EE 325

Econometric Modeling
Model Specification and
Diagnostic Testing

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

Type of specification errors

1. Omission of a relevant variable (s)
2. Inclusion of an unnecessary variable (s)
3. Adoption of the wrong functional form
4. Errors of measurement

Consequences of Model Specification Error

Underfitting a model
(Omitting a relevant variable)

Underfitting a model (Omitting a relevant variable)

Suppose the true model is

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

but for some reason we fit the following model:

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

The consequences of omitting variable X_3 are as follows:

- If the left-out, or omitted, variable X_2 is correlated with the included variable X_3, 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

$$E(\hat{\alpha}_1) \neq \beta_1$$
$$E(\hat{\alpha}_2) \neq \beta_2$$

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, although $\hat{\alpha}_2$ is now unbiased
- The disturbance variance σ^2 is incorrectly estimated
- 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 $\hat{\beta}_2$

- 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*

Inclusion of an irrelevant variable (Overfitting a model)

Underfitting a model (Omitting a relevant variable)

Suppose the true model is

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

but for some reason we fit the following model:

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

The consequences of specification error are as follows:

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

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

- 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 $\hat{\beta}$'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 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 estimated variances of the coefficients are larger, and as a result our probability inferences about the parameters are less precise.

Tests of Specification Errors

Detecting the presence of unnecessary variables (Overfitting a model)

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

$$Y_i = \beta_1 + \beta_2 X_i + \ldots + \beta_k X_{ki} + u_i$$

- t test $\quad t = \dfrac{\hat{\beta}_k}{se(\hat{\beta}_k)}$
- F-test

Tests for Omitted Variables & Incorrect Functional Form

- Examination of residuals
- The Durbin-Watson d Statistic
- Ramsey’s RESET test

Examination of residuals

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

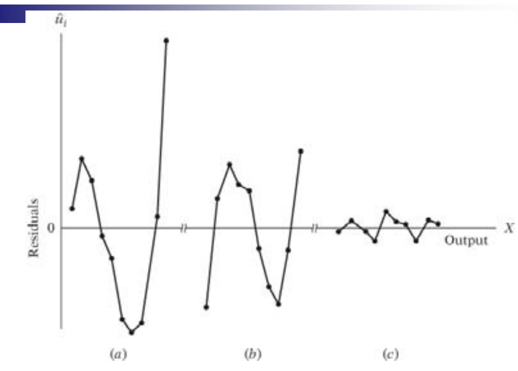
$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i$$

quadratic function:

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

Linear function:

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



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

*$\hat{Y}_i = 166.467 + 19.933X_i$ (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); $R^2 = 0.9983$, $\bar{R}^2 = 0.9975$, $d = 2.70$

The Durbin-Watson d Statistic

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,

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

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

Ramsey's RESET Test

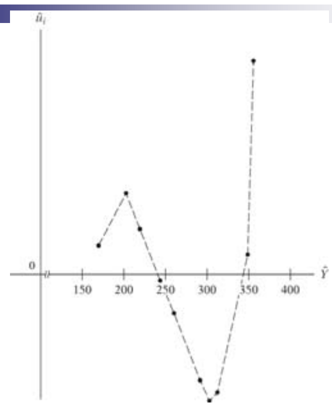
RESET =Regression specification error test

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

Where Y= total cost

X=output

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



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 the increase R^2 is statistically significant, it would suggest that the linear cost function was mis-specified.

The steps involved in RESET are as follows:

1. From the chosen model, obtain R^2
   $$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$
2. Rerun equation introducing $\hat{Y}_i$ in some form as an additional regressor (s).
   $$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \delta_1 \hat{Y}^2 + \delta_2 \hat{Y}^3 + u$$
3. Use F-Test

$$F = \frac{(R^2_{new} - R^2_{old}) / number\ of\ \ new\ regressors}{(1 - R^2_{new}) / (n - number\ of\ \ parameters\ in\ the\ Model)}$$

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

Example

$$\hat{Y}_i = 166.467 + 19.933 X_i$$
$$se = (19.021) \quad (3.066) \qquad R^2 = 0.8409$$

$$\hat{Y}_i = 2140.7223 + 476.6557 X_i - 0.09187 \hat{Y}_i^2 + 0.000119 \hat{Y}_i^3$$
$$se = (132.0044) \quad (33.3951) \quad (0.00620) \quad (0.0000074) \quad R^2 = 0.9983$$

Hypothesis testing

$$H_0 : Model\ is\ not\ mis-specified$$
$$H_1 : otherwise$$

$$F = \frac{(0.9983 - 0.8409)/2}{(1-0.9983)/(10-4)} = 284.4035$$

$$F > critical\ F_{(number\ of\ new\ regressors, n-number\ of\ parameter\ in\ the\ new\ model)}$$

Reject null hypothesis

The computed F value is significant, say, at the 5 percent level, the model is mis-specified

Error of measurement

- Errors of measurement in the Dependent variable Y
- Errors of measurement in the Explanatory variable X

Errors of measurement in the Dependent variable Y

Consider the following model

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

$Y_i^* =$ 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

$$Y_i = Y_i^* + \varepsilon_i$$

where ε_i denote errors of measurement in Y_i^*

$$Y_i = (\alpha + \beta X_i + u_i) + \varepsilon_i$$

$$= \alpha + \beta X_i + (u_i + \varepsilon_i)$$

$$= \alpha + \beta X_i + v_i$$

where $v_i = u_i + \varepsilon_i$ is a composite error term

Assume that

$$E(u_i) = E(\varepsilon_i) = 0$$

$$Cov(X_i, u_i) = 0$$

$$Cov(u_i, \varepsilon_i) = 0$$

With these assumptions, it can be seen that β estimated 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

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

$$\text{var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum x_i^2}$$

$$Y_i = \alpha + \beta X_i + v_i$$

$$\text{var}(\hat{\beta}) = \frac{\sigma_v^2}{\sum x_i^2} = \frac{\sigma_u^2 + \sigma_\varepsilon^2}{\sum x_i^2}$$

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

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

$Y_i =$ current consumption expenditure

$X_i^* =$ permanent income

$u_i =$ disturbance term

Suppose instead of observing X_i^* , we observe

$$X_i = X_i^* + w_i$$

where w_i represents errors of measurement in X_i^*

We estimate

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

Now even if we assume that w_i has a zero mean, is serially independent, and is uncorrelated with u_i , we can no longer assume that the composite error z_i term X_i is independent of the explanatory variable because

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

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

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

where σ_w^2 and $\sigma_{X^*}^2$ are variance of w_i and X_i^*

Even is the sample size increases indefinite $\hat{\beta}$ will not converge to β

Source

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