EE325 Ch.9 Specification Error

Read Guajarati Ch. 13

Outline

Model Selection Criteria

According to Hendry and Richard (1983), a model chosen for empirical analysis should satisfy the following criteria:

- Be data admissible
- Be consistent with theory
- Have weakly exogenous regressors
- Exhibit parameter constancy
- Exhibit data coherency
- Be encompassing

Types of Specification Errors

- Omission of a relevant variable
- Inclusion of an unnecessary variable
- Adoption of the wrong functional form
- Errors of measurement

Consequence of Specification Errors

Underfitting a model

Suppose the true model is

But for some reasons, we fit the following model

The consequences of omitting variable X_3 are:

- If the left-out or omitted variable X_3 is correlated with the included variable X_2, that is, r_{23} or the coefficient of correlation between the two variables is non zero and $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are biased as well as inconsistent.

$$\begin{aligned} E(\hat{\alpha}_1) &\neq \beta_1 \\ E(\hat{\alpha}_2) &\neq \beta_2 \end{aligned}$$

The bias does not disappear as the sample size gets larger

- Even if X_2 and X_3 are not correlated, $\hat{\alpha}_1$ is biased but $\hat{\alpha}_2$ is unbiased.
- The disturbance variance (σ^2) is incorrectly estimated
- The conventionally measured variance of $\hat{\alpha}_2$, calculated from $\frac{\sigma^2}{\sum x_{2i}^2}$ is a biased estimator of the variance of true estimator $\hat{\beta}_2$

- In consequence, the usual confidence interval and hypothesis-testing procedures are likely to give misleading conclusions about the statistical significance of the estimated parameters
- As another consequence, the forecasts based on the incorrect model and the forecast (confidence) intervals will be unreliable

Overfitting a model

Suppose the true model is

But for some reasons, we fit the following model

The consequences of adding unnecessary variables are:

- The OLS estimators of the parameters of the incorrect model are all unbiased and consistent, that is

$$\begin{aligned} E(\hat{\alpha}_1) &= \beta_1 \\ E(\hat{\alpha}_2) &= \beta_2 \\ E(\hat{\alpha}_3) &= \beta_3 = 0 \end{aligned}$$

- The error variance, σ^2, is correctly estimated
- The usual confidence interval and hypothesis-testing procedures remain valid
- However, the estimated α's will be generally inefficient, that is, their variances will be generally larger than those of the β's of the true model

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

If we *include* a 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 estimated variances of the coefficients are larger, and as a result our probability inferences about the parameters are less precise.

Test of Specification Errors

Detecting the presence of unnecessary variables

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

- t-test
- F-test

Detecting incorrect model

- Examination of residuals
- The Durbin-Watson d statistics
- Ramsey's RESET test

Examination of residuals

Let's reconsider the cubic total cost of production function. Assume that the true cost function is described as following,
where Y = total cost and X = output

Quadratic function:

Linear function:

Residuals from (a) linear, (b) quadratic, and (c) cubic total cost functions

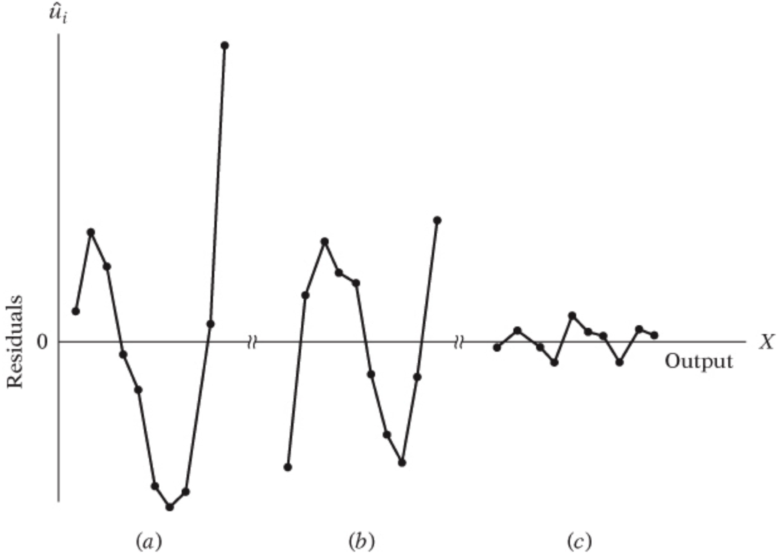


TABLE 13.1
Estimated Residuals from the Linear, Quadratic, and Cubic Total Cost Functions

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

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

$$^{\dagger}\hat{Y}_i = \underset{\substack{(23.488)\\(9.468)}}{222.383} - \underset{\substack{(9.809)\\(-0.818)}}{8.0250X_i} + \underset{\substack{(0.869)\\(2.925)}}{2.542X_i^2} \qquad R^2 = 0.9284,\quad \bar{R}^2 = 0.9079,\quad d = 1.038$$

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

Durbin-Watsin d statistics

Step 1:
From the assumed model, obtain the OLS residuals

Step 2: If it is believed that the assumed model is mis-specified because it excludes a relevant explanatory variable, say X, from the model, order the residuals obtained above in 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

Step 3:
Compute the d statistics from the residuals thus ordered by the usual d formula

Step 4:
From the Durbin-Watson tables, if the estimated d value is significant, then one can accept the hypothesis of model mis-specification

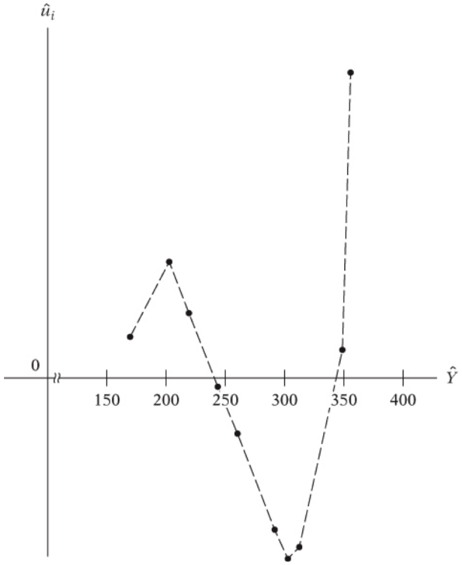
Ramsey's RESET

RESET = Regression Specification Error Test

$$Y_i = \lambda_1 + \lambda_2 X_{2i} + u_{3i}$$

where Y = total cost and X = output

If we plot the residuals ($\hat{u}_i$) obtained from this regression against $\hat{Y}_i$


$\hat{u}_i$
0
150
200
250
300
350
400
$\hat{Y}$

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), it should increase R^2 and if such increase is statistically significant, it would suggest that the linear cost function was mis-specified

The procedure of the RESET are
Step 1:
From the chosen model, obtain R^2

Step 2:
Rerun equation introducing $\hat{Y}_i$ in some forms as an additional regressor(s)

Step 3:
Use F-test

$$F =$$

Step 4:
If the computed F value is significant, say at the 5 percent level, the model is mis-specified.

Example:

1. $\hat{Y}_i = 166.467 + 19.933X_i$
$R^2 = 0.8409$ and $n = 10$

2. $\hat{Y}_i = 2140.7223 + 476.6557X_i - 0.09187\hat{Y}_i^2 + 0.000119\hat{Y}_i^3$
$R^2 = 0.9983$

H_o :
H_a :
3. $F =$
4.

Reject null hypothesis as the computed F value is significant, say at 5 percent level, the model is mis-specified

Errors of Measurement

1. Errors of measurement in the Dependent variable Y
2. Errors of measurement in the Explanatory variable X

Errors of measurement in the dependent variable Y

Consider the following model

$$Y_i^* = \alpha + \beta X_i + u_i$$

where
Y_i^* = permanent consumption expenditure
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

$$Y_i = Y_i^* + \epsilon_i$$

where ϵ_i denotes error of measurement in Y_i^*

$$\begin{aligned} Y_i &= (\alpha + \beta X_i + u_i) + \epsilon_i \\ &= (\alpha + \beta X_i) + (u_i + \epsilon_i) \\ &= (\alpha + \beta X_i) + \nu_i \end{aligned}$$

where $\nu_i = (u_i + \epsilon_i)$ is a composite error term

Assume that

$$\begin{aligned} E(u_i) &= E(\epsilon_i) = 0 \\ Cov(X_i, \epsilon_i) &= 0 \\ Cov(u_i, \epsilon_i) &= 0 \end{aligned}$$

With these assumptions, it can be seen that estimated β will be an unbiased estimator of the true β

The errors of measurement in the dependent variable Y do not destroy the unbiasedness property of OLS estimators

The variances and standard errors of β estimated

$$
\begin{aligned}
Y_i^* &= \alpha + \beta X_i + u_i \\
var(\hat{\beta}) &= \frac{\sigma_u^2}{\sum x_i^2} \\
Y_i &= (\alpha + \beta X_i) + \nu_i \\
var(\hat{\beta}) &= \frac{\sigma_\nu^2}{\sum x_i^2} = \frac{\sigma_u^2 + \sigma_\epsilon^2}{\sum x_i^2}
\end{aligned}
$$

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

Consider the following model

$$Y_i = \alpha + \beta X_i^* + u_i$$

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

Suppose instead of observing X_i^*, we observe

$$X_i = X_i^* + w_i$$

where w_i represents error of measurement in X_i^*

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

where $z_i = (u_i - \beta w_i)$ is a composite error term

Now even if we assume that w_i has a zero mean, is serially independent and 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

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

The explanatory variable and the error term are correlated, which violates the crucial assumption of the CLRM 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

Appendix 13.A.3

$$p\lim \hat{\beta} = \beta[\frac{1}{1+\sigma_w^2/\sigma_{X^*}^2}]$$

where σ_w^2 and $\sigma_{X^*}^2$ are variance of w_i and X_i^*
Even if the sample size increases indefinitely, $\hat{\beta}$ will not converge to β

Gujarati, D.N. (2009) Basic Econometrics. 5th ed. Singapore, McGraw-Hill.