# 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 $\lambda_5$ 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:

$$Y_i^* = \beta_1^* + \beta_2^* X_i^* + \beta_3^* X_i^{*2} + \beta_4^* X_i^{*3} + u_i^* \qquad \textbf{(13.2.7)}$$

where $Y_i^* = Y_i + \varepsilon_i$ and $X_i^* = X_i + w_i$, $\varepsilon_i$ 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:

$$Y_i = \beta X_i u_i \qquad \textbf{(13.2.8)}$$

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

$$Y_i = \alpha X_i + u_i \qquad \textbf{(13.2.9)}$$

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 $\beta$ and the slope coefficient in (13.2.9) by $\alpha$. Now if (13.2.8) is the "correct" or "true" model, would the estimated $\alpha$ provide an unbiased estimate of the true $\beta$? 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 $\sigma^2$ 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 $\beta_3$ or $b_{32}$ or both are zero. We rule out $\beta_3$ 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, $\beta_3$ 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 $\beta_2$ (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 $\sigma^2$ 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 $\sigma^2$ 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 $\sigma^2$ is correctly estimated.

**3.** The usual confidence interval and hypothesis-testing procedures remain valid.

**4.** However, the estimated $\alpha$'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 $\beta_k$ 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.

The primary objective of data mining is to develop the "best" model after several diagnostic tests so that the model finally chosen is a "good" model in the sense that all the estimated coefficients have the "right" signs, they are statistically significant on the basis of the $t$ and $F$ tests, the $R^2$ value is reasonably high and the Durbin–Watson $d$ has acceptable value (around 2), etc. The purists in the profession look down on the practice of data mining. In the words of William Pool, ". . . making an empirical regularity the foundation, rather than an implication of economic theory, is always dangerous."[15] One reason for "condemning" data mining is as follows.

**Nominal versus True Level of Significance in the Presence of Data Mining.** A danger of data mining that the unwary researcher faces is that the conventional levels of significance ($\alpha$) such as 1, 5, or 10 percent are *not the true levels of significance*. Lovell has suggested that if there are $c$ candidate regressors out of which $k$ are finally selected ($k \leq c$) on the basis of data mining, then the true level of significance ($\alpha^*$) is related to the nominal level of significance ($\alpha$) as follows:[16]

$$\alpha^* = 1 - (1 - \alpha)^{c/k} \tag{13.4.2}$$

or approximately as

$$\alpha^* \approx (c/k)\alpha \tag{13.4.3}$$

For example, if $c = 15$, $k = 5$, and $\alpha = 5$ percent, from (13.4.3) the true level of significance is $(15/5)(5) = 15$ percent. Therefore, if a researcher datamines and selects 5 out of 15 regressors and reports only the results of the condensed model at the nominal 5 percent level of significance and declares that the results are statistically significant, one should take this conclusion with a big grain of salt, for we know the (true) level of significance is in fact 15 percent. It should be noted that if $c = k$, that is, there is no data mining, the true and nominal levels of significance are the same. Of course, in practice most researchers report only the results of their "final" regression without necessarily telling about all the data mining, or **pretesting,** that has gone before.[17]

Despite some of its obvious drawbacks, there is increasing recognition, especially among applied econometricians, that the purist (i.e., non–data mining) approach to model building is not tenable. As Zaman notes:

> Unfortunately, experience with real data sets shows that such a [purist approach] is neither feasible nor desirable. It is not feasible because it is a rare economic

[15]William Pool, "Is Inflation Too Low," the *Cato Journal,* vol. 18, no. 3, Winter 1999, p. 456.
[16]M. Lovell, "Data Mining," *Review of Economics and Statistics,* vol. 65, 1983, pp. 1–12.
[17]For a detailed discussion of pretesting and the biases it can lead to, see Wallace, T. D., "Pretest Estimation in Regression: A Survey," *American Journal of Agricultural Economics,* vol. 59, 1977, pp. 431–443.

> theory which leads to a unique model. It is not desirable because a crucial aspect of learning from the data is learning what types of models are and are not supported by data. Even if, by rare luck, the initial model shows a good fit, it is frequently important to explore and learn the types of the models the data does or does not agree with.[18]

A similar view is expressed by Kerry Patterson who maintains that:

> This [data mining] approach suggests that economic theory and empirical specification interact rather than be kept in separate compartments.[19]

Instead of getting caught in the data mining versus the purist approach to model-building controversy, one can endorse the view expressed by Peter Kennedy:

> [that model specification] needs to be a well-thought-out combination of theory and data, and that testing procedures used in specification searches should be designed to minimize the costs of data mining. Examples of such procedures are setting aside data for out-of-sample prediction tests, adjusting significance levels [a la Lovell], and avoiding questionable criteria such as maximizing $R^2$.[20]

If we look at data mining in a broader perspective as a process of discovering empirical regularities that might suggest errors and/or omissions in (existing) theoretical models, it has a very useful role to play. To quote Kennedy again, "The art of the applied econometrician is to allow for data-driven theory while avoiding the considerable dangers in data mining."[21]

### Tests for Omitted Variables and Incorrect Functional Form

In practice we are never sure that the model adopted for empirical testing is "the truth, the whole truth and nothing but the truth." On the basis of theory or introspection and prior empirical work, we develop a model that we believe captures the essence of the subject under study. We then subject the model to empirical testing. After we obtain the results, we begin the postmortem, keeping in mind the criteria of a good model discussed earlier. It is at this stage that we come to know if the chosen model is adequate. In determining model adequacy, we look at some broad features of the results, such as the $\bar{R}^2$ value, the estimated $t$ ratios, the signs of the estimated coefficients in relation to their prior expectations, the Durbin–Watson statistic, and the like. If these diagnostics are reasonably good, we proclaim that the

[18]Asad Zaman, *Statistical Foundations for Econometric Techniques,* Academic Press, New York, 1996, p. 226.
[19]Kerry Patterson, *An Introduction to Applied Econometrics,* St. Martin's Press, New York, 2000, p. 10.
[20]Peter Kennedy, "Sinning in the Basement: What Are the Rules? The Ten Commandments of Applied Econometrics," unpublished manuscript.
[21]Kennedy, op. cit., p. 13.

chosen model is a fair representation of reality. By the same token, if the results do not look encouraging because the $\bar{R}^2$ value is too low or because very few coefficients are statistically significant or have the correct signs or because the Durbin–Watson $d$ is too low, then we begin to worry about model adequacy and look for remedies: Maybe we have omitted an important variable, or have used the wrong functional form, or have not first-differenced the time series (to remove serial correlation), and so on. To aid us in determining whether model inadequacy is on account of one or more of these problems, we can use some of the following methods.

**Examination of Residuals.** As noted in Chapter 12, examination of the residuals is a good visual diagnostic to detect autocorrelation or heteroscedasticity. But these residuals can also be examined, especially in cross-sectional data, for model specification errors, such as omission of an important variable or incorrect functional form. If in fact there are such errors, a plot of the residuals will exhibit distinct patterns.

To illustrate, let us reconsider the cubic total cost of production function first considered in Chapter 7. Assume that the true total cost function is described as follows, where $Y$ = total cost and $X$ = output:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i \tag{13.4.4}$$

but a researcher fits the following quadratic function:

$$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + u_{2i} \tag{13.4.5}$$

and another researcher fits the following linear function:

$$Y_i = \lambda_1 + \lambda_2 X_i + u_{3i} \tag{13.4.6}$$

Although we know that both researchers have made specification errors, for pedagogical purposes let us see how the estimated residuals look in the three models. (The cost-output data are given in Table 7.4.) Figure 13.1 speaks for itself: As we move from left to right, that is, as we approach the truth, not only are the residuals smaller (in absolute value) but also they do not exhibit the pronounced cyclical swings associated with the misfitted models.

The utility of examining the residual plot is thus clear: If there are specification errors, the residuals will exhibit noticeable patterns.

**The Durbin–Watson *d* Statistic Once Again.** If we examine the routinely calculated Durbin–Watson $d$ in Table 13.1, we see that for the linear cost function the estimated $d$ is 0.716, suggesting that there is positive "correlation" in the estimated residuals: for $n = 10$ and $k' = 1$, the 5 percent

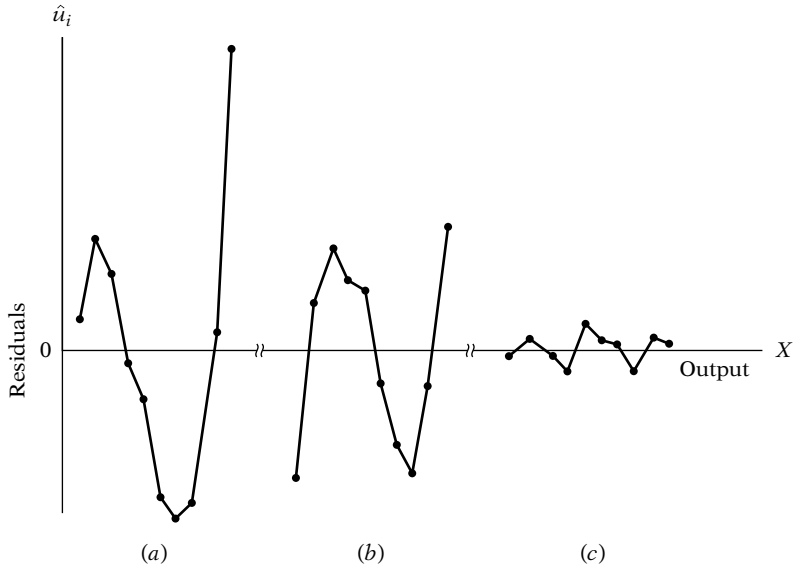


**FIGURE 13.1** Residuals $\hat{u}_i$ from (*a*) linear, (*b*) quadratic, and (*c*) cubic total cost functions.

**TABLE 13.1** ESTIMATED RESIDUALS FROM THE LINEAR, QUADRATIC, AND CUBIC TOTAL COST FUNCTIONS

| Observation number | $\hat{u}_i$, linear model* | $\hat{u}_i$, quadratic model† | $\hat{u}_i$, cubic model** |
|---|---|---|---|
| 1 | 6.600 | −23.900 | −0.222 |
| 2 | 19.667 | 9.500 | 1.607 |
| 3 | 13.733 | 18.817 | −0.915 |
| 4 | −2.200 | 13.050 | −4.426 |
| 5 | −9.133 | 11.200 | 4.435 |
| 6 | −26.067 | −5.733 | 1.032 |
| 7 | −32.000 | −16.750 | 0.726 |
| 8 | −28.933 | −23.850 | −4.119 |
| 9 | 4.133 | −6.033 | 1.859 |
| 10 | 54.200 | 23.700 | 0.022 |

* $\hat{Y}_i = 166.467 + 19.933X_i$ ; (19.021) (3.066); (8.752) (6.502); $R^2 = 0.8409$, $\bar{R}^2 = 0.8210$, $d = 0.716$

† $\hat{Y}_i = 222.383 - 8.0250X_i + 2.542X_i^2$ ; (23.488) (9.809) (0.869); (9.468) (−0.818) (2.925); $R^2 = 0.9284$, $\bar{R}^2 = 0.9079$, $d = 1.038$

** $\hat{Y}_i = 141.767 + 63.478X_i - 12.962X_i^2 + 0.939X_i^3$ ; (6.375) (4.778) (0.9856) (0.0592); (22.238) (13.285) (−13.151) (15.861); $R^2 = 0.9983$, $\bar{R}^2 = 0.9975$, $d = 2.70$

critical $d$ values are $d_L = 0.879$ and $d_U = 1.320$. Likewise, the computed $d$ value for the quadratic cost function is 1.038, whereas the 5 percent critical values are $d_L = 0.697$ and $d_U = 1.641$, indicating indecision. But if we use the modified $d$ test (see Chapter 12), we can say that there is positive "correlation" in the residuals, for the computed $d$ is less than $d_U$. For the cubic cost function, the true specification, the estimated $d$ value does not indicate any positive "correlation" in the residuals.[22]

The observed positive "correlation" in the residuals when we fit the linear or quadratic model is not a measure of (first-order) serial correlation but of (model) specification error(s). The observed correlation simply reflects the fact that some variable(s) that belong in the model are included in the error term and need to be culled out from it and introduced in their own right as explanatory variables: If we exclude the $X_i^3$ from the cost function, then as (13.2.3) shows, the error term in the mis-specified model (13.2.2) is in fact $(u_{1i} + \beta_4 X_i^3)$ and it will exhibit a systematic pattern (e.g., positive autocorrelation) if $X_i^3$ in fact affects $Y$ significantly.

To use the Durbin–Watson test for detecting model specification error(s), we proceed as follows:

**1.** From the assumed model, obtain the OLS residuals.

**2.** If it is believed that the assumed model is mis-specified because it excludes a relevant explanatory variable, say, $Z$ from the model, order the residuals obtained in Step 1 according to increasing values of $Z$. *Note:* The $Z$ variable could be one of the $X$ variables included in the assumed model or it could be some function of that variable, such as $X^2$ or $X^3$.

**3.** Compute the $d$ statistic from the residuals thus ordered by the usual $d$ formula, namely,

$$d = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n}\hat{u}_t^2}$$

*Note:* The subscript $t$ is the index of observation here and does not necessarily mean that the data are time series.

**4.** From the Durbin–Watson tables, if the estimated $d$ value is significant, then one can accept the hypothesis of model mis-specification. If that turns out to be the case, the remedial measures will naturally suggest themselves.

In our cost example, the $Z\,(= X)$ variable (output) was already ordered.[23] Therefore, we do not have to compute the $d$ statistic afresh. As we have seen, the $d$ statistic for both the linear and quadratic cost functions suggests

[22] In the present context, a value of $d = 2$ will mean no specification error. (Why?)

[23] It does not matter if we order $\hat{u}_i$ according to $X_i^2$ or $X_i^3$ since these are functions of $X_i$, which is already ordered.

specification errors. The remedies are clear: Introduce the quadratic and cubic terms in the linear cost function and the cubic term in the quadratic cost function. In short, run the cubic cost model.

**Ramsey's RESET Test.** Ramsey has proposed a general test of specification error called RESET (regression specification error test).[24] Here we will illustrate only the simplest version of the test. To fix ideas, let us continue with our cost-output example and assume that the cost function is linear in output as

$$Y_i = \lambda_1 + \lambda_2 X_i + u_{3i} \qquad \textbf{(13.4.6)}$$

where $Y$ = total cost and $X$ = output. Now if we plot the residuals $\hat{u}_i$ obtained from this regression against $\hat{Y}_i$, the estimated $Y_i$ from this model, we get the picture shown in Figure 13.2. Although $\sum \hat{u}_i$ and $\sum \hat{u}_i \hat{Y}_i$ are necessarily zero

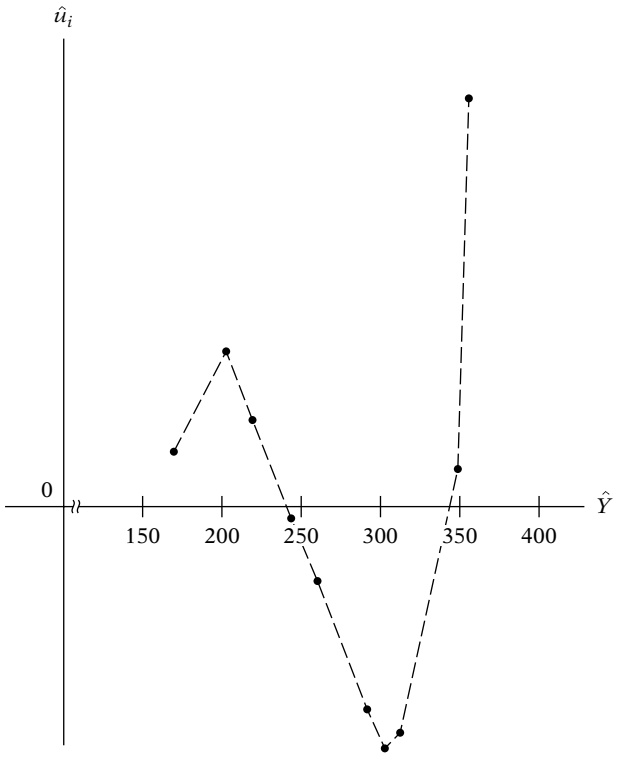


**FIGURE 13.2** Residuals $\hat{u}_i$ and estimated $Y$ from the linear cost function: $Y_i = \lambda_1 + \lambda_2 X_i + u_i$.

[24] J. B. Ramsey, "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis," *Journal of the Royal Statistical Society,* series B, vol. 31, 1969, pp. 350–371.

(why? see Chapter 3), the residuals in this figure show a pattern in which their mean changes systematically with $\hat{Y}_i$. This would suggest that if we introduce $\hat{Y}_i$ in some form as regressor(s) in (13.4.6), it should increase $R^2$. And if the increase in $R^2$ is statistically significant (on the basis of the $F$ test discussed in Chapter 8), it would suggest that the linear cost function (13.4.6) was mis-specified. This is essentially the idea behind RESET. **The steps involved in RESET are as follows:**

**1.** From the chosen model, e.g., (13.4.6), obtain the estimated $Y_i$, that is, $\hat{Y}_i$.

**2.** Rerun (13.4.6) introducing $\hat{Y}_i$ in some form as an additional regressor(s). From Figure 13.2, we observe that there is a curvilinear relationship between $\hat{u}_i$ and $\hat{Y}_i$, suggesting that one can introduce $\hat{Y}_i^2$ and $\hat{Y}_i^3$ as additional regressors. Thus, we run

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 \hat{Y}_i^2 + \beta_4 \hat{Y}_i^3 + u_i \qquad \textbf{(13.4.7)}$$

**3.** Let the $R^2$ obtained from (13.4.7) be $R^2_{\text{new}}$ and that obtained from (13.4.6) be $R^2_{\text{old}}$. Then we can use the $F$ test first introduced in (8.5.18), namely,

$$F = \frac{\left(R^2_{\text{new}} - R^2_{\text{old}}\right)/\text{number of new regressors}}{\left(1 - R^2_{\text{new}}\right)/(n - \text{number of parameters in the new model})} \qquad \textbf{(8.5.18)}$$

to find out if the increase in $R^2$ from using (13.4.7) is statistically significant.

**4.** If the computed $F$ value is significant, say, at the 5 percent level, one can accept the hypothesis that the model (13.4.6) is mis-specified.

Returning to our illustrative example, we have the following results (standard errors in parentheses):

$$\begin{aligned} \hat{Y}_i &= 166.467 \ + 19.933 X_i \\ &\quad\ (19.021) \quad (3.066) \qquad R^2 = 0.8409 \end{aligned} \qquad \textbf{(13.4.8)}$$

$$\begin{aligned} \hat{Y}_i = 2140.7223 \ &+ 476.6557 X_i - \ 0.09187 \hat{Y}_i^2 + \ 0.000119 \hat{Y}_i^3 \\ (132.0044) \quad &(33.3951) \quad (0.00620) \quad (0.0000074) \\ &R^2 = 0.9983 \end{aligned} \qquad \textbf{(13.4.9)}$$

*Note:* $\hat{Y}_i^2$ and $\hat{Y}_i^3$ in (13.4.9) are obtained from (13.4.8).

Now applying the $F$ test we find

$$\begin{aligned} F &= \frac{(0.9983 - 0.8409)/2}{(1 - 0.9983)/(10 - 4)} \\ &= 284.4035 \end{aligned} \qquad \textbf{(13.4.10)}$$

The reader can easily verify that this $F$ value is highly significant, indicating that the model (13.4.8) is mis-specified. Of course, we have reached the same conclusion on the basis of the visual examination of the residuals as well as the Durbin–Watson $d$ value.

One advantage of RESET is that it is easy to apply, for it does not require one to specify what the alternative model is. But that is also its disadvantage because knowing that a model is mis-specified does not help us necessarily in choosing a better alternative.

**Lagrange Multiplier (LM) Test for Adding Variables.** This is an alternative to Ramsey's RESET test. To illustrate this test, we will continue with the preceding illustrative example.

If we compare the linear cost function (13.4.6) with the cubic cost function (13.4.4), the former is a *restricted version* of the latter (recall our discussion of **restricted least-squares** from Chapter 8). The restricted regression (13.4.6) assumes that the coefficients of the squared and cubed output terms are equal to zero. To test this, the LM test proceeds as follows:

**1.** Estimate the restricted regression (13.4.6) by OLS and obtain the residuals, $\hat{u}_i$.

**2.** If in fact the unrestricted regression (13.4.4) is the true regression, the residuals obtained in (13.4.6) should be related to the squared and cubed output terms, that is, $X_i^2$ and $X_i^3$.

**3.** This suggests that we regress the $\hat{u}_i$ obtained in Step 1 on all the regressors (including those in the restricted regression), which in the present case means

$$\hat{u}_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + \alpha_4 X_i^3 + v_i \qquad \textbf{(13.4.11)}$$

where $v$ is an error term with the usual properties.

**4.** For large-sample size, Engle has shown that $n$ (the sample size) times the $R^2$ estimated from the (auxiliary) regression (13.4.11) follows the chi-square distribution with df equal to the number of restrictions imposed by the restricted regression, two in the present example since the terms $X_i^2$ and $X_i^3$ are dropped from the model.[25] Symbolically, we write

$$nR^2 \underset{\text{asy}}{\sim} \chi^2_{(\text{number of restrictions})} \qquad \textbf{(13.4.12)}$$

where asy means asymptotically, that is, in large samples.

**5.** If the chi-square value obtained from (13.4.12) exceeds the critical chi-square value at the chosen level of significance, we reject the restricted regression. Otherwise, we do not reject it.

[25] R. F. Engle, "A General Approach to Lagrangian Multiplier Model Diagnostics," *Journal of Econometrics,* vol. 20, 1982, pp. 83–104.

For our example, the regression results are as follows:

$$\hat{Y}_i = 166.467 + 19.333X_i \qquad \textbf{(13.4.13)}$$

where $Y$ is total cost and $X$ is output. The standard errors for this regression are already given in Table 13.1.

When the residuals from (13.4.13) are regressed as just suggested in Step 3, we obtain the following results:

$$\begin{aligned} \widehat{u_i} &= -24.7 \quad + 43.5443X_i - 12.9615X_i^2 + 0.9396X_i^3 \\ \text{se} &= \quad (6.375) \quad (4.779) \qquad (0.986) \qquad (0.059) \qquad \textbf{(13.4.14)} \\ & \qquad\qquad\qquad\qquad\qquad\qquad R^2 = 0.9896 \end{aligned}$$

Although our sample size of 10 is by no means large, just to illustrate the LM mechanism, we obtain $nR^2 = (10)(0.9896) = 9.896$. From the chi-square table we observe that for 2 df the 1 percent critical chi-square value is about 9.21. Therefore, the observed value of 9.896 is significant at the 1 percent level, and our conclusion would be to reject the restricted regression (i.e., the linear cost function). We reached the similar conclusion on the basis of Ramsey's RESET test.

## 13.5 ERRORS OF MEASUREMENT

All along we have assumed implicitly that the dependent variable $Y$ and the explanatory variables, the $X$'s, are measured without any errors. Thus, in the regression of consumption expenditure on income and wealth of households, we assume that the data on these variables are "accurate"; they are not *guess estimates,* extrapolated, interpolated, or rounded off in any systematic manner, such as to the nearest hundredth dollar, and so on. Unfortunately, this ideal is not met in practice for a variety of reasons, such as nonresponse errors, reporting errors, and computing errors. Whatever the reasons, error of measurement is a potentially troublesome problem, for it constitutes yet another example of specification bias with the consequences noted below.

### Errors of Measurement in the Dependent Variable *Y*

Consider the following model:

$$Y_i^* = \alpha + \beta X_i + u_i \qquad \textbf{(13.5.1)}$$

where $Y_i^*$ = permanent consumption expenditure[26]
$X_i$ = current income
$u_i$ = stochastic disturbance term

[26]This phrase is due to Milton Friedman. See also exercise 13.8.

Since $Y_i^*$ is not directly measurable, we may use an observable expenditure variable $Y_i$ such that

$$Y_i = Y_i^* + \varepsilon_i \tag{13.5.2}$$

where $\varepsilon_i$ denote errors of measurement in $Y_i^*$. Therefore, instead of estimating (13.5.1), we estimate

$$\begin{aligned} Y_i &= (\alpha + \beta X_i + u_i) + \varepsilon_i \\ &= \alpha + \beta X_i + (u_i + \varepsilon_i) \\ &= \alpha + \beta X_i + v_i \end{aligned} \tag{13.5.3}$$

where $v_i = u_i + \varepsilon_i$ is a composite error term, containing the population disturbance term (which may be called the *equation error term*) and the measurement error term.

For simplicity assume that $E(u_i) = E(\varepsilon_i) = 0$, $\text{cov}\,(X_i, u_i) = 0$ (which is the assumption of the classical linear regression), and $\text{cov}\,(X_i, \varepsilon_i) = 0$; that is, the errors of measurement in $Y_i^*$ are uncorrelated with $X_i$, and $\text{cov}\,(u_i, \varepsilon_i) = 0$; that is, the equation error and the measurement error are uncorrelated. With these assumptions, it can be seen that $\beta$ estimated from either (13.5.1) or (13.5.3) will be an unbiased estimator of the true $\beta$ (see exercise 13.7); that is, the errors of measurement in the dependent variable $Y$ do not destroy the unbiasedness property of the OLS estimators. However, the variances and standard errors of $\beta$ estimated from (13.5.1) and (13.5.3) will be different because, employing the usual formulas (see Chapter 3), we obtain

$$\text{Model (13.5.1):} \qquad \text{var}\,(\hat{\beta}) = \frac{\sigma_u^2}{\sum x_i^2} \tag{13.5.4}$$

$$\begin{aligned} \text{Model (13.5.3):} \qquad \text{var}\,(\hat{\beta}) &= \frac{\sigma_v^2}{\sum x_i^2} \\ &= \frac{\sigma_u^2 + \sigma_\varepsilon^2}{\sum x_i^2} \end{aligned} \tag{13.5.5}$$

Obviously, the latter variance is larger than the former.[27] Therefore, **although the errors of measurement in the dependent variable still give unbiased estimates of the parameters and their variances, the estimated variances are now larger than in the case where there are no such errors of measurement.**

[27]But note that this variance is still unbiased because under the stated conditions the composite error term $v_i = u_i + \varepsilon_i$ still satisfies the assumptions underlying the method of least squares.

### Errors of Measurement in the Explanatory Variable *X*

Now assume that instead of (13.5.1), we have the following model:

$$Y_i = \alpha + \beta X_i^* + u_i \tag{13.5.6}$$

where $Y_i$ = current consumption expenditure
$X_i^*$ = permanent income
$u_i$ = disturbance term (equation error)

Suppose instead of observing $X_i^*$, we observe

$$X_i = X_i^* + w_i \tag{13.5.7}$$

where $w_i$ represents errors of measurement in $X_i^*$. Therefore, instead of estimating (13.5.6), we estimate

$$\begin{aligned} Y_i &= \alpha + \beta(X_i - w_i) + u_i \\ &= \alpha + \beta X_i + (u_i - \beta w_i) \\ &= \alpha + \beta X_i + z_i \end{aligned} \tag{13.5.8}$$

where $z_i = u_i - \beta w_i$, a compound of equation and measurement errors.

Now even if we assume that $w_i$ has zero mean, is serially independent, and is uncorrelated with $u_i$, we can no longer assume that the composite error term $z_i$ is independent of the explanatory variable $X_i$ because [assuming $E(z_i) = 0$]

$$\begin{aligned} \operatorname{cov}(z_i, X_i) &= E[z_i - E(z_i)][X_i - E(X_i)] \\ &= E(u_i - \beta w_i)(w_i) \qquad \text{using (13.5.7)} \\ &= E\left(-\beta w_i^2\right) \\ &= -\beta\sigma_w^2 \end{aligned} \tag{13.5.9}$$

Thus, the explanatory variable and the error term in (13.5.8) are correlated, which violates the crucial assumption of the classical linear regression model that the explanatory variable is uncorrelated with the stochastic disturbance term. If this assumption is violated, it can be shown that the *OLS estimators are not only biased but also inconsistent, that is, they remain biased even if the sample size n increases indefinitely.*[28]

[28] As shown in **App. A**, $\hat{\beta}$ is a consistent estimator of $\beta$ if, as $n$ increases indefinitely, the sampling distribution of $\hat{\beta}$ will ultimately collapse to the true $\beta$. Technically, this is stated as $\text{plim}_{n\to\infty}\hat{\beta} = \beta$. As noted in **App. A**, consistency is a large-sample property and is often used to study the behavior of an estimator when its finite or small-sample properties (e.g., unbiasedness) cannot be determined.

For model (13.5.8), it is shown in Appendix 13A, Section 13A.3 that

$$\text{plim}\,\hat{\beta} = \beta\left[\frac{1}{1+\sigma_w^2/\sigma_{X^*}^2}\right] \tag{13.5.10}$$

where $\sigma_w^2$ and $\sigma_{X^*}^2$ are variances of $w_i$ and $X^*$, respectively, and where plim $\hat{\beta}$ means the probability limit of $\beta$.

Since the term inside the brackets is expected to be less than 1 (why?), (13.5.10) shows that even if the sample size increases indefinitely, $\hat{\beta}$ will not converge to $\beta$. Actually, if $\beta$ is assumed positive, $\hat{\beta}$ will underestimate $\beta$, that is, it is biased toward zero. Of course, if there are no measurement errors in $X$ (i.e., $\sigma_w^2 = 0$), $\hat{\beta}$ will provide a consistent estimator of $\beta$.

Therefore, measurement errors pose a serious problem when they are present in the explanatory variable(s) because they make consistent estimation of the parameters impossible. Of course, as we saw, if they are present only in the dependent variable, the estimators remain unbiased and hence they are consistent too. If errors of measurement are present in the explanatory variable(s), what is the solution? The answer is not easy. At one extreme, we can assume that if $\sigma_w^2$ is small compared to $\sigma_{X^*}^2$, for all practical purposes we can "assume away" the problem and proceed with the usual OLS estimation. Of course, the rub here is that we cannot readily observe or measure $\sigma_w^2$ and $\sigma_{X^*}^2$ and therefore there is no way to judge their relative magnitudes.

One other suggested remedy is the use of **instrumental** or **proxy variables** that, although highly correlated with the original $X$ variables, are uncorrelated with the equation and measurement error terms (i.e., $u_i$ and $w_i$). If such proxy variables can be found, then one can obtain a consistent estimate of $\beta$. But this task is much easier said than done. In practice it is not easy to find good proxies; we are often in the situation of complaining about the bad weather without being able to do much about it. Besides, it is not easy to find out if the selected instrumental variable is in fact independent of the error terms $u_i$ and $w_i$.

In the literature there are other suggestions to solve the problem.[29] But most of them are specific to the given situation and are based on restrictive assumptions. There is really no satisfactory answer to the measurement errors problem. That is why it is so crucial to measure the data as accurately as possible.

[29]See Thomas B. Fomby, R. Carter Hill, and Stanley R. Johnson, *Advanced Econometric Methods,* Springer-Verlag, New York, 1984, pp. 273–277. See also Kennedy, op. cit., pp. 138–140, for a discussion of weighted regression as well as instrumental variables.

AN EXAMPLE

We conclude this section with an example constructed to highlight the preceding points.

Table 13.2 gives hypothetical data on true consumption expenditure $Y^*$, true income $X^*$, measured consumption $Y$, and measured income $X$. The table also explains how these variables were measured.[30]

### Measurement Errors in the Dependent Variable *Y* Only

Based on the given data, the true consumption function is

$$\begin{aligned} \hat{Y}_i^* &= 25.00 + 0.6000X_i^* \\ &\quad (10.477) \quad (0.0584) \\ t &= (2.3861) \quad (10.276) \\ R^2 &= 0.9296 \end{aligned} \tag{13.5.11}$$

whereas, if we use $Y_i$ instead of $Y_i^*$, we obtain

$$\begin{aligned} \hat{Y}_i &= 25.00 + 0.6000X_i^* \\ &\quad (12.218) \quad (0.0681) \\ t &= (2.0461) \quad (8.8118) \\ R^2 &= 0.9066 \end{aligned} \tag{13.5.12}$$

As these results show, and according to the theory, the estimated coefficients remain the same. The only effect of errors of measurement in the dependent variable is that the estimated standard errors of the coefficients tend to be larger [see (13.5.5)], which is clearly seen in (13.5.12). In passing, note that the regression coefficients in (13.5.11) and (13.5.12) are the same because the sample was generated to match the assumptions of the measurement error model.

### Errors of Measurement in *X*

We know that the true regression is (13.5.11). Suppose now that instead of using $X_i^*$, we use $X_i$. (*Note:* In reality $X_i^*$ is rarely observable.) The regression results are as follows:

$$\begin{aligned} \hat{Y}_i^* &= 25.992 + 0.5942X_i \\ &\quad (11.0810) \quad (0.0617) \\ t &= (2.3457) \quad (9.6270) \\ R^2 &= 0.9205 \end{aligned} \tag{13.5.13}$$

These results are in accord with the theory—when there are measurement errors in the explanatory variable(s), the estimated coefficients are biased. Fortunately, in this example the bias is rather small—from (13.5.10) it is evident that the bias depends on $\sigma_w^2/\sigma_{X^*}^2$, and in generating the data it was assumed that $\sigma_w^2 = 36$ and $\sigma_{X^*}^2 = 3667$, thus making the bias factor rather small, about 0.98 percent ($= 36/3667$).

We leave it to the reader to find out what happens when there are errors of measurement in both $Y$ and $X$, that is, if we regress $Y_i$ on $X_i$ rather than $Y_i^*$ on $X_i^*$ (see exercise 13.23).

**TABLE 13.2**
HYPOTHETICAL DATA ON $Y^*$ (TRUE CONSUMPTION EXPENDITURE), $X^*$ (TRUE INCOME), $Y$ (MEASURED CONSUMPTION EXPENDITURE), AND $X$ (MEASURED INCOME); ALL DATA IN DOLLARS

| $Y^*$ | $X^*$ | $Y$ | $X$ | $\varepsilon$ | $w$ | $u$ |
|---|---|---|---|---|---|---|
| 75.4666 | 80.00 | 67.6011 | 80.0940 | −7.8655 | 0.0940 | 2.4666 |
| 74.9801 | 100.00 | 75.4438 | 91.5721 | 0.4636 | −8.4279 | −10.0199 |
| 102.8242 | 120.00 | 109.6956 | 112.1406 | 6.8714 | 2.1406 | 5.8242 |
| 125.7651 | 140.00 | 129.4159 | 145.5969 | 3.6509 | 5.5969 | 16.7651 |
| 106.5035 | 160.00 | 104.2388 | 168.5579 | −2.2647 | 8.5579 | −14.4965 |
| 131.4318 | 180.00 | 125.8319 | 171.4793 | −5.5999 | −8.5207 | −1.5682 |
| 149.3693 | 200.00 | 153.9926 | 203.5366 | 4.6233 | 3.5366 | 4.3693 |
| 143.8628 | 220.00 | 152.9208 | 222.8533 | 9.0579 | 2.8533 | −13.1372 |
| 177.5218 | 240.00 | 176.3344 | 232.9879 | −1.1874 | −7.0120 | 8.5218 |
| 182.2748 | 260.00 | 174.5252 | 261.1813 | −7.7496 | 1.1813 | 1.2748 |

*Note:* The data on $X^*$ are assumed to be given. In deriving the other variables the assumptions made were as follows: (1) $E(u_i) = E(\varepsilon_i) = E(w_i) = 0$; (2) cov $(X, u)$ = cov $(X, \varepsilon)$ = cov $(u, \varepsilon)$ = cov $(w, u)$ = cov $(\varepsilon, w) = 0$; (3) $\sigma_u^2 = 100$, $\sigma_s^2 = 36$, and $\sigma_w^2 = 36$; and (4) $Y_i^* = 25 + 0.6X_i^* + u_i$, $Y_i = Y_i^* + \varepsilon_i$, and $X_i = X_i^* + w_i$.

[30] I am indebted to Kenneth J. White for constructing this example. See his *Computer Handbook Using SHAZAM,* for use with Damodar Gujarati, *Basic Econometrics,* September 1985, pp. 117–121.

## 13.6 INCORRECT SPECIFICATION OF THE STOCHASTIC ERROR TERM

A common problem facing a researcher is the specification of the error term $u_i$ that enters the regression model. Since the error term is not directly observable, there is no easy way to determine the form in which it enters the model. To see this, let us return to the models given in (13.2.8) and (13.2.9). For simplicity of exposition, we have assumed that there is no intercept in the model. We further assume that $u_i$ in (13.2.8) is such that $\ln u_i$ satisfies the usual OLS assumptions.

If we assume that (13.2.8) is the "correct" model but estimate (13.2.9), what are the consequences? It is shown in Appendix 13.A, Section 13A.4, that if $\ln u_i \sim N(0,\ \sigma^2)$, then

$$u_i \sim \text{log normal}\left[e^{\sigma^2/2},\ e^{\sigma^2}\left(e^{\sigma^2}-1\right)\right] \tag{13.6.1}$$

as a result:

$$E(\hat{\alpha}) = \beta e^{\sigma^2/2} \tag{13.6.2}$$

where $e$ is the base of the natural logarithm.

As you can see, $\hat{\alpha}$ is a biased estimator, as its average value is not equal to the true $\beta$.

We will have more to say about the specification of the stochastic error term in the chapter on nonlinear-in-the-parameter regression models.

## 13.7 NESTED VERSUS NON-NESTED MODELS

In carrying out specification testing, it is useful to distinguish between **nested and non-nested models.** To distinguish between the two, consider the following models:

$$\text{Model A:} \quad Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$$

$$\text{Model B:} \quad Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

We say that Model B is nested in Model A because it is a special case of Model A: If we estimate Model A and test the hypothesis that $\beta_4 = \beta_5 = 0$ and do not reject it on the basis of, say, the $F$ test,[31] Model A reduces to Model B. If we add variable $X_4$ to Model B, then Model A will reduce to Model B if $\beta_5$ is zero; here we will use the *t test* to test the hypothesis that the coefficient of $X_5$ is zero.

Without calling them such, the specification error tests that we have discussed previously and the restricted $F$ test that we discussed in Chapter 8 are essentially tests of nested hypothesis.

[31]More generally, one can use the likelihood ratio test, or the Wald test or the Lagrange Multiplier test, which were discussed briefly in Chap. 8.

Now consider the following models:

$$\text{Model C:} \quad Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + u_i$$

$$\text{Model D:} \quad Y_i = \beta_1 + \beta_2 Z_{2i} + \beta_3 Z_{3i} + v_i$$

where the $X$'s and $Z$'s are different variables. We say that Models C and D are **non-nested** because one cannot be derived as a special case of the other. In economics, as in other sciences, more than one competing theory may explain a phenomenon. Thus, the monetarists would emphasize the role of money in explaining changes in GDP, whereas the Keynesians may explain them by changes in government expenditure.

It may be noted here that one can allow Models C and D to contain regressors that are common to both. For example, $X_3$ could be included in Model D and $Z_2$ could be included in Model C. Even then these are non-nested models, because Model C does not contain $Z_3$ and Model D does not contain $X_2$.

Even if the same variables enter the model, the functional form may make two models non-nested. For example, consider the model:

$$\text{Model E:} \quad Y_i = \beta_1 + \beta_2 \ln Z_{2i} + \beta_3 \ln Z_{3i} + w_i$$

Models D and E are non-nested, as one cannot be derived as a special case of the other.

Since we already have looked at tests of nested models ($t$ and $F$ tests), in the following section we discuss some of the tests of non-nested models, which earlier we called model mis-specification errors.

## 13.8 TESTS OF NON-NESTED HYPOTHESES

According to Harvey,[32] there are two approaches to testing non-nested hypotheses: (1) the **discrimination approach,** where given two or more competing models, one chooses a model based on some criteria of goodness of fit, and (2) the **discerning approach** (my terminology) where, in investigating one model, we take into account information provided by other models. We consider these approaches briefly.

### The Discrimination Approach

Consider Models C and D above. Since both models involve the same dependent variable, we can choose between two (or more) models based on some goodness-of-fit criterion, such as $R^2$ or adjusted $R^2$, which we have already discussed. But keep in mind that in comparing two or more models,

[32]Andrew Harvey, *The Econometric Analysis of Time Series,* 2d ed., The MIT Press, Cambridge, Mass., 1990, Chap. 5.

the regressand must be the same. Besides these criteria, there are other criteria that are also used. These include **Akaike's information criterion (AIC), Schwarz's information criterion (SIC),** and **Mallows's $C_p$ criterion.** We discuss these criteria in Section 13.9. Most modern statistical software packages have one or more of these criteria built into their regression routines. In the last section of this chapter, we will illustrate these criteria using an extended example. On the basis of one or more of these criteria a model is finally selected that has the highest $\bar{R}^2$ or the lowest value of AIC or SIC, etc.

### The Discerning Approach

**The Non-Nested *F* Test or Encompassing *F* Test.** Consider Models C and D introduced earlier. How do we choose between the two models? For this purpose suppose we estimate the following nested or *hybrid* model:

$$\text{Model F:} \qquad Y_i = \lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \lambda_4 Z_{2i} + \lambda_5 Z_{3i} + u_i$$

Notice that Model F *nests or encompasses* models C and D. But note that C is not nested in D and D is not nested in C, so they are non-nested models.

Now if Model C is correct, $\lambda_4 = \lambda_5 = 0$, whereas Model D is correct if $\lambda_2 = \lambda_3 = 0$. This testing can be done by the usual $F$ test, hence the name non-nested $F$ test.

However, there are problems with this testing procedure. *First,* if the $X$'s and the $Z$'s are highly correlated, then, as noted in the chapter on multicollinearity, it is quite likely that one or more of the $\lambda$'s are individually statistically insignificant, although on the basis of the $F$ test one can reject the hypothesis that all the slope coefficients are simultaneously zero. In this case, we have no way of deciding whether Model C or Model D is the correct model. *Second,* there is another problem. Suppose we choose Model C as the *reference hypothesis* or model, and find that all its coefficients are significant. Now we add $Z_2$ or $Z_3$ or both to the model and find, using the $F$ test, that their incremental contribution to the explained sum of squares (ESS) is statistically insignificant. Therefore, we decide to choose Model C.

But suppose we had instead chosen Model D as the reference model and found that all its coefficients were statistically significant. But when we add $X_2$ or $X_3$ or both to this model, we find, again using the $F$ test, that their incremental contribution to ESS is insignificant. Therefore, we would have chosen model D as the correct model. Hence, "the choice of the reference hypothesis could determine the outcome of the choice model,"[33] especially if severe multicollinearity is present in the competing regressors. *Finally,* the artificially nested model $F$ may not have any economic meaning.

[33]Thomas B. Fomby, R. Carter Hill, and Stanley R. Johnson, *Advanced Econometric Methods,* Springer Verlag, New York, 1984, p. 416.

AN ILLUSTRATIVE EXAMPLE: THE ST. LOUIS MODEL

To determine whether changes in nominal GNP can be explained by changes in the money supply (monetarism) or by changes in government expenditure (Keynesianism), we consider the following models:

$$\begin{aligned}\dot{Y}_t &= \alpha + \beta_0 \dot{M}_t + \beta_1 \dot{M}_{t-1} + \beta_2 \dot{M}_{t-2} + \beta_3 \dot{M}_{t-3} + \beta_4 \dot{M}_{t-4} + u_{1t} \\ &= \alpha + \sum_{i=0}^{4} \beta_i \dot{M}_{t-i} + u_{1t} \end{aligned} \tag{13.8.1}$$

$$\begin{aligned}\dot{Y}_t &= \gamma + \lambda_0 \dot{E}_t + \lambda_1 \dot{E}_{t-1} + \lambda_2 \dot{E}_{t-2} + \lambda_3 \dot{E}_{t-3} + \lambda_4 \dot{E}_{t-4} + u_{2t} \\ &= \gamma + \sum_{i=0}^{4} \lambda_i \dot{E}_{t-i} + u_{2t} \end{aligned} \tag{13.8.2}$$

where $\dot{Y}_t$ = rate of growth in nominal GNP at time $t$
$\dot{M}_t$ = rate of growth in the money supply ($M_1$ version) at time $t$
$\dot{E}_t$ = rate of growth in full, or high, employment government expenditure at time $t$

In passing, note that (13.8.1) and (13.8.2) are examples of **distributed lag models,** a topic thoroughly discussed in Chapter 17. For the time being, simply note that the effect of a unit change in the money supply or government expenditure on GNP is distributed over a period of time and is not instantaneous.

Since a priori it may be difficult to decide between the two competing models, let us enmesh the two models as shown below:

$$\dot{Y}_t = \text{constant} + \sum_{i=0}^{4} \beta_i \dot{M}_{t-i} + \sum_{i=0}^{4} \lambda_i \dot{E}_{t-i} + u_{3t} \tag{13.8.3}$$

This nested model is one form in which the famous (Federal Reserve Bank of) St. Louis model, a pro-monetary-school bank, has been expressed and estimated. The results of this model for the period 1953–I to 1976–IV for the United States are as follows ($t$ ratios in parentheses):[34]

| Coefficient | Estimate | Coefficient | Estimate | |
|---|---|---|---|---|
| $\beta_0$ | 0.40 (2.96) | $\lambda_0$ | 0.08 (2.26) | |
| $\beta_1$ | 0.41 (5.26) | $\lambda_1$ | 0.06 (2.52) | |
| $\beta_2$ | 0.25 (2.14) | $\lambda_2$ | 0.00 (0.02) | |
| $\beta_3$ | 0.06 (0.71) | $\lambda_3$ | −0.06 (−2.20) | **(13.8.4)** |
| $\beta_4$ | −0.05 (−0.37) | $\lambda_4$ | −0.07 (−1.83) | |
| $\sum_{i=0}^{4} \beta_i$ | 1.06 (5.59) | $\sum_{i=0}^{4} \lambda_i$ | 0.03 (0.40) | |
| | | | | $R^2 = 0.40$ |
| | | | | $d = 1.78$ |

What do these results suggest about the superiority of one model over the other? If we consider the cumulative effect of a unit change in $\dot{M}$ and $\dot{E}$ on $\dot{Y}$, we obtain, respectively, $\sum_{i=0}^{4} \beta_i = 1.06$ and $\sum_{i=0}^{4} \lambda_i = 0.03$, the former being statistically significant and the latter not. This comparison would tend to support the monetarist claim that it is changes in the money supply that determine changes in the (nominal) GNP. It is left as an exercise for the reader to evaluate critically this claim.

[34]See Keith M. Carlson, "Does the St. Louis Equation Now Believe in Fiscal Policy?" *Review, Federal Reserve Bank of St. Louis,* vol. 60, no. 2, February 1978, p. 17, table IV.

**Davidson–MacKinnon *J* Test.**[35] Because of the problems just listed in the non-nested *F* testing procedure, alternatives have been suggested. One is the *Davidson–MacKinnon J* test. To illustrate this test, suppose we want to compare hypothesis or Model C with hypothesis or Model D. The ***J* test** proceeds as follows:

**1.** We estimate Model D and from it we obtain the estimated $Y$ values, $\hat{Y}_i^D$.

**2.** We add the predicted $Y$ value in Step 1 as an additional regressor to Model C and estimate the following model:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 \hat{Y}_i^D + u_i \qquad \textbf{(13.8.5)}$$

where the $\hat{Y}_i^D$ values are obtained from Step 1. This model is an example of the **encompassing principle,** as in the Hendry methodology.

**3.** Using the $t$ test, test the hypothesis that $\alpha_4 = 0$.

**4.** If the hypothesis that $\alpha_4 = 0$ is not rejected, we can accept (i.e., not reject) Model C as the true model because $\hat{Y}_i^D$ included in (13.8.5), which represent the influence of variables not included in Model C, have no additional explanatory power beyond that contributed by Model C. In other words, Model C *encompasses* Model D in the sense that the latter model does not contain any additional information that will improve the performance of Model C. By the same token, if the null hypothesis is rejected, Model C cannot be the true model (why?).

**5.** Now we reverse the roles of hypotheses, or Models C and D. We now estimate Model C first, use the estimated $Y$ values from this model as regressor in (13.8.5), repeat Step 4, and decide whether to accept Model D over Model C. More specifically, we estimate the following model:

$$Y_i = \beta_1 + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 \hat{Y}_i^C + u_i \qquad \textbf{(13.8.6)}$$

where $\hat{Y}_i^C$ are the estimated $Y$ values from Model C. We now test the hypothesis that $\beta_4 = 0$. If this hypothesis is not rejected, we choose Model D over C. If the hypothesis that $\beta_4 = 0$ is rejected, choose C over D, as the latter does not improve over the performance of C.

Although it is intuitively appealing, the $J$ test has some problems. Since the tests given in (13.8.5) and (13.8.6) are performed independently, we have the following likely outcomes:

| | **Hypothesis: $\alpha_4 = 0$** | |
|---|---|---|
| **Hypothesis: $\beta_4 = 0$** | **Do not reject** | **Reject** |
| **Do not reject** | Accept both C and D | Accept D, reject C |
| **Reject** | Accept C, reject D | Reject both C and D |

[35] R. Davidson and J. G. MacKinnon, "Several Tests for Model Specification in the Presence of Alternative Hypotheses," *Econometrica,* vol. 49, 1981, pp. 781–793.

As this table shows, we will not be able to get a clear answer if the *J* testing procedure leads to the acceptance or rejection of both models. In case both models are rejected, neither model helps us to explain the behavior of *Y*. Similarly, if both models are accepted, as Kmenta notes, "the data are apparently not rich enough to discriminate between the two hypotheses [models]."[36]

Another problem with the *J* test is that when we use the *t* statistic to test the significance of the estimated *Y* variable in models (13.8.5) and (13.8.6), the *t* statistic has the standard normal distribution only asymptotically, that is, in large samples. Therefore, the *J* test may not be very powerful (in the statistical sense) in small samples because it tends to reject the true hypothesis or model more frequently than it ought to.

AN ILLUSTRATIVE EXAMPLE

To illustrate the *J* test, consider the data given in Table 13.3. This table gives data on per capita personal consumption expenditure (PPCE) and per capita disposable personal income (PDPI), both measured in 1987 dollars, for the United States for the period 1970–1991. Now consider the following rival models:

$$\text{Model A:} \quad \text{PPCE}_t = \alpha_1 + \alpha_2 \text{PDPI}_t + \alpha_3 \text{PDPI}_{t-1} + u_t \qquad \textbf{(13.8.7)}$$

$$\text{Model B:} \quad \text{PPCE}_t = \beta_1 + \beta_2 \text{PDPI}_t + \beta_3 \text{PPCE}_{t-1} + u_t \qquad \textbf{(13.8.8)}$$

Model A states that PPCE depends on PDPI in the current and previous time period; this model is an example of what is known as the **distributed lag model** (see Chapter 17). Model B postulates that PPCE depends on current PDPI as well as PPCE in the previous time period; this model represents what is known as the **autoregressive model** (see Chapter 17). The

**TABLE 13.3**
PER CAPITA PERSONAL CONSUMPTION EXPENDITURE (PPCE) AND PER CAPITA PERSONAL DISPOSABLE INCOME (PDPI), 1987 DOLLARS, U.S., 1970–1991

| Year | PPCE | PDPI | Year | PPCE | PDPI |
|---|---|---|---|---|---|
| 1970 | 8,842 | 9,875 | 1981 | 10,770 | 12,156 |
| 1971 | 9,022 | 10,111 | 1982 | 10,782 | 12,146 |
| 1972 | 9,425 | 10,414 | 1983 | 11,179 | 12,349 |
| 1973 | 9,752 | 11,013 | 1984 | 11,617 | 13,029 |
| 1974 | 9,602 | 10,832 | 1985 | 12,015 | 13,258 |
| 1975 | 9,711 | 10,906 | 1986 | 12,336 | 13,552 |
| 1976 | 10,121 | 11,192 | 1987 | 12,568 | 13,545 |
| 1977 | 10,425 | 11,406 | 1988 | 12,903 | 13,890 |
| 1978 | 10,744 | 11,851 | 1989 | 13,029 | 14,005 |
| 1979 | 10,876 | 12,039 | 1990 | 13,044 | 14,068 |
| 1980 | 10,746 | 12,005 | 1991 | 12,824 | 13,886 |

*Source: Economic Report of the President, 1993,* Table B-5, p. 355.

(*Continued*)

[36]Jan Kmenta, op. cit., p. 597.

AN ILLUSTRATIVE EXAMPLE (*Continued*)

reason for introducing the lagged value of PPCE in this model is to reflect inertia or habit persistence.

The results of estimating these models separately were as follows:

$$\text{Model A:} \quad \widehat{\text{PPCE}}_t = -1{,}299.0536 + 0.9204\,\text{PDPI}_t + 0.0931\,\text{PDPI}_{t-1}$$
$$t = \quad (-4.0378) \quad (6.0178) \quad (0.6308) \qquad \textbf{(13.8.9)}$$
$$R^2 = 0.9888 \quad d = 0.8092$$

$$\text{Model B:} \quad \widehat{\text{PPCE}}_t = -841.8568 + 0.7117\,\text{PDPI}_t + 0.2954\,\text{PPCE}_{t-1}$$
$$t = \quad (-2.4137) \quad (5.4634) \quad (2.3681) \qquad \textbf{(13.8.10)}$$
$$R^2 = 0.9912 \quad d = 1.0144$$

If one were to choose between these two models on the basis of the discrimination approach, using, say, the highest $R^2$ criterion, one would choose (13.8.10); besides, in (13.8.10) both variables seem to be individually statistically significant, whereas in (13.8.9) only the current PDPI is statistically significant (but beware of the collinearity problem!).

But choosing (13.8.10) over (13.8.9) may not be appropriate because for predictive purposes there is not much difference in the two estimated $R^2$ values.

To apply the *J* test, suppose we assume Model A is the null hypothesis, that is, the maintained model, and Model B is the alternative hypothesis. Now following the *J* test steps discussed earlier we use the estimated PPCE values from model (13.8.10) as an additional regressor in Model A, giving the following outcome:

$$\widehat{\text{PPCE}}_t = 1{,}322.7958 - 0.7061\text{PDPI}_t - 0.4357\text{PDPI}_{t-1} + 2.1335\widehat{\text{PPCE}}_t^{\,B}$$
$$t = \quad (1.5896) \quad (-1.3958) \quad (-2.1926) \quad (3.3141) \qquad \textbf{(13.8.11)}$$
$$R^2 = 0.9932 \quad d = 1.7115$$

where $\widehat{\text{PPCE}}_t^{\,B}$ on the right side of (13.8.11) are the estimated PPCE values from model B, (13.8.10). Since the coefficient of this variable is statistically significant (at the two-tail 0.004 level), following the *J* test procedure, we have to reject Model A in favor of Model B.

Now assuming Model B as the maintained hypothesis and Model A as the alternative hypothesis, and following exactly the same procedure as before, we obtain the following results:

$$\widehat{\text{PPCE}}_t = -6{,}549.8659 + 5.1176\text{PDPI}_t + 0.6302\text{PPCE}_{t-1} - 4.6776\widehat{\text{PPCE}}_t^{\,A}$$
$$t = \quad (-2.4976) \quad (2.5424) \quad (3.4141) \quad (-2.1926) \qquad \textbf{(13.8.12)}$$
$$R^2 = 0.9920 \quad d = 1.7115$$

where $\widehat{\text{PPCE}}^{A}$ on the right side of (13.8.12) is obtained from the Model A, (13.8.9). But in this regression, the coefficient of $\widehat{\text{PPCE}}_t^{\,A}$ on the right side is also statistically significant (at the two-tail 0.0425 level). This result would suggest that we should now reject Model B in favor of Model A!

All this tells us is that neither model is particularly useful in explaining the behavior of per capita personal consumption expenditure in the United States over the period 1970–1991.

Of course, we have considered only two competing models. In reality, there may be more than two models. The *J* test procedure can be extended to multiple model comparisons, although the analysis can become quickly complex.

This example shows very vividly why the CLRM assumes that the regression model used in the analysis is correctly specified. Obviously it is very crucial in developing a model to pay very careful attention to the phenomenon being modeled.

**Other Tests of Model Selection.** The *J* test just discussed is only one of a group of tests of model selection. There is the **Cox test,** the **JA test,** the **P test, Mizon–Richard encompassing test,** and variants of these tests. Obviously, we cannot hope to discuss these specialized tests, for which the reader may want to consult the references cited in the various footnotes.[37]

## 13.9 MODEL SELECTION CRITERIA

In this section we discuss several criteria that have been used to choose among competing models and/or to compare models for forecasting purposes. Here we distinguish between **in-sample** forecasting and **out-of-sample** forecasting. In-sample forecasting essentially tells us how the chosen model fits the data in a given sample. Out-of-sample forecasting is concerned with determining how a fitted model forecasts future values of the regressand, given the values of the regressors.

Several criteria are used for this purpose. In particular, we discuss these criteria: (1) $R^2$, (2) adjusted $R^2 (= \bar{R}^2)$, (3) Akaike information criterion (AIC), (4) Schwarz Information criterion (SIC), (5) Mallow's $C_p$ criterion, and (6) forecast $\chi^2$ (chi-square). All these criteria aim at minimizing the residual sum of squares (RRS) (or increasing the $R^2$ value). However, except for the first criterion, criteria (2), (3), (4), and (5) impose a penalty for including an increasingly large number of regressors. Thus there is a tradeoff between goodness of fit of the model and its complexity (as judged by the number of regressors).

### The $R^2$ Criterion

We know that one of the measures of goodness of fit of a regression model is $R^2$, which, as we know, is defined as:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{13.9.1}$$

$R^2$, thus defined, of necessity lies between 0 and 1. The closer it is to 1, the better is the fit. But there are problems with $R^2$. *First*, it measures *in-sample* goodness of fit in the sense of how close an estimated $Y$ value is to its actual value in the given sample. There is no guarantee that it will forecast well *out-of-sample* observations. *Second,* in comparing two or more $R^2$'s, the dependent variable, or regressand, must be the same. *Third,* and more importantly, an $R^2$ cannot fall when more variables are added to the model. Therefore, there is every temptation to play the game of "maximizing the $R^2$" by simply adding more variables to the model. Of course, adding more variables to the model may increase $R^2$ but it may also increase the variance of forecast error.

[37] See also Badi H. Baltagi, *Econometrics,* Springer, New York, 1998, pp. 209–222.

### Adjusted $R^2$

As a penalty for adding regressors to increase the $R^2$ value, Henry Theil developed the adjusted $R^2$, denoted by $\bar{R}^2$, which we studied in Chapter 7. Recall that

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k} \tag{13.9.2}$$

As you can see from this formula, $\bar{R}^2 \leq R^2$, showing how the adjusted $R^2$ penalizes for adding more regressors. As we noted in Chapter 8, unlike $R^2$, the adjusted $R^2$ will increase only if the absolute $t$ value of the added variable is greater than 1. For comparative purposes, therefore, $\bar{R}^2$ is a better measure than $R^2$. But again keep in mind that the regressand must be the same for the comparison to be valid.

### Akaike Information Criterion (AIC)

The idea of imposing a penalty for adding regressors to the model has been carried further in the AIC criterion, which is defined as:

$$\text{AIC} = e^{2k/n}\frac{\sum \hat{u}_i^2}{n} = e^{2k/n}\frac{\text{RSS}}{\text{n}} \tag{13.9.3}$$

where $k$ is the number of regressors (including the intercept) and $n$ is the number of observations. For mathematical convenience, (13.9.3) is written as

$$\ln \text{AIC} = \left(\frac{2k}{n}\right) + \ln\left(\frac{\text{RSS}}{n}\right) \tag{13.9.4}$$

where ln AIC = natural log of AIC and $2k/n$ = penalty factor. Some textbooks and software packages define AIC only in terms of its log transform so there is no need to put ln before AIC. As you see from this formula, AIC imposes a harsher penalty than $\bar{R}^2$ for adding more regressors. In comparing two or more models, the model with the lowest value of AIC is preferred. One advantage of AIC is that it is useful for not only in-sample but also out-of-sample forecasting performance of a regression model. Also, it is useful for both nested and non-nested models. It has been also used to determine the lag length in an AR($p$) model.

### Schwarz Information Criterion (SIC)

Similar in spirit to the AIC, the SIC criterion is defined as:

$$\text{SIC} = n^{k/n}\frac{\sum \hat{u}^2}{n} = n^{k/n}\frac{\text{RSS}}{n} \tag{13.9.5}$$

or in log-form:

$$\ln \text{SIC} = \frac{k}{n} \ln n + \ln \left( \frac{\text{RSS}}{\text{n}} \right) \tag{13.9.6}$$

where $[(k/n) \ln n]$ is the penalty factor. SIC imposes a harsher penalty than AIC, as is obvious from comparing (13.9.6) to (13.9.4). Like AIC, the lower the value of SIC, the better the model. Again, like AIC, SIC can be used to compare in-sample or out-of-sample forecasting performance of a model.

### Mallows's $C_p$ Criterion

Suppose we have a model consisting of $k$ regressors, including the intercept. Let $\hat{\sigma}^2$ as usual be the estimator of the true $\sigma^2$. But suppose that we only choose $p$ regressors ($p \leq k$) and obtain the RSS from the regression using these $p$ regressors. Let $\text{RSS}_p$ denote the residual sum of squares using the $p$ regressors. Now C. P. Mallows has developed the following criterion for model selection, known as the $C_p$ criterion:

$$C_p = \frac{\text{RSS}_p}{\hat{\sigma}^2} - (n - 2p) \tag{13.9.7}$$

where $n$ is the number of observations.

We know that $E(\hat{\sigma}^2)$ is an unbiased estimator of the true $\sigma^2$. Now, if the model with $p$ regressors is adequate in that it does not suffer from lack of fit, it can be shown[38] that $E(\text{RSS}_p) = (n - p)\sigma^2$. In consequence, it is true *approximately* that

$$E(C_p) \approx \frac{(n - p)\sigma^2}{\sigma^2} - (n - 2p) \approx p \tag{13.9.8}$$

In choosing a model according to the $C_p$ criterion, we would look for a model that has a low $C_p$ value, about equal to $p$. In other words, following the principle of parsimony, we will choose a model with $p$ regressors ($p < k$) that gives a fairly good fit to the data.

In practice, one usually plots $C_p$ computed from (13.9.7) against $p$. An "adequate" model will show up as a point close to the $C_p = p$ line, as can be seen from Figure 13.3. As this figure shows, Model A may be preferable to Model B, as it is closer to the $C_p = p$ line than Model B.

### A Word of Caution about Model Selection Criteria

We have discussed several model selection criteria. But one should look at these criteria as an adjunct to the various specification tests we have

[38] Norman D. Draper and Harry Smith, *Applied Regression Analysis,* 3d ed., John Wiley & Sons, New York, 1998, p. 332. See this book for some worked examples of $C_p$.

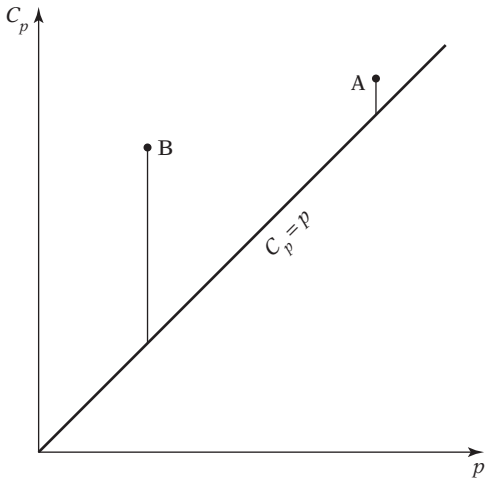


**FIGURE 13.3** Mallows's $C_p$ plot.

discussed in this chapter. Some of the criteria discussed above are purely descriptive and may not have strong theoretical properties. Some of them may even be open to the charge of data mining. Nonetheless, they are so frequently used by the practitioner that the reader should be aware of them. No one of these criteria is necessarily superior to the others.[39] Most modern software packages now include $R^2$, adjusted $R^2$, AIC, and SIC. Mallows's $C_p$ is not routinely given, although it can be easily computed from its definition.

### Forecast Chi-Square ($\chi^2$)

Suppose we have a regression model based on $n$ observations and suppose we want to use it to forecast the (mean) values of the regressand for an additional $t$ observations. As noted elsewhere, it is a good idea to save part of the sample data to see how the estimated model forecasts the observations not included in the sample, the postsample period.

Now the forecast $\chi^2$ test is defined as follows:

$$\text{Forecast, } \chi^2 = \frac{\sum_{n+1}^{n+t} \hat{u}_i^2}{\hat{\sigma}^2} \tag{13.9.9}$$

where $\hat{u}_i$ is the forecast error made for period $i$ $(= n+1, n+2, \ldots, +n+t)$, using the parameters obtained from the fitted regression and the values of the regressors in the postsample period. $\hat{\sigma}^2$ is the usual OLS estimator of $\sigma^2$ based on the fitted regression.

[39]For a useful discussion on this topic, see Francis X. Diebold, *Elements of Forecasting*, 2d ed., South Western Publishing, 2001, pp. 83–89. On balance, Diebold recommends the SIC criterion.

If we hypothesize that the parameter values have not changed between the sample and postsample periods, it can be shown that the statistic given in (13.9.9) follows the chi-square distribution with $t$ degrees of freedom, where $t$ is the number of periods for which the forecast is made. As Charemza and Deadman note, the forecast $\chi^2$ test has *weak statistical power*, meaning that the probability that the test will correctly reject a false null hypothesis is low and therefore the test should be used as a signal rather than a definitive test.[40]

## 13.10 ADDITIONAL TOPICS IN ECONOMETRIC MODELING

As noted in the introduction to this chapter, the topic of econometric modeling and diagnostic testing is so vast and evolving that specialized books are written on this topic. In the previous section we have touched on some major themes in this area. In this section we consider a few additional features that researchers may find useful in practice. In particular, we consider these topics: (1) **outliers, leverage, and influence;** (2) **recursive least squares,** and (3) **Chow's prediction failure test.** Of necessity the discussion of each of these topics will be brief.

### Outliers, Leverage, and Influence[41]

Recall that, in minimizing the residual sum of squares (RSS), OLS gives equal weight to every observation in the sample. But every observation may not have equal impact on the regression results because of the presence of three types of special data points called **outliers, leverage points,** and **influence points.** It is important that we know what they are and how they influence regression analysis.

In the regression context, an **outlier** may be defined as an observation with a "large residual." Recall that $\hat{u}_i = (Y_i - \hat{Y}_i)$, that is, the residual represents the difference (positive or negative) between the actual value of the regressand and its value estimated from the regression model. When we say that a residual is large, it is in comparison with the other residuals and very often such a large residual catches our attention immediately because of its rather large vertical distance from the estimated regression line. Note that in a data set there may be more than one outlier. We have already encountered an example of this in exercise 11.22, where you were asked to regress percent change in stock prices ($Y$) on percent change in consumer prices ($X$) for a sample of 20 countries. One observation, that relating to Chile, was an outlier.

[40] Wojciech W. Charemza and Derek F. Deadman, *New Directions in Econometric Practice: A General to Specific Modelling, Cointegration and Vector Autoregression,* 2d ed., Edward Elgar Publishers, 1997, p. 30. See also pp. 250–252 for their views on various model selection criteria.

[41] The following discussion is influenced by Chandan Mukherjee, Howard White, and Marc Wyuts, *Econometrics and Data Analysis for Developing Countries,* Routledge, New York, 1998, pp. 137–148.

A data point is said to exert (high) **leverage** if it is disproportionately distant from the bulk of the values of a regressor(s). Why does a leverage point matter? It matters because it is capable of pulling the regression line toward itself, thus distorting the slope of the regression line. If this actually happens, then we call such a leverage (data) point an **influential point.** The removal of such a data point from the sample can dramatically affect the regression line. Returning to exercise 11.22, you will see that if you regress $Y$ on $X$ including the observation for Chile, the slope coefficient is positive and "highly statistically significant." But if you drop the observation for Chile, the slope coefficient is practically zero. Thus the Chilean observation has leverage and is also an influential observation.

To further clarify the nature of outliers, leverage and influence points, consider the diagram in Figure 13.4, which is self-explanatory.[42]

How do we handle such data points? Should we just drop them and confine our attention to the remaining data points? According to Draper and Smith:

> Automatic rejection of outliers is not always a wise procedure. Sometimes the outlier is providing information that other data points cannot due to the fact that it arises from an unusual combination of circumstances which may be of vital interest and requires further investigation rather than rejection. As a general rule, outliers should be rejected out of hand only if they can be traced to causes such as errors of recording the observations or setting up the apparatus [in a physical experiment]. Otherwise, careful investigation is in order.[43]

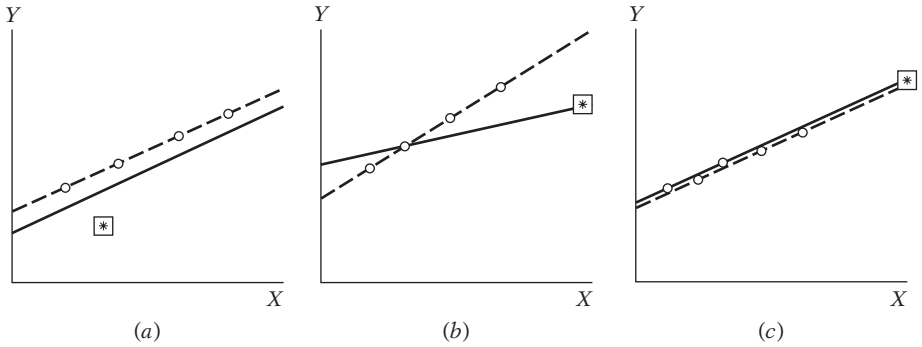


**FIGURE 13.4** In each subfigure, the solid line gives the OLS line for all the data and the broken line gives the OLS line with the outlier, denoted by an ⊡, omitted. In (*a*), the outlier is near the mean value of $X$ and has low leverage and little influence on the regression coefficients. In (*b*), the outlier is far away from the mean value of $X$ and has high leverage as well as substantial influence on the regression coefficients. In (*c*), the outlier has high leverage but low influence on the regression coefficients because it is in line with the rest of the observations.

*Source:* Adapted from John Fox, op. cit., p. 268.

[42]Adapted from John Fox, *Applied Regression Analysis, Linear Models, and Related Methods,* Sage Publications, California, 1997, p. 268.

[43]Norman R. Draper and Harry Smith, op. cit., p. 76.

What are some of the tests that one can use to detect outliers and leverage points? There are several tests discussed in the literature, but we will not discuss them here because that will take us far afield.[44] Software packages such as Shazam and Microfit have routines to detect outliers, leverage, and influential points.

### Recursive Least Squares

In Chapter 8 we examined the question of the structural stability of a regression model involving time series data and showed how the **Chow test** can be used for this purpose. Specifically, you may recall that in that chapter we discussed a simple savings function (savings as a function of income) for the United States for the period 1970–1995. There we saw that the savings income relationship probably changed around 1982. Knowing the point of the structural break we were able to confirm it with the Chow test.

But what happens if we do not know the point of the structural break (or breaks)? This is where one can use **recursive least squares (RELS).** The basic idea behind RELS is very simple and can be explained with the savings–income regression.

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

where $Y$ = savings and $X$ = income and where the sample is for the period 1970–1995. (See the data in Table 8.9.)

Suppose we first use the data for 1970–1974 and estimate the savings function, obtaining the estimates of $\beta_1$ and $\beta_2$. Then we use the data for 1970–1975 and again estimate the savings function and obtain the estimates of the two parameters. Then we use the data for 1970–1976 and re-estimate the savings model. In this fashion we go on adding an additional data point on $Y$ and $X$ until we exhaust the entire sample. As you can imagine, each regression run will give you a new set of estimates of $\beta_1$ and $\beta_2$. If you plot the estimated values of these parameters against each iteration, you will see how the values of estimated parameters change. If the model under consideration is structurally stable, the changes in the estimated values of the two parameters will be small and essentially random. However, if the estimated values of the parameters change significantly, it would indicate a structural break. RELS is thus a useful routine with time series data since time is ordered chronologically. It is also a useful diagnostic tool in cross-sectional data where the data are ordered by some "size" or "scale" variable, such as

[44]Here are some accessible sources: Alvin C. Rencher, *Linear Models in Statistics,* John Wiley & Sons, New York, 2000, pp. 219–224; A. C. Atkinson, *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis,* Oxford University Press, New York, 1985, Chap. 3; Ashis Sen and Muni Srivastava, *Regression Analysis: Theory, Methods, and Applications,* Springer-Verlag, New York, 1990, Chap. 8; and John Fox, op. cit., Chap. 11.

the employment or asset size of the firm. In exercise 13.30 you are asked to apply RELS to the savings data given in Table 8.9.

Software packages such as Shazam, Eviews, and Microfit now do recursive least-squares estimates routinely. RELS also generates **recursive residuals** on which several diagnostic tests have been based.[45]

### Chow's Prediction Failure Test

We have already discussed Chow's test of structural stability in Chapter 8. Chow has shown that his test can be modified to test the predictive power of a regression model. Again, we will revert to the U.S. savings–income regression for the period 1970–1995.

Suppose we estimate the savings–income regression for the period 1970–1981, obtaining $\hat{\beta}_{1,70-81}$ and $\hat{\beta}_{2,70-81}$, which are the estimated intercept and slope coefficients based on the data for 1970–1981. Now using the actual values of income for period 1982–1995 and the intercept and slope values for the period 1970–1981, we predict the values of savings for each of 1982–1995 years. The logic here is that if there is no serious structural change in the parameter values, the values of savings estimated for 1982–1995 based on the parameter estimates for the earlier period, should not be very different from the actual values of savings prevailing in the latter period. Of course, if there is a vast difference between the actual and predicted values of savings for the latter period, it will cast doubts on the stability of the savings–income relation for the entire data period.

Whether the difference between the actual and estimated savings value is large or small can be tested by the $F$ test as follows:

$$F = \frac{\left(\sum \hat{u}_t^{*2} - \sum \hat{u}_t^2\right)/n_2}{\left(\sum \hat{u}_t^2\right)/(n_1 - k)} \qquad \textbf{(13.10.1)}$$

where $n_1$ = number of observations in the first period (1970–1981) on which the initial regression is based, $n_2$ = number of observations in the second or forecast period, $\sum \hat{u}_t^{*2}$ = RSS when the equation estimated for all the observations $(n_1 + n_2)$, and $\sum \hat{u}_t^2$ = RSS when the equation is estimated for the first $n_1$ observations and $k$ is the number of parameters estimated (two in the present instance). If the errors are independent, and identically, normally distributed, the $F$ statistic given in (13.10.1) follows the $F$ distribution with $n_2$ and $n_1$ df, respectively. In exercise 13.31 you are asked to apply Chow's predictive failure test to find out if the savings–income relation has in fact changed. In passing, note the similarity between this test and the forecast $\chi^2$ test discussed previously.

[45]For details, see Jack Johnston and John DiNardo, *Econometric Methods,* 4th ed., McGraw-Hill, New York, 1997, pp. 117–121.

## 13.11 A CONCLUDING EXAMPLE

EXAMPLE: A MODEL OF HOURLY WAGE DETERMINATION

To determine what factors determine hourly wages, the following model was considered:

$$\text{Hwage} = \beta_1 + \beta_2\text{Edu}_i + \beta_3\text{Gender}_i + \beta_4\text{Hispanic}_i + \beta_5\text{Lfexp}_i + \beta_6\text{Mstatus}_i + \beta_7\text{Race}_i + \beta_8\text{Region}_i + \beta_9\text{Union}_i + u_i \quad \textbf{(13.11.1)}$$

where
Hwage = hourly wage ($)
Edu = education in years
Gender = 1 if female, 0 otherwise
Hispanic = 1 if Hispanic, 0 otherwise
Race = 1 if nonwhite and non-Hispanic, 0 otherwise
Lfexp = potential labor market experience in years
Mstatus = marital status, 1 if married, 0 otherwise
Region = region of residence, 1 if south, 0 otherwise
Union = union status, 1 if in union job and 0 otherwise

The origin of the wage function (13.11.1) can be traced to Jacob Mincer.[46] As you can see, the wage function includes quantitative as well as qualitative or dummy variables. A priori, all these variables seem logical. Notice that the race variable has three categories: Hispanic, non-Hispanic whites, and non-Hispanic nonwhites (largely black or African–American); hence, there are two dummies. The left-out or reference category thus is non-Hispanic whites.

The data consist of 528 persons interviewed in 1985 as a part of the current population survey (CPS) periodically conducted by the U.S. Census Bureau. These data were originally collected by Berndt and were adapted by Arthur Goldberg. We have already discussed this source in Chapter 2. Keep in mind that the data are cross sectional.

A priori, hourly wage is expected to be positively related to education, life experience, marital status and union status and negatively related to Hispanic, race, gender, and region; again note that all comparisons are in relation to non-Hispanic whites. Consult any book on labor economics to learn more about the various determinants of hourly wages.[47]

Using the data, I asked my students to estimate the model (13.11.1). The regression results are given in Table 13.4. As you can see, all the variables in (13.11.1) have the expected signs, although not all variables are individually statistically significant. The $R^2$ value of about 0.2826 might seem low, but such low $R^2$ values are typically observed in cross-sectional data with a large number of observations. But this $R^2$ value is statistically significant, since the computed $F$ value of about 25.56 is highly significant, as its $p$ value is almost zero: Remember that the $F$ statistic tests the hypothesis that all the

[46]See J. Mincer, *School, Experience and Earnings,* Columbia University Press, New York, 1974.

[47]See, for example, George Borjas, *Labor Economics,* 2d. ed., McGraw-Hill, New York, 2000.

**TABLE 13.4** REGRESSION RESULTS BASED ON (13.11.1)

Dependent Variable: HWAGE
Sample: 1 528

| Variable | Coefficient | Std. Error | *t*-Statistic | Prob. |
|---|---|---|---|---|
| *C* | -4.182714 | 1.275908 | -3.278227 | 0.0011 |
| EDUCATION | 0.937130 | 0.082625 | 11.34194 | 0.0000 |
| GENDER | -2.140661 | 0.391546 | -5.467200 | 0.0000 |
| HISPANIC | -0.512385 | 0.911056 | -0.562408 | 0.5741 |
| LFEXP | 0.098486 | 0.017494 | 5.629597 | 0.0000 |
| MSTATUS | 0.485134 | 0.418881 | 1.158167 | 0.2473 |
| RACE | -0.942389 | 0.583578 | -1.614849 | 0.1070 |
| REGION | -0.771424 | 0.430173 | -1.793287 | 0.0735 |
| UNION | 1.468088 | 0.512735 | 2.863248 | 0.0044 |

| | | | |
|---|---|---|---|
| *R*-squared | 0.282693 | Mean dependent var | 9.047538 |
| Adjusted *R*-squared | 0.271636 | S.D. dependent var | 5.144082 |
| S.E. of regression | 4.390177 | Akaike info criterion | 5.813515 |
| Sum squared resid | 10003.03 | Schwarz criterion | 5.886283 |
| Log likelihood | -1525.768 | *F*-statistic | 25.56745 |
| Durbin-Watson stat | 1.857457 | Prob(*F*-statistic) | 0.000000 |

slope coefficients are simultaneously zero; that is, all the explanatory values jointly have no impact on the regressand.

Noting the individual statistical insignificance of the variables Hispanic, marital status, and race, but noting that the region variable is "reasonably" statistically significant, some of my students dropped the first three of these variables and obtained the results shown in Table 13.5. Now all the variables are individually statistically significant at a 5 percent or better level (i.e., the *p* value less than 5 percent). The interpretation of the various coefficients is

**TABLE 13.5**

Dependent Variable: HWAGE
Sample: 1 528

| Variable | Coefficient | Std. Error | *t*-Statistic | Prob. |
|---|---|---|---|---|
| *C* | -4.289796 | 1.258229 | -3.409392 | 0.0007 |
| EDUCATION | 0.953006 | 0.082184 | 11.59596 | 0.0000 |
| GENDER | -2.134171 | 0.391740 | -5.447929 | 0.0000 |
| LFEXP | 0.104037 | 0.016888 | 6.160545 | 0.0000 |
| REGION | -0.840832 | 0.427621 | -1.966303 | 0.0498 |
| UNION | 1.427421 | 0.509978 | 2.798988 | 0.0053 |

| | | | |
|---|---|---|---|
| *R*-squared | 0.276707 | Mean dependent var | 9.047538 |
| Adjusted *R*-squared | 0.269779 | S.D. dependent var | 5.144082 |
| S.E. of regression | 4.395772 | Akaike info criterion | 5.810462 |
| Sum squared resid | 10086.51 | Schwarz criterion | 5.858974 |
| Log likelihood | -1527.962 | *F*-statistic | 39.93978 |
| Durbin-Watson stat | 1.858629 | Prob(*F*-statistic) | 0.000000 |

straightforward. For example, the value of −0.8408 of the region dummy suggests that holding all the other variables constant, on average, workers in the South earn about 84 cents less per hour than their counterparts elsewhere, perhaps because of the low cost of living in the South and/or the fact the South is less unionized. Similarly, on average, women earn less than their male counterparts, by about $2.13, holding all other factors constant. Whether this amounts to gender discrimination cannot be told from the statistical analysis alone.

As expected, the "short" regression (omitting Hispanic, marital status, and race variables) has a lower adjusted $R^2$ than the "long" regression (i.e., the regression that includes all the variables), as one would expect. But notice the **Akaike** and **Schwarz** statistics: They are both lower for the short regression compared to the long regression, showing how they penalize for introducing more regressors in the model. Since the values of both statistics are so close that one can choose either of the statistics, the Durbin–Watson $d$ value in both models is sufficiently close to 2 to suggest any "autocorrelation" or specification errors.

Since the data underlying regression (13.11.1) are given in the data disk, you may want to "experiment" with the data. It is quite possible that there might be some interaction between the gender and education dummies or gender and marital status dummies. It is also possible that the relationship between hourly wage and labor market experience is nonlinear, necessitating the introduction of the squared education term in the regression model. As you can see, even with a given data set, there are several possibilities. This might sound like data mining, but we have already noted that data mining may have some role to play in econometric modeling. Of course, you should keep in mind the true level of significance in carrying out data mining.

## 13.12 A WORD TO THE PRACTITIONER

We have covered a lot of ground in this chapter. There is no question that model building is an art as well as a science. A practical researcher may be bewildered by theoretical niceties and an array of diagnostic tools. But it is well to keep in mind Martin Feldstein's caution that "The applied econometrician, like the theorist, soon discovers from experience that a useful model is not one that is 'true' or 'realistic' but one that is parsimonious, plausible and informative."[48]

Peter Kennedy of Simon Fraser University in Canada advocates the following "Ten Commandments of Applied Econometrics"[49]:

**1.** Thou shalt use common sense and economic theory.

**2.** Thou shalt ask the right questions (i.e., put relevance before mathematical elegance).

[48]Martin S. Feldstein, "Inflation, Tax Rules and Investment: Some Econometric Evidence," *Econometrica*, vol. 30, 1982, p. 829.

[49]Peter Kennedy, op. cit., pp. 17–18.

**3.** Thou shalt know the context (do not perform ignorant statistical analysis).

**4.** Thou shalt inspect the data.

**5.** Thou shalt not worship complexity. Use the **KISS principle,** that is, *keep it stochastically simple.*

**6.** Thou shalt look long and hard at thy results.

**7.** Thou shalt beware the costs of data mining.

**8.** Thou shalt be willing to compromise (do not worship textbook prescriptions).

**9.** Thou shalt not confuse significance with substance (do not confuse statistical significance with practical significance).

**10.** Thou shalt confess in the presence of sensitivity (that is, anticipate criticism).

You may want to read Kennedy's paper fully to appreciate the conviction with which he advocates the above ten commandments. Some of these commandments may sound tongue-in-cheek, but there may be a grain of truth in each.

## 13.13 SUMMARY AND CONCLUSIONS

**1.** The assumption of the CLRM that the econometric model used in analysis is correctly specified has two meanings. One, there are no **equation specification errors,** and two, there are no **model specification errors.** In this chapter the major focus was on equation specification errors.

**2.** The equation specification errors discussed in this chapter were (1) omission of important variable(s), (2) inclusion of superfluous variable(s), (3) adoption of the wrong function form, (4) incorrect specification of the error term $u_i$, and (5) errors of measurement in the regressand and regressors.

**3.** When legitimate variables are omitted from a model, the consequences can be very serious: The OLS estimators of the variables retained in the model not only are biased but are inconsistent as well. Additionally, the variances and standard errors of these coefficients are incorrectly estimated, thereby vitiating the usual hypothesis-testing procedures.

**4.** The consequences of including irrelevant variables in the model are fortunately less serious: The estimators of the coefficients of the relevant as well as "irrelevant" variables remain unbiased as well as consistent, and the error variance $\sigma^2$ remains correctly estimated. The only problem is that the estimated variances tend to be larger than necessary, thereby making for less precise estimation of the parameters. That is, the confidence intervals tend to be larger than necessary.

**5.** To detect equation specification errors, we considered several tests, such as (1) examination of residuals, (2) the Durbin–Watson $d$ statistic, (3) Ramsey's RESET test, and (4) the Lagrange multiplier test.

**6.** A special kind of specification error is errors of measurement in the values of the regressand and regressors. If there are errors of measurement in the regressand only, the OLS estimators are unbiased as well as consistent but they are less efficient. If there are errors of measurement in the regressors, the OLS estimators are biased as well as inconsistent.

**7.** Even if errors of measurement are detected or suspected, the remedies are often not easy. The use of instrumental or proxy variables is theoretically attractive but not always practical. Thus it is very important in practice that the researcher be careful in stating the sources of his/her data, how they were collected, what definitions were used, etc. Data collected by official agencies often come with several footnotes and the researcher should bring those to the attention of the reader.

**8.** Model mis-specification errors can be as serious as equation specification errors. In particular, we distinguished between nested and nonnested models. To decide on the appropriate model we discussed the nonnested, or encompassing, $F$ test and the Davidson–MacKinnon $J$ test and pointed out the limitation of each test.

**9.** In choosing an empirical model in practice researchers have used a variety of criteria. We discussed some of these, such as the Akaike and Schwarz information criteria, Mallows's $C_p$ criterion, and forecast $\chi^2$ criterion. We discussed the advantages and disadvantages of these criteria and also warned the reader that these criteria are not absolute but are adjunct to a careful specification analysis.

**10.** We also discussed these additional topics: (1) outliers, leverage, and influence; (2) recursive least squares; and (3) Chow's prediction failure test. We discussed the role of each in applied work.

**11.** We concluded this chapter by discussing Peter Kennedy's "ten commandments of applied econometrics." The point of these commandments is to ask the researcher to look beyond the purely technical aspects of econometrics.

## EXERCISES

### Questions

**13.1.** Refer to the demand function for chicken estimated in Eq. (8.7.23). Considering the attributes of a good model discussed in Section 13.1, could you say that this demand function is "correctly" specified?

**13.2.** Suppose that the true model is

$$Y_i = \beta_1 X_i + u_i \tag{1}$$

but instead of fitting this regression through the origin you routinely fit the usual intercept-present model:

$$Y_i = \alpha_0 + \alpha_1 X_i + v_i \tag{2}$$

Assess the consequences of this specification error.

**13.3.** Continue with exercise 13.2 but assume that it is model (2) that is the truth. Discuss the consequences of fitting the mis-specified model (1).

**13.4.** Suppose that the "true" model is

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_t \tag{1}$$

but we add an "irrelevant" variable $X_3$ to the model (irrelevant in the sense that the true $\beta_3$ coefficient attached to the variable $X_3$ is zero) and estimate

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

**a.** Would the $R^2$ and the adjusted $R^2$ for model (2) be larger than that for model (1)?
**b.** Are the estimates of $\beta_1$ and $\beta_2$ obtained from (2) unbiased?
**c.** Does the inclusion of the "irrelevant" variable $X_3$ affect the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$?

**13.5.** Consider the following "true" (Cobb–Douglas) production function:

$$\ln Y_i = \alpha_0 + \alpha_1 \ln L_{1i} + \alpha_2 \ln L_{2i} + \alpha_3 \ln K_i + u_i$$

where $Y =$ output
$L_1 =$ production labor
$L_2 =$ nonproduction labor
$K =$ capital

But suppose the regression actually used in empirical investigation is

$$\ln Y_i = \beta_0 + \beta_1 \ln L_{1i} + \beta_2 \ln K_i + u_i$$

On the assumption that you have cross-sectional data on the relevant variables,
**a.** Will $E(\hat{\beta}_1) = \alpha_1$ and $E(\hat{\beta}_2) = \alpha_3$?
**b.** Will the answer in **a** hold if it is known that $L_2$ is an *irrelevant* input in the production function? Show the necessary derivations.

**13.6.** Refer to Eqs. (13.3.4) and (13.3.5). As you can see, $\hat{\alpha}_2$, although biased, has a smaller variance than $\hat{\beta}_2$, which is unbiased. How would you decide on the tradeoff between bias and smaller variance? *Hint:* The MSE (mean-square error) for the two estimators is expressed as

$$\begin{aligned}\text{MSE}(\hat{\alpha}_2) &= \left(\sigma^2 \Big/ \sum x_{2i}^2\right) + \beta_3^2 b_{3\,2}^2 \\ &= \text{sampling variance} + \text{square of bias} \\ \text{MSE}(\hat{\beta}_2) &= \sigma^2 \Big/ \sum x_2^2 (1 - r_{2\,3}^2)\end{aligned}$$

On MSE, see **Appendix A.**

**13.7.** Show that $\beta$ estimated from either (13.5.1) or (13.5.3) provides an unbiased estimate of true $\beta$.

**13.8.** Following Friedman's permanent income hypothesis, we may write

$$Y_i^* = \alpha + \beta X_i^* \tag{1}$$

where $Y_i^*$ = "permanent" consumption expenditure and $X_i^*$ = "permanent" income. Instead of observing the "permanent" variables, we observe

$$Y_i = Y_i^* + u_i$$
$$X_i = X_i^* + v_i$$

where $Y_i$ and $X_i$ are the quantities that can be observed or measured and where $u_i$ and $v_i$ are measurement errors in $Y^*$ and $X^*$, respectively.

Using the observable quantities, we can write the consumption function as

$$\begin{aligned} Y_i &= \alpha + \beta(X_i - v_i) + u_i \\ &= \alpha + \beta X_i + (u_i - \beta v_i) \end{aligned} \tag{2}$$

Assuming that (1) $E(u_i) = E(v_i) = 0$, (2) $\text{var}(u_i) = \sigma_u^2$ and $\text{var}(v_i) = \sigma_v^2$, (3) $\text{cov}(Y_i^*, u_i) = 0$, $\text{cov}(X_i^*, v_i) = 0$, and (4) $\text{cov}(u_i, X_i^*) = \text{cov}(v_i, Y_i^*) = \text{cov}(u_i, v_i) = 0$, show that in large samples $\beta$ estimated from (2) can be expressed as

$$\text{plim}(\hat{\beta}) = \frac{\beta}{1 + (\sigma_v^2 / \sigma_{X^*}^2)}$$

**a.** What can you say about the nature of the bias in $\hat{\beta}$?
**b.** If the sample size increases indefinitely, will the estimated $\beta$ tend to equality with the true $\beta$?

**13.9. Capital asset pricing model.** The capital asset pricing model (CAPM) of modern investment theory postulates the following relationship between the average rate of return of a security (common stock), measured over a certain period, and the volatility of the security, called the *beta coefficient* (volatility is measure of risk):

$$\bar{R}_i = \alpha_1 + \alpha_2(\beta_i) + u_i \tag{1}$$

where $\bar{R}_i$ = average rate of return of security $i$
$\beta_i$ = true beta coefficient of security $i$
$u_i$ = stochastic disturbance term

The true $\beta_i$ is not directly observable but is measured as follows:

$$r_{it} = \alpha_1 + \beta^* r_{m_t} + e_t \tag{2}$$

where $r_{it}$ = rate of return of security $i$ for time $t$
$r_{m_t}$ = market rate of return for time $t$ (this rate is the rate of return on some broad market index, such as the S&P index of industrial securities)
$e_t$ = residual term

and where $\beta^*$ is an estimate of the "true" beta coefficient. In practice, therefore, instead of estimating (1), one estimates

$$\bar{R}_i = \alpha_1 + \alpha_2(\beta_i^*) + u_i \tag{3}$$

where $\beta_i^*$ are obtained from the regression (2). But since $\beta_i^*$ are estimated, the relationship between true $\beta$ and $\beta^*$ can be written as

$$\beta_i^* = \beta_i + v_i \tag{4}$$

where $v_i$ can be called the *error of measurement*.

**a.** What will be the effect of this error of measurement on the estimate of $\alpha_2$?

**b.** Will the $\alpha_2$ estimated from (3) provide an unbiased estimate of true $\alpha_2$? If not, is it a consistent estimate of $\alpha_2$? If not, what remedial measures do you suggest?

**13.10.** Consider the model

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{1}$$

To find out whether this model is mis-specified because it omits the variable $X_3$ from the model, you decide to regress the residuals obtained from model (1) on the variable $X_3$ only (*Note:* There is an intercept in this regression). The Lagrange multiplier (LM) test, however, requires you to regress the residuals from (1) on both $X_2$ and $X_3$ and a constant. Why is your procedure likely to be inappropriate?*

**13.11.** Consider the model

$$Y_i = \beta_1 + \beta_2 X_i^* + u_i$$

In practice we measure $X_i^*$ by $X_i$ such that

**a.** $X_i = X_i^* + 5$

**b.** $X_i = 3X_i^*$

**c.** $X_i = (X_i^* + \varepsilon_i)$, where $\varepsilon_i$ is a purely random term with the usual properties

What will be the effect of these measurement errors on estimates of true $\beta_1$ and $\beta_2$?

**13.12.** Refer to the regression Eqs. (13.3.1) and (13.3.2). In a manner similar to (13.3.3) show that

$$E(\hat{\alpha}_1) = \beta_1 + \beta_3(\bar{X}_3 - b_{32}\bar{X}_2)$$

where $b_{32}$ is the slope coefficient in the regression of the omitted variable $X_3$ on the included variable $X_2$.

**13.13.** Critically evaluate the following view expressed by Leamer†:

> My interest in metastatistics [i.e., theory of inference actually drawn from data] stems from my observations of economists at work. The opinion that econometric theory is irrelevant is held by an embarrassingly large share of the economic profession. The wide gap between econometric theory and econometric practice might be expected to cause professional tension. In fact, a calm equilibrium permeates our

*See Maddala, op. cit., p. 477.

†Edward E. Leamer, *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, John Wiley & Sons, New York, 1978, p. vi.

journals and our [professional] meetings. We comfortably divide ourselves into a celibate priesthood of statistical theorists, on the one hand, and a legion of inveterate sinner-data analysts, on the other. The priests are empowered to draw up lists of sins and are revered for the special talents they display. Sinners are not expected to avoid sins; they need only confess their errors openly.

**13.14.** Evaluate the following statement made by Henry Theil[*]:

> Given the present state of the art, the most sensible procedure is to interpret confidence coefficients and significance limits liberally when confidence intervals and test statistics are computed from the final regression of a regression strategy in the conventional way. That is, a 95 percent confidence coefficient may actually be an 80 percent confidence coefficient and a 1 percent significance level may actually be a 10 percent level.

**13.15.** Commenting on the econometric methodology practiced in the 1950s and early 1960s, Blaug stated[†]:

> . . . much of it [i.e., empirical research] is like playing tennis with the net down: instead of attempting to refute testable predictions, modern economists all too frequently are satisfied to demonstrate that the real world conforms to their predictions, thus replacing falsification [à la Popper], which is difficult, with verification, which is easy.

Do you agree with this view? You may want to peruse Blaug's book to learn more about his views.

**13.16.** According to Blaug, "There is no logic of proof but there is logic of disproof."[‡] What does he mean by this?

**13.17.** Refer to the St. Louis model discussed in the text. Keeping in mind the problems associated with the nested $F$ test, critically evaluate the results presented in regression (13.8.4).

**13.18.** Suppose the true model is

$$Y_i = \beta_1 + \beta_2 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$$

but you estimate

$$Y_i = \alpha_1 + \alpha_2 X_i + v_i$$

If you use observations of $Y$ at $X = -3, -2, -1, 0, 1, 2, 3$, and estimate the "incorrect" model, what bias will result in these estimates?[§]

**13.19.** To see if the variable $X_i^2$ belongs in the model $Y_i = \beta_1 + \beta_2 X_i + u_i$, Ramsey's RESET test would estimate the linear model, obtaining the estimated $Y_i$ values from this model [i.e., $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$] and then

[*]Henry Theil, *Principles of Econometrics,* John Wiley & Sons, New York, 1971, pp. 605–606.
[†]M. Blaug, *The Methodology of Economics. Or How Economists Explain,* Cambridge University Press, New York, 1980, p. 256.
[‡]Ibid., p. 14.
[§]Adapted from G. A. F., *Linear Regression Analysis*, John Wiley & Sons, New York, 1977, p. 176.

estimating the model $Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 \hat{Y}_i^2 + v_i$ and testing the significance of $\alpha_3$. Prove that, if $\hat{\alpha}_3$ turns out to be statistically significant in the preceding (RESET) equation, it is the same thing as estimating the following model directly: $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i$. (*Hint:* Substitute for $\hat{Y}_i$ in the RESET regression[*]).

**13.20.** State with reason whether the following statements are true or false.[†]

**a.** An observation can be influential but not an outlier.

**b.** An observation can be an outlier but not influential.

**c.** An observation can be both influential and an outlier.

**d.** If in the model $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i$ $\hat{\beta}_3$ turns out to be statistically significant, we should retain the linear term $X_i$ even if $\hat{\beta}_2$ is statistically insignificant.

**e.** If you estimate the model $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ or $Y_i = \alpha_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ by OLS, the estimated regression line is the same, where $x_{2i} = (X_{2i} - \bar{X}_2)$ and $x_{3i} = (X_{3i} - \bar{X}_3)$.

## Problems

**13.21.** Use the data for the demand for chicken given in exercise 7.19. Suppose you are told that the true demand function is

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + \beta_6 \ln X_{6t} + u_t \qquad \textbf{(1)}$$

but you think differently and estimate the following demand function:

$$\ln Y_t = \alpha_1 + \alpha_2 \ln X_{2t} + \alpha_3 \ln X_{3t} + v_t \qquad \textbf{(2)}$$

where $Y$ = per capita consumption of chickens (lb)
$X_2$ = real disposable per capita income
$X_3$ = real retail price of chickens
$X_6$ = composite real price of chicken substitutes

**a.** Carry out RESET and LM tests of specification errors, assuming the demand function (1) just given is the truth.

**b.** Suppose $\hat{\beta}_6$ in (1) turns out to be statistically insignificant. Does that mean there is no specification error if we fit (2) to the data?

**c.** If $\hat{\beta}_6$ turns out to be insignificant, does that mean one should not introduce the price of a substitute product(s) as an argument in the demand function?

**13.22.** Continue with exercise 13.21. Strictly for pedagogical purposes, assume that model (2) is the true demand function.

**a.** If we now estimate model (1), what type of specification error is committed in this instance?

**b.** What are the theoretical consequences of this specification error? Illustrate with the data at hand.

**13.23.** The true model is

$$Y_i^* = \beta_1 + \beta_2 X_i^* + u_i \qquad \textbf{(1)}$$

[*]Adapted from Kerry Peterson, op. cit., pp. 184–185.
[†]Adapted from Norman R. Draper and Harry Smith, op. cit., pp. 606–607.

but because of errors of measurement you estimate

$$Y_i = \alpha_1 + \alpha_2 X_i + v_i \tag{2}$$

where $Y_i = Y_i^* + \varepsilon_i$ and $X_i = X_i^* + w_i$, where $\varepsilon_i$ and $w_i$ are measurement errors.

Using the data given in Table 13.2, document the consequences of estimating (2) instead of the true model (1).

**13.24.** In exercise 6.14 you were asked to estimate the elasticity of substitution between labor and capital using the CES (constant elasticity of substitution) production function. But the function shown there is based on the assumption that there is perfect competition in the labor market. If competition is imperfect, the correct formulation of the model is

$$\log\left(\frac{V}{L}\right) = \log \beta_1 + \beta_2 \log W + \beta_3 \log\left(1 + \frac{1}{E}\right)$$

where $(V/L)$ = value added per unit of labor
$L$ = labor input
$W$ = real wage rate
$E$ = elasticity of supply of labor

**a.** What kind of specification error is involved in the original CES estimation of the elasticity of substitution if in fact the labor market is imperfect?
**b.** What are the theoretical consequences of this error for $\beta_2$, the elasticity of substitution parameter?
**c.** Assume that the labor supply elasticities in the industries shown in exercise 6.23 were as follows: 2.0, 1.8, 2.5, 2.3, 1.9, 2.1, 1.7, 2.7, 2.2, 2.1, 2.9, 2.8, 3.2, 2.9, and 3.1. Using these data along with those given in exercise 6.14, estimate the foregoing model and comment on your results in light of the theory of specification errors.

**13.25. Monte Carlo experiment*:** Ten individuals had weekly permanent income as follows: \$200, 220, 240, 260, 280, 300, 320, 340, 380, and 400. Permanent consumption ($Y_i^*$) was related to permanent income $X_i^*$ as

$$Y_i^* = 0.8 X_i^* \tag{1}$$

Each of these individuals had transitory income equal to 100 times a random number $u_i$ drawn from a normal population with mean = 0 and $\sigma^2 = 1$ (i.e., standard normal variable). Assume that there is no transitory component in consumption. Thus, measured consumption and permanent consumption are the same.

**a.** Draw 10 random numbers from a normal population with zero mean and unit variance and obtain 10 numbers for measured income $X_i$ $(= X_i^* + 100u_i)$.
**b.** Regress permanent (= measured) consumption on measured income using the data obtained in **a** and compare your results with those

*Adapted from Christopher Dougherty, *Introduction to Econometrics,* Oxford University Press, New York, 1992, pp. 253–256.

shown in (1). A priori, the intercept should be zero (why?). Is that the case? Why or why not?

**c.** Repeat **a** 100 times and obtain 100 regressions as shown in **b** and compare your results with the true regression (1). What general conclusions do you draw?

**13.26.** Refer to exercise 8.26. With the definitions of the variables given there, consider the following two models to explain $Y$:

$$\text{Model A:} \quad Y_t = \alpha_1 + \alpha_2 X_{3t} + \alpha_3 X_{4t} + \alpha_4 X_{6t} + u_t$$

$$\text{Model B:} \quad Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{5t} + \beta_4 X_{6t} + u_t$$

Using the nested $F$ test, how will you choose between the two models?

**13.27.** Continue with exercise 13.26. Using the $J$ test, how would you decide between the two models?

**13.28.** Refer to exercise 7.19, which is concerned with the demand for chicken in the United States. There you were given five models.

**a.** What is the difference between model 1 and model 2? If model 2 is correct and you estimate model 1, what kind of error is committed? Which test would you apply—equation specification error or model selection error? Show the necessary calculations.

**b.** Between models 1 and 5, which would you choose? Which test(s) do you use and why?

**13.29.** Refer to Table 8.9, which gives data on personal savings ($Y$) and personal disposable income ($X$) for the period 1970–1995. Now consider the following models:

$$\text{Model A:} \quad Y_t = \alpha_1 + \alpha_2 X_t + \alpha_3 X_{t-1} + u_t$$

$$\text{Model B:} \quad Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t$$

How would you choose between these two models? State clearly the test procedure(s) you use and show all the calculations. Suppose someone contends that the interest rate variable belongs in the savings function. How would you test this? Collect data on 3-month treasury bill rate as a proxy for the interest and demonstrate your answer.

**13.30.** Use the data in exercise 13.29. To familiarize yourself with recursive least squares, estimate the savings functions for 1970–1981, 1970–1985, 1970–1990, and 1970–1995. Comment on the stability of estimated coefficients in the savings functions.

**13.31.** Continue with exercise 13.30. Suppose you estimate the savings function for 1970–1981. Using the parameters thus estimated and the personal disposable income data from 1982–1995, estimate the predicted savings for the latter period and use Chow's prediction failure test to find out if it rejects the hypothesis that the savings function between the two time periods has not changed.

**13.32.** *Omission of a variable in the K-variable regression model.* Refer to Eq. (13.3.3), which shows the bias in omitting the variable $X_3$ from the model $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$. This can be generalized as follows: In the $k$-variable model $Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$, suppose we omit the variable $X_k$. Then it can be shown that the omitted variable bias

of the slope coefficient of included variable $X_j$ is:

$$E(\hat{\beta}_j) = \beta_j + \beta_k b_{kj} \qquad j = 2, 3, \ldots, (k-1)$$

where $b_{kj}$ is the (partial) slope coefficient of $X_j$ in the auxiliary regression of the excluded variable $X_k$ on all the explanatory variables included in the model.*

Refer to exercise 13.21. Find out the bias of the coefficients in Eq. (1) if we excluded the variable $\ln X_6$ from the model. Is this exclusion serious? Show the necessary calculations.

## APPENDIX 13A

### 13A.1 THE PROOF THAT $E(b_{12}) = \beta_2 + \beta_3 b_{32}$ [EQUATION (13.3.3)]

In the deviation form the three-variable population regression model can be written as

$$y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + (u_i - \bar{u}) \tag{1}$$

First multiplying by $x_2$ and then by $x_3$, the usual normal equations are

$$\sum y_i x_{2i} = \beta_2 \sum x_{2i}^2 + \beta_3 \sum x_{2i} x_{3i} + \sum x_{2i}(u_i - \bar{u}) \tag{2}$$

$$\sum y_i x_{3i} = \beta_2 \sum x_{2i} x_{3i} + \beta_3 \sum x_{3i}^2 + \sum x_{3i}(u_i - \bar{u}) \tag{3}$$

Dividing (2) by $\sum x_{2i}^2$ on both sides, we obtain

$$\frac{\sum y_i x_{2i}}{\sum x_{2i}^2} = \beta_2 + \beta_3 \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2} + \frac{\sum x_{2i}(u_i - \bar{u})}{\sum x_{2i}^2} \tag{4}$$

Now recalling that

$$b_{12} = \frac{\sum y_i x_{2i}}{\sum x_{2i}^2}$$

$$b_{32} = \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2}$$

Eq. (4) can be written as

$$b_{12} = \beta_2 + \beta_3 b_{32} + \frac{\sum x_{2i}(u_i - \bar{u})}{\sum x_{2i}^2} \tag{5}$$

*This can be generalized to the case where more than one relevant $X$ variable is excluded from the model. On this, see Chandan Mukherjee et al., op. cit., p. 215.

Taking the expected value of (5) on both sides, we finally obtain

$$E(b_{12}) = \beta_2 + \beta_3 b_{32} \tag{6}$$

where use is made of the facts that (*a*) for a given sample, $b_{32}$ is a known fixed quantity, (*b*) $\beta_2$ and $\beta_3$ are constants, and (*c*) $u_i$ is uncorrelated with $X_{2i}$ (as well as $X_{3i}$).

## 13A.2 THE CONSEQUENCES OF INCLUDING AN IRRELEVANT VARIABLE: THE UNBIASEDNESS PROPERTY

For the true model (13.3.6), we have

$$\hat{\beta}_2 = \frac{\sum yx_2}{\sum x_2^2} \tag{1}$$

and we know that it is unbiased.

For the model (13.3.7), we obtain

$$\hat{\alpha}_2 = \frac{(\sum yx_2)(\sum x_3^2) - (\sum yx_3)(\sum x_2x_3)}{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2} \tag{2}$$

Now the true model in deviation form is

$$y_i = \beta_2 x_2 + (u_i - \bar{u}) \tag{3}$$

Substituting for $y_i$ from (3) into (2) and simplifying, we obtain

$$\begin{aligned} E(\hat{\alpha}_2) &= \beta_2 \frac{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2}{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2} \\ &= \beta_2 \end{aligned} \tag{4}$$

that is, $\hat{\alpha}_2$ remains unbiased.

We also obtain

$$\hat{\alpha}_3 = \frac{(\sum yx_3)(\sum x_2^2) - (\sum yx_2)(\sum x_2x_3)}{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2} \tag{5}$$

Substituting for $y_i$ from (3) into (5) and simplifying, we obtain

$$\begin{aligned} E(\hat{\alpha}_3) &= \beta_2 \frac{[(\sum x_2x_3)(\sum x_2^2) - (\sum x_2x_3)(\sum x_2^2)]}{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2} \\ &= 0 \end{aligned} \tag{6}$$

which is its value in the true model since $X_3$ is absent from the true model.

### 13A.3 THE PROOF OF EQUATION (13.5.10)

We have

$$Y = \alpha + \beta X_i^* + u_i \tag{1}$$

$$X_i = X_i^* + w_i \tag{2}$$

Therefore, in deviation form we obtain

$$y_i = \beta x_i^* + (u_i - \bar{u}) \tag{3}$$

$$x_i = x_i^* + (w_i - \bar{w}) \tag{4}$$

Now when we use

$$Y_i = \alpha + \beta X_i + u_i \tag{5}$$

we obtain

$$\begin{aligned}
\hat{\beta} &= \frac{\sum yx}{\sum x^2} \\
&= \frac{\sum[\beta x^* + (u - \bar{u})][x^* + (w - \bar{w})]}{\sum[x^* + (w - \bar{w})]^2} \qquad \text{using (3) and (4)} \\
&= \frac{\beta \sum x^{*2} + \beta \sum x^*(w - \bar{w}) + \sum x^*(u - \bar{u}) + \sum(u - \bar{u})(w - \bar{w})}{\sum x^{*2} + 2\sum x^*(w - \bar{w}) + \sum(w - \bar{w})^2}
\end{aligned}$$

Since we cannot take expectation of this expression because the expectation of the ratio of two variables is not equal to the ratio of their expectations (*note:* the expectations operator $E$ is a linear operator), first we divide each term of the numerator and the denominator by $n$ and take the probability limit, plim (see **Appendix A** for details of plim), of

$$\hat{\beta} = \frac{(1/n)\left[\beta \sum x^{*2} + \beta \sum x^*(w - \bar{w}) + \sum x^*(u - \bar{u}) + \sum(u - \bar{u})(w - \bar{w})\right]}{(1/n)\left[\sum x^{*2} + 2\sum x^*(w - \bar{w}) + \sum(w - \bar{w})^2\right]}$$

Now the probability limit of the ratio of two variables is the ratio of their probability limits. Applying this rule and taking plim of each term, we obtain

$$\text{plim}\, \hat{\beta} = \frac{\beta \sigma_{X^*}^2}{\sigma_{X^*}^2 + \sigma_w^2}$$

where $\sigma_{X^*}^2$ and $\sigma_w^2$ are variances of $X^*$ and $w$ as sample size increases indefinitely and where we have used the fact that as the sample size increases indefinitely there is no correlation between the errors $u$ and $w$ as well as

between them and the true $X^*$. From the preceding expression, we finally obtain

$$\text{plim}\,\hat{\beta} = \beta\left[\frac{1}{1+(\sigma_w^2/\sigma_{X^*}^2)}\right]$$

which is the required result.

### 13A.4 THE PROOF OF EQUATION (13.6.2)

Since there is no intercept in the model, the estimate of $\alpha$, according to the formula for the regression through the origin, is as follows:

$$\hat{\alpha} = \frac{\sum X_i Y_i}{\sum X_i^2} \tag{1}$$

Substituting for $Y$ from the true model (13.2.8), we obtain

$$\hat{\alpha} = \frac{\sum X_i(\beta X_i u_i)}{\sum X_i^2} = \beta\frac{\sum X_i^2 u_i}{\sum X_i^2} \tag{2}$$

Statistical theory shows that if $\ln u_i \sim N(0,\ \sigma^2)$ then

$$u_i = \text{log normal}\left[e^{\sigma^2/2},\ e^{\sigma^2}\left(e^{\sigma^2-1}\right)\right] \tag{3}$$

Therefore,

$$\begin{aligned} E(\hat{\alpha}) &= \beta E\left(\frac{\sum X_i^2 u_i}{\sum X_i^2}\right) \\ &= \beta\left(E\frac{(X_1^2u_1 + X_2^2u_2 + \cdots + X_n^2u_n)}{\sum X_i^2}\right) \\ &= \beta e^{\sigma^2/2}\left(\frac{\sum X_i^2}{\sum X_i^2}\right) = \beta e^{\sigma^2/2} \end{aligned}$$

where use is made of the fact that the $X$'s are nonstochastic and each $u_i$ has an expected value of $e^{\sigma^2/2}$.

Since $E(\hat{\alpha}) \neq \beta,\ \hat{\alpha}$ is a biased estimator of $\beta$.