13

ECONOMETRIC MODELING: MODEL SPECIFICATION AND DIAGNOSTIC TESTING

> Applied econometrics cannot be done mechanically; it needs understanding, intuition and skill.[1]
>
> . . . we generally drive across bridges without worrying about the soundness of their construction because we are reasonably sure that someone rigorously checked their engineering principles and practice. Economists must do likewise with models or else attach the warning 'not responsible if attempted use leads to collapse'.[2]
>
> Economists' search for "truth" has over the years given rise to the view that economists are people searching in a dark room for a non-existent black cat; econometricians are regularly accused of finding one.[3]

One of the assumptions of the classical linear regression model (CLRM), Assumption 9, is that the regression model used in the analysis is "correctly" specified: If the model is not "correctly" specified, we encounter the problem of **model specification error** or **model specification bias.** In this chapter we take a close and critical look at this assumption, because searching for the correct model is like searching for the Holy Grail. In particular we examine the following questions:

1. How does one go about finding the "correct" model? In other words, what are the criteria in choosing a model for empirical analysis?

[1]Keith Cuthbertson, Stephen G. Hall, and Mark P. Taylor, *Applied Econometrics Techniques,* Michigan University Press, 1992, p. X.
[2]David F. Hendry, *Dynamic Econometrics,* Oxford University Press, U.K., 1995, p. 68.
[3]Peter Kennedy, *A Guide to Econometrics,* 3d ed., The MIT Press, Cambridge, Mass., 1992, p. 82.

2. What types of model specification errors is one likely to encounter in practice?

3. What are the consequences of specification errors?

4. How does one detect specification errors? In other words, what are some of the diagnostic tools that one can use?

5. Having detected specification errors, what remedies can one adopt and with what benefits?

6. How does one evaluate the performance of competing models?

The topic of model specification and evaluation is vast, and very extensive empirical work has been done in this area. Not only that, but there are philosophical differences on this topic. Although we cannot do full justice to this topic in one chapter, we hope to bring out some of the essential issues involved in model specification and model evaluation.

13.1 MODEL SELECTION CRITERIA

According to Hendry and Richard, a model chosen for empirical analysis should satisfy the following criteria[4]:

1. *Be data admissible;* that is, predictions made from the model must be logically possible.

2. *Be consistent with theory;* that is, it must make good economic sense. For example, if Milton Friedman's **permanent income hypothesis** holds, the intercept value in the regression of permanent consumption on permanent income is expected to be zero.

3. *Have weakly exogenous regressors;* that is, the explanatory variables, or regressors, must be uncorrelated with the error term.

4. *Exhibit parameter constancy;* that is, the values of the parameters should be stable. Otherwise, forecasting will be difficulty. As Friedman notes, "The only relevant test of the validity of a hypothesis [model] is comparison of its predictions with experience."[5] In the absence of parameter constancy, such predictions will not be reliable.

5. *Exhibit data coherency;* that is, the residuals estimated from the model must be purely random (technically, white noise). In other words, if the regression model is adequate, the residuals from this model must be white noise. If that is not the case, there is some specification error in the model. Shortly, we will explore the nature of specification error(s).

6. *Be encompassing;* that is, the model should *encompass* or include all the rival models in the sense that it is capable of explaining their results. In short, other models cannot be an improvement over the chosen model.

---

[4]D. F. Hendry and J. F. Richard, "The Econometric Analysis of Economic Time Series," *International Statistical Review,* vol. 51, 1983, pp. 3–33.

[5]Milton Friedman, "The Methodology of Positive Economics," in *Essays in Positive Economics,* University of Chicago Press, Chicago, 1953, p. 7.

It is one thing to list criteria of a "good" model and quite another to actually develop it, for in practice one is likely to commit various model specification errors, which we discuss in the next section.

13.2 TYPES OF SPECIFICATION ERRORS

Assume that on the basis of the criteria just listed we arrive at a model that we accept as a good model. To be concrete, let this model be

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_{1i} \tag{13.2.1}$$

where $Y =$ total cost of production and $X =$ output. Equation (13.2.1) is the familiar textbook example of the cubic total cost function.

But suppose for some reason (say, laziness in plotting the scattergram) a researcher decides to use the following model:

$$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + u_{2i} \tag{13.2.2}$$

Note that we have changed the notation to distinguish this model from the true model.

Since (13.2.1) is assumed true, adopting (13.2.2) would constitute a specification error, the error consisting in **omitting a relevant variable** (X_i^3). Therefore, the error term u_{2i} in (13.2.2) is in fact

$$u_{2i} = u_{1i} + \beta_4 X_i^3 \tag{13.2.3}$$

We shall see shortly the importance of this relationship.

Now suppose that another researcher uses the following model:

$$Y_i = \lambda_1 + \lambda_2 X_i + \lambda_3 X_i^2 + \lambda_4 X_i^3 + \lambda_5 X_i^4 + u_{3i} \tag{13.2.4}$$

If (13.2.1) is the "truth," (13.2.4) also constitutes a specification error, the error here consisting in **including an unnecessary or irrelevant variable** in the sense that the true model assumes λ_5 to be zero. The new error term is in fact

$$\begin{aligned} u_{3i} &= u_{1i} - \lambda_5 X_i^4 \\ &= u_{1i} \qquad \text{since } \lambda_5 = 0 \text{ in the true model} \qquad \text{(Why?)} \end{aligned} \tag{13.2.5}$$

Now assume that yet another researcher postulates the following model:

$$\ln Y_i = \gamma_1 + \gamma_2 X_i + \gamma_3 X_i^2 + \gamma_4 X_i^3 + u_{4i} \tag{13.2.6}$$

In relation to the true model, (13.2.6) would also constitute a specification bias, the bias here being the use of the **wrong functional form:** In (13.2.1) Y appears linearly, whereas in (13.2.6) it appears log-linearly.

Finally, consider the researcher who uses the following model:

$$Y_i^* = \beta_1^* + \beta_2^* X_i^* + \beta_3^* X_i^{*2} + \beta_4^* X_i^{*3} + u_i^* \tag{13.2.7}$$

where $Y_i^* = Y_i + \varepsilon_i$ and $X_i^* = X_i + w_i$, ε_i and w_i being the errors of measurement. What (13.2.7) states is that instead of using the true Y_i and X_i we use their proxies, Y_i^* and X_i^*, which may contain errors of measurement. Therefore, in (13.2.7) we commit the **errors of measurement bias.** In applied work data are plagued by errors of approximations or errors of incomplete coverage or simply errors of omitting some observations. In the social sciences we often depend on secondary data and usually have no way of knowing the types of errors, if any, made by the primary data-collecting agency.

Another type of specification error relates to the way the stochastic error u_i (or u_t) enters the regression model. Consider for instance, the following bivariate regression model without the intercept term:

$$Y_i = \beta X_i u_i \tag{13.2.8}$$

where the stochastic error term enters multiplicatively with the property that $\ln u_i$ satisfies the assumptions of the CLRM, against the following model

$$Y_i = \alpha X_i + u_i \tag{13.2.9}$$

where the error term enters additively. Although the variables are the same in the two models, we have denoted the slope coefficient in (13.2.8) by β and the slope coefficient in (13.2.9) by α. Now if (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.

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. Adopting the wrong functional form
4. Errors of measurement
5. Incorrect specification of the stochastic error term

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, there 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.

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,* $\hat{\alpha}_1$ 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 $\hat{\alpha}_2$ $(= \sigma^2 / \sum x_{2i}^2)$ 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]But see exercise 13.32.

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

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

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 (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 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.

ILLUSTRATIVE EXAMPLE: CHILD MORTALITY REVISITED

Regressing child mortality (CM) on per capita GNP (PGNP) and 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 (7.6.2) as the correct model, then (7.7.2) is a misspecified 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 (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).

[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.

[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.

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, $\text{var}(\hat{\alpha}_2)$ will be different from $\text{var}(\hat{\beta}_2)$. But we know that $\text{var}(\hat{\beta}_2)$ is unbiased (why?). Therefore, $\text{var}(\hat{\alpha}_2)$ 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 $\text{var}(\hat{\alpha}_2) < \text{var}(\hat{\beta}_2)$. 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 tradeoff 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 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 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 precision (i.e., reduce 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 tradeoff in this situation between bias and precision can be substantial. As you can see from this discussion, the tradeoff will depend on the relative importance of the various regressors.

[10]To bypass the tradeoff 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, **App. A**. See also exercise 13.6.

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 (13.3.3) that $\hat{\alpha}_2$ is now unbiased.[11] Also, it seems from (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 (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.

From the usual OLS formula we know that

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

[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.

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 addition of unnecessary variables will lead to loss in 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

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

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]

The practical question then is not why specification errors are made, for they generally are, but how to detect them. Once it is found that specification errors have been made, the remedies often suggest themselves. If, for example, it can be shown that a variable is inappropriately omitted from a model, the obvious remedy is to include that variable in the analysis, assuming, of course, the data on that variable are available.

In this section we discuss some tests that one may use to detect specification errors.

Detecting the Presence of Unnecessary Variables (Overfitting a Model)

Suppose we develop a k-variable model to explain a phenomenon:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \qquad \textbf{(13.4.1)}$$

However, we are not totally sure that, say, the variable X_k really belongs in the model. One simple way to find this out is to test the significance of the estimated β_k with the usual t test: $t = \hat{\beta}_k / \text{se}\,(\hat{\beta}_k)$. But suppose that we are not sure whether, say, X_3 and X_4 legitimately belong in the model. This can be easily ascertained by the F test discussed in Chapter 8. Thus, detecting the presence of an irrelevant variable (or variables) is not a difficult task.

It is, however, very important to remember that in carrying out these tests of significance we have a specific model in mind. We accept that model as the **maintained hypothesis** or the "truth," however tentative it may be. Given that model, then, we can find out whether one or more regressors are really relevant by the usual t and F tests. But note carefully that we should not use the t and F tests to build a model *iteratively,* that is, we should not say that initially Y is related to X_2 only because $\hat{\beta}_2$ is statistically significant and then expand the model to include X_3 and decide to keep that variable in the model if $\hat{\beta}_3$ turns out to be statistically significant, and so on. This strategy of building a model is called the **bottom-up approach** (starting with a smaller model and expanding it as one goes along) or by the somewhat pejorative term, **data mining** (other names are **regression fishing, data grubbing, data snooping,** and **number crunching**).

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