EE 425(1/2011)

Specification and Model Selection Criteria

- One of the assumptions of CLRM is that the model is **“correctly”** specified. If the model is not correctly specified, we encounter ***model specification error*** or ***model specification bias.***

- *We will examine the following issues:*
- *1.Model selection criteria*
- *2.Types of specification errors*
- *3.Consequences of specification errors*
- *4.Detection of specification errors*

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. It may be added that in some situations the exogenous regressors may be **strictly exogenous.** A strictly exogenous variable is independent of current, future, and past values of the error term.

4. *Exhibit parameter constancy;* that is, the values of the parameters should be stable. Otherwise, forecasting will be difficult. 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.

1. Model Selection Criteria

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.

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.

2. Consequences of Model Specification Errors

(1) Underfitting a Model

Suppose the true model is:

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

but for some reason we fit the following model:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i \qquad (13.3.2)$$

The consequences of omitting variable X_3 are as follows:

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

2. Consequences of Model Specification Errors

(2) Inclusion of an Irrelevant Variable

Now let us assume that

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

is the truth, but we fit the following model:

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

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

The consequences of this specification error are as follows:

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

2. Consequences of Model Specification Errors

(2) Inclusion of an Irrelevant Variable

From the usual OLS formula we know that

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

and

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

Therefore,

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

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

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

3. Detection of Specification Errors

Detecting the Presence of Unnecessary Variables

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

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

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

It is, however, very important to remember that in carrying out these tests of significance we have a specific model in mind. We accept that model as the **maintained hypothesis** or the "truth," however tentative it may be. Given that model, then, we can find out whether one or more regressors are really relevant by the usual t and F tests.

3. Detection of Specification Errors

Test 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 post-mortem, 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 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.

3. Tests of Specification Errors

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:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i \qquad (13.4.4)$$

but a researcher fits the following quadratic function:

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

and another researcher fits the following linear function:

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

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.

3. Tests of Specification Errors

Test for Omitted Variables and Incorrect Functional Form

The Durbin–Watson d Statistic Once Again

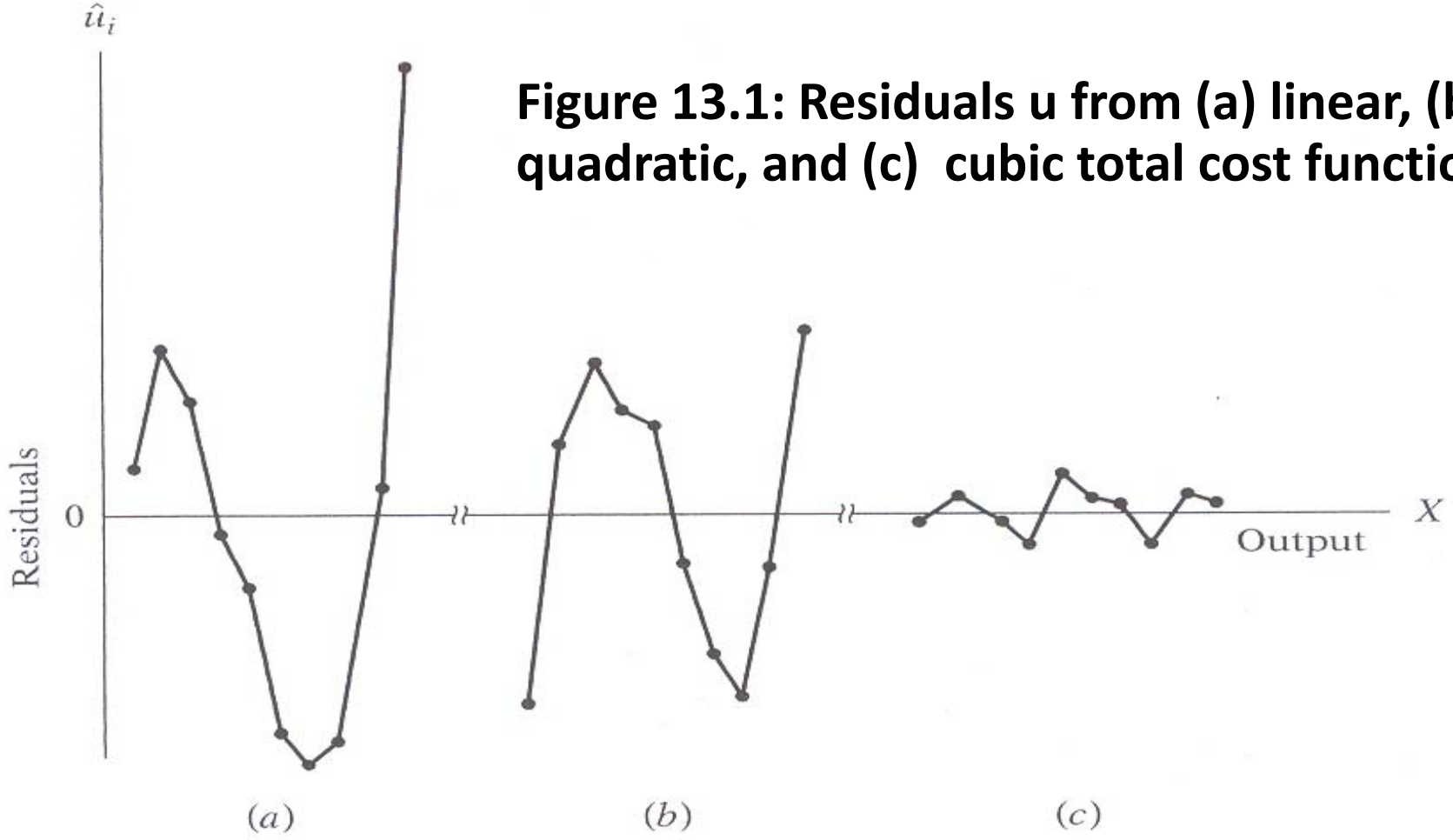


Figure 13.1: Residuals u from (a) linear, (b) quadratic, and (c) cubic total cost functions.

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)
$R^2 = 0.9983$, $\bar{R}^2 = 0.9975$, $d = 2.70$

3. Tests of Specification Errors

Test for Omitted Variables and Incorrect Functional Form

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 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 belongs in the model is included in the error term and needs to be culled out from it and introduced in its own right as an explanatory variable: If we exclude the X_i^3 from the cost function, then as Eq. (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.

3. Tests of Specification Errors

Test for Omitted Variables and Incorrect Functional Form

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

1. From the assumed model, obtain the ordinary least squares (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.

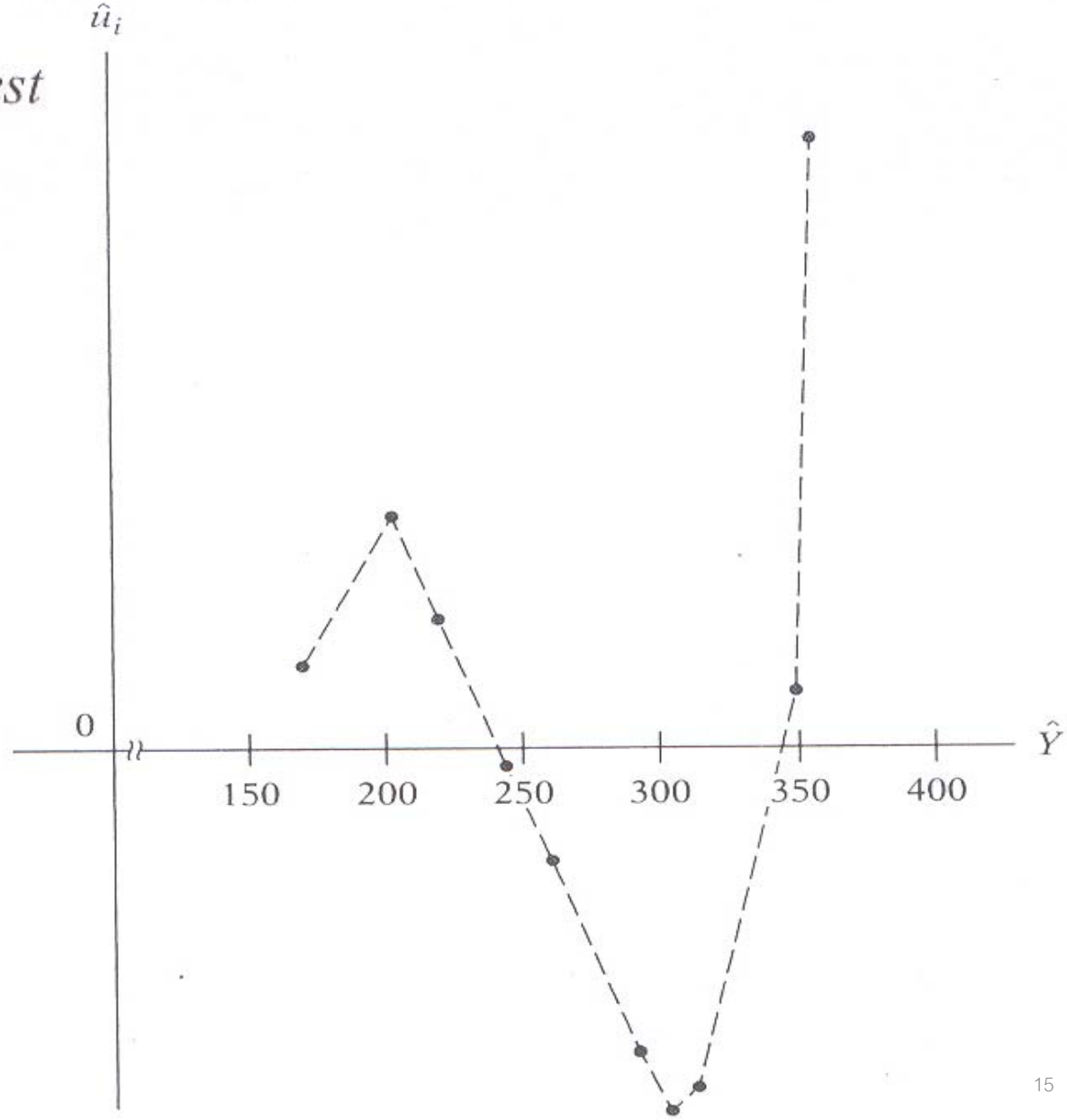
4. From the Durbin–Watson tables, if the estimated d value is significant, then one can accept the hypothesis of model mis-specification. If that turns out to be the case, the remedial measures will naturally suggest themselves.

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 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.

Tests of Specification Errors

Ramsey's RESET Test

Figure 13.2: Residuals u and estimated Y from the linear cost function



Test for Omitted Variables and Incorrect Functional Form

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

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

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 (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 a regressor(s) in Eq. (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:

Test for Omitted Variables and Incorrect Functional Form

Ramsey's RESET Test

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

2. Rerun Eq. (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

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

3. Let the R^2 obtained from Eq. (13.4.7) be R^2_{new} and that obtained from Eq. (13.4.6) be R^2_{old}. Then we can use the F test first introduced in Eq. (8.4.18), namely,

$$F = \frac{\left(R^2_{\text{new}} - R^2_{\text{old}}\right)/\text{number of new regressors}}{\left(1 - R^2_{\text{new}}\right)/(n - \text{number of parameters in the new model})} \qquad (8.4.18)$$

to find out if the increase in R^2 from using Eq. (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.

Test for Omitted Variables and Incorrect Functional Form

Ramsey's RESET Test

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

$$\hat{Y}_i = 166.467 + 19.933X_i \qquad (13.4.8)$$

$$(19.021) \quad (3.066) \qquad R^2 = 0.8409$$

$$\hat{Y}_i = 2140.7223 + 476.6557X_i - 0.09187\hat{Y}_i^2 + 0.000119\hat{Y}_i^3 \qquad (13.4.9)$$

$$(132.0044) \quad (33.3951) \quad (0.00620) \quad (0.0000074)$$

$$R^2 = 0.9983$$

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

Now applying the F test we find

$$F = \frac{(0.9983 - 0.8409)/2}{(1 - 0.9983)/(10 - 4)} \qquad (13.4.10)$$

$$= 284.4035$$

Test for Omitted Variables and Incorrect Functional Form

F is highly significant, indicating that the model (13.4.8) is mis-specified.

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.

In practice, the RESET test's usefulness lies in acting as a general indicator that something is wrong. For this reason, a test such as RESET is sometimes described as a test of misspecification, as opposed to a test of specification. A misspecification test, can detect a range of alternatives and indicate that something is wrong under the null, without necessarily giving clear guidance as to what alternative hypothesis is appropriate.

Test for Omitted Variables and Incorrect Functional Form

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 Eq. (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

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

where v is an error term with the usual properties.

Test for Omitted Variables and Incorrect Functional Form

Lagrange Multiplier (LM) Test for Adding Variables

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.[26] Symbolically, we write

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

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

5. If the chi-square value obtained from Eq. (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.

For our example, the regression results are as follows:

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

Test for Omitted Variables and Incorrect Functional Form

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 Eq. (13.4.13) are regressed as just suggested in Step 3, we obtain the following results:

$$\begin{aligned} \widehat{\hat{u}_i} &= -24.7 \quad + 43.5443X_i - 12.9615X_i^2 + 0.9396X_i^3 \\ \text{se} &= \quad (6.375) \quad (4.779) \quad (0.986) \quad (0.059) \end{aligned} \tag{13.4.14}$$

$$R^2 = 0.9896$$

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 a similar conclusion on the basis of Ramsey's RESET test.

Errors of Measurement

Errors of Measurement in the Dependent Variable Y

Consider the following model:

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

where Y_i^* = permanent consumption expenditure[27]
X_i = current income
u_i = stochastic disturbance term

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

$$Y_i = Y_i^* + \varepsilon_i \qquad (13.5.2)$$

where ε_i denote errors of measurement in Y_i^*. Therefore, instead of estimating Eq. (13.5.1), we estimate

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

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.

Errors of Measurement in the Dependent Variable Y

For simplicity assume that $E(u_i) = E(\varepsilon_i) = 0$, cov $(X_i, u_i) = 0$ (which is the assumption of the classical linear regression), and cov $(X_i, \varepsilon_i) = 0$; that is, the errors of measurement in Y_i^* are uncorrelated with X_i, and 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 β estimated from either Eq. (13.5.1) or Eq. (13.5.3) will be an unbiased estimator of the true β (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 β estimated from Eqs. (13.5.1) and (13.5.3) will be different because, employing the usual formulas (see Chapter 3), we obtain

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

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

Obviously, the latter variance is larger than the former.[28] 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.**

Errors of Measurement in the Explanatory Variable X

Now assume that instead of Eq. (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 Eq. (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.

Errors of Measurement in the Explanatory Variable X

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 $E[z_i] = 0$)

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

Thus, the explanatory variable and the error term in Eq. (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.*[29]

Incorrect Specification of the Stochastic Error Terms

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 Eqs. (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 Eq. (13.2.8) is such that ln u_i satisfies the usual OLS assumptions.

If we assume that Eq. (13.2.8) is the "correct" model but estimate Eq. (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.

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's information criterion (AIC), (4) Schwarz's information criterion (SIC), (5) Mallows's C_p criterion, and (6) forecast χ^2 (chi-square). All these criteria aim at minimizing the residual sum of squares (RSS) (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 trade-off between goodness of fit of the model and its complexity (as judged by the number of regressors).

Model Selection Criteria

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:

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

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.

Model Selection Criteria

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

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

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.

Model Selection Criteria

Akaike's 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:

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

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

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

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 also has been used to determine the lag length in an AR(p) model.

Model Selection Criteria

Schwarz's 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 Eq. (13.9.6) to Eq. (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.

Model Selection Criteria

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 σ^2. But suppose that we only choose p regressors ($p \leq k$) and obtain the RSS from the regression using these p regressors. Let 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:

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

where n is the number of observations.

Model Selection Criteria

We know that $E(\hat{\sigma}^2)$ is an unbiased estimator of the true σ^2. Now, if the model with p regressors is adequate in that it does not suffer from lack of fit, it can be shown[39] that $E(\text{RSS}_p) = (n - p)\sigma^2$. In consequence, it is true *approximately* that

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

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 Eq. (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.