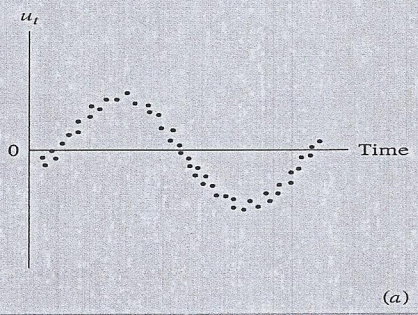
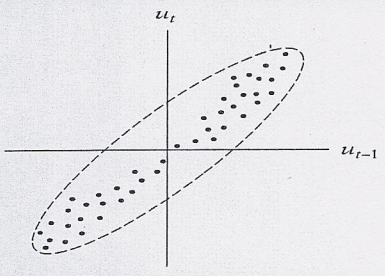
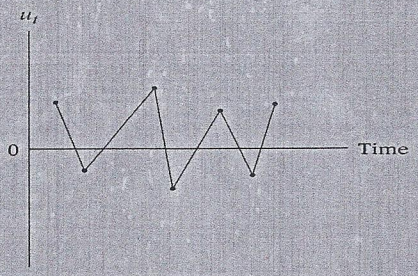
BASIC ECONOMETRICS

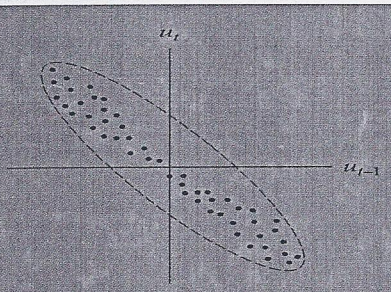


(a)





(b)



Fifth Edition

Damodar N. Gujarati
Dawn C. Porter

in Eq. (13.2.9) by α. Now if Eq. (13.2.8) is the "correct" or "true" model, would the estimated α provide an unbiased estimate of the true β? That is, will $E(\hat{\alpha}) = \beta$? If that is not the case, improper stochastic specification of the error term will constitute another source of specification error.

A specification error that is sometimes overlooked is the **interaction among the regressors,** that is, the **multiplicative effect of one or more regressors on the regressand.** To illustrate, consider the following simplified wage function:

$$\ln W_i = \beta_1 + \beta_2 \text{Education}_i + \beta_3 \text{Gender}_i + \beta_4 (\text{Education})(\text{Gender}) + u \qquad (13.2.10)$$

In this model, the change in the relative wages with respect to education depends not only on education but also on the gender ($\frac{\partial \ln W}{\partial \text{Education}} = \beta_2 + \beta_4 \text{Gender}$). Likewise, the change in relative wages with respect to gender depends not only on gender but also on education.

To sum up, in developing an empirical model, one is likely to commit one or more of the following specification errors:

1. Omission of a relevant variable(s).
2. Inclusion of an unnecessary variable(s).
3. Adoption of the wrong functional form.
4. Errors of measurement.
5. Incorrect specification of the stochastic error term.
6. Assumption that the error term is normally distributed.

Before turning to an examination of these specification errors in some detail, it may be fruitful to distinguish between **model specification errors** and **model mis-specification errors.** The first four types of error discussed above are essentially in the nature of model specification errors in that we have in mind a "true" model but somehow we do not estimate the correct model. In model mis-specification errors, we do not know what the true model is to begin with. In this context one may recall the controversy between the Keynesians and the monetarists. The monetarists give primacy to money in explaining changes in GDP, whereas the Keynesians emphasize the role of government expenditure to explain changes in GDP. So to speak, these are two competing models.

In what follows, we will first consider model specification errors and then examine model mis-specification errors.

13.3 Consequences of Model Specification Errors

Whatever the sources of specification errors, what are the consequences? To keep the discussion simple, we will answer this question in the context of the three-variable model and consider in this section the first two types of specification errors discussed earlier, namely, (1) **underfitting a model,** that is, omitting relevant variables, and (2) **overfitting a model,** that is, including unnecessary variables. Our discussion here can be easily generalized to more than two regressors, but with tedious algebra;[6] matrix algebra becomes almost a necessity once we go beyond the three-variable case.

[6]But see Exercise 13.32.

Underfitting a Model (Omitting a Relevant Variable)

Suppose the true model is:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{13.3.1}$$

but for some reason we fit the following model:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i \tag{13.3.2}$$

The consequences of omitting variable X_3 are as follows:

1. If the left-out, or omitted, variable X_3 is correlated with the included variable X_2, that is, r_{23}, the correlation coefficient between the two variables is *nonzero* and $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are *biased as well as inconsistent.* That is, $E(\hat{\alpha}_1) \neq \beta_1$ and $E(\hat{\alpha}_2) \neq \beta_2$, and the bias does not disappear as the sample size gets larger.
2. Even if X_2 and X_3 are not correlated, $\hat{\alpha}_1$ is biased, although $\hat{\alpha}_2$ is now unbiased.
3. The disturbance variance σ^2 is incorrectly estimated.
4. The conventionally measured variance of $\hat{\alpha}_2 \left(= \sigma^2 / \sum x_{2i}^2\right)$ is a *biased* estimator of the variance of the true estimator $\hat{\beta}_2$.
5. In consequence, the usual confidence interval and hypothesis-testing procedures are likely to give misleading conclusions about the statistical significance of the estimated parameters.
6. As another consequence, the forecasts based on the incorrect model and the forecast (confidence) intervals will be unreliable.

Although proofs of each of the above statements will take us far afield,[7] it is shown in Appendix 13A, Section 13A.1, that

$$E(\hat{\alpha}_2) = \beta_2 + \beta_3 b_{32} \tag{13.3.3}$$

where b_{32} is the slope in the regression of the excluded variable X_3 on the included variable X_2 ($b_{32} = \sum x_{3i} x_{2i} / \sum x_{2i}^2$). As Eq. (13.3.3) shows, $\hat{\alpha}_2$ is biased, unless β_3 or b_{32} or both are zero. We rule out β_3 being zero, because in that case we do not have specification error to begin with. The coefficient b_{32} will be zero if X_2 and X_3 are uncorrelated, which is unlikely in most economic data.

Generally, however, the extent of the bias will depend on the *bias term* $\beta_3 b_{32}$. If, for instance, β_3 is positive (i.e., X_3 has a positive effect on Y) and b_{32} is positive (i.e., X_2 and X_3 are positively correlated), $\hat{\alpha}_2$, on average, will overestimate the true β_2 (i.e., positive bias). But this result should not be surprising, for X_2 represents not only its *direct effect* on Y but also its *indirect effect* (via X_3) on Y. In short, X_2 gets credit for the influence that is rightly attributable to X_3, the latter being prevented from showing its effect explicitly because it is not "allowed" to enter the model. As a concrete example, consider the example discussed in Chapter 7 (Example 7.1).

[7] For an algebraic treatment, see Jan Kmenta, *Elements of Econometrics,* Macmillan, New York, 1971, pp. 391–399. Those with a matrix algebra background may want to consult J. Johnston, *Econometrics Methods,* 4th ed., McGraw-Hill, New York, 1997, pp. 119–112.

EXAMPLE 13.1
Illustrative Example: Child Mortality Revisited

Regressing child mortality (CM) on per capita GNP (PGNP) and the female literacy rate (FLR), we obtained the regression results shown in Eq. (7.6.2), giving the partial slope coefficient values of the two variables as −0.0056 and −2.2316, respectively. But if we now drop the FLR variable, we obtain the results shown in Eq. (7.7.2). If we regard Eq. (7.6.2) as the correct model, then Eq. (7.7.2) is a mis-specified model in that it omits the relevant variable FLR. Now you can see that in the correct model the coefficient of the PGNP variable was −0.0056, whereas in the "incorrect" model (7.7.2) it is now −0.0114.

In absolute terms, now PGNP has a greater impact on CM as compared with the true model. But if we regress FLR on PGNP (regression of the excluded variable on the included variable), the slope coefficient in this regression (b_{32} in terms of Eq. [13.3.3]) is 0.00256.[8] This suggests that as PGNP increases by a unit, on average, FLR goes up by 0.00256 units. But if FLR goes up by these units, its effect on CM will be $(-2.2316)(0.00256) = \hat{\beta}_3 b_{32} = -0.00543$.

Therefore, from Eq. (13.3.3) we finally have $(\hat{\beta}_2 + \hat{\beta}_3 b_{32}) = [-0.0056 + (-2.2316)(0.00256)] \approx -0.0111$, which is about the value of the PGNP coefficient obtained in the incorrect model (7.7.2).[9] As this example illustrates, the true impact of PGNP on CM is much less (−0.0056) than that suggested by the incorrect model (7.7.2), namely, (−0.0114).

Something to be aware of: Please note that the "homoskedasticity-only" formula for the variance of Beta-hat is being used in equations 13.3.4 and 13.3.5 to illustrate the "trade-off" between "bias" and "variance." You can illustrate the same "trade-off" if you were to use the "heteroskedasticity-robust" variance formula.

Now let us examine the variances of $\hat{\alpha}_2$ and $\hat{\beta}_2$

$$\text{var}(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2} \tag{13.3.4}$$

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} = \frac{\sigma^2}{\sum x_{2i}^2} \text{VIF} \tag{13.3.5}$$

where VIF (a measure of collinearity) is the variance inflation factor $[= 1/(1 - r_{23}^2)]$ discussed in Chapter 10 and r_{23} is the correlation coefficient between variables X_2 and X_3; Eqs. (13.3.4) and (13.3.5) are familiar to us from Chapters 3 and 7.

As formulas (13.3.4) and (13.3.5) are not the same, in general, var $(\hat{\alpha}_2)$ will be different from var $(\hat{\beta}_2)$. But we know that var $(\hat{\beta}_2)$ is unbiased (why?). Therefore, var $(\hat{\alpha}_2)$ is biased, thus substantiating the statement made in point 4 earlier. Since $0 < r_{23}^2 < 1$, it would *seem* that in the present case var $(\hat{\alpha}_2)$ < var $(\hat{\beta}_2)$. Now we face a dilemma: Although $\hat{\alpha}_2$ is biased, its variance is smaller than the variance of the unbiased estimator $\hat{\beta}_2$ (of course, we are ruling out the case where $r_{23} = 0$, since in practice there is some correlation between regressors). So, there is a trade-off involved here.[10]

The story is not complete yet, however, for the σ^2 estimated from model (13.3.2) and that estimated from the true model (13.3.1) are not the same because the residual sum of squares (RSS) of the two models as well as their degrees of freedom (df) are different. You may recall that we obtain an estimate of σ^2 as $\hat{\sigma}^2 = \text{RSS/df}$, which depends on the number of regressors included in the model as well as the df ($= n$, number of parameters

[8]The regression results are:

$$\widehat{\text{FLR}} = 47.5971 + 0.00256\text{PGNP}$$
$$\text{se} = (3.5553) \quad (0.0011) \qquad r^2 = 0.0721$$

[9]Note that in the true model $\hat{\beta}_2$ and $\hat{\beta}_3$ are unbiased estimates of their true values.

[10]To bypass the trade-off between bias and efficiency, one could choose to minimize the mean square error (MSE), since it accounts for both bias and efficiency. On MSE, see the statistical appendix, **Appendix A.** See also Exercise 13.6.

estimated). Now if we add variables to the model, the RSS generally decreases (recall that as more variables are added to the model, the R^2 increases), but the degrees of freedom also decrease because more parameters are estimated. The net outcome depends on whether the RSS decreases sufficiently to offset the loss of degrees of freedom due to the addition of regressors. It is quite possible that if a regressor has a strong impact on the regressand—for example, it may reduce RSS more than the loss in degrees of freedom as a result of its addition to the model—inclusion of such variables will not only reduce the bias but will also increase the precision (i.e., reduce the standard errors) of the estimators.

On the other hand, if the relevant variables have only a marginal impact on the regressand, and if they are highly correlated (i.e., VIF is larger), we may reduce the bias in the coefficients of the variables already included in the model, but increase their standard errors (i.e., make them less efficient). Indeed, the trade-off in this situation between bias and precision can be substantial. As you can see from this discussion, the trade-off will depend on the relative importance of the various regressors.

To conclude this discussion, let us consider the special case where $r_{23} = 0$, that is, X_2 and X_3 are uncorrelated. This will result in b_{32} being zero (why?). Therefore, it can be seen from Eq. (13.3.3) that $\hat{\alpha}_2$ is now unbiased.[11] Also, it seems from Eqs. (13.3.4) and (13.3.5) that the variances of $\hat{\alpha}_2$ and $\hat{\beta}_2$ are the same. Is there no harm in dropping the variable X_3 from the model even though it may be relevant theoretically? The answer generally is no, for in this case, as noted earlier, var $(\hat{\alpha}_2)$ estimated from Eq. (13.3.4) is still biased and therefore our hypothesis-testing procedures are likely to remain suspect.[12] Besides, in most economic research X_2 and X_3 will be correlated, thus creating the problems discussed previously. **The point is clear: Once a model is formulated on the basis of the relevant theory, one is ill-advised to drop a variable from such a model.**

Inclusion of an Irrelevant Variable (Overfitting a Model)

Now let us assume that

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{13.3.6}$$

is the truth, but we fit the following model:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + v_i \tag{13.3.7}$$

and thus commit the specification error of including an unnecessary variable in the model.

The consequences of this specification error are as follows:

1. The OLS estimators of the parameters of the "incorrect" model are all *unbiased and consistent*, that is, $E(\alpha_1) = \beta_1$, $E(\hat{\alpha}_2) = \beta_2$, and $E(\hat{\alpha}_3) = \beta_3 = 0$.
2. The error variance σ^2 is correctly estimated.
3. The usual confidence interval and hypothesis-testing procedures remain valid.
4. However, the estimated α's will be generally inefficient, that is, their variances will be generally larger than those of the $\hat{\beta}$'s of the true model. The proofs of some of these statements can be found in Appendix 13A, Section 13A.2. The point of interest here is the relative inefficiency of the $\hat{\alpha}$'s. This can be shown easily.

[11] Note, though, $\hat{\alpha}_1$ is still biased, which can be seen intuitively as follows: We know that $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$, whereas $\hat{\alpha}_1 = \bar{Y} - \hat{\alpha}_2 \bar{X}_2$, and even if $\hat{\alpha}_2 = \hat{\beta}_2$, the two intercept estimators will not be the same.

[12] For details, see Adrian C. Darnell, *A Dictionary of Econometrics*, Edward Elgar Publisher, 1994, pp. 371–372.

From the usual OLS formula we know that

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2} \tag{13.3.8}$$

and

$$\text{var}(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1 - r_{23}^2)} \tag{13.3.9}$$

Therefore,

$$\frac{\text{var}(\hat{\alpha}_2)}{\text{var}(\hat{\beta}_2)} = \frac{1}{1 - r_{23}^2} \tag{13.3.10}$$

Since $0 \leq r_{23}^2 \leq 1$, it follows that $\text{var}(\hat{\alpha}_2) \geq \text{var}(\hat{\beta}_2)$; that is, the variance of $\hat{\alpha}_2$ is generally greater than the variance of $\hat{\beta}_2$ even though, on average, $\hat{\alpha}_2 = \beta_2$ [i.e., $E(\hat{\alpha}_2) = \beta_2$].

The implication of this finding is that the inclusion of the unnecessary variable X_3 makes the variance of $\hat{\alpha}_2$ larger than necessary, thereby making $\hat{\alpha}_2$ less precise. This is also true of $\hat{\alpha}_1$.

Notice the **asymmetry** in the two types of specification biases we have considered. If we exclude a relevant variable, the coefficients of the variables retained in the model are generally biased as well as inconsistent, the error variance is incorrectly estimated, and the usual hypothesis-testing procedures become invalid. On the other hand, including an irrelevant variable in the model still gives us unbiased and consistent estimates of the coefficients in the true model, the error variance is correctly estimated, and the conventional hypothesis-testing methods are still valid; the only penalty we pay for the inclusion of the superfluous variable is that the estimated variances of the coefficients are larger, and as a result our probability inferences about the parameters are less precise. An unwanted conclusion here would be that it is better to include irrelevant variables than to omit the relevant ones. But this philosophy is not to be espoused because the addition of unnecessary variables will lead to a loss in the efficiency of the estimators and may also lead to the problem of multicollinearity (why?), not to mention the loss of degrees of freedom. Therefore,

> In general, the best approach is to include only explanatory variables that, on theoretical grounds, *directly* influence the dependent variable and that are not accounted for by other included variables.[13]

13.4 Tests of Specification Errors

Knowing the consequences of specification errors is one thing but finding out whether one has committed such errors is quite another, for we do not deliberately set out to commit such errors. Very often specification biases arise inadvertently, perhaps from our inability to formulate the model as precisely as possible because the underlying theory is weak or because we do not have the right kind of data to test the model. As Davidson notes, "Because of the non-experimental nature of economics, we are never sure how the observed data were generated. The test of any hypothesis in economics always turns out to depend on additional assumptions necessary to specify a reasonably parsimonious model, which may or may not be justified."[14]

[13]Michael D. Intriligator, *Econometric Models, Techniques and Applications,* Prentice Hall, Englewood Cliffs, NJ, 1978, p. 189. Recall the Occam's razor principle.

[14]James Davidson, *Econometric Theory,* Blackwell Publishers, Oxford, U.K., 2000, p. 153.