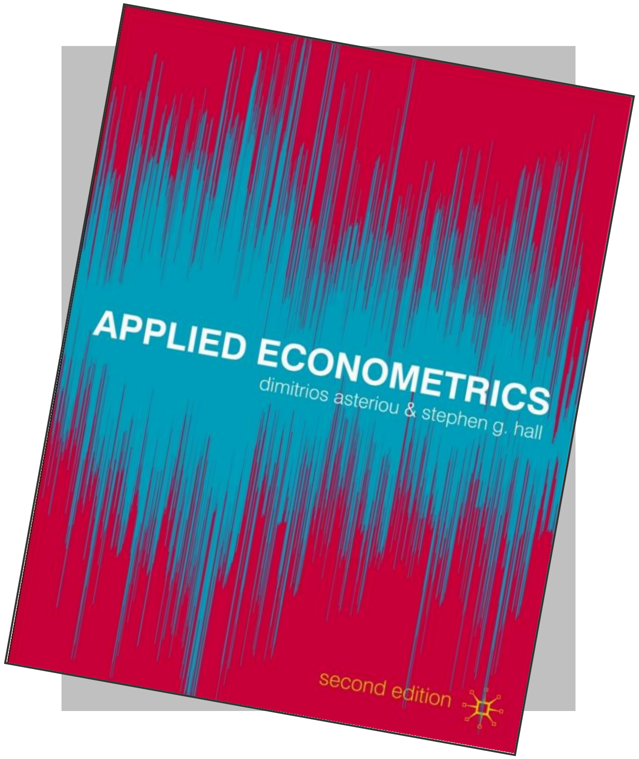
Applied Econometrics
Second edition

Dimitrios Asteriou and
Stephen G. Hall

Chapter 8:
Misspecification



palgrave
macmillan

Misspecification

1. Omitting influential or including non-influential explanatory variables
2. Various functional forms
3. Measurement errors
4. Tests for misspecification
5. Approaches in choosing an appropriate model

Learning Objectives

1. Various forms of possible misspecification in the CLRM
2. Appreciate the importance and learn the consequences of omitting influential variables in the CLRM
3. Distinguish among wide range of functional forms and understand meaning & interpretation of their coefficients
4. Understand importance of measurement errors in data
5. Perform misspecification tests using econometric software
6. Understand meaning of nested and non-nested models
7. Be familiar with the concept of data mining and choose an appropriate econometric model.

Omitting Influential Variables

Omitting influential variables from a regression model causes these variables to become part of the error term. Therefore one or more of the assumptions of the CLRM will be violated.

Consider the population regression function:

$$Y=\beta_1+\beta_2 X_2+ \beta_3 X_3+u$$

where $\beta_2 \neq 0$ and $\beta_3 \neq 0$, and assume this as correct.

Omitting Influential Variables (2)

However, we estimate the following:

$$Y=\beta_1+\beta_2 X_2+u$$

where X_3 is wrongfully omitted.

Then, the error term of this equation is:

$$u=\beta_3 X_3+e$$

It is clear that the assumption that the error term has a zero mean is now violated:

$$E(u)=E(\beta_3 X_3+e)=E(\beta_3 X_3)+E(e)=E(\beta_3 X_3)\neq 0$$

Omitting Influential Variables (3)

Furthermore, if the excluded variable X_3 happens to be correlated with X_2 then the error term is no longer independent of X_2.

This results in estimators of β_2 and β_3 to be biased and inconsistent.

This is called **omitted variable bias**.

Including Non-influential Variables

This is the opposite. The correct model is:

$$Y=\beta_1+\beta_2X_2+u$$

and we estimated:

$$Y=\beta_1+\beta_2X_2+ \beta_3X_3+e$$

where X_3 is wrongly included in the model.

Including Non-influential Variables (2)

As X_3 does not belong to the correct model, its population coefficient should be equal to zero (i.e. $\beta_3=0$).

If $\beta_3=0$ then none of the CLRM assumptions is violated and OLS estimators are both unbiased and consistent.

However, it is unlikely that they are efficient.

If X_2 is correlated with X_3 then an additional unnecessary element of multicollinearity will be introduced.

Omission and Inclusion at the same time

In this case the correct model is:

$$Y=\beta_1+\beta_2X_2+\beta_3X_3+v$$

and we estimate:

$$Y=\beta_1+\beta_2X_2+\beta_4X_4+w$$

Now we understand the problems that this double mistake causes.

The Plug-in Solution

Sometimes it is possible to encounter omitted variable bias when a key variable that affects Y is not available.

For example, consider a model where the monthly salary of an individual is associated with:

- whether or not he/she is male/female
- years he/she has spent in education

The Plug-in Solution (2)

Both these factors can be quantified and included in the model.

However, if we also assume that the salary level can be affected by the socio-economic environment in which each person was brought up, then this is hard to measure in order to be included in the model:

$$(salary) = \beta_1 + \beta_2(sex) + \beta_3(educ) + \beta_3(background) + u$$

The Plug-in Solution (3)

Not including the *background* variable in the model leads to biased estimates of β_1 and β_2.

Main aim, however, is to get appropriate estimates for those two coefficients (i.e. not so concerned with β_3 because we will never get the appropriate coefficient for that).

A way to resolve that is to include an alternative proxy variable for the omitted variable.

The Plug-in Solution (4)

For example, family income.

Family income is not, of course, exactly what we mean by *background* but is definitely a variable that is highly correlated with that.

The Plug-in Solution (5)

To illustrate this, consider the model:

$$Y=\beta_1+\beta_2X_2+\beta_3X_3+\beta_4X^*_4+u$$

where X_2 and X_3 are observed, X^*_4 is unobserved.

We know that:

$$X^*_4=\delta_1+\delta_2X_4+e$$

where error term e should be included because not exactly the same and δ_1 is also included to allow them to be measured on a different scale. Need variables that are positively correlated (i.e. $\delta_2>0$)

The Plug-in Solution (6)

So we estimate:

$$Y=\beta_1+\beta_2X_2+\beta_3X_3+\beta_4(\delta_1+\delta_2X_4+e)+u$$

$$=(\beta_1+\beta_4\delta_1)+\beta_2X_2+\beta_3X_3+\beta_4\delta_2X_4+(\beta_4e+u)$$

$$=\quad a_1 \qquad +\beta_2X_2+\beta_3X_3+a_4X_4+w$$

By estimating this model we do not get unbiased estimates for β_1 and β_4, but unbiased estimators for a_1, β_2, β_3 and a_4

Various Functional Forms

- Linear $Y=\beta_1+\beta_2X_2$
- Linear-log $Y=\beta_1+\beta_2 lnX_2$
- Reciprocal $Y=\beta_1+\beta_2\,(1/X_2)$
- Quadratic $Y=\beta_1+\beta_2X_2+\beta_3X^2{}_2$
- Interaction $Y=\beta_1+\beta_2X_2+\beta_3X_2Z$
- Log-linear $lnY=\beta_1+\beta_2X_2$
- Double log $lnY=\beta_1+\beta_2 lnX_2$

Box-Cox Transformation

The choice of functional form plays an important role; thus, we need a formal test of comparing alternative models (functional forms).

If we have the same dependent variable things are easy: estimate both models and choose the one with the higher R^2.

However, if the dependent variables are different an immediate comparison is impossible.

Box-Cox Transformation (2)

Assume we have the two models:

$$Y=\beta_1+\beta_2 X_2 \qquad \text{and} \qquad \ln Y=\beta_1+\beta_2 \ln X_2$$

In these cases we need to scale the Y variable in such a way that we will be able to compare the two models.

The procedure to do that is called the Box-Cox Transformation.

Box-Cox Transformation (3)

Step 1: Obtain the geometric mean of the sample Y values:

$$Y' = (Y_1 Y_2 Y_3 \ldots Y_n)^{1/n} = \exp[(1/n)\Sigma \ln Y)$$

Step 2: Transform the sample *Y* values by dividing each of them by *Y'* obtained from step 1 to get:

$$Y^* = Y_i / Y'$$

Step 3: Estimate both models with Y^* as the dependent variable. The equation with the lower *RSS* should be preferred.

Step 4: To check whether it is significantly better, calculate $(1/2\ n)\ln(RSS_2/RSS_1)$ and check with the chi-square distribution. RSS_2 is the one with the lower.

Measurement Errors

Sometimes data not measured appropriately.

Can have measurement errors either in the dependent, or the explanatory variables, or both.

If in the dependent, there are larger variances of the OLS coefficients. Unavoidable.

If in the explanatory variables, there are biased and inconsistent estimators. Totally wrong results.

Tests for Misspecification

We have the following tests:

- Test for normality of the residuals
- The Ramsey RESET test
- Tests for Non-nested Models

Normality of Residuals

Step 1: Calculate the Jarque-Berra (*JB*) Statistic (given in EViews)

Step 2: Find the chi-square critical value from the corresponding tables

Step 3: If *JB*>chi-square critical reject the null hypothesis of normality

Ramsey Reset Test

Step 1: Estimate the model you think correct and obtain fitted values of *Y*, call them *Y'*

Step 2: Estimate the model of step 1 again, this time including Y'^2 and Y'^3 as additional explanatory variables

Step 3: Model in step 1 is restricted model and in step 2 unrestricted model. Calculate F-statistic for these two models

Step 4: Compare F-statistic with F-critical and conclude (if F-stat>F-crit reject the null of correct specification)

Tests for Non-nested Models

To test models which are not nested do not use the F-statistic approach.

Non-nested are the models in which neither equation is a special case of the other, i.e. we don't have restricted and unrestricted models.

Suppose for example that we have the following:

$$Y=\beta_1+\beta_2X_2+\beta_3X_3+u \qquad (1)$$

$$Y=\beta_1+\beta_2 lnX_2+\beta_3 lnX_3+u \qquad (2)$$

Tests for Non-nested Models (2)

One approach (Mizon and Richard) suggests the estimation of a comprehensive model of the form:

$$Y= \delta_1+ \delta_2 X_2 + \delta_3 X_3+ \delta_4 \ln X_2 +\delta_5 \ln X_3+e$$

and then to apply an F-test for significance of δ_4 and δ_5 having as restricted model equation (1).

Tests for Non-nested Models (3)

A second approach (Davidson and McKinnon) suggests that if model (1) is true then the fitted values of (2) should be insignificant in (1) and vice versa.

So they suggest the estimation of:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \delta Y^* + e$$

where Y^* is the fitted values of model (2).

A simple t-test of the coefficient of Y^* can conclude.

Choosing the Appropriate Model

Two major approaches:

Traditional view: Average Economic Regressions (AER)

- Hendry's General to Specific Approach

Choosing the Appropriate Model (2)

- AER essentially starts with simple model and then 'builds up' the model as the situation demands. Also called simple to specific.
- Two disadvantages:

(a) Suffers from data mining. Only final model presented by the researcher.

(b) Alterations to original model done in arbitrary manner based on beliefs of researcher.

Choosing the Appropriate Model (3)

Hendry approach starts with general model that contains – nested within it as special cases – other simpler models, and then appropriate tests to narrow down the model to simpler ones.

The model should be:

(a) Data admissible

(b) Consistent with the theory

(c) Use regressors not correlated with error term

(d) Exhibit parameter constancy

(e) Exhibit data coherency

(f) Encompassing, meaning include all possible rival models