Multiple Regression

In: Applied Regression

By: Michael S. Lewis-Beck

Pub. Date: 2011
Access Date: September 12, 2019
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780803914940
Online ISBN: 9781412983440
DOI: https://dx.doi.org/10.4135/9781412983440
Print pages: 48-74

Multiple Regression

With multiple regression, we can incorporate more than one independent variable into an equation. This is useful in two ways. First, it almost inevitably offers a fuller explanation of the dependent variable, since few phenomena are products of a single cause. Second, the effect of a particular independent variable is made more certain, for the possibility of distorting influences from the other independent variables is removed. The procedure is a straightforward extension of bivariate regression. Parameter estimation and interpretation follow the same principles. Likewise, the significance test and the R^2 are parallel. Further, the bivariate regression assumptions necessary for BLUE are carried over to the multivariate case. The technique of multiple regression has great range, and its mastery will enable the researcher to analyze virtually any set of quantitative data.

The General Equation

In the general multiple regression equation, the dependent variable is seen as a linear function of more than one independent variable,

$$Y = a_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k + e,$$

where the subscript identifies the independent variables. The elementary three-variable case, which we shall be using below, is written,

$$Y = a_0 + b_1X_1 + b_2X_2 + e,$$

and suggests that Y is determined by X_1 and X_2, plus an error term.

To estimate the parameters we again employ the least squares principle, minimizing the sum of the squares of the prediction errors (SSE):

$$SSE = \Sigma(Y - \hat{Y})^2.$$

For the three-variable model, this least squares equation is,

$$\hat{Y} = a_0 + b_1X_1 + b_2X_2.$$

The least squares combination of values for the coefficients (a_0, b_1, b_2) yields less prediction error than other possible combinations of values. Hence, the least squares equation fits the set of observations better than any other linear equation. However, it can no longer be represented graphically with a simple straight line fitted to a two-dimensional scatter-plot. Rather, we must imagine fitting a *plane* to a three-dimensional scatter of points. The location of this plane, of course, is dictated by the values of a_0, b_1, and b_2, which are given by the calculus. For most of us, it is impossible to visualize the fitting of equations with more than three variables. Indeed, for the general case, with k independent variables, it requires conceiving of adjusting a k-dimensional hyperplane to a (k + 1)-dimensional scatter.

For purposes of illustration, let us look at a simple three-variable model from our Riverside study. On the

basis of our earlier work, we believe income is related to education. But we know that education is not the only factor influencing income. Another factor is undoubtedly seniority. In most occupations, the longer one is on the job, the more money one makes. This seems likely to be so in Riverside city government. Therefore, our explanation for income differences should be improved if we revise our bivariate regression model to this multiple regression model:

$$Y = a_0 + b_1X_1 + b_2X_2 + e.$$

where Y = income (in dollars), X_1 = education (in years), X_2 = seniority (in years), e = error. The least squares estimates for the parameters are as follows:

$$\hat{Y} = 5666 + 432X_1 + 281X_2.$$

Interpreting the Parameter Estimates

The interpretation of the intercept, which merely extends the bivariate case, need not detain us: a_0 = the average value of Y when each independent variable equals zero. The interpretation of the slope, however, requires more attention: b_k = the average change in Y associated with a unit change in X_k, *when the other independent variables are held constant.* By this means of control, we are able to separate out the effect of X_k itself, free of any distorting influences from the other independent variables. Such a slope is called a *partial slope, or partial regression coefficient.* In the above Riverside example, partial slope b_2 estimates that a one-year increase in seniority is associated with an average income rise of $281, even assuming the employee's amount of education remains constant. In other words, a city worker can expect this annual salary increment, independent of any personal effort at educational improvement. Nevertheless, according to b_1, acquiring an additional year of schooling would add to an employee's income, regardless of the years of seniority accumulated. That is, an extra year of education will augment income an average of $432, beyond the benefits that come from seniority.

To appreciate fully the interpretation of the partial slope, one must grasp how multiple regression “holds constant” the other independent variables. First, it involves *statistical control*, rather than *experimental control.* For instance, in our Riverside study, if we were able to exercise experimental control, we might hold everyone's education at a constant value, say 10 years, and then record the effect on income of assigning respondents different amounts of seniority. To assess the effect of education on income, a similar experiment could be carried out. If such manipulation were possible, we could analyze the effects of seniority and education, respectively, by running two separate bivariate regressions, one on each experiment. However, since such experimental control is out of the question, we have to rely on the statistical control multiple regression provides. We can show how this statistical control operates to separate the effect of one independent variable from the others by examining the formula for a partial slope.

We confine ourselves to the following three-variable model, the results of which are generalizable:

$$Y = a_0 + b_1X_1 + b_2X_2 + e.$$

Let us explicate the b, estimation. Assuming $r_{12} \neq 0$, each independent variable can be accounted for, at least

in part, by the other independent variables. That is, for example, X_1 can be written as a linear function of X_2,

$$X_1 = c_1 + c_2X_2 + u.$$

Supposing X_1 is not perfectly predicted by X_2, there is error, u. Hence, the observed X_1 can be expressed as the predicted X_1, plus error:

$$X_1 = \hat{X}_1 + u,$$

where $\circ_1 = c_1 + c_2X_2$. The error, u, is the portion of X_1 which the other independent variable, X_2, cannot explain,

$$u = X_1 - \hat{X}_1.$$

This component, u, thus represents a part of X_1 which is completely separate from X_2.

By the same steps, we can also isolate the portion of Y which is linearly independent of X_2:

$$Y = d_1 + d_2X_2 + v$$

$$= (d_1 + d_2X_2) + v$$

$$Y = \hat{Y} + v.$$

The error, v, is that portion of Y which cannot be accounted for by X_2,

$$v = Y - \hat{Y}.$$

This component, v, then, is that part of Y which is unrelated to X_2.

These two error components, u and v, are joined in the following formula for b_1:

$$b_1 = \frac{\Sigma(u)(v)}{\Sigma u^2} = \frac{\Sigma(X_1 - \hat{X}_1)(Y - \hat{Y})}{\Sigma(X_1 - \hat{X}_1)^2}.$$

In words, *b_1 is determined by X_1 and Y values that have been freed of any linear influence from X_2.* In this way, the effect of X_1 is separated from the effect of X_2. The formula, generally applicable for any partial slope, should be familiar, for we saw a special version of it in the bivariate case, where

$$b = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2}.$$

While the statistical control of multiple regression is weaker than experimental control, it still has great value. The careful introduction of additional variables into an equation permits greater confidence in our findings. For instance, the bivariate regression model of the Riverside study suggested that education is a determinant of income. However, this conclusion is open to challenge. That apparent bivariate relationship could be spurious, a product of the common influence of another variable on education and income. For example, an antagonist might argue that the observed bivariate relationship is actually caused by senority, for those with more years on the job are those with more education, as well as higher pay. An implication is that if seniority were “held constant,” education would be exposed as having no effect on income. Multiple regression permits us to test this hypothesis of spuriousness. From the above least squares estimates, we discovered that education still has an apparent effect, even after taking the influence of seniority into account. Hence, through actually

bringing this third variable into the equation, we are able to rule out an hypothesis of spuriousness, and thereby strengthen our belief that education affects income.

Confidence Intervals and Significance Tests

The procedure for confidence intervals and significance tests carries over from the bivariate case. Suppose we wish to know whether the partial slope estimate, b_1, from our three-variable equation for the Riverside study, is significantly different from zero. Again, we confront the null hypothesis, which says there is no relationship in the population, and the alternative hypothesis, which says there is a relationship in the population. Let us construct a two-tailed, 95% confidence interval around this partial slope estimate, in order to test these hypotheses:

$$(b_1 \pm t_{n-3;.975}\, s_b).$$

Note that the only difference between this formula and the bivariate formula is the number of degrees of freedom. Here, we have one less degree of freedom, (n-3) instead of (n-2), because we have one more independent variable. In general, the degrees of freedom of the t variable equal (n-k-1), where n = sample size and k = number of independent variables. Applying the formula,

$$(432 \pm t_{29;.975}\, s_b) = 432 \pm 2.045\ (144) = (432 \pm 294).$$

The probability is .95 that the value of the partial slope in the population is between $138 and $726. Because the value of zero is not captured within this band, we reject the null hypothesis. We state that the partial slope estimate, bi, is significantly different from zero, at the .05 level.

A second approach to the significance testing of b_1 would be examination of the t ratio,

$$b_1/s_{b_1} = 432/144 = 3.01.$$

We observe that the value of this t ratio exceeds the t distribution value, $t_{n-3;975}$. That is,

$$3.01 > 2.045.$$

Therefore, we conclude that b_1 is statistically significant at the .05 level.

The most efficient means of significance testing is to use the rule of thumb, which claims statistical significance at the .05 level, two-tailed, for any coefficient whose t ratio exceeds 2 in absolute value. Below is the three-variable Riverside equation, with the t ratios in parentheses:

$$\hat{Y} = \underset{}{5666} + \underset{(3.01)}{432X_1} + \underset{(3.04)}{281X_2}.$$
$$\quad\ (4.22)$$

An examination of these t ratios, with this rule of thumb in mind, instantly reveals that all the parameter estimates of the model (a_0, b_1, b_2) are significant at the .05 level.

The R^2

To assess the goodness of fit of a multiple regression equation, we employ the R^2, now referred to as the *coefficient of multiple determination.* Once again,

$$R^2 = \frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2} = \frac{\text{regression (explained) sum of squares}}{\text{total sum of squares}}.$$

The R^2 for a multiple regression equation indicates the proportion of variation in Y "explained" by all the independent variables. In the above three-variable Riverside model, $R^2 = .67$, indicating that education and seniority together account for 67% of the variance in income. This multiple regression model clearly provides a more powerful explanation for income differences than the bivariate regression model, where $R^2 = .56$.

Obviously, it is desirable to have a high R^2, for it implies a more complete explanation of the phenomenon under study. Nevertheless, if a higher R^2 were the only goal, then one could simply add independent variables to the equation. That is, an additional independent variable cannot lower the R^2, and is virtually certain to increase it at least somewhat. In fact, if independent variables are added until their number equals n-1, then $R^2 = 1.0$. This "perfect" explanation is of course nonsense, and amounts to no more than a mathematical necessity, which occurs because the degrees of freedom have been exhausted. In sum, rather than entering variables primarily to enhance R^2, the analyst must be guided by theoretical considerations in deciding which variables to include.

Predicting Y

A multiple regression equation is used for prediction as well as explanation. Let us predict the income of a Riverside city employee who has 10 years of education and has been on the job 5 years:

$$\hat{Y} = 5666 + 432X_1 + 281X_2$$
$$= 5666 + 432(10) + 281(5)$$
$$= 5666 + 4320 + 1405$$
$$\hat{Y} = 11{,}391.$$

In order to get some notion of the accuracy of this prediction, we can construct a confidence interval around it, utilizing the standard error of estimate of Y, S_e:

$$(\hat{Y} \pm 2s_e) = \hat{Y} \pm 2(2529) = 11{,}391 \pm 5058.$$

This confidence interval indicates there is a 95% chance that a municipal employee with 10 years of education and 5 years of seniority will earn between $6333 and $16449. While this prediction is more accurate than that generated by the bivariate regression equation, it is still far from precise.

The model is even less useful for forecasting beyond its range of experience. Certainly, we could plug in any values for X_1 and X_2 and produce a prediction for Y. However, the worth of the forecast diminishes as these

X_1 and X_2 values depart from the actual range of variable values in the data. For instance, it would be risky to predict the income of a city worker with two years of education and 35 years of seniority, for no one in the data-set registered such extreme scores. Possibly, at such extreme values, the linearity of the relationships would no longer exist. Then, any prediction based on our linear model would be quite wide of the mark.

The Possibility of Interaction Effects

Thus far, we have assumed that effects are *additive.* That is, Y is determined, in part, by X_1*plus* X_2, not X_1*times* X_2. This additivity assumption dominates applied regression analysis and is frequently justified. However, it is not a necessary assumption. Let us explore an example.

We have mentioned the variable of sex of respondent as a candidate for inclusion in the Riverside income equation. The question is, should the sex variable enter additively or as an interaction. It might be argued that sex is involved interactively with education. In general, an *interaction effect* exists when the impact of one independent variable depends on the value of another independent variable. Specifically, perhaps the effect of education is dependent on the sex of the employee, with education yielding a greater financial return for men.

Formally, this particular interaction model is as follows (we ignore the seniority variable for the moment):

$$\hat{Y} = a_0 + b_1X_1 + b_2(X_1X_2) + e,$$

where Y = income (in dollars); X_1 = education (in years); X_2 = sex of respondent (0 = female, 1 = male); X_1X_2 = an interaction variable created by multiplying X_1 times X_2. The least squares estimates for this model are,

$$\hat{Y} = \underset{}{5837} + \underset{(4.20)}{556X_1} + \underset{(4.44)}{202\ (X_1X_2)} \qquad R^2 = .65,$$
$$\qquad\qquad\quad (2.70)$$

where the figures in parentheses are t ratios. These results indicate that education, while increasing the income of both sexes, provides a greater income increase for men. This becomes clearer when we separate out the prediction equations for men and women.

Prediction equation for women:

$$\hat{Y} = a_0 + b_1X_1 + b_2X_1(0)$$
$$= a_0 + b_1X_1$$
$$\hat{Y} = 5837 + 556X_1.$$

Prediction equation for men:

$$\hat{Y} = a_0 + b_1X_1 + b_2X_1(1)$$
$$= a_0 + (b_1 + b_2)X_1$$
$$\hat{Y} = 5837 + 758X_1.$$

We observe that, for men, the slope for the education variable is greater. Further, this slope difference is statistically significant (see the t ratio for b_2).

The rival, strictly additive, model, is,

$$Y = a_0 + b_1X_1 + b_2X_2 + e,$$

where the variables are defined as before. Estimating this model yields,

$$\hat{Y} = \underset{(3.64)}{4995} + \underset{(5.54)}{633X_1} + \underset{(2.60)}{2555X_2} \qquad R^2 = .65,$$

where the values in parentheses are t ratios. These estimates suggest that education and sex have significant, independent effects on income.

The data are congruent with both the interaction model and the additive model. The coefficients are all statistically significant, and the R^2 is the same in both. Which model is correct? The answer must base itself on theoretical considerations and prior research, since the empirical evidence does not permit us to decide between them. The additive model seems more in keeping with a “discrimination” theory of income determination; that is, other things being equal, society pays women less solely because they are women. The interaction model appears to square better with an “individual failure” theory of income determination; that is, women are paid less because they are less able to make education work to their advantage. On the basis of prior theorizing and research, I favor the “discrimination” interpretation and therefore choose to allow the sex variable to enter the larger income equation additively. (A resolution of the two models might come from estimation of an equation which allows sex to have additive and interactive effects:

$$Y = a_0 + b_1X_1 + b_2X_2 + b_3(X_1X_2) + e.$$

Unfortunately, the estimates from this model are made unreliable by severe multicollinearity, a problem not uncommon with interaction models. We consider multicollinearity at length below.)

A Four-Variable Model: Overcoming Specification Error

Incorporating the sex variable additively into our model for income differences in Riverside leads to the following equation:

$$Y = a_0 + b_1X_1 + b_2X_2 + b_3X_3 + e,$$

where Y = income (in dollars), X_1 = education (in years), X_2 = seniority (in years), X_3 = sex (0 = female, 1 = male), e = error. Theoretically, this four-variable model is much more complete than the earlier two-variable model. It asserts that income is a linear additive function of three factors: education, seniority, and sex.

Estimating this multiple regression model with least squares yields,

$$\hat{Y} = \underset{(4.44)}{5526} + \underset{(2.86)}{385X_1} + \underset{(2.84)}{247X_2} + \underset{(2.40)}{2140X_3}$$

$$R^2 = .73 \qquad n = 32 \qquad s_e = 2344,$$

where the values in parentheses are the t ratios, R^2 = coefficient of multiple determination, n = sample size, s_e = standard error of estimate of Y.

These estimates tell us a good deal about what affects income in Riverside city government. The pay of a municipal employee is significantly influenced by years of education, amount of seniority, and sex. (Each t ratio exceeds 2, indicating statistical significance at the .05 level.) These three factors largely determine income differences within this population. In fact, almost three-quarters of the variation in income is explained by these variables (R^2 = .73). The differences caused are not inconsequential. For each year of education, $385 is added to income, on the average. An extra year of seniority contributes another $247. Male workers can expect $2140 more than females workers, even if the women have the same education and seniority. The cumulative impact of these variables can create sizable income disparities. For example, a male with a college education and 10 years seniority would expect to make $16,296; in contrast, a female with a high school degree and just starting work could only expect to earn $10,146.

Inclusion of relevant variables, that is, seniority and sex, beyond the education variable, has markedly diminished specification error, helping ensure that our estimates are BLU. (To refresh yourself on the meaning of specification error, review the discussion of assumptions in Chapter 2.) In particular, the estimate of the education coefficient, which equaled 732 in the bivariate model, has been sharply reduced. The comparable estimate in this four-variable model, b_1 = 385, indicates that the true impact of education is something like one-half that estimated in the original bivariate equation.

For certain models, it is fairly easy to detect the direction of bias resulting from the exclusion of a relevant variable. Suppose the real world is congruent with this model;

$$\mathbf{Y = a_0 + b_1X_1 + b_2X_2 + e \quad (correct\ model),}$$

but we mistakenly estimate,

$$\mathbf{Y = a_0 + b_1X_1 + e^* \quad (incorrect\ model),}$$

where $e^* = (b_2X_2 + e)$. By excluding X_2 from our estimation, we have committed specification error. Assuming that X_1 and X_2 are correlated, as they almost always are, the slope estimate, b_1, will be biased. This bias is inevitable, for the independent variable, X_1, and the error term, e^*, are correlated, thus violating an assumption necessary for regression to yield desirable estimators. (We see that $r_{x1e^*} \neq 0$, because $r_{x1d2} \neq 0$, and X_2 is a component of e^*.) The direction of the bias of b_1 in the estimated model is determined by: (1) the sign of b_1 and (2) the sign of the correlation, r_{12}. If b_2 and r_{12} have the same sign, then the bias of b_1 is positive; if not, then the bias is negative.

It happens that the direction of bias in the somewhat more complicated Riverside case accords with these rules. As noted, the bias of b_1 in the bivariate equation of the Riverside study is positive, accepting the specification and estimation of the four-variable model. The presence of this positive bias follows the above guidelines: (1) the sign of b_2 (and b_3) is positive and (2) the sign of r_{12} (and r_{13}) is positive; therefore, the bivariate estimate of b_1 must be biased upward. Part of the variance in Y that X_1 is accounting for should be explained by X_2 and X_3, but these variables are not in the equation. Thus, some of the impact of X_2 and X_3 on Y is erroneously assigned to X_1.

The formulation of rules for the detection of bias implies that it is possible to predict the consequences of a given specification error. For instance, the analyst is able to foresee the direction of bias coming from the exclusion of a certain variable. With simpler models, such as those treated here, such insight might be attainable. However, for models which include several variables, and face several candidates for inclusion, the direction of bias is not readily foreseeable. In this more complex situation, the analyst is better served by immediate attention to proper specification of the model.

The Multicollinearity Problem

For multiple regression to produce the "best linear unbiased estimates," it must meet the bivariate regression assumptions, plus one additional assumption: the absence of *perfect multicollinearity.* That is, none of the independent variables is perfectly correlated with another independent variable or linear combination of other independent variables. For example, with the following multiple regression model,

$$Y = a_0 + b_1X_1 + b_2X_2 + e,$$

perfect multicollinearity would exist if,

$$X_2 = c_0 + c_1X_1,$$

for X_2 is a perfect linear function of X_1 (that is, $R^2 = 1.0$). When perfect multicollinearity exists, it is impossible to arrive at a unique solution for the least squares parameter estimates. Any effort to calculate the partial regression coefficients, by computer or by hand, will fail. Thus, the presence of perfect multicollinearity is immediately detectable. Further, in practice, it is obviously quite unlikely to occur. However, *high multicollinearity* commonly perplexes the users of multiple regression.

With nonexperimental social science data, the independent variables are virtually always intercorrelated, that is, multicollinear. When this condition becomes extreme, serious estimation problems often arise. The general difficulty is that parameter estimates become unreliable. The magnitude of the partial slope estimate in the present sample may differ considerably from its magnitude in the next sample. Hence, we have little confidence that a particular slope estimate accurately reflects the impact of X on Y in the population. Obviously, because of such imprecision, this partial slope estimate cannot be usefully compared to other partial slope estimates in the equation, in order to arrive at a judgment of the relative effects of the independent variables. Finally, an estimated regression coefficient may be so unstable that it fails to achieve statistical significance, even though X is actually associated with Y in the population.

High multicollinearity creates these estimation problems because it produces large variances for the slope estimates and, consequently, large standard errors. Recalling the formula for a confidence interval (95%, two-tailed),

$$(b \pm t_{n-k-1;.975}s_b),$$

we recognize that a larger standard error, s_b, will widen the range of values that b might take on. Reviewing the formula for the t ratio,

$$b/s_b,$$

we observe that a larger s_b makes it more difficult to achieve statistical significance (e.g., more difficult to exceed the value of 2, which indicates statistical significance at the .05 level, two-tailed).

We can see how large variances occur with high multicollinearity by examining this variance formula,

$$\text{variance } b_i = s^2_{b_i} = s^2_u / v^2_i$$

where s^2_u is the variance of the error term in the multiple regression model, and v^2_i is the squared residual from the regression of the i^{th} independent variable, X_i, on the rest of the independent variables in the model. Hence,

$$v_i = X_i - \hat{X}_i.$$

If these other independent variables are highly predictive of X_i, then X_i and $\circ_i$ will be very close in value, and so v_i will be small. Therefore, the denominator in the above variance formula will be small, yielding a large variance estimate for b_i.

Of course, when analysts find a partial regression coefficient is statistically insignificant, they cannot simply dismiss the result on grounds of high multicollinearity. Before such a claim can be made, high multicollinearity must be demonstrated. Let us first look at common symptoms of high multicollinearity, which may alert the researcher to the problem. Then, we will proceed to a technique for diagnosis. One rather sure symptom of high multicollinearity is a substantial R^2 for the equation, but statistically insignificant coefficients. A second, weaker, signal is regression coefficients which change greatly in value when independent variables are dropped or added to the equation. A third, still less sure, set of symptoms involves suspicion about the magnitudes of the coefficients. A coefficient may be regarded as unexpectedly large (small), either in itself, or relative to another coefficient in the equation. It may even be so large (or small) as to be rejected as nonsensical. A fourth alert is a coefficient with the “wrong” sign. Obviously, this last symptom is feeble, for knowledge of the “right” sign is often lacking.

The above symptoms might provide the watchful analyst hints of a multicollinearity problem. However, by themselves, they cannot establish that the problem exists. For diagnosis, we must look directly at the inter-correlation of the independent variables. A frequent practice is to examine the bivariate correlations among the independent variables, looking for coefficients of about .8, or larger. Then, if none is found, one goes on to conclude that multicollinearity is not a problem. While suggestive, this approach is unsatisfactory, for it fails to take into account the relationship of an independent variable with *all* the other independent variables. It is possible, for instance, to find no large bivariate correlations, although one of the independent variables is a nearly perfect linear combination of the remaining independent variables. This possibility points to the preferred method of assessing multicollinearity: *Regress each independent variable on all the other independent variables.* When any of the R^2 from these equations is near 1.0, there is high multicollinearity. In fact, the largest of these R^2 serves as an indicator of the amount of multicollinearity which exists.

Let us apply what we have learned about multicollinearity to the four-variable Riverside model,

$$Y = a_0 + b_1X_1 + b_2X_2 + b_3X_3 + e,$$

where Y = income, X_1 = education, X_2 = seniority, X_3 = sex, e = error. The estimates for this model, which we have already examined, reveal no symptoms of a multicollinearity problem. That is, the coefficients are all significant, and their signs and magnitudes are reasonable. Therefore, we would anticipate that the above multicollinearity test would produce R^2_{xi} far from unity. Regressing each independent variable on all the others yields,

$$\hat{X}_1 = 7.02 + .42X_2 + .96X_3 \qquad R^2 = .49$$
$$\hat{X}_2 = -2.15 + 1.00X_1 + 1.68X_3 \qquad R^2 = .49$$
$$\hat{X}_3 = .066 + .022X_1 + .016X_2 \qquad R^2 = .14.$$

These R^2_{xi} show that these independent variables are intercorrelated in the Riverside sample, as we would expect with data of this type. But, we observe that the largest coefficient of multiple determination, $R^2 = .49$, lies a good distance from 1.0. Our conclusion is that multicollinearity is not a problem for the partial slope estimates in the Riverside multiple regression model.

The results do not always turn out so well. What can we do if high multicollinearity is detected? Unfortunately, none of the possible solutions is wholly satisfactory. In general, we must make the best of a bad situation. The standard prescription is to increase our information by enlarging the sample. As noted in an earlier chapter, the bigger the sample size, the greater the chance of finding statistical significance, other things being equal. Realistically, however, the researcher is usually unable to increase the sample. Also, multicollinearity may be severe enough that even a large n will not provide much relief.

Assuming the sample size is fixed, other strategies have to be implemented. One is to combine those independent variables that are highly intercorrelated into a single indicator. If this approach makes conceptual sense, then it can work well. Suppose, for example, a model which explains political participation (Y) as a function of income (X_1), race (X_2), radio listening (X_3), television watching (X_4), and newspaper reading (X_5). On the one hand, it seems sensible to combine the highly intercorrelated variables (X_3, X_4, X_5) into an index of media involvement. On the other hand, it is not sensible to combine the income and race variables, even if they are highly related.

Suppose our variables are “apples and oranges,” making it impractical to combine them. In the face of high multicollinearity, we cannot reliably separate the effects of the involved variables. Still, the equation may have value if its use is restricted to prediction. That is, it might be employed to predict Y for a given set of values on *all* the X's (e.g., when $X_1 = 2$, $X_2 = 4$, … $X_k = 3$), but not to interpret the independent effect on Y of a change in the value of a *single* X. Usually, this prediction strategy is uninteresting, for the goal is generally explanation, in which we talk about the impact of a particular X on Y.

A last technique for combatting multicollinearity is to discard the offending variable(s). Let us explore an example. Suppose we specify the following elementary multiple regression model,

$$Y = a_0 + b_1X_1 + b_2X_2 + e \qquad \text{Model I.}$$

Lamentably, however, we find that X_1 and X_2 are so highly related ($r_{12} = .9$), that the least squares estimates are unable reliably to assess the effect of either. An alternative is to drop one of the variables, say X_2, from the equation, and simply estimate this model:

$$Y = a_0 + b_1X_1 + e^* \qquad \text{Model II.}$$

A major problem with this procedure, of course, is its willful commission of specification error. Assuming Model I is the correct explanatory model, we know the estimate for b_1 in Model II will be biased. A revision which makes this technique somewhat more acceptable is to estimate yet another equation, now discarding the other offending variable (X_1),

$$Y = a_0 + b_2X_2 + e^{**} \qquad \text{Model III.}$$

If the Model II and Model III estimates are evaluated, along with those of Model I, then the damage done by the specification error can be more fully assessed.

High Multicollinearity: An Example

In order to grasp more completely the influences of high multicollinearity, it is helpful to explore a real data example. First, we present research findings reported by sociologist Gino Germani (1973). Then, we examine these findings with an eye to the multicollinearity issue.[4] Germani wishes to explain the vote support Juan Peron garnered in the 1946 presidential election in Argentina. His special interest is in assessing the backing Peron received from workers and internal migrants. To do so, he formulates a multiple regression model, arriving at the following estimates,

$$\hat{Y} = \underset{}{.52} + \underset{(.43)}{.18X_1} - \underset{(.41)}{.10X_2} - \underset{(.43)}{.57X_3} - \underset{(2.54)}{3.57X_4} + \underset{(.07)}{.29^*X_5}$$

$$R^2 = .24 \qquad n = 181 \qquad s_e = .11,$$

where Y = the percentage of the county's 1946 presidential vote going to Peron; X_1 = urban blue-collar workers (as a percentage of the economically active population in the county); X_2 = rural blue-collar workers (as a percentage of the economically active population in the county); X_3 = urban white-collar workers (as a percentage of the economically active population in the county); X_4 = rural white-collar workers (as a percentage of the economically active population in the county); X_5 = internal migrants (as a percentage of Argentinian-born males); the figures in parentheses are the standard errors of the slope estimates; the asterisk, *, indicates a coefficient statistically significant at the .05 level, two-tailed; R^2 = coefficient of multiple determination; n = 181 counties that contained a city of at least 5000 people; s_e = the standard error of estimate of Y.

These results suggest that only the presence of internal migrants significantly affected Peron support. We are pushed to the conclusion that the workers were not an influential factor in the election of Juan Peron. Such a conclusion becomes much less certain when we inspect the multicollinearity in the data. Let us diagnose the level of multicollinearity by regressing each independent variable on the remaining independent variables.

This yields the following R^2_{xi}, in order of magnitude: $R^2_{x2} = .99$, $R^2_{x3} = .98$, $R^2_{x1} = .98$, $R^2_{x4} = .75$, $R^2_{x5} = .32$.

Obviously, extreme multicollinearity is present. How might it be corrected? Further observations cannot be gathered. It is not sensible to combine any of the variables into an index. The purpose of the equation is not prediction. (If it were, the low R^2_y would inhibit it.) We are left with the strategy of discarding offending variables. An examination of the R^2_{xi} shows that the largest is R^2_{x2}. The variable, X^2, is an almost perfect linear function of all the other independent variables (X_1, X_3, X_4, X_5). Suppose we remove X_2 from the equation and reestimate:

$$\hat{Y} = \underset{}{.42} + \underset{(.07)}{.28*X_1} - \underset{(.10)}{.47*X_3} - \underset{(1.41)}{3.07*X_4} + \underset{(.07)}{.30*X_5}$$

$$R^2 = .24 \qquad n = 181 \qquad s_e = .11,$$

where definitions are the same as above.

According to these new estimates, *all* the variables have a statistically significant impact. Contrary to the earlier conclusion, workers do appear to have contributed to the election of Peron. How reliable are these new estimates? One check is to recalculate the level of multicollinearity. Regressing each independent variable on the remaining variables in the revised equation yields, $R^2_{x3} = .38$, $R^2_{x5} = .30$, $R^2_{x1} = .29$, $R^2_{x4} = .20$. We observe that all of these R^2_{xi} are quite far from unity, indicating that multicollinearity has ceased to be problematic. The revised parameter estimates would appear much more reliable than the contrary ones generated with the offending X_2 in the equation. Hopefully, this rather dramatic example brings home the perils of high multicollinearity.

The Relative Importance of the Independent Variables

We sometimes want to evaluate the relative importance of the independent variables in determining Y. An obvious procedure is to compare the magnitudes of the partial slopes. However, this effort is often thwarted by the different measurement units and variances of the variables. Suppose, for example, the following multiple regression equation predicting dollars contributed to political campaigns as a function of an individual's age and income,

$$\hat{Y} = 8 + 2X_1 + .010X_2,$$

where Y = campaign contributions (in dollars), X_1 = age (in years), X_2 = income (in dollars).

The relative influence of income and age on campaign contributions is difficult to assess, for the measurement units are not comparable, that is, dollars versus years. One solution is to *standardize* the variables, re-estimate, and evaluate the new coefficients. (Some computing routines for regression, such as that of SPSS,

automatically provide the standardized coefficients along with the unstandardized coefficients.) Any variable is standardized by converting its scores into standard deviation units from the mean. For the above variables, then,

$$Y^* = \frac{Y - \bar{Y}}{s_y}, \quad X_1^* = \frac{X_1 - \bar{X}_1}{s_{x_1}}, \quad X_2^* = \frac{X_2 - \bar{X}_2}{s_{x_2}},$$

where the asterisk, *, indicates the variable is standardized. Reformulating the model with these variables yields,

$$\hat{Y}^* = \beta_1 X_1^* + \beta_2 X_2^*.$$

(Note that standardization forces the intercept to zero.) The standardized partial slope is often designated with "β," and referred to as a *beta weight*, or *beta coefficient.* (Do not confuse this β with the symbol for the population slope.)

The beta weight corrects the unstandardized partial slope by the ratio of the standard deviation of the independent variable to the standard deviation of the dependent variable:

$$\beta_i = b_i \frac{s_{x_i}}{s_y}.$$

In the special case of the bivariate regression model, the beta weight equals the simple correlation between the two variables. That is, assuming the model,

$$Y = a + bX + e,$$

then,

$$\beta = b \frac{s_x}{s_y} = r.$$

However, this equality does not hold for a multiple regression model. (Only in the unique circumstance of *no* multicollinearity would β = r with a multiple regression model.)

The standardized partial slope estimate, or beta weight, indicates *the average standard deviation change in Y associated with a standard deviation change in X, when the other independent variables are held constant.* Suppose the beta weights for the above campaign contribution equation are as follows:

$$\hat{Y}^* = .15X_1^* + .45X_2^*.$$

For example, $\beta_2 = .45$ says that a one standard deviation change in income is associated with a .45 standard deviation change in campaign contributions, on the average, with age held constant. Let us consider the meaning of this interpretation more fully. Assuming X_2 is normally distributed, then a one standard deviation income rise for persons at, say, the mean income would move them into a high income bracket, above which only about 16% of the population resided. We see that this strong manipulation of X does not result in as strong a response in Y, for β_2 is far from unity. Still, campaign contributions do tend to climb by almost one-half

of a standard deviation. In contrast, a considerable advance in age (a full one standard deviation increase) elicits a very modest increment in contributions (only .15 of a standard deviation). We conclude that the impact of income, as measured in standard deviation units, is greater than the impact of age, likewise measured. Indeed, it seems that the effect of income on campaign contributions is three times that of age (.45/.15 = 3).

The ability of standardization to assure the comparability of measurement units guarantees its appeal, when the analyst is interested in the relative effects of the independent variables. However, difficulties can arise if one wishes to make comparisons across samples. This is because, in estimating the same equation across samples, the value of the beta weight, unlike the value of the unstandardized slope, can change merely because the variance of X changes. In fact, the larger (smaller) the variance in X, the larger (smaller) the beta weight, other things being equal. (To understand this, consider again the beta weight formula,

$$\beta_i = b_i \frac{s_{x_i}}{s_y}.$$

We see that, as s_{xi}, the numerator of the fraction, increases, the magnitude of β_i must necessarily increase.)

As an example, suppose that the above campaign contributions model was developed from a U.S. sample, and we wished to test it for another Western democracy, say Sweden. Our beta weights from this hypothetical sample of the Swedish electorate might be,

$$\hat{Y}^* = .18X_1^* + .22X_2^*,$$

where the variables are defined as above. Comparing β_2 (United States) = .45 to β_2 (Sweden) = .22, we are tempted to conclude that the effect of income in Sweden is about one-half its effect in the United States. However, this inference may well be wrong, given that the standard deviation of of income in the United States is greater than the standard deviation of income in Sweden. That is, the wider spread of incomes in the United States may be masking the more equal effect a unit income change actually has in both countries, that is, b_2 (United States) ? b_2 (Sweden). To test for this possibility, we must of course examine the unstandardized partial slopes, which we suppose to be the following:

$$\hat{Y} = 9 + 1.7X_1 + .012X_2.$$

When these unstandardized Swedish results are compared to the unstandardized United States results, they suggest that, in reality, the effect of income on campaign contributions is essentially the same in both countries (.010 ? .012). In general, when the variance in X diverges from one sample to the next, it is preferable to base any cross-sample comparisons of effect on the unstandardized partial slopes.

Extending the Regression Model: Dummy Variables

Regression analysis encourages the use of variables whose amounts can be measured with numeric precision, that is, *interval variables.* A classic example of such a variable is income. Individuals can be ordered numerically according to their quantity of income, from the lowest to the highest. Thus, we can say that John's

income of $12,000 is larger than Bill's income of $6,000; in fact, it is exactly twice as large. Of course, not all variables are measured at a level which allows such precise comparison. Nevertheless, these noninterval variables are candidates for incorporation into a regression framework, through the employment of *dummy variables.*

Many noninterval variables can be considered *dichotomies*, e.g., sex (male, female), race (Black, White), marital status (single, married). Dichotomous independent variables do not cause the regression estimates to lose any of their desirable properties. Because they have two categories, they manage to “trick” least squares, entering the equation as an interval variable with just two values. It is useful to examine how such “dummy” variables work. Suppose we argue that a person's income is predicted by race in this bivariate regression,

$$\hat{Y} = a + bX,$$

where Y = income, X = race (0 = Black, 1 = White). If X = 0, then

$$\hat{Y} = a,$$

the prediction of the mean income for Blacks. If X = 1, then

$$\hat{Y} = a + b,$$

the prediction of the mean income for Whites. Therefore, the slope estimate, b, indicates the difference between the mean incomes of Blacks and Whites. As always, the t ratio of b measures its statistical significance. We have already observed such a dummy variable in action, in the four-variable Riverside equation, which included sex as an independent variable (0 = female, 1 = male). There, the partial regression coefficient, b_3, reports the difference in average income between men and women, after the influences of education and seniority have been accounted for. As noted, this difference is statistically and substantively signficant.

Obviously, not all noninterval variables are dichotomous. Noninterval variables with multiple categories are of two basic types: *ordinal* and *nominal.* With an ordinal variable, cases can be ordered in terms of amount, but not with numeric precision. Attitudinal variables are commonly of this kind. For example, in a survey of the electorate, respondents may be asked to evaluate their political interest, ranking themselves as “not interested,” “somewhat interested,” or “very interested.” We can say that Respondent A, who chooses “very interested,” is more interested in politics than Respondent B, who selects “not interested,” but we cannot say numerically how much more. Ordinal variables, then, only admit of a ranking from “less to more.” The categories of a nominal variable, in contrast, cannot be so ordered. The variable of religious affiliation is a good example. The categories of Protestant, Catholic, or Jew represent personal attributes which yield no meaningful ranking.

Noninterval variables with multiple categories, whether ordinal or nominal, can be incorporated into the multiple regression model through the dummy variable technique. Let us explore an example. Suppose the dollars an individual contributes to a political campaign are a function of the above-mentioned ordinal variable, political interest. Then, a correct model would be

$$Y = a_0 + b_1X_1 + b_2X_2 + e,$$

where Y = campaign contributions (in dollars); X_1 = a dummy variable, scored 1 if “somewhat interested,” 0 if otherwise; X_2 = a dummy variable, scored 1 if “very interested,” 0 if otherwise; e = error.

Observe that there are only *two* dummy variables to represent the trichotomous variable of political interest. If there were three dummy variables, then the parameters could not be uniquely estimated. That is, a third dummy, X_3 (scored 1 if “not interested,” 0 if otherwise), would be an exact linear function of the others, X_1 and X_2. (Consider that when the score of any respondent on X_1 and X_2 is known, it would always be possible to predict his or her X_3 score. For example, if a respondent has values of 0 on X_1 and 0 on X_2, then he or she is necessarily “not interested” in politics, and would score 1 on X_3.) This describes a situation of perfect multicollinearity, in which estimation cannot proceed. To avoid such a trap, which is easy to fall into, we memorize this rule: *When a noninterval variable has G categories, use G − I dummy variables to represent it.*

A question now arises as to how to estimate the campaign contributions of this excluded group, those who responded “not interested.” Their average campaign contribution is estimated by the intercept of the equation. That is, for someone who is “not interested,” the prediction equation reduces to,

$$\hat{Y} = a_0 + b_1X_1 + b_2X_2$$

$$= a_0 + b_1(0) + b_2(0)$$

$$\hat{Y} = a_0.$$

Thus, the intercept estimates the average campaign contribution of someone who is “not interested” in politics.

This estimated contribution, a_0, for the “not interested” category serves as a base for comparing the effects of the other categories of political interest. The prediction equation for someone in the category, “somewhat interested,” reduces to

$$\hat{Y} = a_0 + b_1X_1 + b_2X_2$$

$$= a_0 + b_1(1) + b_2(0)$$

$$\hat{Y} = a_0 + b_1.$$

Hence, the partial slope estimate, b_1, indicates the difference in mean campaign contributions between those “somewhat interested” and those “not interested,” that is, $(a_0 + b_1) - a_0 = b_1$.

For the last category, “very interested,” the prediction equation reduces to

$$\hat{Y} = a_0 + b_1X_1 + b_2X_2$$

$$= a_0 + b_1(0) + b_2(1)$$

$$\hat{Y} = a_0 + b_2.$$

Thus, the partial slope estimate, b_2, points out the difference in average campaign contributions between the “very interested” and the “not interested.” Given the hypothesis that heightened political interest increases campaign contributions, we would expect that $b_2 > b_1$.

A data example will increase our appreciation of the utility of dummy variables. Suppose, with the Riverside study, it occurs to us that the income received from working for city government might be determined in part by the employee's political party affiliation (Democrat, Republican, or independent). In that case, the proper specification of the model becomes,

$$Y = a_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e,$$

where Y = income; X_1 = education; X_2 = seniority; X_3 = sex; X_4 = a dummy variable scored 1 if independent, 0 otherwise; X_5 = a dummy variable scored 1 if Republican, 0 otherwise; e = error.

The variable, political party, has three categories. Thus, applying the G − 1 rule, we had to formulate 3 − 1 = 2 dummy variables. We chose to construct one for independents (X_4) and one for Republicans (X_5), which left Democrats as the base category. The selection of a base category is entirely up to the analyst. Here, we selected Democrats as the standard for comparison because we guessed they would have the lowest income, with independents and Republicans having successively higher incomes.

Least squares yields the following parameter estimates,

$$\begin{array}{llllll} \hat{Y} = 5496 & + 382X_1 & + 250X_2 & + 2134X_3 & - 572X_4 & + 386X_5 \\ \quad (3.90) & (2.74) & (2.78) & (2.33) & (-.48) & (.41) \end{array}$$

$$R^2 = .73 \qquad n = 32 \qquad s_e = 2403,$$

where the variables are defined as above, the values in parentheses are t ratios, the R^2 = the coefficient of multiple determination, n = sample size, s_c = standard error of estimate for Y.

First, we note that the estimates from our prior specification remain virtually unchanged. Further, from the t ratio, we see that the average income of independents is not significantly different (.05 level) from the average income of Democrats, once the effects of education, seniority, and sex are controlled. (Put another way, b_4 does not add significantly to the intercept, a_0.) Likewise, the average income of Republicans is found not to differ significantly from that of the Democrats. We must conclude that, contrary to our expectation, political party affiliation does not influence the income of Riverside municipal employees. Our original four-variable model remains the preferred specification.

Through use of the dummy variable technique, the inclusion into our multiple regression equation of the noninterval variable, political party, poses no problem. Some researchers would argue that this variable could be inserted into our regression equation directly, bypassing the dummy variable route. The argument is that an ordinal variable is a candidate for regression, even though the distances between the categories are not exactly equal. This is a controversial point of view. In brief, the advocates' primary defense is that, in practice, the conclusions are usually equivalent to those generated by more correct techniques (i.e., the application of dummy variable regression or ordinal-level statistics). A secondary argument is that multiple regression

analysis is so powerful, compared to ordinal-level techniques, that the risk of error is acceptable. We cannot resolve this debate here. However, we can provide a practical test by incorporating political party into the Riverside equation as an ordinal variable.

At first blush, political party affiliation may appear as strictly nominal. Nevertheless, political scientists commonly treat it as ordinal. We can say, for example, that an independent is “more Republican” than a Democrat, who is “least Republican” of all. Hence, we can order the categories in terms of their “distance” from Republicans. This order is indicated in the following numeric code, Democrat = 0, independent = 1, Republican = 2, which ranks the categories along this dimension of “Republicanism.” This code provides each respondent a score on a political party variable, X_4, which we now enter into the Riverside equation. Least squares yields the following estimates,

$$\hat{Y} = \underset{(3.87)}{5314} + \underset{(2.85)}{392X_1} + \underset{(2.74)}{243X_2} + \underset{(2.36)}{2137X_3} + \underset{(.40)}{186X_4}$$

$$R^2 = .73 \qquad n = 32 \qquad s_e = 2380,$$

where Y = income; X_1 = education; X_2 = seniority; X_3 = sex; X_4 = political party affiliation, scored 0 = Democrat, 1 = independent, 2 = Republican; and the statistics are defined as above.

The estimates for the coefficients of our original variables are essentially unchanged. Also, political party affiliation is shown to have no statistically significant impact on employee's income ($t < 2$). Thus, in this particular case, regression analysis with an ordinal variable arrives at the same conclusion as the more proper regression analysis with dummy variables.

Determinants of Coal Mining Fatalities: A Multiple Regression Example

Let us pick up the explanation of coal mining fatalities begun earlier. It is now clear that our bivariate model is incomplete. On the basis of theoretical considerations, prior research, and indicator availability, we formulate the following explanatory model:

$$Y = a_0 + b_1X_1 + b_2X_2 + b_3X_3 + e,$$

where Y = the annual coal mining fatality rate (measured as deaths per million hours worked); X_1 = the natural logarithm of the annual federal coal mining safety budget (measured in thousands of constant dollars, 1967 = 100); X_2 = the percentage of miners working underground; X_3 = a dummy variable for the President's political party, scored 0 when the President that year is Republican and 1 when Democrat; e = error.

We have already argued that the coal mining fatality rate falls in response to more vigorous safety enforcement, measured by the Bureau of Mines safety budget, X_1. Further, we contend that when the percentage of miners working underground (as opposed to strip mining) advances, the fatality rate rises. Last, we believe that the political party in the White House, X_3, makes a difference, with Democrats more likely than Republicans to take measures to reduce fatalities. Let us test these hypotheses.

Least squares yields these estimates (the data sources are those mentioned previously):

$$\hat{Y} = \underset{(1.75)}{1.23} - \underset{(-6.48)}{.189X_1} + \underset{(3.06)}{.019X_2} + \underset{(.84)}{.046X_3}$$

$$R^2 = .83 \qquad n = 44 \qquad s_e = .13$$

where the values in parentheses are the t ratios, the R^2 = the coefficient of multiple determination, n = 44 annual observations from 1932–1975 (the 1976 figure was not available for X_2), s_e = the standard error of estimate for Y.

These results suggest that federal safety enforcement, X_1, and the extent to which mining is carried on underground, X_2, significantly influence the fatality rate. However, the President's party, X_3, appears to have no significant impact on the fatality rate. (The t ratio for b_3 is quite far from the value of 2.) But before rejecting our hypothesis on the effect of the President's party, we should perhaps check for a multicollinearity problem. After all, it may simply be multicollinearity that is causing b_3 to fall short of statistical significance. Regressing each independent variable on the others in the equation yields $R^2_{x1} = .63$, $R^2_{x2} = .45$, $R^2_{x3} = .46$. The presidential party variable, X_3, when regressed on X_1 and X_2, produces an R^2 which is a long way from unity. Further, according to the R^2 for the other independent variables, they manifest at least the same degree of multicollinearity, but their regression coefficients still manage to attain statistical significance. In sum, it seems unlikely that multicollinearity is the cause of a lack of statistical significance for b_3.

We can conclude, with greater confidence, that the coal mining fatality rate is unaltered by political party changes in the White House. This causes us to revise our model specification and reestimate our equation, as follows:

$$\hat{Y} = \underset{(2.80)}{1.58} - \underset{(-9.58)}{.206X_1} + \underset{(3.00)}{.017X_2}$$

$$R^2 = .83 \qquad n = 44 \qquad s_e = .13,$$

where the terms are defined as above.

This multiple regression model improves our explanation of the coal mining fatality rate, over our earlier bivariate regression model. The R^2, which is somewhat greater, indicates that fully 83% of the variance is being accounted for. Further, the more adequate specification has reduced the bias of the slope estimate for the safety expenditures variable, X_1. In the bivariate equation, this slope = −.247, which exaggerates the ability of safety budget increases to lower fatalities. Because X_2 was excluded, X_1 was permitted to account for a part of Y which should be explained by X_2. Inclusion of X_2 in our multiple regression equation shrunk the effect of safety expenditures to its proper size (b_1 = −.206).

Is this newly incorporated variable, the percentage of miners underground, even more important a determinant of the coal mining fatality rate than the safety budget variable? Evaluation of the beta weights

provides one answer to this question. Standardizing the variables and re-estimating the equation yields,

$$\hat{Y}^* = -.75X_1^* + .24X_2^*,$$

where the variables are defined as above, and standardized, as indicated by the asterisk, *. These beta weights suggest that the safety budget is a more important influence on the fatality rate than is the percentage of miners underground. In fact, a standard deviation change in the safety budget variable has about three times the impact of a comparable change in the percentage of miners underground.

What Next?

Comprehension of the material in this monograph should permit the reader to use regression analysis widely and easily. Of course, in so few pages, not everything can be treated exhaustively. There are topics which merit further study. Nonlinearity is one such topic. While relationships among social science variables are often linear, it is not uncommon for nonlinearity to occur. We spelled out the consequences of violating the linearity assumption, and provided an example of how a nonlinear relationship was straightened out by a logarithmic transformation. Other such linearizing transformations are available, whose appropriateness depends on the shape of the particular curve. Popular ones are the reciprocal,

$$Y = a_0 + b_1 \frac{1}{X} + e,$$

and the second-order polynomial.

$$Y = a_0 + b_1X + b_2X^2 + e.$$

(For good discussions of these and other transformations, see Kelejian and Oates, 1974, pp. 92–102, 167–175; Tufte, 1974, pp. 108–130.)

Another topic which we only touched on was the use of time-series. As noted, autocorrelation is frequently a problem in the analysis of time-series data. Take, for example, the model,

$$Y_t = a + bX_t + e_t,$$

where the subscript i has been replaced with t in order to indicate “time,” Y_t = annual federal government expenditures, X_t = annual presidential budget request, e_t = error term. When we think of e_t as including omitted explanatory variables, autocorrelation appears quite likely. Suppose, for instance, that one of these omitted variables is annual gross national product (GNP); clearly, GNP from the previous year (GNP_{t-1}) is correlated with GNP from the current year (GNP_t); hence, $rc_tc_t \neq 0$. This error process, in which error from the immediately prior time (e_{t-1}) is correlated with error at the present time (e_t), describes a *first-order autoregressive process.* This process can be easily detected (e.g., with the Durbin-Watson test) and corrected (e.g., with the Cochrane-Orcutt technique).[5] Other error processes are more difficult to diagnose and cure. (The problems and opportunities of time-series are introduced in Ostrom, 1978.)

In our exposition of regression, we have consciously stressed verbal interpretation rather than mathematical derivation. Given it is an introduction, such emphasis seems proper. At this point, the serious student might

wish to work through the material using the calculus and matrix algebra. (For this purpose, consult the relevant sections of Kmenta, 1971, and Pindyck and Rubinfeld, 1976.)

Throughout, we have formulated *single-equation models*, either bivariate or multivariate. We could also propose *multiequation models.* These models, known technically as *simultaneous-equation models*, become important when we believe causation is two way, rather than one way. For example, a simple regression model assumes that X causes Y, but not vice versa, that is, X → Y. Perhaps, though, X causes Y, and Y causes X, that is, X r? Y. This is a case of reciprocal causation, where we have two equations,

$$Y = a + bX + e$$

$$X = a + bY + e.$$

The temptation is to estimate each with the ordinary least squares procedure we have learned here. Unfortunately, in the face of reciprocal causation, ordinary least squares will generally produce biased parameter estimates. Therefore, we must modify our procedure, probably applying *two-stage least squares.* Reciprocal causation and the ensuing problems of estimation form the core issues of causal modeling (Asher, 1976, provides a useful treatment of this topic). Happily, a firm grasp of regression analysis will speed the student's mastery of causal modeling, as well as a host of other quantitative techniques.

http://dx.doi.org/10.4135/9781412983440.n3