section 14.3.

[2] A proof of the Gauss-Markov theorem appears in Section 9.3.2.

[3] See the discussion of Box-Cox transformations in Section 12.5.1, however.

[4] Different residuals are correlated because the off-diagonal entries of the hat-matrix (i.e., h_{ij} for $i \neq j$) are generally nonzero; see Section 11.8.

[5] See Section 12.1.1.

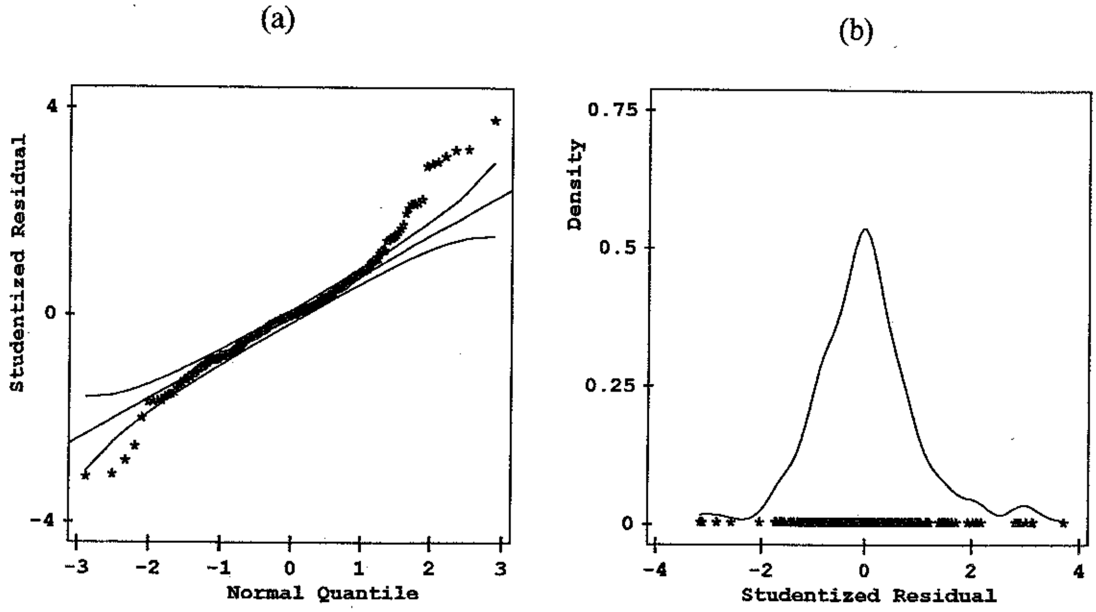

Figure 12.1. The distribution of the studentized residuals from Ornstein's interlocking-directorate regression. A normal quantile comparison plot is shown in (a). The 95% confidence envelope is based on the standard errors of the order statistics for an independent normal sample. A nonparametric density estimate is shown in (b).

The quantile comparison plot is particularly effective in displaying the tail behavior of the residuals: Outliers, skewness, heavy tails, or light tails all show up clearly. Other univariate graphical displays effectively supplement the quantile comparison plot. In large samples, a histogram with many bars conveys a good impression of the shape of the residual distribution, and generally reveals multiple modes more clearly than the quantile comparison plot does. In smaller samples, a more stable impression is formed by smoothing the histogram of the residuals with a nonparametric density estimator.

Figure 12.1 shows plots of the studentized residuals for a regression model fit to Ornstein's interlocking-directorate data, first discussed in Chapter 3. The dependent variable in the regression is the number of executive and director interlocks maintained by each of 248 dominant Canadian firms with other companies in this group. The independent variables are the assets of the firm (in millions of dollars), the industrial sector in which the firm operates (10 categories), and the nation in which the firm is controlled (four categories). The results of the regression are shown on the left of Table 12.1.[6] Because the residual degrees of freedom are relatively large (234), the studentized residuals are plotted against the normal distribution in the quantile comparison plot of Figure 12.1(a). The quantile comparison plot suggests that the distribution of the residuals has heavy tails—particularly the upper tail. The density estimate [in Figure 12.1(b)] suggests that there may be two groups of observations somewhat separated from the others, one group at the low end of the residual distribution, another at the high end.

[6] The square root of assets is used in place of assets to make the regression more nearly linear. See Section 12.3 for a discussion of nonlinearity.

TABLE 12.1 Regression of Number of Interlocking Directorate and Executive Positions Maintained by 248 Dominant Canadian Corporations on Corporate Assets, Sector, and Nation of Control. The baseline category for Sector is *Heavy Manufacturing*; for Nation of Control, *Canada*

	Interlocks		$\sqrt{\mathit{Interlocks}+1}$	
Regressor	*B*	$\widehat{\text{SE}}$	*B*	$\widehat{\text{SE}}$
Constant	4.19	1.85	2.33	0.23
$\sqrt{\text{Assets}}$	0.252	0.019	0.0260	0.0023
Sector				
Wood, paper	5.15	2.68	0.786	0.335
Mining, metals	0.342	2.01	0.356	0.252
Transport	−0.381	2.82	0.354	0.353
Merchandizing	−0.867	2.63	0.148	0.329
Agriculture, food, Light industry	−1.20	2.04	−0.0567	0.255
Holding companies	−2.43	4.01	−0.245	0.502
Construction	−5.13	4.70	−0.740	0.588
Other financials	−5.70	2.93	−0.0880	0.366
Banking	−14.4	5.58	−2.25	0.697
Nation of Control				
Other	−1.16	2.66	−0.114	0.333
Britain	−4.44	2.65	−0.527	0.331
United States	−8.09	1.48	−1.11	0.185
R^2	.655		.580	

A positive skew in the residuals can usually be corrected by moving the dependent variable down the ladder of powers and roots. In the present case, *both* tails of the residual distribution are heavy, but I decided to try power transformations because (1) the upper tail appears heavier than the lower tail; (2) the distribution of number of interlocks (i.e., the dependent variable) is positively skewed; and (3) transformations down the ladder of powers, particularly square root and log, often are effective when the dependent variable is (as here) a count. Because some of the firms maintained 0 interlocks, I used a start of 1 for the transformations. Trial and error suggests that the square-root transformation of *number of interlocks*+1 renders the distribution of the residuals close to normal, as shown in Figure 12.2.

The results of Ornstein's regression using $\sqrt{\text{interlocks}+1}$ as the dependent variable is shown on the right of Table 12.1. Although we cannot compare the coefficients directly across these two models—the scale of the dependent variable is different in the two cases—the general character of the results does not change much: In both models, assets has a substantial impact on interlocks; the rankings of the nation-of-control categories are identical in the two models; and the rankings of the sectors are nearly the same.[7]

[7] Exercise 12.1 suggests a more precise comparison of the two sets of results using "adjusted" means.

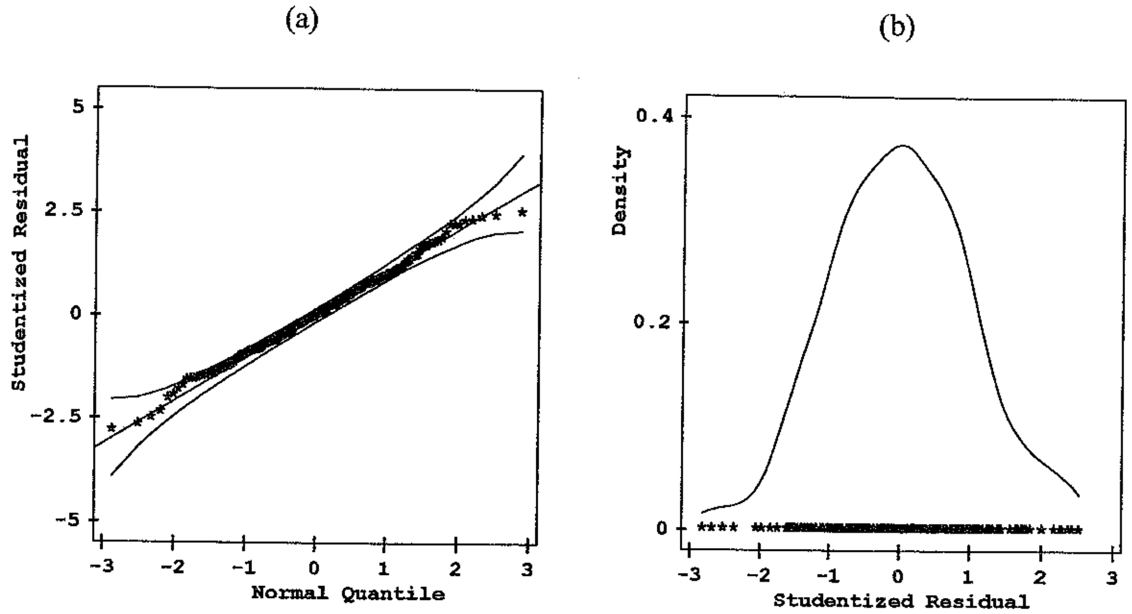

Figure 12.2. The distribution of the studentized residuals from Ornstein's interlocking-directorate regression, after transforming the dependent variable. A normal quantile comparison plot is shown in (*a*), a nonparametric density estimate in (*b*).

> Heavy-tailed errors threaten the efficiency of least-squares estimation; skewed and multimodal errors compromise the interpretation of the least-squares fit. Nonnormality can often be detected by examining the distribution of the least-squares residuals, and frequently can be corrected by transforming the data.

12.1.1 Confidence Envelopes by Simulated Sampling*

Atkinson (1985) has suggested the following procedure for constructing an approximate confidence "envelope" in a quantile comparison plot, taking into account the correlational structure of the independent variables. Atkinson's procedure employs simulated sampling, and uses the assumption of normally distributed errors.[8]

1. Fit the regression model as usual, obtaining fitted values $\hat{Y}_i$ and the estimated standard error S_E.
2. Construct m samples, each consisting of n simulated Y-values; for the jth such sample, the simulated value for observation i is

$$Y_{ij}^s = \hat{Y}_i + S_E Z_{ij}$$

[8] The notion of simulated sampling from a population constructed from the observed data is the basis of "bootstrapping," discussed in Section 16.1. Atkinson's procedure described here is an example of the parametric bootstrap.

where Z_{ij} is a random draw from the unit-normal distribution. In other words, we sample from a "population" in which the expectation of Y_i is $\hat{Y}_i$; the true standard deviation of the errors is S_E; and the errors are normally distributed.

3. Regress the n simulated observations for sample j on the X's in the original sample, obtaining simulated studentized residuals, $E^*_{1j}, E^*_{2j}, \ldots, E^*_{nj}$. Because this regression employs the original X-values, the simulated studentized residuals reflect the correlational structure of the X's.
4. Order the studentized residuals for sample j from smallest to largest, as required by a quantile comparison plot: $E^*_{(1)j}, E^*_{(2)j}, \ldots, E^*_{(n)j}$.
5. To construct an estimated $(100-a)\%$ confidence interval for $E^*_{(i)}$ (the ith ordered studentized residual), find the $a/2$ and $1 - a/2$ empirical quantiles of the m simulated values $E^*_{(i)1}, E^*_{(i)2}, \ldots, E^*_{(i)m}$. For example, if $m = 20$ and $a = .05$, then the smallest and largest[9] of the $E^*_{(1)j}$ provide a 95% confidence interval for $E^*_{(1)}$: $(E^*_{(1)(1)}, E^*_{(1)(20)})$. The confidence limits for the n ordered studentized residuals are graphed as a confidence envelope on the quantile comparison plot, along with the studentized residuals themselves.

A weakness of Atkinson's procedure is that the probability of *some* studentized residual straying outside of the confidence limits by chance is greater than a, which is the probability that an *individual* studentized residual falls outside of its confidence interval. Because the joint distribution of the studentized residuals is complicated, however, to construct a correct joint-confidence envelope would require even more calculation. As well, in small samples, where there are few residual degrees of freedom, even radical departures from normally distributed errors can give rise to apparently normally distributed residuals; Andrews (1979) presents an example of this phenomenon, which is sometimes termed "supernormality."

EXERCISE

12.1 Use adjusted means (see Exercises 7.5, 7.10, 8.8, and 8.12) to compare the two regressions for Ornstein's interlocking-directorate data summarized in Table 12.1. For the first regression, in which interlocks is the dependent variable, calculate the adjusted mean number of interlocks for each sector and nation of control. For the second regression, in which $\sqrt{\text{interlocks} + 1}$ is the dependent variable, first calculate the adjusted dependent-variable mean for each sector and nation of control, and then translate back to the interlocks scale by squaring and subtracting 1 from each of these quantities. The partial relationship between interlocks and assets for each model

[9] Selecting the smallest and largest of the 20 simulated values corresponds to our simple convention that the proportion of the data below the jth of m order statistics is $(j - 1/2)/m$. Here, $(1 - 1/2)/20 = .025$ and $(20 - 1/2)/20 = .975$, defining 95% confidence limits. Atkinson uses a slightly different convention. To estimate the confidence limits more accurately, it would help to make m larger, and perhaps to use a more sophisticated version of the bootstrap (see Section 16.1), but the approximate nature of the entire enterprise makes it difficult to justify the additional computation that would be required.

can be displayed graphically by setting each dummy variable to its mean (i.e., the proportion in the corresponding category of sector or nation of control) and substituting these values into the regression equation. Then, letting assets run over its range of values (roughly \$50 million to \$150,000 million), substitute $\sqrt{\text{assets}}$ into the regression equation to calculate the corresponding fitted values for the dependent variable, connecting the fitted values by a smooth curve (as in Figure 12.7 on page 314). For the second regression, remember to translate back to the interlocks scale to facilitate the comparison of the two results. (Ornstein's data are in `ornstein.dat`.)

12.2 Nonconstant Error Variance

As we know, one of the assumptions of the regression model is that the variation of the dependent variable around the regression surface—the error variance—is everywhere the same:

$$V(\varepsilon) = V(Y|x_1, \ldots, x_k) = \sigma_\varepsilon^2$$

Nonconstant error variance is sometimes termed "heteroscedasticity." Although the least-squares estimator is unbiased and consistent even when the error variance is not constant, the efficiency of the least-squares estimator is impaired, and the usual formulas for coefficient standard errors are inaccurate—the degree of the problem depending on the degree to which error variances differ. In this section, I shall describe graphical methods for detecting nonconstant error variances, and methods for dealing with the problem when it is detected.[10]

12.2.1 Residual Plots

Because the regression surface is k-dimensional and embedded in a space of $k+1$ dimensions, it is generally impractical to assess the assumption of constant error variance by direct graphical examination of the data when k is larger than 1 or 2. Nevertheless, it is common for error variance to increase as the expectation of Y grows larger, or there may be a systematic relationship between error variance and a particular X. The former situation can often be detected by plotting residuals against fitted values, and the latter by plotting residuals against each X.[11]

Plotting residuals against Y (as opposed to $\hat{Y}$) is generally unsatisfactory, because the plot is "tilted": Because $Y = \hat{Y} + E$, the linear correlation[12] between Y and E is $\sqrt{1 - R^2}$. In contrast, the least-squares fit ensures that the correlation between $\hat{Y}$ and E is precisely 0, producing a plot that is much easier to examine for evidence of nonconstant spread.

[10] Tests for heteroscedasticity are discussed in Section 12.4 on discrete data, and in Section 12.5 on maximum-likelihood methods.

[11] These displays are not infallible, however: See Cook (1994), and the discussion in Section 12.6.

[12] See Exercise 12.2.

Because the least-squares *residuals* have unequal variances even when the assumption of constant *error* variance is correct, it is preferable to plot studentized residuals against fitted values. A pattern of changing spread is often more easily discerned in a plot of absolute studentized residuals, $|E_i^*|$, or squared studentized residuals, E_i^{*2}, against $\hat{Y}$. Finally, if the values of $\hat{Y}$ are all positive, then we can plot $\log|E_i^*|$ (log spread) against $\log \hat{Y}$ (log level). A line, with slope b fit to this plot, suggests the variance-stabilizing transformation[13] $Y^{(p)}$, with $p = 1 - b$.

Recall Ornstein's interlocking-directorate regression, described in the previous section. Figure 12.3(a) shows the plot of studentized residuals against fitted values for this regression. Although the substantial positive skew in the fitted values makes the plot difficult to examine, there appears to be a tendency for the residual scatter to get wider at larger values[14] of $\hat{Y}$. The log-spread versus log-level plot for the regression, in Figure 12.3(b), is easier to examine. Because there are negative fitted values, I used $\log(\hat{Y} + 2)$ to construct the plot.[15] The least-squares line fit to the plot has slope $b = 0.497$, suggesting the variance-stabilizing transformation $p = 1 - 0.497 = 0.503$. The positive trend in the spread-versus-level plot translates into a transformation *down* the ladder of powers and roots.

In the previous section, the transformation $\sqrt{\text{interlocks} + 1}$—that is, $p = 0.5$—made the distribution of the studentized residuals more nearly normal. The same transformation nearly stabilizes the residual variance, as illustrated in the spread-versus-level plot shown in Figure 12.4.[16] This outcome is not surprising, because the heavy right tail of the residual distribution and nonconstant spread are both common consequences of the lower bound of 0 for the dependent variable.

Transforming Y changes the shape of the error distribution, but it also alters the shape of the regression of Y on the X's. At times, eliminating nonconstant spread also makes the relationship of Y to the X 's more nearly linear, but this is not a necessary consequence of stabilizing the error variance, and it is important to check for nonlinearity following transformation of the dependent variable. Of course, because there is generally no reason to suppose that the regression is linear prior to transforming Y, we should check for nonlinearity in any event.[17]

Nonconstant residual spread sometimes is symptomatic of the omission of important effects from the model. Suppose, for example, that there is an omitted

[13] This is an application of Tukey's rule, presented in Section 4.4. Other analytic methods for choosing a variance-stabilizing transformation are discussed in Section 12.5.

[14] Part of the tendency for the residual spread to increase with $\hat{Y}$ is due to the lower bound of 0 for Y: Because $E = Y - \hat{Y}$, the smallest possible residual corresponding to a particular $\hat{Y}$ is $E = 0 - \hat{Y} = -\hat{Y}$; the boundary $E = -\hat{Y}$ is a line with slope -1 at the lower left of the residual versus fitted-value plot. When there are many observations with 0 values, it may be more appropriate to use a Poisson regression model, as described in Section 15.4.

[15] Several observations have negative fitted values, the smallest of which is -1.57.

[16] Figure 12.4 still shows some relationship between spread and level, but the log transformation substantially overcorrects the original problem, inducing a negative association between spread and level. The start of 1 is not really required here, because the square-root transformation is defined for $Y = 0$. In this dataset, using $\sqrt{\text{interlocks}}$, which is a slightly more powerful transformation than $\sqrt{\text{interlocks} + 1}$, nearly perfectly stablilizes the variance of the residuals.

[17] See Section 12.3.

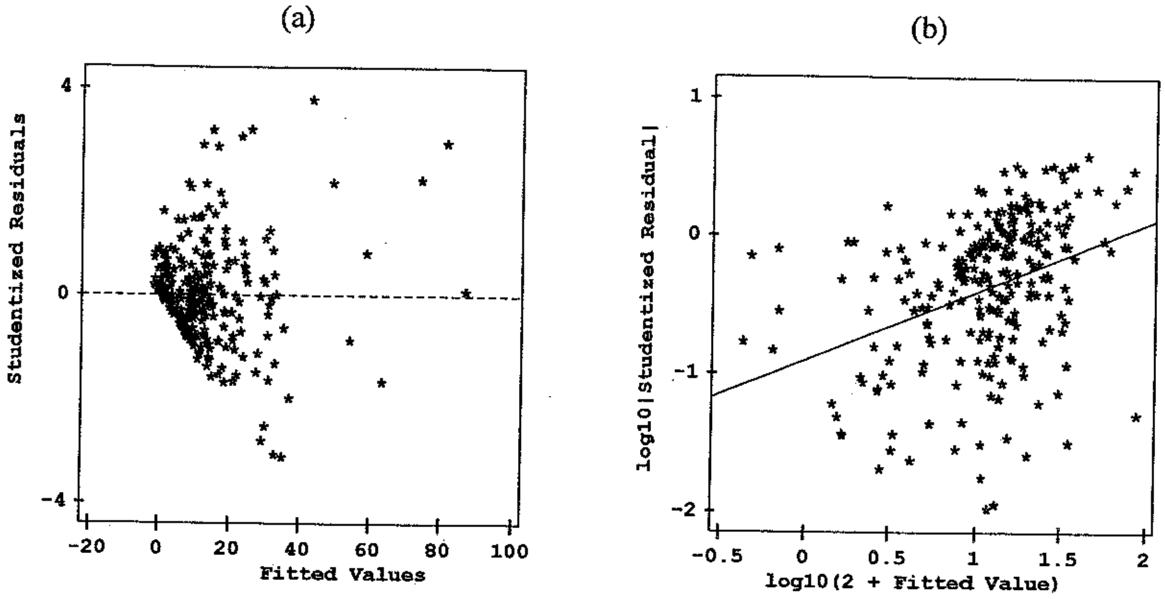

Figure 12.3. Detecting nonconstant spread in Ornstein's interlocking-directorate regression. (*a*) A plot of studentized residuals versus fitted values. (*b*) A plot of log spread (log absolute studentized residuals) versus log level (log fitted values). The least-squares line is shown on the plot.

categorical independent variable, such as regional location, that interacts with assets in affecting interlocks; in particular, suppose that the assets slope, although positive in every region, is steeper in some regions than in others. Then the omission of region and its interaction with assets could produce a fan-shaped residual plot even if the errors from the correct model have constant variance.[18] The detection of this type of specification error requires insight into the process generating the data and cannot rely on diagnostics alone.

12.2.2 Weighted-Least-Squares Estimation*

Weighted-least-squares regression provides an alternative approach to estimation in the presence of nonconstant error variance. Suppose that the errors from the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ are independent and normally distributed, with zero means but *different* variances: $\varepsilon_i \sim N(0, \sigma_i^2)$. Suppose further that the variances of the errors are known up to a constant of proportionality σ_ε^2, so that $V(\varepsilon_i) = \sigma_i^2 = \sigma_\varepsilon^2/w_i^2$. Then, the likelihood for the model is[19]

$$L(\boldsymbol{\beta}, \sigma_\varepsilon^2) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))'\boldsymbol{\Sigma}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the errors,

$$\boldsymbol{\Sigma} = \sigma_\varepsilon^2 \times \text{diag}\{1/w_1^2, \ldots, 1/w_n^2\} \equiv \sigma_\varepsilon^2 \times \mathbf{W}^{-1}$$

[18] See Exercise 12.3 for an illustration of this phenomenon.

[19] See Exercise 12.4 for this and other results pertaining to weighted-least-squares estimation.

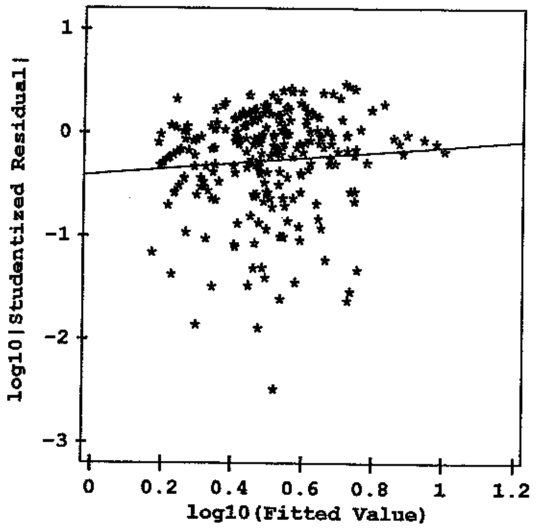

Figure 12.4. Plot of log spread versus log level for Ornstein's interlocking-directorate regression, after transforming the dependent variable. The least-squares line is shown on the plot.

The maximum-likelihood estimators of $\boldsymbol{\beta}$ and σ_ε^2 are

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

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum(E_i/w_i)^2}{n}$$

where the residuals E_i are defined in the usual manner. This procedure is equivalent to minimizing the weighted sum of squares $\sum w_i^2 E_i^2$, according greater weight to observations with smaller variance—hence the term *weighted least squares (WLS)*. The estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\mathscr{V}}(\hat{\beta}) = \hat{\sigma}_\varepsilon^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$$

In practice, we would need to estimate the weights w_i or know that the error variance is systematically related to some observable variable. In the first instance, for example, we could use the residuals from a preliminary *ordinary-least-squares (OLS)* regression to obtain estimates of the error variance within different categories of the data, partitioned by one or more categorical variables. Basing the weights on a preliminary estimate of error variances can, however, seriously bias the estimated covariance matrix $\hat{\mathscr{V}}(\hat{\beta})$, because the sampling error in the estimates should reflect the additional source of uncertainty.[20]

In the second instance, suppose that inspection of a residual plot for the preliminary OLS fit suggests that the magnitude of the errors is proportional to the first independent variable, X_1. We can then use $1/X_{i1}$ as the weights w_i. Dividing both sides of the regression equation by X_{i1} produces

$$\frac{Y_i}{X_{i1}} = \alpha\frac{1}{X_{i1}} + \beta_1 + \beta_2\frac{X_{i2}}{X_{i1}} + \cdots + \beta_k\frac{X_{ik}}{X_{i1}} + \frac{\varepsilon_i}{X_{i1}} \qquad [12.1]$$

[20] In this case, it is probably better to obtain an honest estimate of the coefficient covariance matrix from the bootstrap, described in Section 16.1.

Because the standard deviations of the errors are proportional to X_1, the "new" errors $\varepsilon_i' \equiv \varepsilon_i / X_{i1}$ have constant variance, and Equation 12.1 can be estimated by OLS regression of Y/X_1 on $1/X_1, X_2/X_1, \ldots, X_k/X_1$. Notice that the constant from this regression estimates β_1, while the coefficient of $1/X_1$ estimates α; the remaining coefficients are straightforward.[21]

> It is common for the variance of the errors to increase with the level of the dependent variable. This pattern of nonconstant error variance ("heteroscedasticity") can often be detected in a plot of residuals against fitted values. Strategies for dealing with nonconstant error variance include transformation of the dependent variable to stabilize the variance; the substitution of weighted-least-squares estimation for ordinary least squares; and the correction of coefficient standard errors for heteroscedasticity. A rough rule is that nonconstant error variance seriously degrades the least-squares estimator only when the ratio of the largest to smallest variance is about 10 or more.

12.2.3 Correcting OLS Standard Errors for Nonconstant Variance*

The covariance matrix of the ordinary-least-squares estimator is

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad [12.2]$$

Under the standard assumptions, including the assumption of constant error variance, $V(\mathbf{y}) = \sigma_\varepsilon^2 \mathbf{I}_n$, Equation 12.2 simplifies to the usual formula, $V(\mathbf{b}) = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}$. If, however, the errors are heteroscedastic but independent, then $\mathbf{\Sigma} \equiv V(\mathbf{y}) = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$, and

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Because $E(\varepsilon_i) = 0$, the variance of the ith error is $\sigma_i^2 = E(\varepsilon_i^2)$, which suggests the possibility of estimating $V(\mathbf{b})$ by

$$\tilde{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Sigma}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad [12.3]$$

with $\hat{\mathbf{\Sigma}} = \text{diag}\{E_1^2, \ldots, E_n^2\}$, and where E_i is the OLS residual for observation i. White (1980) shows that Equation 12.3 provides a consistent estimator[22] of $V(\mathbf{b})$.

[21] An application of WLS regression to Ornstein"s interlocking-directorate data is given in Exercise 12.6.

[22] See Exercise 12.7 for an application of White's heteroscedasticity correction to Ornstein's interlocking-directorate regression.

An advantage of White's approach is that knowledge of the *pattern* of nonconstant error variance (e.g., increased variance with the level of Y or with an X) is not required. If, however, the heteroscedasticity problem is severe, and the corrected coefficient standard errors therefore are substantially larger than those produced by the usual formula, then discovering the pattern of nonconstant variance and correcting for it—by a transformation or WLS estimation—offers the possibility of more efficient estimation. In any event, as the next section shows, unequal error variance is worth correcting only when the problem is severe.

12.2.4 How Nonconstant Error Variance Affects the OLS Estimator*

The impact of nonconstant error variance on the efficiency of the ordinary least-squares estimator and on the validity of least-squares inference depends on several factors, including the sample size, the degree of variation in the σ_i^2, the configuration of the X-values, and the relationship between the error variance and the X's. It is therefore not possible to develop wholly general conclusions concerning the harm produced by heteroscedasticity, but the following simple case is nevertheless instructive.

Suppose that $Y_i = \alpha + \beta X_i + \varepsilon_i$, where the errors are independent and normally distributed, with zero means but with different standard deviations proportional to X, so that $\sigma_i = \sigma_\varepsilon X_i$. Then the OLS estimator B is less efficient than the WLS estimator $\hat{\beta}$, which, under these circumstances, is the most efficient unbiased estimator[23] of β.

Formulas for the sampling variances of B and $\hat{\beta}$ are easily derived.[24] The efficiency of the OLS estimator relative to the optimal WLS estimator is given by $V(\hat{\beta})/V(B)$, and the relative precision of the OLS estimator is the square root of this ratio, that is, $\text{SE}(\hat{\beta})/\text{SE}(B)$.

Now suppose that X is uniformly distributed over the interval $[x_0, ax_0]$, where both x_0 and a are positive numbers, so that a is the ratio of the largest to the smallest value of X (and, consequently, of the largest to the smallest σ_i). The relative precision of the OLS estimator stabilizes quickly as the sample size grows, and exceeds 90% when $a = 2$, and 85% when $a = 3$, even when n is as small as 20. For $a = 10$, the penalty for using OLS is greater, but even here the relative precision of OLS exceeds 65% for $n \geq 20$.

The validity of statistical inferences based on OLS estimation is even less sensitive to common patterns of nonconstant error variance. Here, we need to compare the expectation of the usual estimator of $V(B)$, which is typically biased when the error variance is not constant, with the true sampling variance of B. The square root of $E[\widehat{V(B)}]/V(B)$ expresses the result in relative standard-error terms. For the illustration, where the standard deviation of the errors is proportional to X, and where X is uniformly distributed, this ratio is 98% when $a = 2$; 97% when $a = 3$; and 93% when $a = 10$; all for $n \geq 20$.

The results in this section suggest that nonconstant error variance is a serious problem only when the magnitude (i.e., the standard deviation) of the

[23] This property of the WLS estimator requires the assumption of normality. Without normal errors, the WLS estimator is still the most efficient *linear* unbiased estimator—an extension of the Gauss-Markov theorem. See Exercise 12.5.

[24] See Exercise 12.8 for this and other results described in this section.

errors varies by more than a factor of about 3—that is, when the largest error variance is more than about 10 times the smallest.

EXERCISES

12.2 *Show that the correlation between the least-squares residuals E_i and the dependent-variable values Y_i is $\sqrt{1-R^2}$. (*Hint*: Use the geometric vector representation of multiple regression, examining the plane in which the $\mathbf{e}$, $\mathbf{y}^*$, and $\hat{\mathbf{y}}^*$ vectors lie.)

12.3 Nonconstant variance and specification error: Generate 100 observations according to the following model:

$$Y = 10 + (1 \times X) + (1 \times D) + (2 \times X \times D) + \varepsilon$$

where $\varepsilon \sim N(0, 10^2)$; the values of X are $1, 2, \ldots, 50, 1, 2, \ldots, 50$; the first 50 values of D are 0; and the last 50 values of D are 1. Then regress Y on X alone (i.e., omitting D and XD), $Y = A + BX + E$. Plot the residuals E from this regression against the fitted values $\hat{Y}$. Is the variance of the residuals constant? How do you account for the pattern in the plot?

12.4 *Weighted-least-squares estimation: Suppose that the errors from the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ are independent and normally distributed, but with different variances, $\varepsilon_i \sim N(0, \sigma_i^2)$, and that $\sigma_i^2 = \sigma_\varepsilon^2 / w_i^2$. Show that:

(a) The likelihood for the model is

$$L(\boldsymbol{\beta}, \sigma_\varepsilon^2) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]$$

where

$$\boldsymbol{\Sigma} = \sigma_\varepsilon^2 \times \text{diag}\{1/w_1^2, \ldots, 1/w_n^2\} \equiv \sigma_\varepsilon^2 \times \mathbf{W}^{-1}$$

(b) The maximum-likelihood estimators of $\boldsymbol{\beta}$ and σ_ε^2 are

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

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum (E_i / w_i)^2}{n}$$

where $\mathbf{e} = \{E_i\} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

(c) The MLE is equivalent to minimizing the weighted sum of squares $\sum w_i^2 E_i^2$.

(d) The estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\mathscr{V}}(\hat{\beta}) = \hat{\sigma}_\varepsilon^2 (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$$

12.5 *Show that when the covariance matrix of the errors is

$$\boldsymbol{\Sigma} = \sigma_\varepsilon^2 \times \text{diag}\{1/w_1^2, \ldots, 1/w_n^2\} \equiv \sigma_\varepsilon^2 \times \mathbf{W}^{-1}$$

the weighted-least-squares estimator

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \\ &= \mathbf{M}\mathbf{y}\end{aligned}$$

is the minimum-variance linear unbiased estimator of $\boldsymbol{\beta}$. (*Hint*: Adapt the proof of the Gauss-Markov theorem for OLS estimation given in Section 9.3.2.)

12.6 *Apply weighted-least-squares estimation to Ornstein's regression of number of interlocking directorates on square-root assets, sector, and nation of control, supposing that the standard deviation of the errors is proportional to the square root of assets. (The OLS regression is reported in Table 12.1; the data are in `ornstein.dat`.) How do the results of WLS estimation compare with those of OLS estimation? With OLS following the square-root transformation of number of interlocks (also shown in Table 12.1)?

12.7 *Using White's correction for nonconstant variance, recalculate coefficient standard errors for Ornstein's OLS regression of number of interlocking directorates on square-root assets, sector, and nation of control (given in Table 12.1). How do the corrected standard-error estimates compare with those computed by the usual approach?

12.8 *The impact of nonconstant error variance on OLS estimation: Suppose that $Y_i = \alpha + \beta x_i + \varepsilon_i$, with independent errors, $\varepsilon_i \sim N(0, \sigma_i^2)$, and $\sigma_i = \sigma_\varepsilon x_i$. Let B represent the OLS estimator and $\hat{\beta}$ the WLS estimator of β.

(a) Show that the sampling variance of the OLS estimator is

$$V(B) = \frac{\sum(X_i - \overline{X})^2 \sigma_i^2}{\left[\sum(X_i - \overline{X})^2\right]^2}$$

and that the sampling variance of the WLS estimator is

$$V(\hat{\beta}) = \frac{\sigma_\varepsilon^2}{\sum w_i^2 (X_i - \tilde{X})^2}$$

where $\tilde{X} \equiv (\sum w_i^2 X_i)/(\sum w_i^2)$. (*Hint*: Write each slope estimator as a linear function of the Y_i.)

(b) Now suppose that x is uniformly distributed over the interval $[x_0, ax_0]$, where $x_0 > 0$ and $a > 0$, so that a is the ratio of the largest to the smallest σ_i. The efficiency of the OLS estimator relative to the optimal WLS estimator is $V(\hat{\beta})/V(B)$, and the relative precision of the OLS estimator is the square root of this ratio, that is, $\text{SE}(\hat{\beta})/\text{SE}(B)$. Calculate the relative precision of the OLS estimator for all combinations of a = 2, 3, 5, 10, and n = 5, 10, 20, 50, 100. For example, when $a = 3$ and $n = 10$, you can take the x-values as 1, 1.222, 1.444, ..., 2.778, 3. Under what circumstances is the OLS estimator substantially less precise than the WLS estimator?

(c) The usual variance estimate for the OLS slope (assuming constant error variance) is

$$\widehat{V(B)} = \frac{S_E^2}{\sum(X_i - \overline{X})^2}$$

where $S_E^2 = \sum E_i^2/(n-2)$. Kmenta (1986, Section 8.2) shows that the expectation of this variance estimator (under nonconstant error variance σ_i^2) is

$$E[\widehat{V(B)}] = \frac{\overline{\sigma}^2}{\sum(X_i - \overline{X})^2} - \frac{\sum(X_i - \overline{X})^2(\sigma_i^2 - \overline{\sigma}^2)}{(n-2)[\sum(X_i - \overline{X})^2]^2}$$

where $\overline{\sigma}^2 \equiv \sum \sigma_i^2/n$. (*Prove this result.) Kmenta also shows that the true variance of the OLS slope estimator, $V(B)$ [derived in part (a)], is generally different from $E[\widehat{V(B)}]$. If $\sqrt{E[\widehat{V(B)}]/V(B)}$ is substantially below 1, then the usual formula for the standard error of B will lead us to believe that the OLS estimator is more precise than it really is. Calculate $\sqrt{E[\widehat{V(B)}]/V(B)}$ under the conditions of part (b), for $a = 5, 10, 20, 50$, and $n = 5, 10, 20, 50, 100$. What do you conclude about the robustness of validity of OLS inference with respect to nonconstant error variance?

12.3 Nonlinearity

The assumption that the average error, $E(\varepsilon)$, is everywhere 0 implies that the specified regression surface accurately reflects the dependency of the conditional average value of Y on the X's. Conversely, violating the assumption of linearity implies that the model fails to capture the systematic pattern of relationship between the dependent and independent variables. The term "nonlinearity," therefore, is not used in the narrow sense here, although it includes the possibility that a partial relationship assumed to be linear is, in fact, nonlinear: If, for example, two independent variables specified to have additive effects instead interact, then the average error is not 0 for all combinations of X-values.

If nonlinearity, in the broad sense, is slight, then the fitted model can be a useful approximation even though the regression surface $E(Y|X_1, \ldots X_k)$ is not captured precisely. In other instances, however, the model can be seriously misleading.

The regression surface is generally high dimensional, even after accounting for regressors (such as dummy variables, interactions, and polynomial terms) that are functions of a smaller number of fundamental independent variables.[25] As in the case of nonconstant error variance, therefore, it is necessary to focus on particular patterns of departure from linearity. The graphical diagnostics discussed in this section are two-dimensional (and three-dimensional) projections of the $(k+1)$-dimensional point cloud of observations $\{Y_i, X_{i1}, \ldots, X_{ik}\}$.

[25] Polynomial regression—for example, the model $Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$—is discussed in Sections 12.4 and 14.2.1. In this simple quadratic model, there are two regressors, but only one independent variable.

12.3.1 Partial-Residual Plots

Although it is useful in multiple regression to plot Y against each X (e.g., in one row of a scatterplot matrix), these plots often do not tell the whole story—and can be misleading—because our interest centers on the *partial* relationship between Y and each X (controlling for the other X's), not on the *marginal* relationship between Y and an individual X (ignoring the other X's). Residual-based plots are consequently more promising in the specific context of multiple regression.

Plotting residuals or studentized residuals against each X, perhaps augmented by a nonparametric-regression smooth, is frequently helpful for detecting departures from linearity. As Figure 12.5 illustrates, however, simple residual plots cannot distinguish between monotone and nonmonotone nonlinearity. This distinction is lost in the residual plots because the least-squares fit ensures that the residuals are linearly uncorrelated with each X. The distinction is important because monotone nonlinearity frequently can be "corrected" by simple transformations.[26] In Figure 12.5, for example, case (b) might be modeled by $Y = \alpha + \beta\sqrt{X} + \varepsilon$, while case ($a$) cannot be linearized by a power transformation of X, and might instead be dealt with by the quadratic regression,[27] $Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$.

In contrast to simple residual plots, partial-regression plots, introduced in the previous chapter for detecting influential data, can reveal nonlinearity and suggest whether a relationship is monotone. These plots are not always useful for locating a transformation, however: The partial-regression plot adjusts X_j for the other X's, but it is the unadjusted X_j that is transformed in respecifying the model. The similarly named *partial-residual plots*, also called *component-plus-residual plots*, are often an effective alternative. Partial-residual plots are not as suitable as partial-regression plots for revealing leverage and influence.[28]

Define the partial residual for the jth independent variable as

$$E_i^{(j)} = E_i + B_j X_{ij}$$

In words, add back the linear component of the partial relationship between Y and X_j to the least-squares residuals, which may include an unmodeled nonlinear component. Then plot $E^{(j)}$ versus X_j. By construction, the multiple-regression coefficient B_j is the slope of the simple linear regression of $E^{(j)}$ on X_j, but nonlinearity may be apparent in the plot as well. Again, a nonparametric regression may help in interpreting the plot.

The partial-residual plots in Figure 12.6 are for a regression of the rated prestige (P) of 102 Canadian occupations on the average education (S—"schooling") in years, average income (I) in dollars, and percentage of women

[26] Recall the material in Section 4.3 on linearizing transformations.

[27] Case (b) could, however, be accommodated by a more complex transformation of X, of the form $Y = \alpha + \beta(X - \gamma)^\lambda + \varepsilon$. In the illustration, γ could be taken as $\overline{X}$, and λ as 2. More generally, γ and λ could be estimated from the data, for example, by nonlinear least squares (as described in Section 14.2.3). I shall not pursue this approach here.

[28] The argument that partial-residual plots are more suitable than partial-regression plots for diagnosing nonlinearity reflects common experience and advice, but it does not hold in every instance. See Cook (1996).

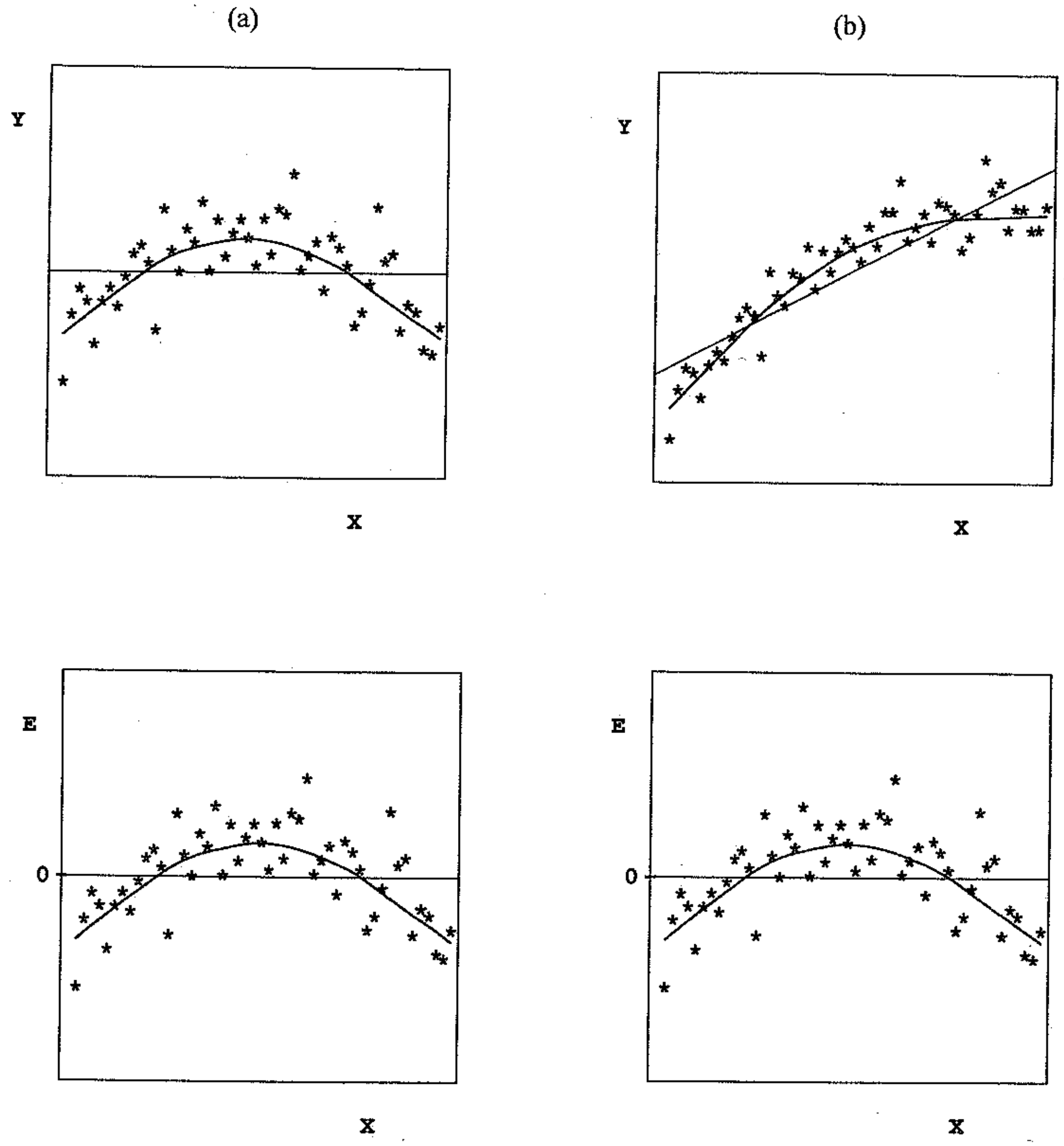

Figure 12.5. The residual plots of E versus X (in the lower panels) are identical, even though the regression of Y on X in (a) is nonmonotone while that in (b) is monotone.

(W) in the occupations in 1971.[29] A nonparametric-regression smooth is shown on each of the plots. The results of the regression are as follows:

$$\begin{array}{ccccc} \hat{P} = -6.79 & + 4.19S & + 0.00131I & - 0.00891W \\ (3.24) & (0.39) & (0.00028) & (0.0304) \end{array}$$

$$R^2 = 0.80 \qquad S_E = 7.85$$

There is apparent monotone nonlinearity in the partial-residual plots for education [Figure 12.6(a)] and, much more strongly, income [Figure 12.6(b)]; there is also a small apparent tendency [in Figure 12.6(c)] for occupations with

[29] This regression was first fit in Section 5.2.2.

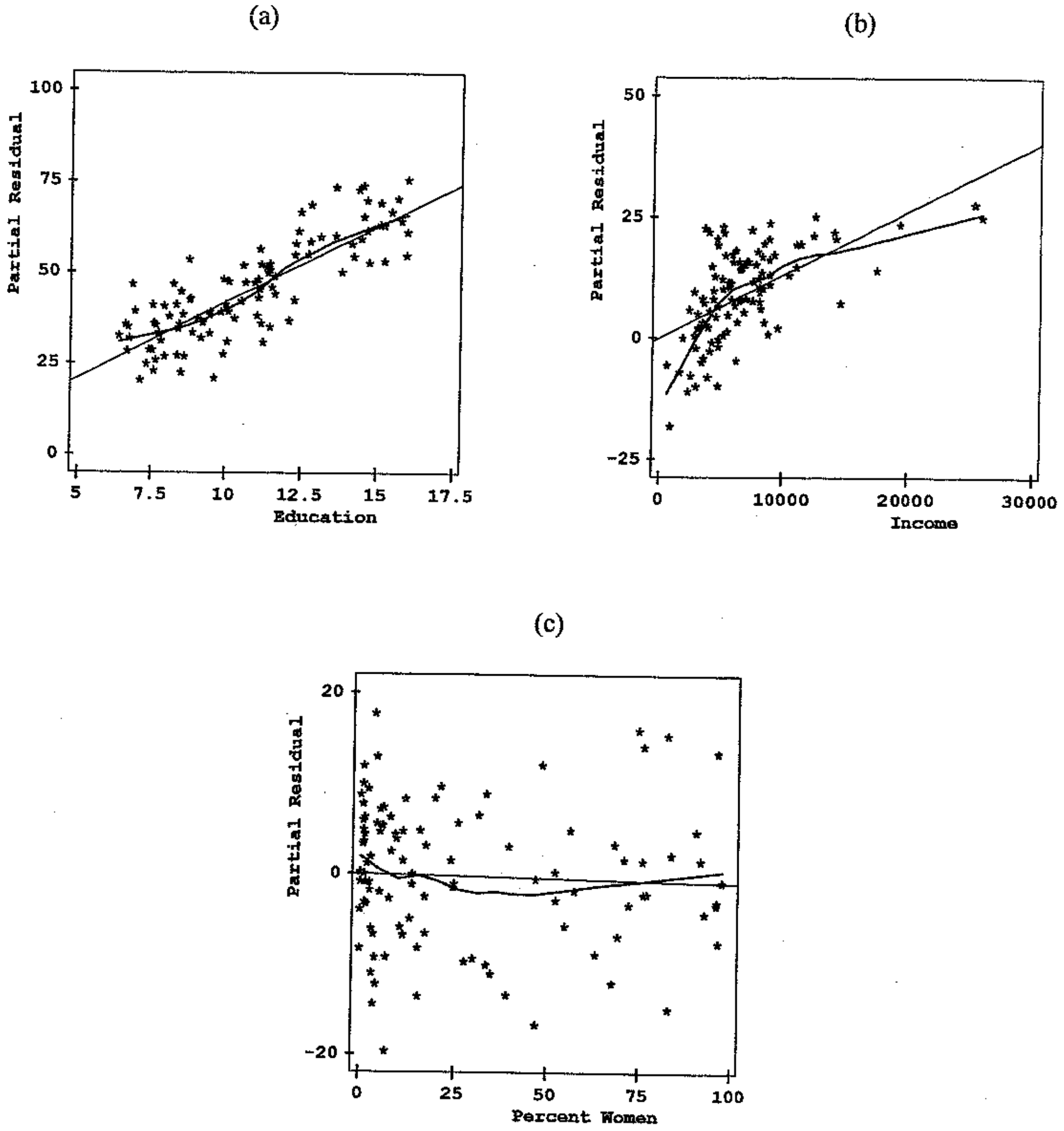

Figure 12.6. Partial-residual plots for the regression of occupational prestige on (*a*) education, (*b*) income, and (*c*) percentage of women. The data are for 102 Canadian occupations in 1971. The least-squares line and a nonparametric-regression smooth are shown on each plot.

intermediate percentages of women to have lower prestige, controlling for education and income, as if occupations with a gender mix pay a small penalty in prestige. To my eye, the patterns in the partial-residual plots for education and percentage of women are not easily discernible without the nonparametric-regression smooth: The departures from linearity in these plots are not great.

The nonlinear pattern for income is simple as well as monotone, suggesting a power transformation; because the bulge points upward and toward the left, we can try to transform prestige *up* the ladder of powers and roots or income *down*. In multiple regression, we are generally loath to transform Y (as opposed to an X), because the relationship between Y and every X would be affected—unless, of course, there is a similar nonlinear pattern to the relationship between

Y and all of the X's. Some experimentation indicates that a log transformation of income straightens its partial relationship to prestige.

The nonlinear pattern for education, in contrast, is monotone (or nearly monotone) but not simple, making a power transformation of education unpromising.[30] The S-shaped pattern in Figure 12.6(a) can be captured, however, by a cubic regression in education. The nonlinear pattern for percentage of women in Figure 12.6(c) is simple but not monotone, suggesting a quadratic regression rather than a power transformation of percentage of women.

The revised fit is as follows:

$$\begin{aligned}\hat{P} = \underset{(56.9)}{20.8} + \underset{(1.27)}{8.78}\log_2 I - \underset{(0.085)}{0.179}W + \underset{(0.00092)}{0.00250}W^2\\ -\underset{(15.3)}{29.9}S + \underset{(1.41)}{2.91}S^2 - \underset{(0.042)}{0.0807}S^3\end{aligned}$$

$$R^2 = 0.86 \qquad S_E = 6.72$$

- The quadratic term for percentage of women is "statistically significant," but the partial effect of this independent variable is relatively small, ranging from a minimum of −3.2 prestige points, for a hypothetical occupation with 32% women, to a maximum of 7.1 points, for a hypothetical occupation consisting entirely of women.[31]
- The partial effect of education is substantial, in contrast, but the departure from linearity is not great, except at very low levels of education (and the coefficient for S^3 is not quite statistically significant at the 5% level, two tailed). Figure 12.7 traces the partial effect of education on prestige, setting the other two independent variables to their average levels, illustrating how a nonlinear relationship can be presented graphically. The curve plotted in the figure is

$$\begin{aligned}\hat{P} = 20.8 + 8.78\,\log_2 6798 - 0.179 \times 29.0 + 0.00250 \times 29.0^2 \quad [12.4]\\ - 29.9S + 2.91S^2 - 0.0807S^3\end{aligned}$$

  where 6798 and 29.0 are, respectively, the means of income and percentage of women. The points in the plot are partial residuals from the cubic education fit, obtained by adding the least-squares residuals to the fitted values determined by Equation 12.4.
- Income also has a large partial effect: Doubling income is associated, on average, with about a 9-point increment in prestige.

In summary, although the small quadratic effect of percentage of women is substantively interesting, not much is gained by including percentage of women in the model. Likewise, little is gained by modeling the effect of education as a third-degree polynomial, even though the leveling off of prestige at low education is potentially interesting. The transformation of income, however, is compelling.

[30] Because, for most of the data, the "bulge" points down and to the right, the transformation $S \to S^2$ does help to straighten the regression.

[31] These numbers are determined by finding the minimum and maximum of $f(W) = -0.179W + 0.00250W^2$ for $0 \leq W \leq 100$.

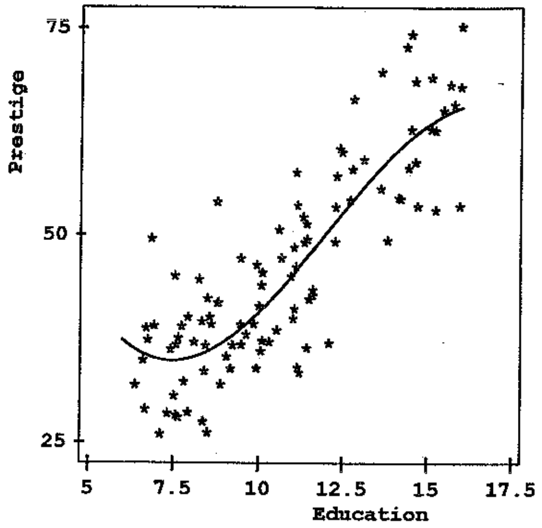

Figure 12.7. The partial relationship between prestige and education, holding income and percentage of women at their average levels. The curve shows the cubic fit for education. The points are partial residuals, obtained by adding the least-squares residuals to the education fit.

12.3.2 When Do Partial-Residual Plots Work?

Circumstances under which regression plots, including partial-residual plots, are informative about the structure of data are an active area of statistical research.[32] It is unreasonable to expect that lower-dimensional displays can always uncover structure in a higher-dimensional problem. We may, for example, discern an interaction between two independent variables in a three-dimensional scatterplot, but could not in two separate two-dimensional plots, one for each independent variable.

It is important, therefore, to understand when graphical displays work and why they sometimes fail: First, understanding the circumstances under which a plot is effective may help us to produce those circumstances. Second, understanding why plots succeed and why they fail may help us to construct more effective displays. Both of these aspects will be developed below.

To provide a point of departure for this discussion, imagine that the following model accurately describes the data:

$$Y_i = \alpha + f(X_{i1}) + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \qquad [12.5]$$

That is, the partial relationship between Y and X_1 is (potentially) nonlinear, characterized by the function $f(X_1)$, while the other independent variables, $X_2, \ldots, X_k$, enter the model linearly.

We do not know in advance the shape of the function $f(X_1)$, and indeed do not know that the partial relationship between Y and X_1 is nonlinear. Instead

[32] Much of this work is due to Cook and his colleagues; see, in particular, Cook (1993), on which the current section is based, and Cook (1994). Cook and Weisberg (1994) provide an accessible summary.

of fitting the true model (Equation 12.5) to the data, therefore, we fit the "working model":

$$Y_i = \alpha' + \beta'_1 X_{i1} + \beta'_2 X_{i2} + \cdots + \beta'_k X_{ik} + \varepsilon'_i$$

The primes indicate that the estimated coefficients for this model do not, in general, estimate the parameters of the true model (Equation 12.5), nor is the "error" of the working model the same as the error of the true model.

Suppose, now, that we construct a partial-residual plot for the working model. The partial residuals estimate

$$\varepsilon_i^{(1)} = \beta'_1 X_{i1} + \varepsilon'_i \qquad [12.6]$$

What we would really like to estimate, however, is $f(X_{i1})+\varepsilon_i$, which, apart from random error, will tell us the partial relationship between Y and X_1. Cook (1993) shows that $\varepsilon_i^{(1)} = f(X_{i1}) + \varepsilon_i$, as desired, under either of two circumstances:

1. The function $f(X_1)$ is linear after all, in which case the population analogs of the partial residuals in Equation 12.6 are appropriately linearly related to X_1.
2. The *other* independent variables $X_2, \ldots, X_k$ are each linearly related to X_1. That is,

   $$E(X_{ij}) = \alpha_{j1} + \beta_{j1} X_{i1} \quad \text{for } j = 2, \ldots, k \qquad [12.7]$$

   If, in contrast, there are *nonlinear* relationships between the other X's and X_1, then the partial-residual plot for X_1 may not reflect the true partial regression $f(X_1)$.[33]

The second result suggests a practical procedure for improving the chances that partial-residual plots will provide accurate evidence of nonlinearity: If possible, transform the independent variables to linearize the relationships among them. Evidence suggests that weak nonlinearity is not especially problematic, but strong nonlinear relationships among the independent variables can invalidate the partial-residual plot as a useful diagnostic display.[34]

> Simple forms of nonlinearity can often be detected in partial-residual plots. Once detected, nonlinearity can frequently be accommodated by variable transformations or by altering the form of the model (to include a quadratic term in an independent variable, for example).. Partial-residual plots adequately reflect nonlinearity when the independent variables are themselves linearly related.

[33] Notice that each of the other X's is regressed on X_1, not vice versa.

[34] See Exercise 12.12.

Mallows (1986) has suggested a variation on the partial-residual plot that sometimes reveals nonlinearity more clearly. I shall focus on X_1, but the spirit of Mallows's suggestion is to construct a plot for each X in turn. First, construct a working model with a quadratic term in X_1 along with the usual linear term:

$$Y_i = \alpha' + \beta'_1 X_{i1} + \gamma_1 X_{i1}^2 + \beta'_2 X_{i2} + \cdots + \beta'_k X_{ik} + \varepsilon'_i$$

Then, after fitting the model, form the "augmented" partial residual

$$E_i'^{(1)} = E'_i + B'_1 X_{i1} + C_1 X_{i1}^2$$

Note that B'_1 generally differs from the regression coefficient for X_1 in the original model, which does not include the squared term. Finally, plot $E'^{(1)}$ versus X_1.

The circumstances under which the augmented partial residuals accurately capture the true partial-regression function $f(X_1)$ are closely analogous to the linear case (see Cook, 1993); either

1. the function $f(X_1)$ is a quadratic in X_1, or
2. the regressions of the other independent variables on X_1 are quadratic:

$$E(X_{ij}) = \alpha_{j1} + \beta_{j1} X_{i1} + \gamma_{j1} X_{i1}^2 \quad \text{for } j = 2, \ldots, k \qquad [12.8]$$

This is a potentially useful result if we cannot transform away nonlinearity among the independent variables—as is the case, for example, when the relationships among the independent variables are not monotone.

The premise of this discussion, expressed in Equation 12.5, is that Y is a nonlinear function of X_1, but linearly related to the other X's. In real applications of partial-residual plots, however, it is quite possible that there is more than one nonlinear partial relationship, and we typically wish to examine each independent variable in turn. Suppose, for example, that the relationship between Y and X_1 is linear; that the relationship between Y and X_2 is nonlinear; and that X_1 and X_2 are correlated. The partial-residual plot for X_1 can, in this situation, show apparent nonlinearity—sometimes termed a "leakage" effect. If more than one partial-residual plot shows evidence of nonlinearity, it may, therefore, be advisable to refit the model and reconstruct the partial-residual plots after correcting the most dramatic instance of nonlinearity.[35]

[35] Exercise 12.9 applies this procedure to the Canadian occupational prestige regression. An iterative formalization of the procedure provides the basis for nonparametric additive regression models, discussed in Section 14.4.2.

CERES Plots*

Cook (1993) provides a still more general procedure, which he calls *CERES* (for "Combining conditional Expectations and RESiduals"): Let

$$\hat{X}_{ij} = \hat{g}_{j1}(X_{i1})$$

represent the estimated regression of X_j on X_1, for $j = 2, \ldots, k$. These regressions may be linear (as in Equation 12.7), quadratic, (as in Equation 12.8), or they may be nonparametric. Of course, the function $\hat{g}_{j1}(X_1)$ will generally be different for different X_j's. Once the regression functions for the other independent variables are found, form the working model

$$Y_i = \alpha'' + \beta_2'' X_{i2} + \cdots + \beta_k'' X_{ik} + \gamma_{12}\hat{X}_{i2} + \cdots + \gamma_{1k}\hat{X}_{ik} + \varepsilon_i''$$

The residuals from this model are then combined with the estimates of the γ's:

$$E_i''^{(1)} = E_i'' + C_{12}\hat{X}_{i2} + \cdots + C_{1k}\hat{X}_{ik}$$

and plotted against X_1.

EXERCISES

12.9 The partial-residual plot for the Canadian occupational prestige data showing the most severe nonlinearity is the plot for income (see Figure 12.6). Reconstruct the three partial-residual plots *after* transforming income. Do the resulting plots for education and percentage of women differ substantially from those shown in Figure 12.6(*a*) and (*c*)? (The data are in `prestige.dat`.)

12.10 Apply Mallows's procedure to construct augmented partial-residual plots for the Canadian occupational prestige regression. (The data are in `prestige.dat`.) *Then apply Cook's CERES procedure to this regression. Compare the results of these two procedures with each other and with the ordinary partial-residual plots shown in Figure 12.6. Do the more complex procedures give clearer indications of nonlinearity in this case?

12.11 Consider the following alternative analysis of the Canadian occupational prestige data: Regress prestige on income, education, percentage of women, and on dummy regressors for type of occupation (professional and managerial, white collar, blue collar); include interactions between type of occupation and each of income, education, and percentage of women. Why is it that the interaction between income and type of occupation can induce a nonlinear relationship between prestige and income when the interaction is ignored? (*Hint*: Construct a scatterplot of prestige versus income, labeling the points in the plot by occupational type, and plotting the separate regression line for each occupational type.)

12.12 Experimenting with partial-residual plots: Generate random samples of 100 observations according to each of the following schemes. In each case, construct the partial-residual plots for X_1 and X_2. Do these plots accurately capture the partial relationships between Y and each of X_1 and X_2? Whenever they appear, E and U are $N(0, 1)$ and independent of each other and of the other variables.

(a) Independent X's and a linear regression: X_1 and X_2 independent and uniformly distributed on the interval $[0, 1]$; $Y = X_1 + X_2 + 0.1E$.

(b) Linearly related X's and a linear regression: X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = X_1 + 0.1U$; $Y = X_1 + X_2 + 0.1E$.

(c) Independent X's and a nonlinear regression on one X: X_1 and X_2 independent and uniformly distributed on the interval $[0, 1]$; $Y = 2(X_1 - 0.5)^2 + X_2 + 0.1E$.

(d) Linearly related X's and a nonlinear regression on one X: X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = X_1 + 0.1U$; $Y = 2(X_1 - 0.5)^2 + X_2 + 0.1E$. (Note the "leakage" here from X_1 to X_2.)

(e) Nonlinearly related X's and a linear regression: X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = |X_1 - 0.5|$; $Y = X_1 + X_2 + 0.02E$.

(f) Nonlinearly related X's and a linear regression on one X: X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = |X_1 - 0.5|$; $Y = 2(X_1 - 0.5)^2 + X_2 + 0.02E$. (Note how strong a nonlinear relationship between the X's, and how small an error variance in the regression, are required for the effects in this example to be noticeable.)

12.4 Discrete Data

As explained in Chapter 3, discrete independent and dependent variables often lead to plots that are difficult to interpret, a problem that can be partially rectified by "jittering" the plotted points.[36] A discrete *dependent* variable also violates the assumption that the errors in a linear model are normally distributed. This problem, like that of a limited dependent variable (i.e., one that is bounded below or above), is only serious in extreme cases—for example, when there are very few response categories, or where a large proportion of the data is in a small number of categories, conditional on the values of the independent variables. In these cases, it is best to use statistical models for categorical dependent variables.[37]

Discrete *independent* variables, in contrast, are perfectly consistent with the general linear model, which makes no distributional assumptions about the X's, other than independence between the X's and the errors. Indeed, because it partitions the data into groups, a discrete X (or combination of X's) facilitates straightforward tests of nonlinearity and nonconstant error variance.

12.4.1 Testing for Nonlinearity ("Lack of Fit")

Recall the data on vocabulary and education collected in the 1989 General Social Survey.[38] Years of education in this dataset range between 0 and 20.

[36] See Section 3.2.

[37] See Chapter 15.

[38] This dataset is described in Section 3.2.

TABLE 12.2 Analysis of Variance for Vocabulary-Test Scores, Showing the Incremental F-Test for Nonlinearity of the Relationship Between Vocabulary and Education

Source	*df*	*SS*	*F*	*p*
Education (*Model 12.10*)	19	1261.7	18.1	<<.0001
Linear (*Model 12.9*)	1	1175.1	320.0	<<.0001
Nonlinear (*"lack of fit"*)	18	86.58	1.31	.17
Error (*"pure error"*)	948	3472.8		
Total	967	4734.5		

Suppose that we model the relationship between vocabulary score and education in two ways:

1. Fit a linear regression of vocabulary on education:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad [12.9]$$

2. Model education with a set of dummy regressors. Although there are 21 conceivable education scores, none of the individuals in the sample has 2 years of education, yielding 20 categories and 19 dummy regressors (treating 0 years of education as the baseline category):

$$Y_i = \alpha' + \gamma_1 D_{i1} + \gamma_3 D_{i3} + \cdots + \gamma_{20} D_{i,20} + \varepsilon'_i \qquad [12.10]$$

Contrasting the two models produces a test for nonlinearity, because the model in Equation 12.9, specifying a linear relationship between vocabulary and education, is a special case of the model given in Equation 12.10, which can capture *any* pattern of relationship between $E(Y)$ and X. The resulting incremental F-test for nonlinearity appears in Table 12.2. There is, therefore, very strong evidence of a *linear* relationship between vocabulary and education, but little evidence of nonlinearity.

The incremental F-test for nonlinearity can easily be extended to a discrete independent variable—say X_1—in a multiple-regression model. Here, we need to contrast the general model:

$$Y_i = \alpha + \gamma_1 D_{i1} + \cdots + \gamma_{m-1} D_{i,m-1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

with the model specifying a linear effect of X_1:

$$Y_i = \alpha \ + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

where $D_1, \ldots, D_{m-1}$ are dummy regressors constructed to represent the m distinct values of X_1.

Another approach to testing for nonlinearity exploits the fact that a polynomial of degree $m - 1$ can perfectly capture the relationship between Y and a discrete X with m categories, regardless of the specific form of this relationship. We remove one term at a time from the model

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_{m-1} X_i^{m-1} + \varepsilon_i$$

beginning with X^{m-1}. If the decrement in the regression sum of squares is nonsignificant (by an incremental F-test on 1 degree of freedom), then we proceed to remove X^{m-2}, and so on.[39] This approach has the potential advantage of parsimony, because we may well require more than one term (i.e., a linear relationship) but fewer than $m-1$ (i.e., a relationship of arbitrary form). High-degree polynomials, however, are usually difficult to interpret.[40]

12.4.2 Testing for Nonconstant Error Variance

A discrete X (or combination of X's) partitions the data into m groups (as in analysis of variance). Let Y_{ij} denote the ith of n_j dependent-variable scores in group j. If the error variance is constant across groups, then the within-group sample variances

$$S_j^2 = \frac{\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2}{n_j - 1}$$

should be similar. Tests that examine the S_j^2 directly, such as Bartlett's (1937) classic (and commonly employed) test, do not maintain their validity well when the distribution of the errors is nonnormal.

Many alternative tests have been proposed. In a large-scale simulation study, Conover et al. (1981) found that the following simple F-test (called "Levine's test") is both robust and powerful: Calculate the values

$$Z_{ij} \equiv |Y_{ij} - \tilde{Y}_j|$$

where $\tilde{Y}_j$ is the median dependent-variable value in group j. Then perform a one-way analysis of variance of the Z_{ij} over the m groups. If the error variance is not constant across the groups, then the group means $\overline{Z}_j$ will tend to differ, producing a large value of the F-test statistic.[41]

For the vocabulary data, for example, where education partitions the 968 observations into $m = 20$ groups, this test gives $F_0 = 1.48$, with 19 and 948

[39] As usual, the estimate of error variance in the denominator of these F-tests is taken from the full model with all $m - 1$ terms.

[40] *There is a further, technical difficulty with this procedure: The several powers of X are usually highly correlated, sometimes to the point that least-squares calculations break down. A solution is to orthogonalize the power regressors prior to fitting the model: Let X^{2*} represent the residual from the regression (i.e., orthogonal projection) of X^2 on X; let X^{3*} represent the residual from the regression of X^3 on X and X^{2*}; and so on. The set of regressors $X, X^{2*}, \ldots, X^{m-1*}$ is orthogonal and spans the same subspace as the original set of powers. Because the new regressors are orthogonal, it is no longer necessary to fit successively smaller models sequentially; t-tests for the individual terms in the full model provide the same results as sequential incremental F-tests.

[41] This test ironically exploits the robustness of the F-test in one-way ANOVA. (The irony lies in the common use of tests of constant variance as a preliminary to tests of differences in means.)

degrees of freedom, for which $p = .08$. There is, therefore, weak evidence of nonconstant spread in vocabulary across the categories of education.

> Discrete independent variables divide the data into groups. A simple incremental F-test for nonlinearity compares the sum of squares accounted for by the linear regression of Y on X with the sum of squares accounted for by differences in the group means. Likewise, tests of nonconstant variance can be based on comparisons of spread in the different groups.

EXERCISE

12.13 Recall (from Section 8.2) Moore and Krupat's analysis of variance of conformity by authoritarianism and partner's status. The data are in Table 8.3 and `moore.dat`.

(a) Treating the three categories of authoritarianism as evenly spaced, fit a model to the data that incorporates the linear effect of this factor (e.g., coding the categories as 1, 2, and 3). Include the interaction between authoritarianism and partner's status in the model. Compare this model with the standard two-way ANOVA model to determine whether there is a significant departure from linearity.

(b) Test for nonconstant variance across the six cells of Moore and Krupat's design.

12.5 Maximum-Likelihood Methods*

A statistically sophisticated approach to selecting a transformation of Y or an X is to embed the usual linear model in a more general nonlinear model that contains a parameter for the transformation. If several variables are potentially to be transformed, or if the transformation is complex, then there may be several such parameters.[42]

Suppose that the transformation is indexed by a single parameter λ (e.g., $Y \rightarrow Y^{\lambda}$), and that we can write down the likelihood for the model as a function of the transformation parameter and the usual regression parameters: $L(\lambda, \alpha, \beta_1, \ldots, \beta_k, \sigma_\varepsilon^2)$. Maximizing the likelihood yields the maximum-likelihood estimate of λ along with the MLEs of the other parameters. Now

[42] Models of this type are fundamentally nonlinear, and can be treated by the general methods of Section 14.2.3, as well as by the methods described in the present section.

suppose that $\lambda = \lambda_0$ represents *no* transformation (e.g., $\lambda_0 = 1$ for the power transformation Y^λ). A likelihood-ratio test, Wald test, or score test of H_0: $\lambda = \lambda_0$ assesses the evidence that a transformation is required.

A disadvantage of the likelihood-ratio and Wald tests in this context is that they require finding the MLE, which usually necessitates iteration (i.e., a repetitive process of successively closer approximations). In contrast, the slope of the log likelihood at λ_0—on which the score test depends—generally can be assessed or approximated without iteration, and therefore is faster to compute.

Often, the score test can be formulated as the t-statistic for a new regressor, called a *constructed variable*, to be added to the linear model. A partial-regression plot for the constructed variable then can reveal whether one or a small group of observations is unduly influential in determining the transformation, or, alternatively, whether evidence for the transformation is spread throughout the data.

12.5.1 Box-Cox Transformation of Y

Box and Cox (1964) have suggested a power transformation of Y with the object of normalizing the error distribution, stabilizing the error variance, and straightening the relationship of Y to the X's.[43] The general Box-Cox model is

$$Y_i^{(\lambda)} = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

where the errors ε_i are independently $N(0, \sigma_\varepsilon^2)$, and

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log_e Y_i & \text{for } \lambda = 0 \end{cases}$$

Note that all of the Y_i must be positive.[44]

For a particular choice of λ, the conditional maximized log likelihood is[45]

$$\begin{aligned} \log_e L(\alpha, \beta_1, \ldots, \beta_k, \sigma_\varepsilon^2 | \lambda) = &-\frac{n}{2}(1 + \log_e 2\pi) \\ &-\frac{n}{2}\log_e \hat{\sigma}_\varepsilon^2(\lambda) + (\lambda - 1)\sum_{i=1}^{n} \log_e Y_i \end{aligned}$$

where $\hat{\sigma}_\varepsilon^2(\lambda) = \sum E_i^2(\lambda)/n$, and where the $E_i(\lambda)$ are the residuals from the least-squares regression of $Y^{(\lambda)}$ on the X's. The least-squares coefficients from this regression are the maximum-likelihood estimates of α and the β's, conditional on the value of λ.

[43] Subsequent work (Hernandez and Johnson, 1980) suggests that Box and Cox's method principally serves to normalize the error distribution.

[44] Strictly speaking, the requirement that the Y_i are positive precludes the possibility that they are normally distributed (because the normal distribution is unbounded), but this is not a serious practical difficulty unless many Y-values stack up near 0.

[45] See Exercise 12.14.

A simple procedure for finding the maximum-likelihood estimator $\hat{\lambda}$, then, is to evaluate the maximized $\log_e L$ (called the "profile log likelihood") for a range of values of λ, say between -2 and $+2$. If this range turns out not to contain the maximum of the log likelihood, then the range can be expanded. To test H_0: $\lambda = 1$, calculate the likelihood-ratio statistic

$$G_0^2 = -2[\log_e L(\lambda = 1) - \log_e L(\lambda = \hat{\lambda})]$$

which is asymptotically distributed as χ^2 with 1 degree of freedom under H_0. Alternatively (but equivalently), a 95% confidence interval for λ includes those values for which

$$\log_e L(\lambda) > \log_e L(\lambda = \hat{\lambda}) - 1.92$$

The figure 1.92 comes from $1/2 \times \chi^2_{1,.05} = 1/2 \times 1.96^2$.

Figure 12.8 shows a plot of the maximized log likelihood against λ for Ornstein's interlocking-directorate regression. The maximum-likelihood estimate of λ is $\hat{\lambda} = 0.31$, and a 95% confidence interval, marked out by the intersection of the line near the top of the graph with the log likelihood, runs from 0.20 to 0.41.[46]

Atkinson (1985) has proposed an approximate score test for the Box-Cox model, based on the constructed variable

$$G_i = Y_i\left[\log_e\left(\frac{Y_i}{\tilde{Y}}\right) - 1\right]$$

where $\tilde{Y}$ is the *geometric mean* of Y:[47]

$$\tilde{Y} \equiv (Y_1 \times Y_2 \times \cdots \times Y_n)^{1/n}$$

This constructed variable is obtained by a linear approximation to the Box-Cox transformation $Y^{(\lambda)}$ evaluated at $\lambda = 1$. The augmented regression, including the constructed variable, is then

$$Y_i = \alpha' + \beta_1' X_{i1} + \cdots + \beta_k' X_{ik} + \phi G_i + \varepsilon_i'$$

The t-test of H_0: $\phi = 0$, that is, $t_0 = \hat{\phi}/\widehat{\mathrm{SE}}(\hat{\phi})$, assesses the need for a transformation. The quantities $\hat{\phi}$ and $\widehat{\mathrm{SE}}(\hat{\phi})$ are obtained from the least-squares regression of Y on $X_1, \ldots, X_k$ and G. An estimate of λ (though not the MLE) is given by $\tilde{\lambda} = 1 - \hat{\phi}$; and the partial-regression plot for the constructed variable G shows influence and leverage on $\hat{\phi}$, and hence on the choice of λ.

Atkinson's constructed-variable plot for the interlocking-directorate regression is shown in Figure 12.9. Although the trend in the plot is not altogether

[46] Recall that we previously employed a square-root transformation for these data to make the residual distribution more nearly normal and to stabilize the error variance.

[47] It is more practical to compute the geometric mean as $\tilde{Y} = \exp[(\sum \log_e Y_i)/n]$.

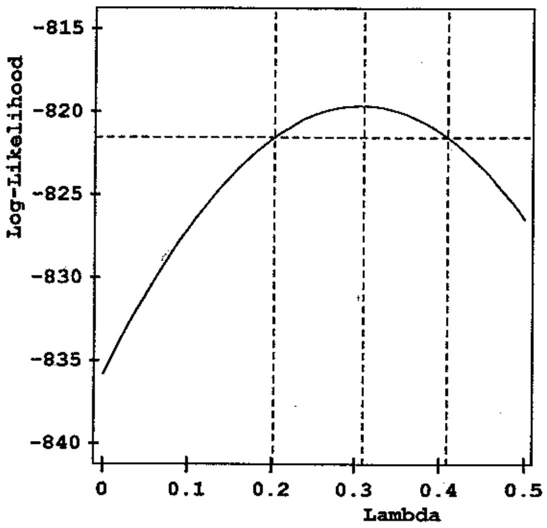

Figure 12.8. Box-Cox transformations for Ornstein's interlocking-directorate regression. The maximized log likelihood is plotted against the transformation parameter λ. The intersection of the line near the top of the graph with the log likelihood curve marks off a 95% confidence interval for λ. The maximum of the log likelihood corresponds to the MLE of λ.

linear, it appears that evidence for the transformation of Y is spread throughout the data and does not depend unduly on a small number of observations. The coefficient of the constructed variable in the regression is $\hat{\phi} = 0.585$, with $\widehat{\text{SE}}(\hat{\phi}) = 0.031$, providing very strong evidence of the need to transform Y. The suggested transformation, $\tilde{\lambda} = 1 - 0.585 = 0.415$, is close to the MLE (but just at the boundary of the narrow 95% confidence interval constructed around the MLE).

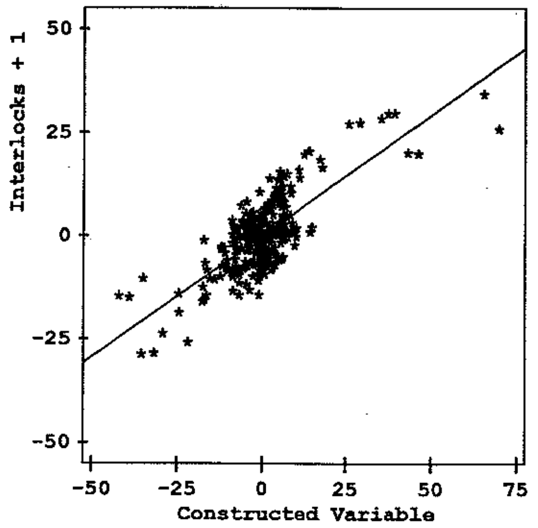

Figure 12.9. Constructed-variable plot for the Box-Cox transformation of Ornstein's interlocking-directorate regression. The least-squares line is shown on the plot.

12.5.2 Box-Tidwell Transformation of the X's

Now, consider the model

$$Y_i = \alpha + \beta_1 X_{i1}^{\gamma_1} + \cdots + \beta_k X_{ik}^{\gamma_k} + \varepsilon_i$$

where the errors are independently distributed as $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and all of the X_{ij} are positive. The parameters of this model—α, $\beta_1, \ldots, \beta_k$, $\gamma_1, \ldots, \gamma_k$, and σ_ε^2—could be estimated by general nonlinear least squares, but Box and Tidwell (1962) suggest instead a computationally more efficient procedure that also yields a constructed-variable diagnostic:[48]

1. Regress Y on $X_1, \ldots, X_k$, obtaining $A, B_1, \ldots, B_k$.
2. Regress Y on $X_1, \ldots, X_k$ and the constructed variables $X_1 \log_e X_1, \ldots, X_k \log_e X_k$, obtaining $A', B'_1, \ldots, B'_k$ and $D_1, \ldots, D_k$. Because of the presence of the constructed variables in this second regression, in general $A \neq A'$ and $B_j \neq B'_j$. As in the Box-Cox model, the constructed variables result from a linear approximation[49] to $X_j^{\gamma_j}$ evaluated at $\gamma_j = 1$.
3. The constructed variable $X_j \log_e X_j$ can be used to assess the need for a transformation of X_j by testing the null hypothesis H_0: $\delta_j = 0$, where δ_j is the population coefficient of $X_j \log_e X_j$ in step 2. Partial-regression plots for the constructed variables are useful for assessing leverage and influence on the decision to transform the X's.
4. A preliminary estimate of the transformation parameter γ_j (not the MLE) is given by

$$\tilde{\gamma}_j = 1 + \frac{D_j}{B_j}$$

Recall that B_j is from the *initial* (i.e., step 1) regression (not from step 2).

This procedure can be iterated through steps 1, 2, and 4 until the estimates of the transformation parameters stabilize, yielding the MLEs $\hat{\gamma}_j$.

For the Canadian occupational prestige data, leaving the regressors for percentage of women (W and W^2) untransformed, the coefficients of $S \log_e S$ (education) and $I \log_e I$ (income) are, respectively, $D_S = 5.30$ with $\widehat{\text{SE}}(D_S) = 2.20$, and $D_I = -0.00243$ with $\widehat{\text{SE}}(D_I) = 0.00046$. There is, consequently, much stronger evidence of the need to transform income than education.

The first-step estimates of the transformation parameters are

$$\tilde{\gamma}_S = 1 + \frac{D_S}{B_S} = 1 + \frac{5.30}{4.26} = 2.2$$

$$\tilde{\gamma}_I = 1 + \frac{D_I}{B_I} = 1 + \frac{-0.00243}{0.00127} = -0.91$$

The fully iterated MLEs of the transformation parameters are $\hat{\gamma}_S = 2.2$ and $\hat{\gamma}_I = -0.038$. Compare these values with the square and log transformations

[48] Nonlinear least-squares regression is described in Section 14.2.3.

[49] See Exercise 12.15.

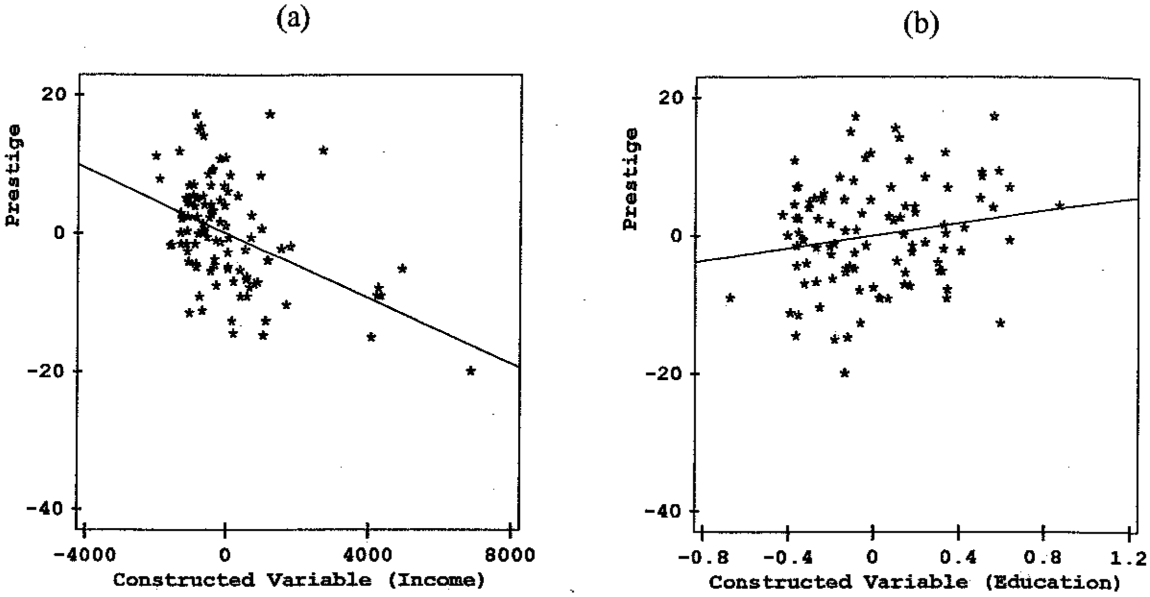

Figure 12.10. Constructed-variable plots for the Box-Tidwell transformation of (a) income and (b) education in the regression of occupational prestige on income, education, and percentage of women.

discovered following trial and error in Section 12.3.1.[50] Constructed-variable plots for the transformation of education and income, shown in Figure 12.10, suggest that there is general evidence for these transformations, although there are some high-leverage in-line observations in the income plot.

A statistically sophisticated general approach to selecting a transformation of Y or an X is to embed the linear-regression model in a more general model that contains a parameter for the transformation. The Box-Cox procedure selects a power transformation of Y to normalize the errors. The Box-Tidwell procedure selects power transformations of the X's to linearize the regression of Y on the X's. In both cases, "constructed-variable" plots help us to decide whether individual observations are unduly influential in determining the transformation parameters.

12.5.3 Nonconstant Error Variance Revisited

Breusch and Pagan (1979) develop a score test for heteroscedasticity based on the specification:

$$\sigma_i^2 \equiv V(\varepsilon_i) = g(\gamma_0 + \gamma_1 Z_{i1} + \cdots + \gamma_p Z_{ip})$$

[50] Recall from Section 12.3.1 that the power transformation of education is not wholly appropriate because the partial relationship between prestige and education did not appear to be simple.

where $Z_1, \ldots, Z_p$ are known variables, and where the function $g(\cdot)$ is quite general (and need not be explicitly specified). The same test was independently derived by Cook and Weisberg (1983). The score statistic for the hypothesis that the σ_i^2 are all the same, which is equivalent to H_0: $\gamma_1 = \cdots = \gamma_p = 0$, can be formulated as an auxiliary-regression problem.

Let $U_i \equiv E_i^2/\hat{\sigma}_\varepsilon^2$, where $\hat{\sigma}_\varepsilon^2 = \sum E_i^2/n$ is the MLE of the error variance.[51] The U_i are a type of standardized squared residual. Regress U on the Z's:

$$U_i = \eta_0 + \eta_1 Z_{i1} + \cdots + \eta_p Z_{ip} + \omega_i \qquad [12.11]$$

Breusch and Pagan (1979) show that the score statistic

$$S_0^2 = \frac{\sum(\hat{U}_i - \overline{U})^2}{2}$$

is asymptotically distributed as χ^2 with p degrees of freedom under the null hypothesis of constant error variance. Here, the $\hat{U}_i$ are fitted values from the regression of U on the Z's, and thus S_0^2 is half the regression sum of squares from fitting Equation 12.11.

To apply this result, it is, of course, necessary to select Z's, the choice of which depends on the suspected pattern of nonconstant error variance. If several patterns are suspected, then several score tests can be performed. Employing $X_1, \ldots, X_k$ in the auxiliary regression (Equation 12.11), for example, permits detection of a tendency of the error variance to increase with the values of one or more of the independent variables in the main regression.

Likewise, Cook and Weisberg (1983) suggest regressing U on the fitted values from the main regression (i.e., fitting the auxiliary regression $U_i = \eta_0 + \eta_1 \hat{Y}_i + \omega_i$), producing a 1-degree-of-freedom score test to detect the common tendency of the error variance to increase with the level of the dependent variable. When the error variance follows this pattern, the auxiliary regression of U on $\hat{Y}$ provides a more powerful test than the more general regression of U on the X's. A similar, but more complex, procedure is described by Anscombe (1961), who suggests correcting detected heteroscedasticity by transforming Y to $Y^{(\tilde{\lambda})}$ with $\tilde{\lambda} = 1 - \frac{1}{2}\hat{\eta}_1 \overline{Y}$.

Finally, White (1980) proposes a score test based on a comparison of his heteroscedasticity-corrected estimator of coefficient sampling variance with the usual estimator of coefficient variance.[52] If the two estimators are sufficiently different, then doubt is cast on the assumption of constant error variance. White's test can be implemented as an auxiliary regression of the squared residuals from the main regression, E_i^2, on all of the X's together with all of the squares and pairwise products of the X's. Thus, for $k = 2$ independent variables in the main regression, we would fit the model

$$E_i^2 = \delta_0 + \delta_1 X_{i1} + \delta_2 X_{i2} + \delta_{11} X_{i1}^2 + \delta_{22} X_{i2}^2 + \delta_{12} X_{i1} X_{i2} + \upsilon_i$$

[51] Note the division by n rather than by $n - 1$ in $\hat{\sigma}_\varepsilon^2$. See Section 9.3.3.

[52] White's coefficient-variance estimator is described in Section 12.2.3.

In general, there will be $p = k(k+3)/2$ terms in the auxiliary regression, plus the constant.

The score statistic for testing the null hypothesis of constant error variance is $S_0^2 = nR^2$, where R^2 is the squared multiple correlation from the auxiliary regression. Under the null hypothesis, S_0^2 follows an asymptotic χ^2 distribution with p degrees of freedom.

Because all of these score tests are potentially sensitive to violations of model assumptions other than constant error variance, it is important, in practice, to supplement the tests with graphical diagnostics, as suggested by Cook and Weisberg (1983). When there are several Z's, a simple diagnostic is to plot U_i against $\hat{U}_i$, the fitted values from the auxiliary regression. We can also construct partial-regression plots for the Z's in the auxiliary regression. When U_i is regressed on $\hat{Y}_i$, these plots convey essentially the same information as the plot of studentized residuals against fitted values proposed in Section 12.2.

Simple score tests are available to determine the need for a transformation and to test for nonconstant error variance.

Applied to Ornstein's interlocking-directorate data, an auxiliary regression of U on $\hat{Y}$ yields $\hat{U} = 0.134 + 0.0594\hat{Y}$, and $S_0^2 = 147.6/2 = 73.8$ on 1 degree of freedom. There is, consequently, very strong evidence that the error variance increases with the level of the dependent variable. The suggested variance-stabilizing transformation using Anscombe's rule is $\tilde{\lambda} = 1 - \frac{1}{2}(0.0594)(14.58) = 0.57$. Compare this value with those produced by the Box-Cox model ($\hat{\lambda} = 0.3$, in Section 12.5.1) and by trial and error ($\lambda = 0.5$, in Section 12.2).

An auxiliary regression of U on the independent variables in the main regression yields $S_0^2 = 172.6/2 = 86.3$ on $k = 13$ degrees of freedom, and thus also provides strong evidence against constant error variance. Examination of the coefficients from the auxiliary regression (not shown here) indicates, in particular, a tendency of the error variance to increase with assets. The score statistic for the more general test is not much larger than that for the regression of U on $\hat{Y}$, however, suggesting that the pattern of nonconstant error variance is indeed for the spread of the errors to increase with the level of Y. Assets are, of course, an important component of $\hat{Y}$. Because White's test requires 104 regressors for this problem, it was not performed.

EXERCISES

12.14 *Box-Cox transformations of Y: In matrix form, the Box-Cox regression model given in Section 12.5.1 can be written as

$$\mathbf{y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(a) Show that the probability density for the observations is given by

$$p(y) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\left[-\frac{\sum_{i=1}^n (Y_i^{(\lambda)} - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma_\varepsilon^2}\right] \prod_{i=1}^n Y_i^{\lambda-1}$$

where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$. (*Hint*: $Y_i^{\lambda-1}$ is the Jacobian of the transformation from Y_i to ε_i.)

(b) For a given value of λ, the *conditional* maximum-likelihood estimator of $\boldsymbol{\beta}$ is the least-squares estimator

$$\mathbf{b}_\lambda = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{(\lambda)}$$

(Why?) Show that the maximized log likelihood can be written as

$$\log_e L(\alpha, \beta_1, \ldots, \beta_k, \sigma_\varepsilon^2|\lambda)$$
$$= -\frac{n}{2}(1 + \log_e 2\pi) - \frac{n}{2}\log_e \hat{\sigma}_\varepsilon^2(\lambda) + (\lambda - 1)\sum_{i=1}^n \log_e Y_i$$

as stated in the text.

12.15 *Box-Tidwell transformations of the X's: Recall the Box-Tidwell model

$$Y = \alpha + \beta_1 X_1^{\gamma_1} + \cdots + \beta_k X_k^{\gamma_k} + \varepsilon$$

and focus on the first regressor, X_1. Show that the first-order Taylor series approximation for $X_1^{\gamma_1}$ at $\gamma_1 = 1$ is

$$X_1^{\gamma_1} \simeq X_1 + (\gamma_1 - 1)X_1 \log_e X_1$$

providing the basis for the constructed variable $X_1 \log_e X_1$.

12.6 Structural Dimension*

In discussing the use and potential failure of partial-residual plots as a diagnostic for nonlinearity, I explained that it is unreasonable to expect that a collection of two- or three-dimensional graphs can, in every instance, adequately capture the dependence of Y on the X's: The surface representing this dependence lies, after all, in a space of $k + 1$ dimensions. Relying primarily on Cook (1994), I shall now briefly consider the geometric notion of dimension in regression analysis, along with the implications of this notion for diagnosing problems with regression models that have been fit to data.[53] The *structural dimension* of a regression problem corresponds to the dimensionality of the smallest subspace of the X's required to represent the dependency of Y on the X's.

[53] An extended discussion of structural dimension, at a much simpler level than Cook (1994), may be found in Cook and Weisberg (1994).

Let us initially suppose that the distribution of Y is completely independent of the independent variables $X_1, \ldots, X_k$. Then, in Cook and Weisberg's (1994) terminology, an "ideal summary" of the data is simply the univariate, unconditional distribution of Y—represented, say, by the density $p(y)$. In a sample, we could compute a density estimate, a histogram, or some other univariate display. In this case, the *structural dimension* of the data is 0.

Now suppose that Y depends on the X's only through the regression equation

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

where $E(\varepsilon) = 0$ and the distribution of the error is independent of the X's. Then the expectation of Y conditional on the X's is a linear function of the X's:

$$E(Y|x_1, \ldots, x_k) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

A plot of Y against $\alpha+\beta_1 X_1+\cdots+\beta_k X_k$, therefore, constitutes an ideal summary of the data. This two-dimensional plot shows the systematic component of Y in an edge-on view of the regression hyperplane, and also shows the conditional variation of Y around the hyperplane (i.e., the variation of the errors).

Because the subspace spanned by the linear combination $\alpha + \beta_1 X_1 + \cdots + \beta_k X_k$ is one dimensional, the structural dimension of the data is 1. In a sample, the ideal summary is a two-dimensional scatterplot of Y_i against $\hat{Y}_i = A + B_1 X_{i1} + \cdots + B_k X_{ik}$; the regression line in this plot is an edge-on view of the fitted least-squares surface.

The structural dimension of the data can be 1 even if the regression is nonlinear or if the errors are not identically distributed, as long as the expectation of Y and the distribution of the errors depend only on a single linear combination of the X's—that is, a subspace of dimension 1. The structural dimension is 1, for example, if

$$E(Y|x_1, \ldots, x_k) = f(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k) \qquad [12.12]$$

and

$$V(Y|x_1, \ldots, x_k) = g(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k) \qquad [12.13]$$

where the mean function $f(\cdot)$ and the variance function $g(\cdot)$, though generally different, depend on the *same* linear function of the X's. In this case, a plot of Y against $\alpha + \beta_1 X_1 + \cdots + \beta_k X_k$ is still an ideal summary of the data, showing the nonlinear dependency of the expectation of Y on the X's, along with the pattern of nonconstant error variance.

Similarly, we hope to see these features of the data in a sample plot of Y against $\hat{Y}$ from the *linear* regression of Y on the X's (even though the linear regression does not itself capture the dependency of Y on the X's). It turns out, however, that the plot of Y against $\hat{Y}$ can fail to reflect the mean and variance functions accurately if the X's themselves are not linearly related—even when the true structural dimension is 1 (i.e., when Equations 12.12 and 12.13 hold).[54] This, then, is another context in which linearly related independent variables are desirable.[55] Linearly related independent variables are not required here if the true regression is linear—something that, however, we are typically not in a position to know prior to examining the data.

> The structural dimension of a regression is the dimensionality of the smallest subspace of the independent variables required, along with the dependent variable, to represent the dependence of Y on the X's. When Y is completely independent of the X's, the structural dimension is 0, and an ideal summary of the data is simply the unconditional distribution of Y. When the linear-regression model holds—or when the conditional expectation and variance of Y are a function of a single linear combination of the X's—the structural dimension is 1.

The structural dimension of the data exceeds 1 if Equations 12.12 and 12.13 do not both hold. If, for example, the mean function depends on one linear combination of the X's:

$$E(Y|x_1, \ldots, x_k) = f(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)$$

and the variance function on a different linear combination

$$V(Y|x_1, \ldots, x_k) = g(\gamma + \delta_1 x_1 + \cdots + \delta_k x_k)$$

then the structural dimension is 2.

Correspondingly, if the mean function depends on two different linear combinations of the X's, implying interaction among the X's,

$$E(Y|x_1, \ldots, x_k) = f(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k, \gamma + \delta_1 x_1 + \cdots + \delta_k x_k)$$

while the errors are independent of the X's, then the structural dimension is also 2. When the structural dimension is 2, a plot of Y against $\hat{Y}$ (from the linear regression of Y on the X's) is necessarily incomplete.

[54] See Exercise 12.16.

[55] The requirement of linearity here is, in fact, stronger than pairwise linear relationships among the X's: The regression of any linear function of the X's on any set of linear functions of the X's must be linear. If the X's are multivariate normal, then this condition is necessarily satisfied (although it may be satisfied even if the X's are not normal). It is not possible to check for linearity in this strict sense when there are more than two or three X's, but there is some evidence that checking pairs—and perhaps triples—of X's is usually sufficient. See Cook and Weisberg (1994). Cf. Section 12.3.2 for the conditions under which partial-residual plots are informative.

These observations are interesting, but their practical import—beyond the advantage of linearly related regressors—is unclear: Short of modeling the regression of Y on the X's nonparametrically, we can never be sure that we have captured all of the structure of the data in a lower-dimensional subspace of the independent variables.

There is, however, a further result that does have direct practical application: Suppose that the independent variables are linearly related and that there is one-dimensional structure. Then the *inverse regressions* of each of the independent variables on the dependent variable have the following character:[56]

$$E(X_j|y) = \mu_j + \eta_j m(y) \qquad [12.14]$$
$$V(X_j|y) \simeq \sigma_j^2 + \eta_j^2 v(y)$$

Equation 12.14 has two special features that are useful in checking whether one-dimensional structure is reasonable for a set of data:[57]

1. Most important, the functions $m(\cdot)$ and $v(\cdot)$, through which the means and variances of the X's depend on Y, are the same for all of the X's. Consequently, if the scatterplot of X_1 against Y shows a linear relationship, for example, then the scatterplots of each of $X_2, \ldots, X_k$ against Y must also show linear relationships. If one of these relationships is quadratic, in contrast, then the others must be quadratic. Likewise, if the variance of X_1 increases linearly with the level of Y, then the variances of the other X's must also be linearly related to Y. There is only one exception: The constant η_j can be 0, in which case the mean and variance of the corresponding X_j are *unrelated* to Y.
2. The constant η_j appears in the formula for the conditional mean of X_j, and η_j^2 in the formula for its conditional variance, placing constraints on the patterns of these relationships. If, for example, the mean of X_1 is unrelated to Y, then so should the variance.

The sample inverse regressions of the X's on Y can be conveniently examined in the first column of the scatterplot matrix for $\{Y, X_1, \ldots, X_k\}$.[58]

If the structural dimension is 1, and if the independent variables are linearly related to one another, then the inverse regressions of the independent variables on the dependent variable all have the same general form.

[56] See Exercise 12.17 for illustrative applications.

[57] Equation 12.14 is the basis for formal dimension-testing methods, such as *sliced inverse regression* (Duan and Li, 1991) and related techniques. See Cook and Weisberg (1994) for an introductory treatment of dimension testing and for additional references.

[58] See, for example, Figure 3.15.

EXERCISES

12.16 Experimenting with structural dimension: Generate random samples of 100 observations according to each of the following schemes. In each case, fit the linear regression of Y on X_1 and X_2, and plot the values of Y against the resulting fitted values $\hat{Y}$. Do these plots accurately capture the dependence of Y on X_1 and X_2 ? To decide this question in each case, it may help (1) to draw graphs of $E(Y|x_1, x_2) = f(\alpha + \beta_1 x_1 + \beta_2 x_2)$ and $V(Y|x_1, x_2) = g(\alpha + \beta_1 x_1 + \beta_2 x_2)$ over the observed range of values for $\alpha + \beta_1 X_1 + \beta_2 X_2$; and (2) to plot a nonparametric-regression smooth in the plot of Y against $\hat{Y}$. Whenever they appear, E and U are $N(0, 1)$ and independent of each other and of the other variables.

(a) Independent X's, a linear regression, and constant error variance: X_1 and X_2 independent and uniformly distributed on the interval $[0, 1]$; $E(Y|x_1, x_2) = x_1 + x_2$; $V(Y|x_1, x_2) = 0.1E$.

(b) Independent X's, mean and variance of Y dependent on the same linear function of the X's: X_1 and X_2 independent and uniformly distributed on the interval $[0, 1]$; $E(Y|x_1, x_2) = (x_1+x_2-1)^2$; $V(Y|x_1, x_2) = 0.1 \times |x_1+x_2-1| \times E$.

(c) Linearly related X's, mean and variance of Y dependent on the same linear function of the X's: X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = X_1+0.1U$; $E(Y|x_1, x_2) = (x_1+x_2-1)^2$; $V(Y|x_1, x_2) = 0.1 \times |x_1+x_2-1| \times E$.

(d) Nonlinearly related X's, mean and variance of Y dependent on the same linear function of the X's: X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = |X_1 - 0.5|$; $E(Y|x_1, x_2) = (x_1+x_2-1)^2$; $V(Y|x_1, x_2) = 0.1 \times |x_1+x_2-1| \times E$.

12.17 Dimension checking: Apply the dimension-checking conditions

$$E(X_j|y) = \mu_j + \eta_j m(y)$$

$$V(X_j|y) \simeq \sigma_j^2 + \eta_j^2 v(y)$$

to each of the following regression analyses. In each case, construct the scatterplot matrix for the independent variables and the dependent variable. If the independent variables do not appear to be linearly related, attempt to make their relationship more nearly linear by transforming one or more of the X's. Then examine the column of the scatterplot matrix that shows the relationship of each X to Y. Are these relationships qualitatively similar, as required for one-dimensional structure?

(a) Duncan's regression of prestige on income and education, for 45 U.S. occupations (Table 3.2 and `duncan.dat`).

(b) The regression of prestige on income, education, and percentage of women, for the Canadian occupational prestige data (`prestige.dat`).

(c) Angell's regression of moral integration on ethnic heterogeneity and geographic mobility, for 43 U.S. cities (Table 2.3 and `angell.dat`).

(d) Anscombe's regression of state education expenditures on per-capita income, proportion under 18 years of age, and proportion urban (Table 5.1 and `anscombe.dat`).

12.7 Summary

- Heavy-tailed errors threaten the efficiency of least-squares estimation; skewed and multimodal errors compromise the interpretation of the least-squares fit. Nonnormality can often be detected by examining the distribution of the least-squares residuals, and frequently can be corrected by transforming the data.
- It is common for the variance of the errors to increase with the level of the dependent variable. This pattern of nonconstant error variance ("heteroscedasticity") can often be detected in a plot of residuals against fitted values. Strategies for dealing with nonconstant error variance include transformation of the dependent variable to stabilize the variance; the substitution of weighted-least-squares estimation for ordinary least squares; and the correction of coefficient standard errors for heteroscedasticity. A rough rule is that nonconstant error variance seriously degrades the least-squares estimator only when the ratio of the largest to smallest variance is about 10 or more.
- Simple forms of nonlinearity can often be detected in partial-residual plots. Once detected, nonlinearity can frequently be accommodated by variable transformations or by altering the form of the model (to include a quadratic term in an independent variable, for example). Partial-residual plots adequately reflect nonlinearity when the independent variables are themselves linearly related. More complex versions of these displays, such as augmented partial-residual plots and CERES plots, are more robust.
- Discrete independent variables divide the data into groups. A simple incremental F-test for nonlinearity compares the sum of squares accounted for by the linear regression of Y on X with the sum of squares accounted for by differences in the group means. Likewise, tests of nonconstant variance can be based on comparisons of spread in the different groups.
- A statistically sophisticated general approach to selecting a transformation of Y or an X is to embed the linear-regression model in a more general model that contains a parameter for the transformation. The Box-Cox procedure selects a power transformation of Y to normalize the errors. The Box-Tidwell procedure selects power transformations of the X's to linearize the regression of Y on the X's. In both cases, "constructed-variable" plots help us to decide whether individual observations are unduly influential in determining the transformation parameters.
- Simple score tests are available to determine the need for a transformation and to test for nonconstant error variance.
- The structural dimension of a regression is the dimensionality of the smallest subspace of the independent variables required, along with the dependent variable, to represent the dependence of Y on the X's. When Y is completely independent of the X's, the structural dimension is 0, and an ideal summary of the data is simply the unconditional distribution of Y. When the linear-regression model

holds—or when the conditional expectation and variance of Y are a function of a single linear combination of the X's—the structural dimension is 1. If the structural dimension is 1, and if the independent variables are linearly related to one another, then the inverse regressions of the independent variables on the dependent variable all have the same general form.

EXERCISES

12.18 Use the methods of this chapter to check for nonnormality, nonconstant error variance, and nonlinearity in each of the following regressions. In each case, attempt to correct any problems that are detected. Because many different methods are discussed in this chapter, you might find the following strategy useful: Use relatively simple diagnostics to check for problems and more sophisticated methods to follow up. To check for nonnormality, construct a quantile comparison plot and a histogram of the studentized residuals; to check for nonconstant error variance, plot studentized residuals against fitted values; to check for nonlinearity, examine partial-residual plots.

(a) Angell's regression of moral integration of U.S. cities on ethnic heterogeneity and geographic mobility (Table 2.3 and `angell.dat`).

(b) Anscombe's regression of state education expenditures on income, proportion under 18, and proportion urban (Table 5.1 and `anscombe.dat`).

12.19 Using Leinhardt and Wasserman's data on national infant mortality rates (given in `leinhard.dat`), regress infant mortality on income and dummy regressors for region. Using the methods of this chapter and the previous one, check the adequacy of the model and attempt to correct any problems that you find.

12.8 Recommended Reading

Methods for diagnosing problems in regression analysis and for visualizing regression data are an active area of research in statistics. The following texts summarize the current state of the art and include extensive references to the journal literature.

- Cook and Weisberg (1994) present a lucid and accessible treatment of many of the topics discussed in this chapter. They also describe a computer program, written in Lisp-Stat, that implements the graphical methods presented in their book (and much more). A copy of the program, called R-Code, and many programmed demonstrations, are included with the book.
- Cleveland (1993) describes novel graphical methods for regression data, including two-dimensional, three-dimensional, and higher-dimensional displays.

- Atkinson (1985) has written an interesting, if somewhat idiosyncratic, book which stresses the author's important contributions to regression diagnostics. There is, therefore, an emphasis on diagnostics that yield constructed-variable plots. This text includes a strong treatment of transformations, and a discussion of the extension of least-squares diagnostics to generalized linear models (e.g., logistic regression, as described in Chapter 15).

13

Collinearity and Its Purported Remedies

As I have explained,[1] when there is a perfect linear relationship among the regressors in a linear model, the least-squares coefficients are not uniquely defined. A strong, but less-than-perfect, linear relationship among the X's causes the least-squares coefficients to be unstable: Coefficient standard errors are large, reflecting the imprecision of estimation of the β's; consequently, confidence intervals for the β's are broad. Small changes in the data—even, in extreme cases, due to rounding errors—can substantially alter the least-squares coefficients; and relatively large changes in the coefficients from the least-squares values hardly increase the sum of squared residuals from its minimum (i.e., the least-squares coefficients are not sharply defined).

This chapter describes methods for detecting collinearity and techniques that are often employed for dealing with collinearity when it is present. I need to make three important points at the outset, however:

1. Except in certain specific contexts—such as time series regression[2]—collinearity is a comparatively rare problem in social science applications of linear models: Insufficient variation in independent variables, small samples, and large error variance (i.e., weak relationships) are much more frequently the source of imprecision in estimation.
2. Methods that are commonly employed as cures for collinearity—in particular, biased estimation and variable selection—can easily be worse than the disease.

[1] See Sections 5.2 and 9.2.

[2] See Section 14.1.

A principal goal of this chapter is to explain the substantial limitations of this statistical snake oil.

3. It is not at all obvious that the detection of collinearity in data has practical implications. There are, as mentioned in point 1, several sources of imprecision in estimation, which can augment or partially offset each other. The standard errors of the regression estimates are the bottom line: If these estimates are sufficiently precise, then the degree of collinearity is irrelevant; if the estimates are insufficiently precise, then knowing that the culprit is collinearity is of use only if the study can be redesigned to decrease the correlations among the X's. In observational studies, where the X's are sampled along with Y, it is usually impossible to influence their correlational structure, but it may very well be possible to increase the precision of estimation by increasing the sample size or by decreasing the error variance.[3]

13.1 Detecting Collinearity

We have encountered the notion of collinearity at several points, and it is therefore useful to summarize what we know:

- When there is a perfect linear relationship among the X's,

$$c_1X_{i1} + c_2X_{i2} + \cdots + c_kX_{ik} = c_0$$

1. the least-squares normal equations do not have a unique solution; and
2. the sampling variances of the regression coefficients are infinite.

*Points 1 and 2 follow from the observation that the matrix $\mathbf{X'X}$ of sums of squares and products is singular. Moreover, because the columns of $\mathbf{X}$ are perfectly collinear, the regressor subspace is of deficient dimension.

Perfect collinearity is usually the product of some error in formulating the linear model, such as failing to employ a baseline category in dummy regression.

- When collinearity is less than perfect:

1. The sampling variance of the least-squares slope coefficient B_j is

$$V(B_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\varepsilon^2}{(n-1)S_j^2}$$

where R_j^2 is the squared multiple correlation for the regression of X_j on the other X's, and $S_j^2 = \sum(X_{ij} - \overline{X}_j)^2/(n-1)$ is the variance of X_j. The term $1/(1 - R_j^2)$, called the *variance inflation factor* (VIF), directly and straightforwardly indicates the impact of collinearity on the precision of the estimate B_j. Because the precision of estimation of β_j is most naturally expressed as the width of the confidence interval for this parameter, and because the width of the confidence interval is proportional to the standard error of B_j (not its variance), I recommend examining the square root of the VIF in preference to the VIF itself.

[3] The error variance can sometimes be decreased by improving the procedures of the study or by introducing additional independent variables. The latter remedy may, however, increase collinearity, and may change the nature of the research. It may be possible, in some contexts, to increase precision by increasing the variation of the X's, but only if their values are under the control of the researcher, in which case collinearity could also be reduced. Sometimes, however, researchers may be able to exert indirect control over the variational and correlational structure of the X's by selecting a research setting judiciously or by designing an advantageous sampling procedure.

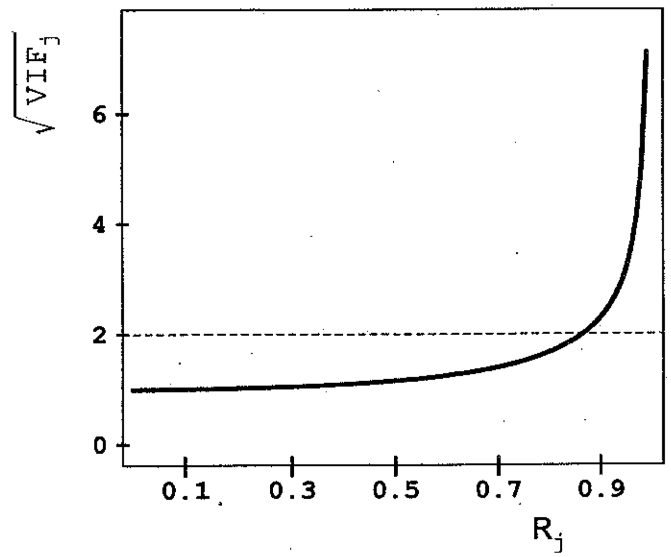

Figure 13.1. Precision of estimation (square root of the variance inflation factor) of β_j as a function of the multiple correlation between X_j and the other independent variables. It is not until the multiple correlation gets very large that the precision of estimation is seriously degraded.

Figure 13.1 reveals that the linear relationship among the X's must be very strong before collinearity seriously impairs the precision of estimation: It is not until R_j approaches .9 that the precision of estimation is halved.

Because of its simplicity and direct interpretation, the VIF (or its square root) is the principal diagnostic for collinearity. It is not, however, applicable to sets of related regressors, such as sets of dummy-variable coefficients, or coefficients for polynomial regressors.[4]

2. When X_1 is strongly collinear with the other regressors, the residuals $X^{(1)}$ from the regression of X_1 on $X_2, \ldots, X_k$ show little variation—most of the variation in X_1 is accounted for by its regression on the other X's. The partial-regression plot graphs the residuals from the regression of Y on $X_2, \ldots, X_k$ against $X^{(1)}$, converting the multiple regression into a simple regression.[5] Because the independent variable in this plot, $X^{(1)}$, is nearly invariant, the slope B_1 is subject to substantial sampling variation.[6]
3. *Confidence intervals for individual regression coefficients are projections of the confidence interval generating ellipse. Because this ellipse is the inverse—that is, the rescaled, 90° rotation—of the data ellipse for the independent variables, the individual confidence intervals for the coefficients are wide. If the correlations among the X's are positive, however, then there is substantial information in the data about the *sum* of the regression coefficients, if not about individual coefficients.[7]

[4] Section 13.1.2 describes a generalization of variance inflation to sets of related regressors.

[5] More precisely, the multiple regression is converted into a sequence of simple regressions, for each X in turn. Partial-regression plots are discussed in Section 11.6.

[6] See Stine (1995) for a nice graphical interpretation of this point.

[7] See Section 9.4.

When the regressors in a linear model are perfectly collinear, the least-squares coefficients are not unique. Strong, but less-than-perfect, collinearity substantially increases the sampling variances of the least-squares coefficients and can render them useless as estimators. The variance inflation factor $\text{VIF}_j = 1/(1 - R_j^2)$ indicates the deleterious impact of collinearity on the precision of the estimate B_j.

Figures 13.2 and 13.3 provide further insight into collinearity, illustrating its effect on estimation when there are two independent variables in a regression.

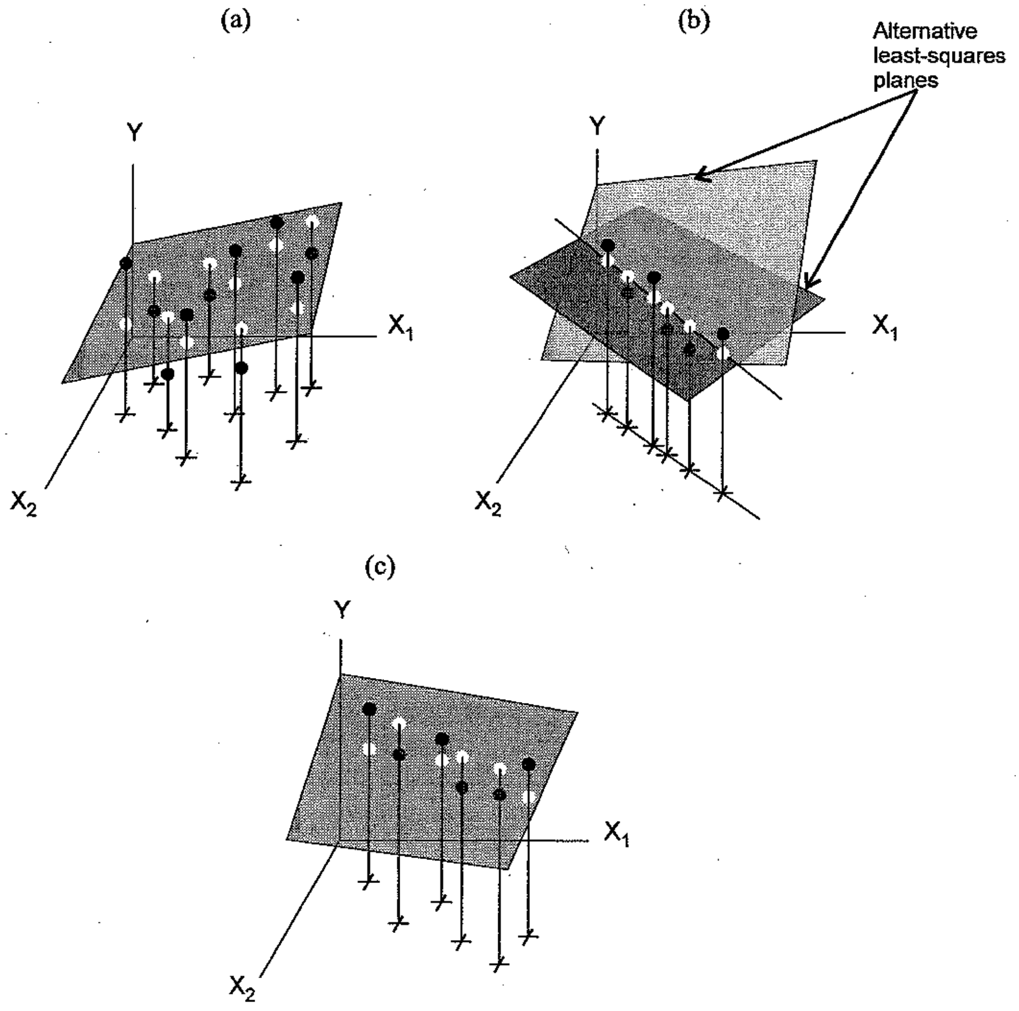

Figure 13.2. The impact of collinearity on the stability of the least-squares regression plane. In (*a*), the correlation between X_1 and X_2 is small, and the regression plane therefore has a broad base of support. In (*b*), X_1 and X_2 are perfectly correlated; the least-squares plane is not uniquely defined. In (*c*), there is a strong, but less-than-perfect, linear relationship between X_1 and X_2; the least-squares plane is uniquely defined, but it is not well supported by the data.

The black and gray dots in Figure 13.2 represent the data points (the gray dots are below the regression plane), while the white dots represent fitted values lying in the regression plane; the +'s show the projection of the data points onto the X_1, X_2 plane. Figure 13.3 shows the sum of squared residuals as a function of the slope coefficients B_1 and B_2. The residual sum of squares is at a minimum,

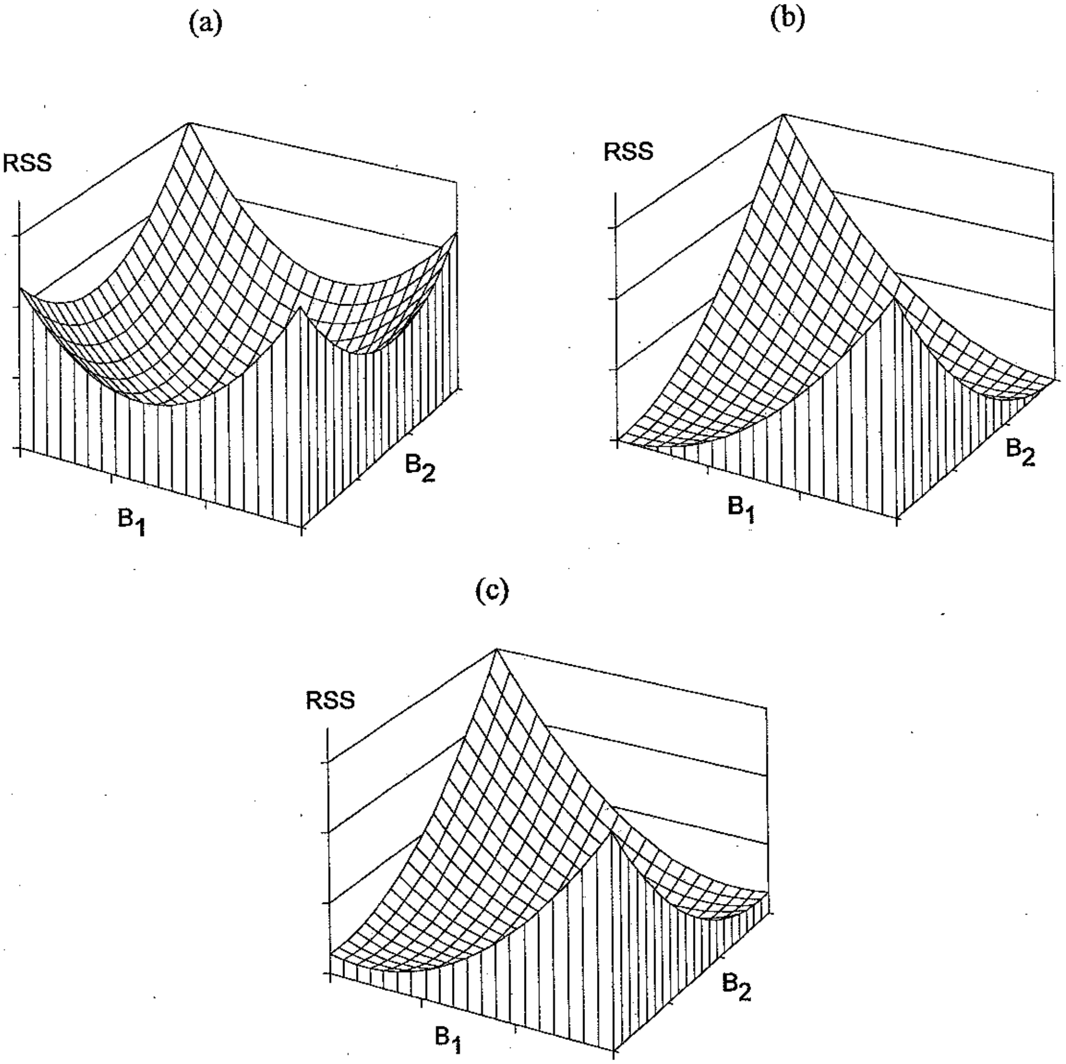

Figure 13.3. The residual sum of squares as a function of the slope coefficients B_1 and B_2. In each graph, the vertical axis is scaled so that the least-squares value of RSS is at the bottom of the axis. When, as in (*a*), the correlation between the independent variables X_1 and X_2 is small, the residual sum of squares has a well-defined minimum, much like a deep bowl. When there is a perfect linear relationship between X_1 and X_2, as in (*b*), the residual sum of squares is flat at its minimum, above a line in the B_1, B_2 plane: The least-squares values of B_1 and B_2 are not unique. When, as in (*c*), there is a strong, but less-than-perfect, linear relationship between X_1 and X_2, the residual sum of squares is nearly flat at its minimum, so values of B_1 and B_2 quite different from the least-squares values are associated with residual sums of squares near the minimum.

of course, when the B's are equal to the least-squares estimates; the vertical axis is scaled so that the minimum is at the "floor" of the figures.[8]

In Figure 13.2(a), the correlation between the independent variables X_1 and X_2 is slight, as indicated by the broad scatter of points in the X_1, X_2 plane. The least-squares regression plane, also shown in this figure, therefore has a firm base of support. Correspondingly, Figure 13.3(a) shows that small changes in the regression coefficients are associated with relatively large increases in the residual sum of squares—the sum-of-squares function is like a deep bowl, with steep sides and a well-defined minimum.

In Figure 13.2(b), X_1 and X_2 are perfectly collinear. Because the independent-variable observations form a line in the X_1, X_2 plane, the least-squares regression plane, in effect, also reduces to a line. The plane can tip about this line without changing the residual sum of squares, as Figure 13.3(b) reveals: The sum-of-squares function is flat at its minimum along a line defining pairs of values for B_1 and B_2—rather like a sheet of paper with two corners raised—and thus there are an infinite number of pairs of coefficients (B_1, B_2) that yield the minimum RSS.

Finally, in Figure 13.2(c), the linear relationship between X_1 and X_2 is strong, though not perfect. The support afforded to the least-squares plane is tenuous, so that the plane can be tipped without causing large increases in the residual sum of squares, as is apparent in Figure 13.3(c)—the sum-of-squares function is like a shallow bowl with a poorly defined minimum.

Consider the regression analysis reported in Table 13.1, from data presented by Ericksen et al. (1989).[9] The object here was to develop a prediction equation to improve estimates of the 1980 U.S. Census undercount. It is well established that the census fails to count all residents of the country, and that the likelihood of being missed is greater for certain categories of individuals, such as nonwhites, the poor, and residents of large cities. The dependent variable in the regression is a preliminary estimate of the undercount for each of 66 areas into which the authors divided the country. The 66 areas include 16 large cities, the remaining portions of the 16 states in which the cities are located, and the other 34 states. The preliminary estimates are regressed on eight predictors thought to influence the undercount:

1. the percentage black or Hispanic ("Minority" in Table 13.1);
2. the rate of serious crimes per 1000 population ("Crime");
3. the percentage poor ("Poverty");
4. the percentage having difficulty speaking or writing English ("Language");
5. the percentage aged 25 or older who have *not* finished high school ("High School");
6. the percentage of housing in small, multiunit buildings ("Housing");
7. a dummy variable coded 1 for cities, 0 for states or state remainders ("City"); and

[8] For each pair of slopes B_1 and B_2, the intercept A is chosen to make the residual sum of squares as small as possible.

[9] The authors employed a weighted-least-squares regression (see Section 12.2.2) to take account of differences in precision of initial estimates of the undercount in the 66 areas. The results reported here, in contrast, are for an ordinary-least-squares regression.

TABLE 13.1 Regression of Estimated 1980 U.S. Census Undercount on Area Characteristics, for 66 Central Cities, State Remainders, and States

Predictor	*Coefficient*	*Standard Error*	$\sqrt{VIF}$
Constant	−1.77	1.38	—
Minority	0.0798	0.0226	2.24
Crime	0.0301	0.0130	1.83
Poverty	−0.178	0.0849	2.15
Language	0.215	0.0922	1.28
High school	0.0613	0.0448	2.15
Housing	−0.0350	0.0246	1.37
City	1.16	0.77	1.88
Conventional	0.0370	0.0093	1.30
R^2	.708		

Source of Data: Ericksen et al. (1989).

8. the percentage of households counted by "conventional" personal enumeration, as opposed to mail-back questionnaire with follow-ups ("Conventional").

Correlations among the eight predictors appear in Table 13.2. Although some of the pairwise correlations are fairly large—the biggest are about .75—none is close to 1. It is apparent from the square-root VIFs shown in Table 13.1, however, that the precision of several of the regression estimates—in particular, the coefficients for Minority, Poverty, and High School—suffer from moderate collinearity. This result illustrates that collinearity in multiple regression is not restricted to pairwise relationships between regressors; sometimes the term *multicollinearity* is employed to emphasize this point.

13.1.1 Principal Components*

The method of principal components, developed in the early part of the 20th century by K. Pearson and H. Hotelling, provides a useful representation of the correlational structure of a set of variables. I shall develop the method briefly here, with particular reference to its application to collinearity in regression; more complete accounts can be obtained from texts on multivariate

TABLE 13.2 Correlations Among Eight Predictors of the 1980 U.S. Census Undercount

Predictor	*Minority*	*Crime*	*Poverty*	*Language*	*High School*	*Housing*	*City*
Crime	.655						
Poverty	.738	.369					
Language	.395	.512	.152				
High school	.535	.067	.751	−.116			
Housing	.357	.532	.335	.340	.235		
City	.758	.729	.538	.480	.315	.566	
Conventional	−.334	−.233	−.157	−.108	−.414	−.086	−.269

statistics (e.g., Morrison, 1976, Chapter 8). Because the material in this section is relatively complex, the section includes a summary; you may, on first reading, wish to pass lightly over most of the section and refer primarily to the summary, and to the two-variable case, which is treated immediately prior to the summary.

We begin with the vectors of standardized regressors, $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$. Because vectors have length equal to the square root of their sum of squared elements, each $\mathbf{z}_j$ has length $\sqrt{n-1}$. As we shall see, the *principal components* $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p$ provide an orthogonal basis for the regressor subspace.[10] The first principal component, $\mathbf{w}_1$, is oriented so as to account for maximum collective variation in the $\mathbf{z}_j$; the second principal component, $\mathbf{w}_2$, is orthogonal to $\mathbf{w}_1$, and—under this restriction of orthogonality—is oriented to account for maximum remaining variation in the $\mathbf{z}_j$; the third component, $\mathbf{w}_3$, is orthogonal to $\mathbf{w}_1$ and $\mathbf{w}_2$; and so on. Each principal component is scaled so that its variance is equal to the combined regressor variance for which it accounts.

There are as many principal components as there are linearly independent regressors: $p \equiv \text{rank}(\mathbf{z}_X)$, where $\mathbf{z}_X \equiv [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k]$. Although the method of principal components is more general, I shall assume throughout most of this discussion that the regressors are not perfectly collinear and, consequently, that $p = k$.

Because the principal components lie in the regressor subspace, each is a linear combination of the regressors. Thus, the first principal component can be written as

$$\underset{(n\times 1)}{\mathbf{w}_1} = A_{11}\mathbf{z}_1 + A_{21}\mathbf{z}_2 + \cdots + A_{k1}\mathbf{z}_k$$
$$= \underset{(n\times k)}{\mathbf{Z}_X}\,\underset{(k\times 1)}{\mathbf{a}_1}$$

The variance of the first component is

$$S^2_{W_1} = \frac{1}{n-1}\mathbf{w}_1'\mathbf{x}_1 = \frac{1}{n-1}\,\mathbf{a}_1'\mathbf{Z}_X'\mathbf{Z}_X\mathbf{a}_1 = \mathbf{a}_1'\mathbf{R}_{XX}\mathbf{a}_1$$

where $\mathbf{R}_{XX} \equiv [1/(n-1)]\mathbf{Z}_X'\mathbf{Z}_X$ is the correlation matrix of the regressors.

We want to maximize $S^2_{W_1}$, but, to make maximization meaningful, it is necessary to constrain the coefficients $\mathbf{a}_1$. In the absence of a constraint, $S^2_{W_1}$ can be made arbitrarily large simply by picking large coefficients. The normalizing constraint

$$\mathbf{a}_1'\mathbf{a}_1 = 1 \qquad [13.1]$$

proves convenient, but any constraint of this general form would do.[11]

[10] It is also possible to find principal components of the *unstandardized* regressors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$, but these are not generally interpretable unless all of the X's are measured on the same scale.

[11] Normalizing the coefficients so that $\mathbf{a}_1'\mathbf{a}_1 = 1$ causes the variance of the first principal component to be equal to the combined variance of the standardized regressors accounted for by this component, as will become clear presently.

We can maximize $S^2_{W_1}$ subject to the restriction of Equation 13.1 by employing a Lagrange multiplier L_1, defining[12]

$$F_1 \equiv \mathbf{a}_1'\mathbf{R}_{XX}\mathbf{a}_1 - L_1(\mathbf{a}_1'\mathbf{a}_1 - 1)$$

Then, differentiating this equation with respect to $\mathbf{a}_1$ and L_1 ,

$$\begin{aligned}\frac{\partial F_1}{\partial \mathbf{a}_1} &= 2\mathbf{R}_{XX}\mathbf{a}_1 - 2L_1\mathbf{a}_1\\ \frac{\partial F_1}{\partial L_1} &= -(\mathbf{a}_1'\mathbf{a}_1 - 1)\end{aligned}$$

Setting the partial derivatives to 0 produces the equations

$$\begin{aligned}(\mathbf{R}_{XX} - L_1\mathbf{I}_k)\mathbf{a}_1 &= \mathbf{0}\\ \mathbf{a}_1'\mathbf{a}_1 &= 1\end{aligned} \qquad [13.2]$$

The first formula in Equation 13.2 has nontrivial solutions for $\mathbf{a}_1$ only when $(\mathbf{R}_{XX}-L_1\mathbf{I}_k)$ is singular—that is, when $|\mathbf{R}_{XX}-L_1\mathbf{I}_k| = 0$. The multiplier L_1, therefore, is an eigenvalue of $\mathbf{R}_{XX}$, and $\mathbf{a}_1$ is the corresponding eigenvector, scaled so that $\mathbf{a}_1'\mathbf{a}_1 = 1$.

There are, however, k solutions to Equation 13.2, corresponding to the k eigenvalue-eigenvector pairs of $\mathbf{R}_{XX}$, so we must decide which solution to choose. From the first formula in Equation 13.2, we have $\mathbf{R}_{XX}\mathbf{a}_1 = L_1\mathbf{a}_1$. Consequently,

$$S^2_{W_1} = \mathbf{a}_1'\mathbf{R}_{XX}\mathbf{a}_1 = L_1\mathbf{a}_1'\mathbf{a}_1 = L_1$$

Because our purpose is to maximize $S^2_{W_1}$ (subject to the constraint on $\mathbf{a}_1$), we must select the *largest* eigenvalue of $\mathbf{R}_{XX}$ to define the first principal component.

The second principal component is derived similarly, under the further restriction that it is orthogonal to the first; the third that it is orthogonal to the first two; and so on.[13] It turns out that the second principal component corresponds to the second-largest eigenvalue of $\mathbf{R}_{XX}$, the third to the third-largest eigenvalue, and so forth. We order the eigenvalues of $\mathbf{R}_{XX}$ so that[14]

$$L_1 \geq L_2 \geq \cdots \geq L_k > 0$$

[12] See Appendix C, Section C.2, for an explanation of the method of Lagrange multipliers for constrained optimization.

[13] See Exercise 13.1.

[14] Recall that we are assuming that $\mathbf{R}_{XX}$ is of full rank, and hence none of its eigenvalues is 0. It is possible, but unlikely, that two or more eigenvalues of $\mathbf{R}_{XX}$ are equal. In this event, the orientation of the principal components corresponding to the equal eigenvalues is not unique, although the subspace spanned by these components—and for which they constitute a basis—is unique.

The matrix of principal-component coefficients

$$\underset{(k\times k)}{\mathbf{A}} \equiv [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k]$$

contains normalized eigenvectors of $\mathbf{R}_{XX}$. This matrix is, therefore, orthonormal: $\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}' = \mathbf{I}_k$.

The principal components

$$\underset{(n\times k)}{\mathbf{W}} = \underset{(n\times k)}{\mathbf{Z}_X}\underset{(k\times k)}{\mathbf{A}} \qquad [13.3]$$

have covariance matrix

$$\frac{1}{n-1}\mathbf{W}'\mathbf{W} = \frac{1}{n-1}\mathbf{A}'\mathbf{Z}_X'\mathbf{Z}_X\mathbf{A}$$
$$= \mathbf{A}'\mathbf{R}_{XX}\mathbf{A} = \mathbf{A}'\mathbf{A}\mathbf{L} = \mathbf{L}$$

where $\mathbf{L} \equiv \text{diag}[L_1, L_2, \ldots, L_k]$ is the matrix of eigenvalues of $\mathbf{R}_{XX}$; the covariance matrix $\mathbf{W}$ of the principal components is, therefore, orthogonal, as required. Furthermore,

$$\text{trace}(\mathbf{L}) = \sum_{j=1}^{k} L_j = k = \text{trace}(\mathbf{R}_{XX})$$

and thus the principal components partition the combined variance of the standardized variables $Z_1, Z_2, \ldots, Z_k$.

Solving Equation 13.3 for $\mathbf{Z}_X$ produces

$$\mathbf{Z}_X = \mathbf{W}\mathbf{A}^{-1} = \mathbf{W}\mathbf{A}'$$

and, consequently,

$$\mathbf{R}_{XX} = \frac{1}{n-1}\mathbf{Z}_X'\mathbf{Z}_X = \frac{1}{n-1}\mathbf{A}\mathbf{W}'\mathbf{W}\mathbf{A}' = \mathbf{A}\mathbf{L}\mathbf{A}'$$

Finally,

$$\mathbf{R}_{XX}^{-1} = (\mathbf{A}')^{-1}\mathbf{L}^{-1}\mathbf{A}^{-1} = \mathbf{A}\mathbf{L}^{-1}\mathbf{A}' \qquad [13.4]$$

We shall use this result presently in our investigation of collinearity.

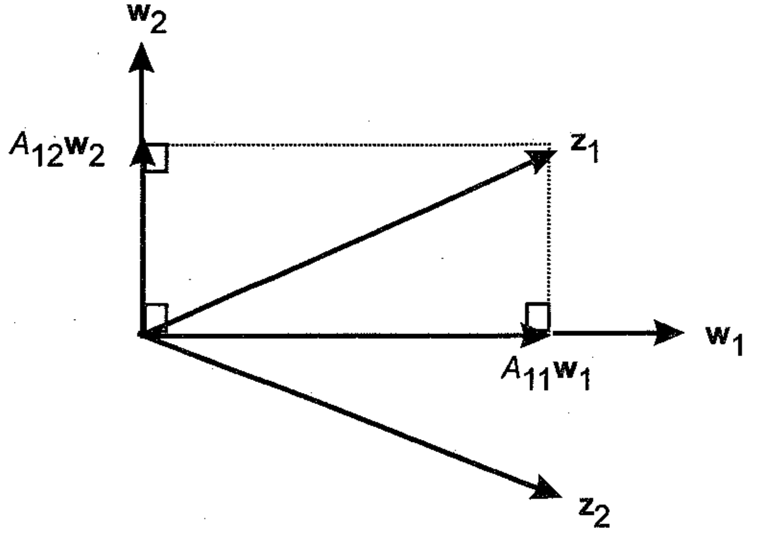

Figure 13.4. Vector geometry of principal components for two, positively correlated, standardized variables $\mathbf{z}_1$ and $\mathbf{z}_2$.

Two Variables

The vector geometry of principal components is illustrated for two variables in Figure 13.4. The symmetry of this figure is peculiar to the two-dimensional case. The length of each principal-component vector is the square root of the sum of squared orthogonal projections of $\mathbf{z}_1$ and $\mathbf{z}_2$ on the component. The direction of $\mathbf{w}_1$ is chosen to maximize the combined length of these projections, and hence to maximize the length of $\mathbf{w}_1$. Because the subspace spanned by $\mathbf{z}_1$ and $\mathbf{z}_2$ is two dimensional, $\mathbf{w}_2$ is simply chosen to be orthogonal to $\mathbf{w}_1$. Note that[15] $\|\mathbf{w}_j\|^2 = L_j(n-1)$.

It is clear from the figure that as the correlation between Z_1 and Z_2 increases, the first principal component grows at the expense of the second; thus, L_1 gets larger and L_2 smaller. If, alternatively, $\mathbf{z}_1$ and $\mathbf{z}_2$ are orthogonal, then $\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = \sqrt{n-1}$, and $L_1 = L_2 = 1$.

The algebra of the two-variable case is also particularly simple. The eigenvalues of $\mathbf{R}_{XX}$ are the solutions of the characteristic equation

$$\begin{vmatrix} 1-L & r_{12} \\ r_{12} & 1-L \end{vmatrix} = 0$$

that is,

$$(1-L)^2 - r_{12}^2 = L^2 - 2L + 1 - r_{12}^2 = 0$$

[15] There is a small subtlety here: The subspace spanned by each component is one dimensional, and the length of each component is fixed by the corresponding eigenvalue, but these factors determine the orientation of the component only up to a rotation of 180°—that is, a change in sign.

Using the quadratic formula to find the roots of the characteristic equation yields

$$L_1 = 1 + \sqrt{r_{12}^2} \qquad [13.5]$$
$$L_2 = 1 - \sqrt{r_{12}^2}$$

And so, consistent with the geometry of Figure 13.4, as the magnitude of the correlation between the two variables increases, the variation attributed to the first principal component also grows. If r_{12} is positive, then solving for $\mathbf{A}$ from the relation $\mathbf{R}_{XX}\mathbf{A} = \mathbf{L}\mathbf{A}$ under the restriction $\mathbf{A}'\mathbf{A} = \mathbf{I}_2$ gives[16]

$$\mathbf{A} = \begin{bmatrix} \dfrac{\sqrt{2}}{2} & \dfrac{\sqrt{2}}{2} \\ \dfrac{\sqrt{2}}{2} & -\dfrac{\sqrt{2}}{2} \end{bmatrix}$$

The generalization to k standardized regressors is straightforward: If the variables are orthogonal, then all $L_j = 1$ and all $\|\mathbf{w}_j\| = \sqrt{n-1}$. As collinearities among the variables increase, some eigenvalues become large while others grow small. Small eigenvalues and the corresponding short principal components represent dimensions along which the regressor subspace has (nearly) collapsed. Perfect collinearities are associated with eigenvalues of 0.

The Data Ellipsoid

The principal components have an interesting interpretation in terms of the standard data ellipsoid for the Z's.[17] The data ellipsoid is given by the equation

$$\mathbf{z}'\mathbf{R}_{XX}^{-1}\mathbf{z} = 1$$

where $\mathbf{z} \equiv (Z_1, \ldots, Z_k)'$ is a vector of values for the k standardized regressors. Because the variables are standardized, the data ellipsoid is centered at the origin, and the shadow of the ellipsoid on each axis is of length 2 (i.e., 2 standard deviations). It can be shown that the principal components correspond to the principal axes of the data ellipsoid, and, further, that the half-length of each axis is equal to the square root of the corresponding eigenvalue[18] L_j of $\mathbf{R}_{XX}$.

These properties are depicted in Figure 13.5 for $k = 2$. When the variables are uncorrelated, the data ellipse becomes circular, and each axis has a half-length of 1.

[16] Exercise 13.2 derives the solution for $r_{12} < 0$.

[17] The standard data ellipsoid was introduced in Section 9.4.

[18] See Exercise 13.3. These relations also hold for *unstandardized* variables. That is, the principal components calculated from the covariance matrix $\mathbf{S}_{XX}$ give the principal axes of the standard data ellipsoid $(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S}_{XX}^{-1}(\mathbf{x}-\bar{\mathbf{x}})$; and the half-length of the jth principal axis of this ellipsoid is equal to the square root of the jth eigenvalue of $\mathbf{S}_{XX}$.

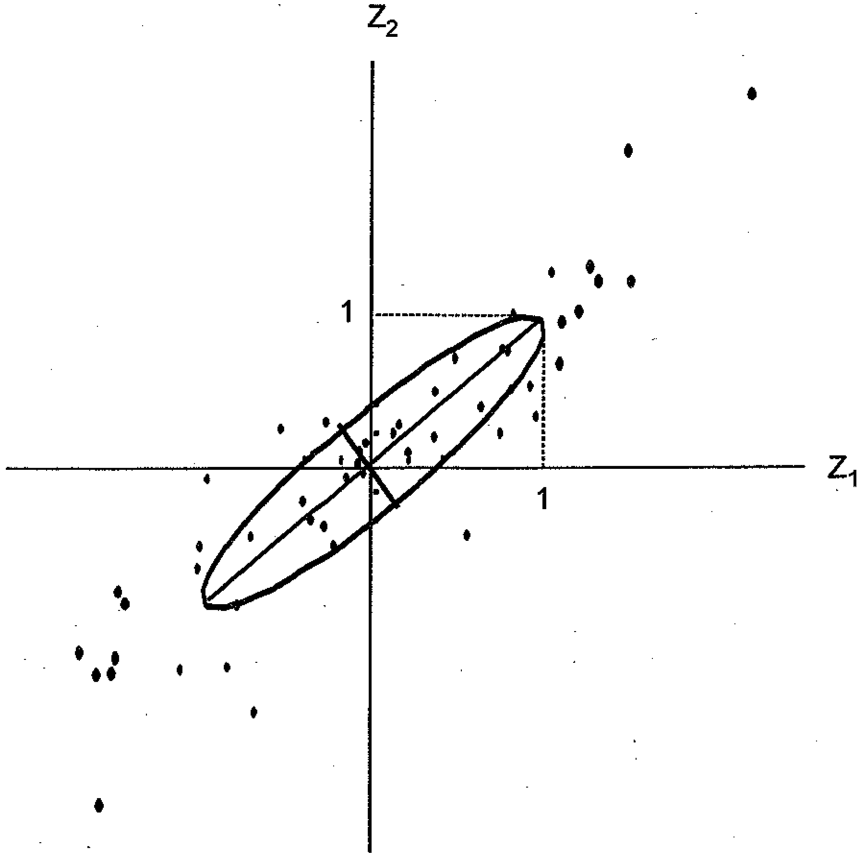

Figure 13.5. The principal components are the principal axes of the standard data ellipse $\mathbf{z}'\mathbf{R}_{XX}^{-1}\mathbf{z} = 1$. The first eigenvalue L_1 of $\mathbf{R}_{XX}$ gives the half-length of the major axis of the ellipse; the second eigenvalue L_2 gives the half-length of the minor axis. In this illustration, the two variables are highly correlated, so L_1 is large and L_2 is small.

Summary

- The principal components of the k standardized regressors $\mathbf{Z}_X$ are a new set of k variables derived from $\mathbf{Z}_X$ by a linear transformation: $\mathbf{W} = \mathbf{Z}_X\mathbf{A}$, where $\mathbf{A}$ is the $(k \times k)$ transformation matrix.
- The transformation $\mathbf{A}$ is selected so that the columns of $\mathbf{W}$ are orthogonal—that is, the principal components are uncorrelated. In addition, $\mathbf{A}$ is constructed so that the first component accounts for maximum variance in the Z's; the second for maximum variance under the constraint that it is orthogonal to the first; and so on. Each principal component is scaled so that its variance is equal to the variance in the Z's for which it accounts. The principal components therefore partition the variance of the Z's.
- The transformation matrix $\mathbf{A}$ contains (by columns) normalized eigenvectors of $\mathbf{R}_{XX}$, the correlation matrix of the regressors. The columns of $\mathbf{A}$ are ordered by their corresponding eigenvalues: The first column corresponds to the largest eigenvalue, and the last column to the smallest. The eigenvalue L_j associated with the jth component represents the variance attributable to that component.
- If there are perfect collinearities in $\mathbf{Z}_X$, then some eigenvalues of $\mathbf{R}_{XX}$ will be 0, and there will be fewer than k principal components, the number of components corresponding to $\text{rank}(\mathbf{Z}_X) = \text{rank}(\mathbf{R}_{XX})$. Near collinearities are associated with small eigenvalues and correspondingly short principal components.

Principal components can be used to explicate the correlational structure of the independent variables in regression. The principal components are a derived set of variables that form an orthogonal basis for the subspace of the standardized *X*'s. The first principal component spans the one-dimensional subspace that accounts for maximum variation in the standardized *X*'s. The second principal component accounts for maximum variation in the standardized *X*'s, under the constraint that it is orthogonal to the first. The other principal components are similarly defined; unless the *X*'s are perfectly collinear, there are as many principal components as there are *X*'s. Each principal component is scaled to have variance equal to the collective variance in the standardized *X*'s for which it accounts. Collinear relations among the independent variables, therefore, correspond to very short principal components, which represent dimensions along which the regressor subspace has nearly collapsed.

Diagnosing Collinearity

I explained earlier that the sampling variance of the regression coefficient B_j is

$$V(B_j) = \frac{\sigma_\varepsilon^2}{(n-1)S_j^2} \times \frac{1}{1-R_j^2}$$

It can be shown that $\text{VIF}_j = 1/(1 - R_j^2)$ is the *j*th diagonal entry of $\mathbf{R}_{XX}^{-1}$ (see Theil, 1971, p. 166). Using Equation 13.4, the variance inflation factors can be expressed as functions of the eigenvalues of $\mathbf{R}_{XX}$ and the principal components; specifically,

$$\text{VIF}_j = \sum_{l=1}^{k} \frac{A_{jl}^2}{L_l}$$

Thus, it is only the small eigenvalues that contribute to large sampling variance, but only for those regressors that have large coefficients associated with the corresponding short principal components. This result is sensible, for small eigenvalues and their short components correspond to collinear relations among the regressors; regressors with large coefficients for these components are the regressors implicated in the collinearities (see below).

The relative size of the eigenvalues serves as an indicator of the degree of collinearity present in the data. The square root of the ratio of the largest to smallest eigenvalue, $K \equiv \sqrt{L_1/L_k}$, called the *condition number*, is a commonly employed standardized index of the global instability of the least-squares regression coefficients: A large condition number (say, 10 or more) indicates that relatively small changes in the data tend to produce large changes in the least-squares solution. In this event, $\mathbf{R}_{XX}$ is said to be *ill conditioned.*

It is instructive to examine the condition number in the simplified context of the two-regressor model. From Equation 13.5,

$$K = \sqrt{\frac{L_1}{L_2}} = \sqrt{\frac{1+\sqrt{r_{12}^2}}{1-\sqrt{r_{12}^2}}}$$

and thus $K = 10$ corresponds to $r_{12}^2 = .9608$, for which VIF $= 26$.

Belsley et al. (1980, Chapter 3) define a *condition index* $K_j \equiv \sqrt{L_1/L_j}$ for each principal component[19] of $\mathbf{R}_{XX}$. Then, the number of large condition indices points to the number of different collinear relations among the regressors.

Chatterjee and Price (1991, Chapter 7) employ the principal-component coefficients to estimate these near collinearities: A component $\mathbf{w}_l$ associated with a very small eigenvalue $L_l \simeq 0$ is itself approximately equal to the zero vector; consequently,

$$A_{1l}\mathbf{z}_1 + A_{2l}\mathbf{z}_2 + \cdots + A_{kl}\mathbf{0}_k \simeq 0$$

and we can use the large A_{jl}'s to specify a linear combination of the Z's that is approximately equal to 0.

13.1.2 Generalized Variance Inflation*

The methods for detecting collinearity described thus far are not fully applicable to models that include related sets of regressors, such as dummy regressors constructed from a polytomous categorical variable or polynomial regressors. The reasoning underlying this qualification is subtle, but can be illuminated by appealing to the vector representation of linear models.

The correlations among a set of dummy regressors are affected by the choice of baseline category. Similarly, the correlations among a set of polynomial regressors in an independent variable X are affected by adding a constant

[19] Primarily for computational accuracy, Belsley et al. (1980, Chapter 3) develop diagnostic methods for collinearity in terms of the *singular-value decomposition* of the regressor matrix, scaled so that each variable has a sum of squares of 1. I employ an equivalent eigenvalue-eigenvector approach because of its conceptual simplicity and broader familiarity. The eigenvectors of $\mathbf{R}_{XX}$, it turns out, are the squares of the singular values of $(1/\sqrt{n-1})\mathbf{Z}_X$. Indeed, the condition number K defined here is actually the condition number of $(1/\sqrt{n-1})\mathbf{Z}_X$ (and hence of $\mathbf{Z}_X$). Information on the singular-value decomposition and its role in linear-model analysis can be found in Belsley et al. (1980, Chapter 3) and in Mandel (1982).

A more substantial difference between my approach and that of Belsley et al. is that they base their analysis not on the correlation matrix of the X's, but rather on $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}}$ is the regressor matrix, including the constant regressor, with columns normed to unit length. Consider an independent variable that is uncorrelated with the others, but which has scores that are far from 0. Belsley et al. would say that this independent variable is "collinear with the constant regressor." This seems to me a corruption of the notion of collinearity, which deals fundamentally with the inability to separate the effects of highly correlated independent variables, and should not change with linear transformations of individual independent variables. See Belsley (1984), and the associated commentary, for various points of view on this issue.

to the X-values. Neither of these changes alters the fit of the model to the data, however, so neither is fundamental. It is, indeed, always possible to select an orthogonal basis for the dummy-regressor or polynomial-regressor subspace (although such a basis does not employ dummy variables or simple powers of X). What is at issue is the subspace itself, and not the arbitrarily chosen basis for it.[20]

We are not concerned, therefore, with the "artificial" collinearity among dummy regressors or polynomial regressors in the same set. We are instead interested in the relationships between the subspaces generated to represent the effects of *different* independent variables. As a consequence, we can legitimately employ variance inflation factors to examine the impact of collinearity on the coefficients of numerical regressors, or on any single-degree-of-freedom effects, even when sets of dummy regressors or polynomial regressors are present in the model.

Fox and Monette (1992) generalize the notion of variance inflation to sets of related regressors. Rewrite the linear model as

$$\mathbf{y} = \alpha\mathbf{1} + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

where the p regressors of interest (e.g., a set of dummy regressors) are in $\mathbf{X}_1$, while the remaining $k - p$ regressors (with the exception of the constant) are in $\mathbf{X}_2$. Fox and Monette (1992) show that the squared ratio of the size of the joint confidence region for $\boldsymbol{\beta}_1$ to the size of the same region for orthogonal but otherwise similar data is

$$\text{GVIF}_1 = \frac{\det \mathbf{R}_{11} \det \mathbf{R}_{22}}{\det \mathbf{R}}$$

Here, $\mathbf{R}_{11}$ is the correlation matrix for $\mathbf{X}_1$; $\mathbf{R}_{22}$ is the correlation matrix for $\mathbf{X}_2$; and $\mathbf{R}$ is the matrix of correlations among all of the variables. The generalized variance inflation factor (GVIF) is independent of the bases selected for the subspaces spanned by the columns of $\mathbf{X}_1$ and $\mathbf{X}_2$. If $\mathbf{X}_1$ contains only one column, then the GVIF reduces to the familiar variance inflation factor.

The notion of variance inflation can be extended to sets of related regressors, such as dummy regressors and polynomial regressors, by considering the size of the joint confidence region for the related coefficients.

[20] A particular basis may be a poor computational choice, however, if it produces numerically unstable results. Consequently, researchers are sometimes advised to pick a category with many cases to serve as the baseline for a set of dummy regressors, or to subtract the mean from X prior to constructing polynomial regressors; the latter procedure is called "centering." Neither of these practices fundamentally alters the model, but may lead to more accurate computations.

EXERCISES

13.1 *The second principal component is

$$\begin{aligned}\underset{(n\times 1)}{\mathbf{w}_2} &= A_{12}\mathbf{z}_1 + A_{22}\mathbf{z}_2 + \cdots + A_{k2}\mathbf{z}_k \\ &= \underset{(n\times k)}{\mathbf{Z}_X}\,\underset{(k\times 1)}{\mathbf{a}_2}\end{aligned}$$

with variance

$$S^2_{W_2} = \mathbf{a}_2'\mathbf{R}_{XX}\mathbf{a}_2$$

We need to maximize this variance subject to the *normalizing constraint* $\mathbf{a}_2'\mathbf{a}_2 = 1$ and the *orthogonality constraint* $\mathbf{w}_1'\mathbf{w}_2 = 0$. Show that the orthogonality constraint is equivalent to $\mathbf{a}_1'\mathbf{a}_2 = 0$. Then, using *two* Lagrange multipliers, one for the normalizing constraint and the other for the orthogonality constraint, show that $\mathbf{a}_2$ is an eigenvector corresponding to the second-largest eigenvalue of $\mathbf{R}_{XX}$. Explain how this procedure can be extended to derive the remaining $k - 2$ principal components.

13.2 *Find the matrix **A** of principal-component coefficients when $k = 2$ and r_{12} is negative.

13.3 *Show that when $k = 2$, the principal components of $\mathbf{R}_{XX}$ correspond to the principal axes of the data ellipse for the standardized regressors Z_1 and Z_2; show that the half-length of each axis is equal to the square root of the corresponding eigenvalue of $\mathbf{R}_{XX}$. Now extend this reasoning to the principal axes of the data ellipsoid for the standardized regressors when $k > 2$.

13.4 *The data that follow were constructed by Mandel (1982) to illustrate the problem of collinearity:

X_1	X_2	Y
16.85	1.46	41.38
24.81	−4.61	31.01
18.85	−0.21	37.41
12.63	4.93	50.05
21.38	−1.36	39.17
18.78	−0.08	38.86
15.58	2.98	46.14
16.30	1.73	44.47

(a) Compute the mean and standard deviation of each variable. Find the correlations among X_1, X_2, and Y, and use these correlations to calculate the standardized coefficients B_1^* and B_2^* for the regression of Y on X_1 and X_2. Find the unstandardized coefficients A, B_1, and B_2.

(b) Perform a principal-components analysis for X_1 and X_2. Draw the geometric vector representation of the principal-components analysis. Find the variance inflation factors for the coefficients B_1 and B_2, and calculate the condition number K for the regression.

(c) Use the second principal component to approximate the near-collinear relation between the standardized regressors Z_1 and Z_2. Express this relation as a linear relationship between the unstandardized regressors X_1 and X_2.

(d) Now regress X_1 on X_2. How does the fitted regression equation compare with the linear relationship found in part (c)?

(e) Draw the data ellipse for X_1 and X_2, and the 95% joint confidence ellipse for B_1 and B_2.

13.5 Time series data on Canadian women's labor-force participation in the first three decades of the postwar period are given in Table 13.3 and `bfox.dat`. B. Fox (1980) was interested in determining how women's labor-force participation rate (L, measured as the percentage of adult women in the work force) responds to a variety of factors indicative of the supply of and demand for women's labor. The independent variables in the analysis include:

- the total fertility rate (F), the expected number of births to a hypothetical cohort of 1000 women proceeding through their child-bearing years at current age-specific fertility rates;
- men's (M) and women's (W) average weekly earnings, in constant 1935 dollars and adjusted for current tax rates;
- per-capita consumer debt (D), in constant dollars; and
- the availability of part-time work (P), measured as the percentage of the active labor force working 34 hours a week or less.

Women's earnings, consumer debt, and the availability of part-time work were expected to affect women's labor-force participation positively. Fertility and men's earnings were expected to have negative effects. Because all of the series, including that for the dependent variable, manifest strong linear trends over the 20-year period of the study, year (T), coded from 1946 to 1975, was also included as an independent variable in the regression.

(a) Regress L on T, F, M, W, D, and P. What do you find: Are the researcher's expectations borne out? Are the estimates sufficiently precise?

(b) Employ the methods of this section to diagnose collinearity in B. Fox's data.

13.2 Coping With Collinearity: No Quick Fix

When X_1 and X_2 are strongly collinear, the data contain little information about the impact of X_1 on Y holding X_2 constant statistically, because there is little variation in X_1 when X_2 is fixed. Of course, the same is true for X_2 fixing X_1. Because B_1 estimates the partial effect of X_1 controlling for X_2, this estimate is imprecise.

Although there are several strategies for dealing with collinear data, none magically extracts nonexistent information from the data. Rather, the research problem is redefined, often subtly and implicitly. Sometimes the redefinition is reasonable; usually it is not. The ideal solution to the problem of collinearity is

TABLE 13.3 B. Fox's Canadian Women's Labor-Force Participation Data. T is year; L is women's labor-force participation rate, in percent; F is the total fertility rate, per 1000; M is men's average weekly wages in 1935 dollars; W is women's average weekly wages; D is per-capita consumer debt; and P is the percentage of part-time workers.

T	L	F	M	W	D	P
1946	25.3	3748	25.35	14.05	18.18	10.28
1947	24.4	3996	26.14	14.61	28.33	9.28
1948	24.2	3725	25.11	14.23	30.55	9.51
1949	24.2	3750	24.45	14.61	35.81	8.87
1950	23.7	3669	26.79	15.26	38.39	8.54
1951	24.2	3682	26.33	14.58	26.52	8.84
1952	24.1	3845	27.89	15.66	45.65	8.60
1953	23.8	3905	29.15	16.30	52.99	5.49
1954	23.6	4047	29.52	16.57	54.84	6.67
1955	24.3	4043	32.05	17.99	65.53	6.25
1956	25.1	4092	32.98	18.33	72.56	6.32
1957	26.2	4168	32.25	17.64	69.49	7.30
1958	26.6	4073	32.52	18.16	71.71	8.65
1959	26.9	4100	33.95	18.58	78.89	8.80
1960	27.9	4119	34.63	18.95	84.99	9.39
1961	29.1	4159	35.14	18.78	87.71	10.23
1962	29.9	4134	34.49	18.74	95.31	10.77
1963	29.8	4017	35.99	19.71	104.40	10.84
1964	30.9	3886	36.68	20.06	116.80	11.70
1965	32.1	3467	37.96	20.94	130.99	12.33
1966	33.2	3150	38.68	21.20	135.25	12.18
1967	34.5	2879	39.65	21.95	142.93	13.67
1968	35.1	2681	41.20	22.68	155.47	13.82
1969	36.1	2563	42.44	23.75	165.04	14.91
1970	36.9	2571	42.02	25.63	164.53	15.52
1971	37.0	2503	45.32	26.79	169.63	15.47
1972	37.9	2302	45.61	27.51	190.62	15.85
1973	40.1	2931	45.59	27.35	209.60	15.40
1974	40.6	1875	48.06	29.64	216.66	16.23
1975	42.2	1866	46.12	29.33	224.34	16.71

to collect new data in such a manner that the problem is avoided—for example, by experimental manipulation of the X's, or by a research setting (or sampling procedure) in which the independent variables of interest are not strongly related. Unfortunately, these solutions are rarely practical.

Several less adequate strategies for coping with collinear data are briefly described in this section. I have devoted most space to variable selection, because selection techniques are commonly abused by social scientists, because the rationale for variable selection is straightforward, and because variable selection is a reasonable approach in certain (limited) circumstances. Variable selection also has applications outside of the context of collinearity.

13.2.1 Model Respecification

Although collinearity is a data problem, not (necessarily) a deficiency of the model, one approach to the problem is to respecify the model. Perhaps, after further thought, several regressors in the model can be conceptualized as alternative indicators of the same underlying construct. Then these measures can be combined in some manner, or one can be chosen to represent the others. In this context, high correlations among the X's in question indicate high reliability—a fact to be celebrated, not lamented. Imagine, for example, an international analysis of factors influencing infant mortality, in which gross national product per capita, energy use per capita, and televisions per capita are among the independent variables and are highly correlated. A researcher may choose to treat these variables as indicators of the general level of economic development.

Alternatively, we can reconsider whether we really need to control for X_2 (for example) in examining the relationship of Y to X_1. Generally, though, respecification of this variety is possible only where the original model was poorly thought out, or where the researcher is willing to abandon some of the goals of the research. For example, suppose that in a time series regression examining determinants of married women's labor-force participation, collinearity makes it impossible to separate the effects of men's and women's wage levels.[21] There may be good theoretical reason to want to know the effect of women's wage level on their labor-force participation, holding men's wage level constant, but the data are simply uninformative about this question. It may still be of interest, however, to determine the partial relationship between general wage level and women's labor-force participation, controlling for other independent variables in the analysis.

13.2.2 Variable Selection

A common, but usually misguided, approach to collinearity is variable selection, where some procedure is employed to reduce the regressors in the model to a less highly correlated set. Forward selection methods add independent variables to the model one at a time. At each step, the variable that yields the largest increment in R^2 is selected. The procedure stops, for example, when the increment is smaller than a preset criterion.[22] Backward elimination methods are similar, except that the procedure starts with the full model and deletes variables one at a time. Forward/backward—or *stepwise*—methods combine the two approaches.

These methods frequently are abused by naive researchers who seek to interpret the order of entry of variables into the regression equation as an index of their "importance." This practice is potentially misleading: For example, suppose that there are two highly correlated independent variables that have nearly identical large correlations with Y; only one of these independent variables will enter the regression equation, because the other can contribute little additional information. A small modification to the data, or a new sample, could easily reverse the result.

A technical objection to stepwise methods is that they can fail to turn up the optimal subset of regressors of a given size (i.e., the subset that maximizes R^2). Advances in computer power and in computing procedures make

[21] See Exercises 13.5 and 13.9.

[22] More commonly, the stopping criterion is calibrated by the incremental F for adding a variable to the model.

it feasible to examine all subsets of regressors even when k is quite large.[23] Aside from optimizing the selection criterion, subset techniques also have the advantage of revealing alternative, nearly equivalent models, and thus avoid the misleading appearance of producing a uniquely "correct" result.[24]

One popular approach to subset selection is based on the total (normed) mean-squared error of estimating $E(Y)$ from $\hat{Y}$—that is, estimating the population regression surface over the observed X's from the fitted regression surface:

$$\begin{aligned}\gamma_p &\equiv \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\mathrm{MSE}(\hat{Y}_i) \qquad [13.6]\\ &= \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{n}\{V(\hat{Y}_i)+[E(\hat{Y}_i)-E(Y_i)]^2\}\end{aligned}$$

where the fitted values $\hat{Y}_i$ are based on a model containing $p \leq k+1$ regressors (counting the constant, which is always included in the model). Using the error in estimating $E(Y)$ as a criterion for model quality is reasonable if the goal is literally to predict Y from the X 's, and if new observations on the X's for which predictions are required will be similar to those included in the data.

The term $[E(\hat{Y}_i)-E(Y_i)]^2$ in Equation 13.6 represents the squared bias of $\hat{Y}_i$ as an estimator of the population regression surface $E(Y_i)$. When collinear regressors are deleted from the model, generally $V(\hat{Y}_i)$ will decrease, but—depending on the configuration of data points and the true β's for the deleted regressors—bias may be introduced into the fitted values. Because the prediction MSE is the sum of variance and squared bias, the essential question is whether the decrease in variance offsets any increase in bias.

Mallows's (1973) C_p-statistic estimates γ_p as

$$\begin{aligned}C_p &= \frac{\sum E_i^2}{\hat{\sigma}_\varepsilon^2}+2p-n\\ &= (k+1-p)(F_p-1)+p\end{aligned}$$

where the residuals are from the subset model in question; the error variance estimate $\hat{\sigma}_\varepsilon^2$ is S_E^2 for the *full* model containing all k independent variables; and F_p is the incremental F-statistic for testing the hypothesis that the regressors omitted from the current subset have population coefficients of 0.[25] If this hypothesis is true, then $E(F_p) \simeq 1$, and thus $E(C_p) \simeq p$. A good model, therefore, has C_p close to or below p. As well, minimizing C_p minimizes the sum of squared residuals, and thus maximizes R^2. For the full model, C_{k+1} necessarily equals $k+1$.

Because a good model has C_p close to p, we can identify good models by plotting C_p against p, labeling each point in the plot with a mnemonic representing the independent variables included in the model, and superimposing the

[23] For k independent variables, the number of subsets, excluding the null subset with no predictors, is 2^k-1. See Exercise 13.6.

[24] There are algorithms available to find the optimal subset of a given size without examining all possible subsets (see, e.g., Furnival and Wilson, 1974). When the data are highly collinear, however, the optimal subset of a given size may be only trivially "better" than many of its competitors.

[25] See Exercise 13.7.

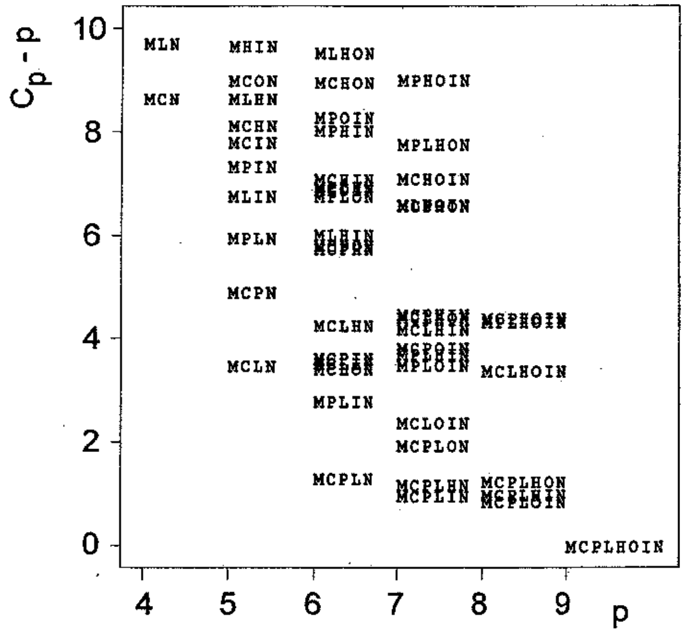

Figure 13.6. Plot of $C_p - p$ against p for the census-undercount regression. Only subsets for which $C_p - p < 10$ are shown. The following capitalized letters are employed to label the predictors in each subset: Minority, Crime, Poverty, Language, High school, hOusing, cIty, and coNventional. Ericksen et al. (1989) selected the predictor subset MCN (i.e., Minority, Crime, and coNventional).

line $C_p = p$ on the plot: Good models are close to or below the reference line. I find that the graph is easier to inspect if it is "detrended" by plotting $C_p - p$ against p (i.e., subtracting the reference line from each point). Now we can look for models with values of $C_p - p$ near or below 0.

An illustrative detrended C_p plot for the census-undercount data is given in Figure 13.6. Only models for which $C_p - p \leq 10$ are shown, including 52 of the $2^8 - 1 = 255$ predictor subsets. Ericksen et al. (1989) employed the subset labeled MCN on the plot (with predictors Minority, Crime, and Conventional).[26] For this subset, $p = 4$ and $C_p = 12.7$, suggesting that there is room for improvement by including more predictors. The regression equation for this subset, and equations for the "best" subsets of four predictors (MCLN, adding Language: $p = 5$ and $C_p = 8.5$) and five predictors (MCPLN, adding Poverty: $p = 6$ and $C_p = 7.3$) appear in Table 13.4. For this dataset, backward and forward/backward stepwise procedures identify the "best" subsets of three, four, and five predictors, but the forward method does not.[27]

In applying variable selection, it is essential to keep the following caveats in mind:

- Most important, variable selection results in a respecified model that usually does not address the research questions that were originally posed. In particular, if the original model is correctly specified, and if the included and omitted variables

[26] Recall, however, that Ericksen et al. (1989) adopted a more complex estimation strategy than ordinary-least-squares regression.

[27] See Exercise 13.8.

TABLE 13.4 "Best" Subset Regression Models for Ericksen et al.'s Census-Undercount Data. Coefficient standard errors are in parentheses.

	Coefficients		
Predictor	$p = 4$	$p = 5$	$p = 6$
Constant	−2.22 (0.56)	−1.98 (0.55)	−0.793 (0.860)
Minority	0.0786 (0.0147)	0.0752 (0.0143)	0.101 (0.020)
Crime	0.0363 (0.0100)	0.0272 (0.0104)	0.0243 (0.0103)
Conventional	0.0280 (0.0081)	0.0273 (0.0078)	0.0293 (0.0077)
Language		0.209 (0.087)	0.184 (0.086)
Poverty			−0.110 (0.062)
R^2	.638	.669	.686
C_p	12.7	8.51	7.32

are correlated, then coefficient estimates following variable selection are biased.[28] Consequently, these methods are most useful for pure prediction problems, in which the values of the regressors for the data to be predicted will be within the configuration of X-values for which selection was employed—as in the census-undercount example. In this case, it is possible to get good estimates of $E(Y)$ even though the regression coefficients themselves are biased. If, however, the X-values for a new observation differ substantially from those used to obtain the estimates, then the predicted Y can be badly biased.

- When regressors occur in sets (e.g., of dummy variables), then these sets should generally be kept together during selection. Likewise, when there are hierarchical relations among regressors, these relations should be respected: For example, an interaction regressor should not appear in a model that does not contain the main effects marginal to the interaction.
- Because variable selection optimizes the fit of the model to the sample data, coefficient standard errors calculated following independent-variable selection—and hence confidence intervals and hypothesis tests—almost surely overstate the precision of results. There is, therefore, a very substantial risk of capitalizing on chance characteristics of the sample.[29]
- Variable selection has applications to statistical modeling even when collinearity is not an issue. It is generally not problematic to eliminate regressors that have small, precisely estimated coefficients, thus producing a more parsimonious model. Indeed, in a very large sample, we may feel justified in deleting regressors with trivially small but "statistically significant" coefficients.

13.2.3 Biased Estimation

Still another general approach to collinear data is biased estimation. The essential idea here is to trade a small amount of bias in the coefficient estimates

[28] See Sections 6.3, 9.6, and 13.2.5.

[29] See Exercise 13.10 and the discussion of cross-validation in Section 16.2.

for a substantial reduction in coefficient sampling variance. The hoped-for result is a smaller mean-squared error of estimation of the β's than is provided by the least-squares estimates. By far the most common biased estimation method is *ridge regression* (due to Hoerl and Kennard, 1970a, 1970b).

Like variable selection, biased estimation is not a magical panacea for collinearity. Ridge regression involves the arbitrary selection of a "ridge constant," which controls the extent to which ridge estimates differ from the least-squares estimates: The larger the ridge constant, the greater the bias and the smaller the variance of the ridge estimator. Unfortunately, but as one might expect, to pick an optimal ridge constant—or even a good one—generally requires knowledge about the unknown β's that we are trying to estimate. My principal reason for mentioning biased estimation here is to caution against its routine use.

*Ridge Regression**

The ridge-regression estimator for the *standardized* regression coefficients is given by

$$\mathbf{b}_d^* \equiv (\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1}\mathbf{r}_{Xy} \qquad [13.7]$$

where $\mathbf{R}_{XX}$ is the correlation matrix for the predictors; $\mathbf{r}_{Xy}$ is the vector of correlations between the predictors and the dependent variable; and $d \geq 0$ is a scalar constant. When $d = 0$, the ridge and least-squares estimators coincide: $\mathbf{b}_0^* = \mathbf{b}^* = \mathbf{R}_{XX}^{-1}\mathbf{r}_{Xy}$. When the data are collinear, some off-diagonal entries of $\mathbf{R}_{XX}$ are generally large, making this matrix ill conditioned. Heuristically, the ridge-regression method improves the conditioning of $\mathbf{R}_{XX}$ by inflating its diagonal entries.

Although the least-squares estimator $\mathbf{b}^*$ is unbiased, its entries tend to be too large in absolute value, a tendency that is magnified as collinearity increases. In practice, researchers working with collinear data often compute wildly large regression coefficients. The ridge estimator may be thought of as a "shrunken" version of the least-squares estimator, correcting the tendency of the latter to produce coefficients that are too far from 0.

The ridge estimator of Equation 13.7 can be rewritten as[30]

$$\mathbf{b}_d^* = \mathbf{U}\mathbf{b}^* \qquad [13.8]$$

where $\mathbf{U} \equiv (\mathbf{I}_k + d\mathbf{R}_{XX}^{-1})^{-1}$. As d increases, the entries of $\mathbf{U}$ tend to grow smaller, and, therefore, $\mathbf{b}_d^*$ is driven toward $\mathbf{0}$. Hoerl and Kennard (1970a) show that for any value of $d > 0$, the squared length of the ridge estimator is less than that of the least-squares estimator: $\mathbf{b}_d^{*\prime}\mathbf{b}_d^* < \mathbf{b}^{*\prime}\mathbf{b}^*$.

The expected value of the ridge estimator can be determined from its relation to the least-squares estimator, given in Equation 13.8; treating the X-values, and hence $\mathbf{R}_{XX}$ and $\mathbf{U}$, as fixed,

$$E(\mathbf{b}_d^*) = \mathbf{U}E(\mathbf{b}^*) = \mathbf{U}\boldsymbol{\beta}^*$$

[30] See Exercise 13.11.

The bias of $\mathbf{b}_d^*$ is, therefore,

$$\text{bias}(\mathbf{b}_d^*) \equiv E(\mathbf{b}_d^*) - \boldsymbol{\beta}^* = (\mathbf{U} - \mathbf{I}_k)\boldsymbol{\beta}^*$$

and because the departure of $\mathbf{U}$ from $\mathbf{I}_k$ increases with d, the bias of the ridge estimator is an increasing function of d.

The variance of the ridge estimator is also simply derived:[31]

$$V(\mathbf{b}_d^*) = \frac{\sigma_\varepsilon^{*2}}{n-1}(\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1}\mathbf{R}_{XX}(\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1} \qquad [13.9]$$

where σ_ε^{*2} is the error variance for the standardized regression. As d increases, the inverted term $(\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1}$ is increasingly dominated by $d\mathbf{I}_k$. The sampling variance of the ridge estimator, therefore, is a decreasing function of d. This result is intuitively reasonable, because the estimator itself is driven toward $\mathbf{0}$.

The mean-squared error of the ridge estimator is the sum of its squared bias and sampling variance. Hoerl and Kennard (1970a) prove that it is always possible to choose a positive value of the ridge constant d so that the mean-squared error of the ridge estimator is less than the mean-squared error of the least-squares estimator. As mentioned, however, the optimal value of d depends on the unknown population regression coefficients.

The central problem in applying ridge regression is to find a value of d for which the trade-off of bias against variance is favorable. In deriving the properties of the ridge estimator, I treated d as fixed. If d is determined from the data, however, it becomes a random variable, casting doubt upon the conceptual basis for the ridge estimator. A number of methods have been proposed for selecting d. Some of these are rough and qualitative, while others incorporate specific formulas or procedures for estimating the optimal value of d. All of these methods, however, have only ad hoc justifications.[32]

There have been many random-sampling simulation experiments exploring the properties of ridge estimation along with other methods meant to cope with collinear data. While these studies are by no means unanimous in their conclusions, the ridge estimator often performs well in comparison with least-squares estimation and in comparison with other biased estimation methods. On the basis of evidence from simulation experiments, it would, however, be misleading to recommend a particular procedure for selecting the ridge constant d, and, indeed, the dependence of the optimal value of d on the unknown regression parameters makes it unlikely that there is a generally best way of finding d. Several authors critical of ridge regression (e.g., Draper and Smith, 1981, p. 324) have noted that simulations supporting the method generally incorporate restrictions on parameter values particularly suited to ridge regression.[33]

[31] See Exercise 13.12.

[32] Exercise 13.13 describes a qualitative method proposed by Hoerl and Kennard in their 1970 papers.

[33] See Section 13.2.5. Simulation studies of ridge regression and other biased estimation methods are too numerous to cite individually here. References to and comments on this literature can be found in many sources, including Draper and Van Nostrand (1979), Vinod (1978), and Hocking (1976). Vinod and Ullah (1981) present an extensive treatment of ridge regression and related methods.

Because the ridge estimator is biased, standard errors based on Equation 13.9 cannot be used in the normal manner for statistical inferences concerning the population regression coefficients. Indeed, as Obenchain (1977) has pointed out, under the assumptions of the linear model, confidence intervals centered at the least-squares estimates paradoxically retain their optimal properties regardless of the degree of collinearity: In particular, they are the *shortest* possible intervals at the stated level of confidence (Scheffé, 1959, Chapter 2). An interval centered at the ridge estimate of a regression coefficient is, therefore, *wider* than the corresponding least-squares interval, even if the ridge estimator has smaller mean-squared error than the least-squares estimator.

13.2.4 Prior Information about the Regression Coefficients

A final approach to estimation with collinear data is to introduce additional prior information (i.e., relevant information external to the data at hand) that reduces the ambiguity produced by collinearity. There are several different ways that prior information can be brought to bear on a regression, including Bayesian analysis, but I shall present a particularly simple case to illustrate the general point. More complex methods are beyond the scope of this discussion and are, in any event, difficult to apply in practice.[34]

Suppose that we wish to estimate the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where Y is savings, X_1 is income from wages and salaries, X_2 is dividend income from stocks, and X_3 is interest income. Imagine that we have trouble estimating β_2 and β_3 because X_2 and X_3 are highly correlated in our data. Suppose further that we have reason to believe that $\beta_2 = \beta_3$, and denote the common quantity β_*. If X_2 and X_3 were not so highly correlated, then we could reasonably test this belief as a hypothesis. In the current situation, we can fit the model

$$Y = \alpha + \beta_1 X_1 + \beta_*(X_2 + X_3) + \varepsilon$$

incorporating our belief in the equality of β_2 and β_3 in the specification of the model, and thus eliminating the collinearity problem (along with the possibility of testing the belief).[35]

13.2.5 Some Comparisons

Although I have presented them separately, the several approaches to collinear data have much in common:

- Model respecification can involve variable selection, and variable selection, in effect, respecifies the model.

[34] See, for example, Belsley et al. (1980, pp. 193–204) and Theil (1971, pp. 346–352).

[35] To test H_0: $\beta_2 = \beta_3$ simply entails contrasting the two models (see Exercise 6.10). In the present context, however, where X_2 and X_3 are very highly correlated, this test has virtually no power: If the second model is wrong, then we cannot, as a practical matter, detect it. We need either to accept the second model on theoretical grounds or to admit that we cannot estimate β_2 and β_3.

- Variable selection implicitly constrains the coefficients of deleted regressors to 0.
- Variable selection produces biased coefficient estimates if the deleted variables have nonzero β's and are correlated with the included variables (as they will be for collinear data).[36] As in ridge regression and similar biased estimation methods, we might hope that the trade-off of bias against variance is favorable, and that, therefore, the mean-squared error of the regression estimates is smaller following variable selection than before. Because the bias, and hence the mean-squared error, depend on the unknown regression coefficients, however, we have no assurance that this will be the case. Even if the coefficients obtained following selection have smaller mean-squared error, their superiority can easily be due to the very large variance of the least-squares estimates when collinearity is high than to acceptably small bias.
- Certain types of prior information (as in the hypothetical example presented in the previous section) result in a respecified model.
- It can be demonstrated that biased-estimation methods like ridge regression place prior constraints on the values of the β's. Ridge regression imposes the restriction $\sum_{j=1}^{k} B_j^{*2} \leq c$, where c is a decreasing function of the ridge constant d; the ridge estimator finds least-squares coefficients subject to this constraint (Draper and Smith, 1981, pp. 320–321). In effect, large absolute standardized coefficients are ruled out a priori, but the specific constraint is imposed implicitly.

The primary lesson to be drawn from these remarks is that mechanical model selection and modification procedures disguise the substantive implications of modeling decisions. Consequently, these methods generally cannot compensate for weaknesses in the data and are no substitute for judgment and thought.

> Several methods have been proposed for dealing with collinear data. Although these methods are sometimes useful, none can be recommended generally: When the X's are highly collinear, the data contain little information about the partial relationship between each X and Y, controlling for the other X's. To resolve the intrinsic ambiguity of collinear data it is necessary either to introduce information external to the data or to redefine the research question asked of the data. Neither of these general approaches should be undertaken mechanically. Methods that are commonly (and, more often than not, unjustifiably) employed with collinear data include: model respecification; variable selection (stepwise and subset methods); biased estimation (e.g., ridge regression); and the introduction of additional prior information. Comparison of the several methods shows that they have more in common than it appears at first sight.

[36] Bias due to the omission of independent variables is discussed in a general context in Sections 6.3 and 9.6.

EXERCISES

13.6 Why are there $2^k - 1$ distinct subsets of k predictors? Evaluate this quantity for $k = 2, 3, \ldots, 15$.

13.7 *Prove that Mallows's C_p-statistic for a subset of p predictors

$$C_p = \frac{\sum E_i^2}{\hat{\sigma}_\varepsilon^2} + 2p - n$$

can also be written as

$$C_p = (k + 1 - p)(F_p - 1) + p$$

Recall that $\hat{\sigma}_\varepsilon^2$ is the estimated error variance based on the model including all k predictors, and that F_p is the incremental F-statistic for testing the hypothesis that the $k - p$ omitted predictors all have zero coefficients. Why is C_p a reasonable estimator of the total normed mean-squared error of prediction?

$$\gamma_p = \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{n} \mathrm{MSE}(\hat{Y}_i)$$

[*Hint*: See Weisberg (1985, Appendix 8A.1).]

13.8 Apply the backward, forward, and forward/backward stepwise-regression methods to Ericksen et al.'s census-undercount data (in `ericksen.dat`). Compare the results of these procedures with those shown in Figure 13.6, based on the application of Mallows's C_p-statistic to all subsets of predictors.

13.9 Apply the variable-selection methods of this section to B. Fox's women's labor-force participation regression (described in Exercise 13.5).

13.10 Cross-validation and variable selection (cf. Hurvich and Tsai, 1990): Perform the following computer-simulation experiment: Independently sample 51 variables, with 200 observations each, from the unit-normal distribution. Call the first variable Y, and the remaining ones $X_1, X_2, \ldots, X_{50}$.

(a) Using all 200 observations, regress Y on $X_1, X_2, \ldots, X_{50}$. Calculate the omnibus F-statistic for the regression, along with the individual t-statistic for each regressor. Construct a quantile comparison plot for the 50 t-statistics, comparing the distribution of the t-values with the normal distribution (or with t for $n - k - 1 = 149$ degrees of freedom). Is the omnibus F statistically significant (say at the 5% level)? How many of the individual B's are statistically significant?

(b) Randomly divide the data in half—placing $n/2 = 100$ observations in each subsample. Employing one or another method of variable selection, and using the data from the first half-sample, find the "best" regression equation that includes $p = 5$ of the $k = 50$ independent variables. Calculate the omnibus F-test and the five individual t-tests for this regression equation. What do you find?

(c) Now recalculate the F- and t-tests in part (b) using the *same* five independent variables but employing the second half-sample. How do these tests compare with those in part (b)?

13.11 *Show that the ridge-regression estimator of the standardized regression coefficients,

$$\mathbf{b}_d^* = (\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1}\mathbf{r}_{Xy}$$

can be written as a linear transformation $\mathbf{b}_d^* = \mathbf{U}\mathbf{b}^*$ of the usual least-squares estimator $\mathbf{b}^* = \mathbf{R}_{XX}^{-1}\mathbf{r}_{Xy}$, where the transformation matrix is $\mathbf{U} \equiv (\mathbf{I}_k + d\mathbf{R}_{XX}^{-1})^{-1}$.

13.12 *Show that the variance of the ridge estimator is

$$V(\mathbf{b}_d^*) = \frac{\sigma_\varepsilon^{*2}}{n-1}(\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1}\mathbf{R}_{XX}(\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1}$$

[*Hint*: Express the ridge estimator as a linear transformation of the standardized dependent-variable values, $\mathbf{b}_d^* = (\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1}[1/(n-1)]\mathbf{Z}_X'\mathbf{z}_y$.]

13.13 *Finding the ridge constant d: Hoerl and Kennard suggest plotting the entries in $\mathbf{b}_d^*$ against values of d ranging between 0 and 1. The resulting graph, called a *ridge trace*, both furnishes a visual representation of the instability due to collinearity and (ostensibly) provides a basis for selecting a value of d. When the data are collinear, we generally observe dramatic changes in regression coefficients as d is gradually increased from 0. As d is increased further, the coefficients eventually stabilize, and then are driven slowly toward $\mathbf{0}$. The estimated error variance, S_E^{*2}, which is minimized at the least-squares solution ($d = 0$), rises slowly with increasing d. Hoerl and Kennard recommend choosing d so that the regression coefficients are stabilized and the error variance is not unreasonably inflated from its minimum value. (A number of other methods have been suggested for selecting d, but none avoids the fundamental difficulty of ridge regression—that good values of d depend on the unknown β's.)

(a) Construct a ridge trace, including the standard error S_E^*, for Ericksen et al.'s census-undercount regression (in `ericksen.dat`). Use this information to select a value of the ridge constant d, and compare the resulting ridge estimates of the regression parameters with the least-squares estimates. Make this comparison for both standardized and unstandardized coefficients.

(b) Repeat part (a) for B. Fox's women's labor-force participation data (in Table 13.3 and `bfox.dat`; see Exercises 13.5 and 13.9). In applying ridge regression to these data, B. Fox selected $d = 0.05$.

13.3 Summary

- When the regressors in a linear model are perfectly collinear, the least-squares coefficients are not unique. Strong, but less-than-perfect, collinearity substantially increases the sampling variances of the least-squares coefficients, and can render them useless as estimators.
- The sampling variance of the least-squares slope coefficient B_j is

$$V(B_j) = \frac{1}{1-R_j^2} \times \frac{\sigma_\varepsilon^2}{(n-1)S_j^2}$$

where R_j^2 is the squared multiple correlation for the regression of X_j on the other X's, and $S_j^2 = \sum(X_{ij} - \overline{X}_j)^2/(n-1)$ is the variance of X_j. The variance inflation factor $\text{VIF}_j = 1/(1-R_j^2)$ indicates the deleterious impact of collinearity on the precision of the estimate B_j. The notion of variance inflation can be extended to sets of related regressors, such as dummy regressors and polynomial regressors, by considering the size of the joint confidence region for the related coefficients.
- Principal components can be used to explicate the correlational structure of the independent variables in regression. The principal components are a derived set of variables that form an orthogonal basis for the subspace of the standardized X's. The first principal component spans the one-dimensional subspace that accounts for maximum variation in the standardized X's. The second principal component accounts for maximum variation in the standardized X's, under the constraint that it is orthogonal to the first. The other principal components are similarly defined; unless the X's are perfectly collinear, there are as many principal components as are there are X's. Each principal component is scaled to have variance equal to the collective variance in the standardized X's for which it accounts. Collinear relations among the independent variables, therefore, correspond to very short principal components, which represent dimensions along which the regressor subspace has nearly collapsed.
- Several methods have been proposed for dealing with collinear data. Although these methods are sometimes useful, none can be recommended generally: When the X's are highly collinear, the data contain little information about the partial relationship between each X and Y, controlling for the other X's. To resolve the intrinsic ambiguity of collinear data, it is necessary either to introduce information external to the data or to redefine the research question asked of the data. Neither of these general approaches should be undertaken mechanically. Methods that are commonly (and, more often than not, unjustifiably) employed with collinear data include: model respecification; variable selection (stepwise and subset methods); biased estimation (e.g., ridge regression); and the introduction of additional prior information. Comparison of the several methods shows that they have more in common than it appears at first sight.