

[Academic Script]

Specification Errors

Subject:

Course:

Paper No. & Title:

Unit No. & Title:

Business Economics

B. A. (Hons.), 5th Semester, Undergraduate

Paper – 531 Elective Paper Q1 – Advanced Econometrics

Unit – 1 Relaxing the Assumptions of The Classical Linear Model

Lecture No. & Title:

Lecture – 6 Specification Errors

Academic Script

1. What are Specifications errors?

Many time it happens that when we do modelling and make applications then there are certain errors which are known as Specification Error. They happen due to inclusion of an unnecessary variable as well as exclusion of an avoidable variable. Similarly sometimes we may use alternative form of model rather than the correct form of the model. All this things are responsible for specification errors. In the same way sometime we have some errors in measurement due to which the model will changed and conclusion will changed so they are known as measurement errors. All this things you will learn in detail in this lecture.

Strictly speaking the term specific error refers to any mistake in the set of assumptions underlying the model and the associated inference procedures, but it has come to be used particularly for errors in specifying the data matrix *x* representing the given set of explanatory variables. Economic theory can normally indicate the set of explanatory variables corresponding to any assumed model (utility maximization, cost minimization, profit maximization, production functional form etc.), but theory cannot usually indicate the precise form of the relationship. Even there may not be clear guideline to the relevant explanatory variables.

Above all, one may not be able to obtain the measurements on appropriate variables and hence have no proxy variables in their place. In turn this leads to the problem of misspecification and hence the specification errors and specification bias.

2 Types of Specification errors

The assumption of the classical general linear model is that the econometric method used in analysis is correctly specified. This has two meanings (1) there are no equation specification errors and (2) there are no model specification errors.

Broadly speaking the problem of specification errors arises due to one or more of the following reasons

(A) Omission of a relevant variable

(B) Inclusion of an unnecessary variable

(C) Adopting the wrong functional form

(D) Errors of measurement

Before going into the details of the above, let us understand them by means of Some illustrations

Illustration (1)

Let us consider the following model

 $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + U_{1i} \quad \dots \quad (1)$

Where *Y* = total cost of production

X = Output

Equation (1) is very commonly known as the cubic cost function. But suppose that the researcher decides to use the following model

 $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + U_{2i} \quad$ (2)

Since theoretically (for a given situation) equation (1) is a very well specified form of the total cost function. Hence if we use equation (2) in place of equation (1), we would constitute a specification error. This is due to omitting the relevant variable X_i^3 .Comparing (1) and (2), actually we have

 $U_{2i} = U_{1i} + \beta_4 X_i^3 \dots (3)$

Illustration (2)

Now suppose that some other researcher proposes the following model (instead of original model given in equation (1) above)

 $Y_{i} = \lambda_{1} + \lambda_{2} X_{i} + \lambda_{3} X_{i}^{2} + \lambda_{4} X_{i}^{3} + \lambda_{5} X_{i}^{4} + U_{3i} \dots$ (4)

Comparing (4) with (1) we find that if model given in equation (1) is the correct one, we commit specification error by means of inclusion of an unnecessary variable

This results in specification error and specification bias.

Illustration (3)

Suppose that someone postulates the following model

Where $L_n Y_i = \text{logarithm of } Y_i$ to the natural base e.

Since equation (1) is assumed to be the true model, we commit a specification error and specification bias resulting from the use of the wrong functional form. This is due to the fact that in equation (1) Y occurs in linear form, whereas in equation (6) Y occurs in logarithmic form.

Illustration (4)

Next suppose that the researcher uses the following model

Here \in_i and w_i are the errors of measurements. In this case, instead of using the actual measurements Y_i and X_i on the variables Y and X, we use their proxies Y_i^* and X_i^* due to the errors of measurement in these variables. Hence we commit here specification error and specification bias due to the errors of measurement. Note that the above 4 types of errors are also called model specification Errors.

From the above illustrations we may summarise that specification errors occur due to the above mentioned reasons and that in turn affects the analysis and inference procedure.

3. Consequences of specification errors

Let us consider here the consequences that occur due to first two types of above specification errors in detail. This is broadly illustrated by means of 3 variables model and it can be extended further for k variables model using matrix algebra.

(I) Omitting a Relevant Variable

This is also called underfitting a model.

Consider the model $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$ (9) Instead of the above, suppose that we fit the following model

 $Y_i = \alpha_1 + \alpha_2 X_{2i} + \vartheta_i \quad \dots \quad (10)$

Here variable X_3 is omitted.

Then $\vartheta_i = \beta_3 X_{3i} + U_i$

In this situation we find the following

(1) If X_3 is correlated with X_2 , that is if $r_{23} \neq 0$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are biased and also inconsistent. The bias does not disappear even with large sample.

(2) Even if X_2 and X_3 are uncorrelated $(r_{23} = 0)$, $\hat{\alpha}_1$ is still biased but $\hat{\alpha}_2$ will be unbiased.

(3) The disturbance variance σ^2 is incorrectly estimated.

(4) We find that
$$V(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2}$$

and $V(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1-r_{22}^2)}$

Both are same if $r_{23} = 0$. Otherwise $V(\hat{\alpha}_2)$ is a biased estimator of the variance of the true estimator of $\hat{\beta}_2$.

(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. As a result, we have a conclusion that once the model is formulated on the basis of the relevant theory, it is incorrect to drop a variable from the model. (II) Inclusion of an Irrelevant variable

This is also called overfitting a model.

Suppose that we have a model expressed as

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

which is the correct model to be considered.

Instead of this let us suppose that we are fitting the model

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

Here we add irrelevant variable X_3 in the model.

This gives rise to specification error.

In the above model as compared with the original one, we have $U_i = \alpha_3 X_{3i} + \vartheta_i$

We have the following Consequences in this presentation of the model

(1) The OLS estimator of the parameters of the incorrect model are all unbiased and consistent, that is $E(\hat{\alpha}_1) = \beta_1$, $E(\hat{\alpha}_2) = \beta_2$ 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) We have the relation, using OLS formula

$$V\left(\hat{\beta}_{2}\right)=\frac{\sigma^{2}}{\sum x_{i}^{2}}$$

and $V(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{22}^2)}$

Hence $\frac{V(\hat{a}_2)}{V(\hat{a}_2)} = \frac{1}{(1-r^2)}$

$$V(\beta_2)$$
 $(1-r_{23}^2)$
Since $0 \le r_{23}^2 \le 1$, $V(\hat{\alpha}_2) \ge V(\hat{\beta}_2)$.

Thus even though \hat{a}_2 is unbiased for β_2 , its variance is greater

than $V(\hat{\beta}_2)$.

Similar result holds good for \hat{a}_1 also.

Hence the estimated α'_{s} will be generally inefficient, that is, their variances will be generally larger than those of the true model. The implication of this discussion 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. Similar thing is about $\hat{\alpha}_{1}$ also.

From this one may think that it is better to include irrelevant variables rather than omitting the relevant ones. This is also not true, because addition of unnecessary variables will lead to loss in efficiency of the estimators and may also lead to problem of multicollinearity.

Hence the best approach can be to include only explanatory variables which on theoretical grounds influence the dependent variable and that are not accounted for by other included variables.

2. How to detect about specification errors?

After understanding the concepts of Specification errors and its consequences, we want to know in brief about the detection of specification phenomena. There are certain tests which may be found to be useful to detect the problem of specification. We study here some tests in brief.

(1) Durbin – Watson <u>d</u> statistic

To use DW test for detecting model specification errors, we adopt the following procedure.

(a) From the assumed model, obtain the residuals using OLS method.

(b) If it is believed that the assumed model is mis-specified because it excludes a relevant explanatory variable, say z, from the model, then order the residuals obtained in step (a) above

according to increasing value of *z*. (Note that *z* variable could be one of the *x* variables included in the assumed model or it could be some function of *x* like x^2, x^3 etc.)

(c) Now computer DW d statistic from the residuals thus ordered by the usual d formula stated as under

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

(d) Use Durbin Watson tables for *d* statistic.

If we find that the estimated *d* value is significant then one can accept the hypothesis of model misclassification. Hence use the remedial measure which can be judged easily from the model presented.

(2) Ramsey's test

It is also called Ramsey's RESET test (RESET means Regression Specification Error Test)

Suppose that we have the following model

Where Y = Total cost, X = output.

We carry out test as under

(a) From the above chosen model run regression and find \hat{Y}_i by OLS method.

(b) Also obtain \hat{v}_i and plot \hat{v}_i values against \hat{Y}_i values. This suggests the type of new relation to be included in the original model as per nature of the diagram.

(c) Suppose that we observe curvilinear regression when we plot the diagram.

(d) We introduce variables which are functions of \hat{Y}_i in the original model.

(e) Suppose that we pose the new model as

(f) Let R^2 obtained from (I) be R_{old}^2 and R^2 obtained from (II) be R_{new}^2 then we compute

$$F = \frac{\frac{(R_{new}^2 - R_{old}^2)}{(Number of new regressors)}}{\frac{(1 - R_{new}^2)}{(n - number of parameters in the new model)}}$$

(g) Carry out *F* test. If *F* is found to be significant, we can accept the hypothesis the given model in (I) is misspecified.

Ramsey's test is simpler to apply as it does not require one to specify what the alternative model is. But this is also a disadvantage because knowing that the model is mis specified does not necessarily help in choosing a better alternative.

(3) Lagrange Multiplier test (LM test)

This test is an alternative to Ramsey's RESET test.

Suppose that the true model is

which is unrestricted regression and We can assume the restricted regression form of the above by writing

The restricted regression in (II) above assumes that the coefficients of squared and cubic terms of *x* are Zero. We want to test this by using LM test which can briefly be summarized as under.

(a) Estimate the restricted regression given in (II) above and find the residuals \hat{u}_i . Using OLS method

(b)If in fact the unrestricted regression given in (I) above is the true regression, the residuals obtained in (II) should be related to the squared and cubic terms that is X_i^2 and X_i^3 .

(c) This suggests that we regress \hat{u}_i obtained in step (a) above against all the regressors (including those in the restricted regression).

Thus we have the model form as

 $\hat{u}_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + \alpha_4 X_i^4 + \vartheta_i \quad \dots \quad (III)$

Where ϑ is the error term having usual properties

(d) For large samples, Engle has shown that if we compute R^2 from auxilliary regression in (III) above then nR^2 follows χ^2 distribution.

Thus $nR^2 \widetilde{asy} \chi_{\vartheta}^2$ where ϑ is degrees of freedom which is equal to number of restrictions. (asy refers to asymptotically, that is in large samples).

(e) If χ^2 obtained above is significant at the chosen Level of significance, we reject the restricted restriction, otherwise we do not reject it (i.e. may accept the case of restricted restriction). Note that LM test is also a very much similar test as Ramsey's RESET test.

3. Measurement Errors

When there remains some errors in the actual measurements of the variables in the system, it affects widely the estimates of the parameters and the predictions based upon the estimates are also affected. Hence it is always desirable to consider such measurement errors and then proceed further for the forecasting work on the basis of the fitted model.

Case of two variables model

Let us define $x = \xi + u$ (1)

 $Y = \eta + \vartheta \quad \tag{2}$

Where ξ = Actual (correct) value of *x*

x = Measurement of the variable *x*

u = Error in the measurement of X

 η = Actual (correct) value of Y

Y = Measurement of Y

 ϑ = Error in the Measurement of Y

We also assume a linear relation between the actual values of x and y given by

 $\eta = \alpha + \beta \xi \dots (3)$ Put η and ξ in (3) from (1) and (2) $Y - \vartheta = \alpha + \beta (X - u)$ Hence $Y = \alpha + \beta X + \vartheta - \beta u$ We write $Y = \alpha + \beta X + e \dots (4)$ Where $e = \vartheta - \beta u$ Note that E(e) = 0 $COV(e, X) = E(u\vartheta) - \beta E(u^2)$ $E(u) = 0, E(u\vartheta) = 0, E(u^2) = V(u)$ $\therefore COV(e, X) = -\beta V(u) \neq 0$

Thus e and x are not independent. Due to this if we estimate the parameters α and β by least squares method, we get only biased estimators and this bias increases with increasing n.

LSE of
$$\beta$$
 is given by $\hat{\beta} = b = \frac{\sum x_i y_i}{\sum x_i^2}$

So that
$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Upon substitution from the earlier equations Since $u \& \vartheta$ are independent and also ξ , η , u and ϑ are mutually independent, as $n \to \infty$, the probability limit of b_n (written $asplimb_n$) is given by the relation after simplification as

$$plimb_n = \frac{\beta}{1 + \sigma_u^2 / \sigma_\xi^2}$$

```
Note that plimb_n \neq \beta
```

Thus the estimator of β is inconsistent. (In fact it will be less than β).

How to estimate parameters in the presence of measurement errors?

There are some methods for estimation which are as listed below

(1) Method of Grouping of Observations

(2) Method of Instrumental Variables

(3) Method of Maximum likelihood

Here we consider briefly the first two methods.

Method of Grouping of Observations

There are two methods shown briefly as under.

(I) Wald's Method

Let there be a set of n observations on the variables X and Y for two variables model.

We assume that n = 2m, so that the observations are divided into two groups, each containing m observations($X_1, X_2, ..., X_m, X_{m+1}, X_{m+2}, ..., X_n$) and ($Y_1, Y_2, ..., Y_m, Y_{m+1}, Y_{m+2}, ..., Y_n$).

We arrange first the observations on the explanatory variable Xin the ascending order, so that we get the ordered set of nvalues on X given by $(X_{(1)}, X_{(2)}, \dots, X_{(m)}, X_{(m+1)}, X_{(m+2)}, \dots, X_{(n)})$ such that $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(m)} \leq X_{(m+1)} \dots \leq X_{(n)}$

Now we can write down the corresponding observations of the variable Y accordingly. These are given by $(Y_1^*, Y_2^*, \dots, Y_m^*, Y_{m+1}^*, \dots, Y_n^*)$ Thus we get two groups as under

Group	Observations	Observations on
	on <i>x</i>	Y

Ι	$X_{(1)}, X_{(2)}, \dots X_{(m)}$	$Y_1^*, Y_2^*, \dots Y_m^*$
II	X _(m+1) ,	$Y_{m+1}^*, Y_{m+2}^*, \dots Y_n^*$
	$X_{(m+2)},\ldots X_{(n)}$	

Now compute

$$\bar{X}_{(1)} = \frac{\sum_{1}^{m} X_{(i)}}{m}$$
$$\bar{X}_{(2)} = \frac{\sum_{m+1}^{n} X_{(i)}}{m}$$
$$\bar{Y}_{1}^{*} = \frac{\sum_{1}^{m} Y_{i}^{*}}{m}$$
$$\bar{Y}_{2}^{*} = \frac{\sum_{m+1}^{n} Y_{i}^{*}}{m}$$

and Let $\bar{X}_0 = \frac{\sum_{1}^{n} X_{(i)}}{n}$ and $\bar{Y}^* = \frac{\sum_{1}^{n} \bar{Y}^*_i}{n}$

Wald suggested the following estimate of β

$$\hat{\beta} = \frac{\bar{Y}_{(1)}^* - \bar{Y}_{(2)}^*}{\bar{X}_{(1)} - \bar{X}_{(2)}}$$

And $\hat{\alpha} = \bar{Y}^* - \hat{\beta}\bar{X}_0$

Which are based upon the above two groups of n = 2m observations

If n = 2m + 1, then first we arrange the observations on x series in order, so that the middle most observation is the median. We can delete this observation and then take the corresponding observations of Y^* values and apply this method with 2m observations.

(I) Bartlett's Method

This method is a slight modification of Wald's method. Here the observations are divided into 3 groups, each containing k

observations, and obtain 3 groups after ordering as shown before

Group	Observations on X	Observations on Y
Ι	$X_{(1)}, X_{(2)}, \dots X_{(K)}$	$Y_1^*, Y_2^*, \dots Y_K^*$
II	$X_{(k+1)}, X_{(k+2)}, \dots X_{(2K)}$	$Y_{K+1}^*, Y_{K+2}^*, \dots Y_{2K}^*$
III	$X_{(2k+1)}, X_{(2k+2)}, \dots X_{(3K)}$	$Y_{2K+1}^*, Y_{2K+2}^*, \dots Y_{3K}^*$

Now compute

$$\begin{split} \bar{X}_{(1)} &= \frac{\sum_{i=1}^{K} X_{(i)}}{k} \\ \bar{X}_{(2)} &= \frac{\sum_{k=1}^{2K} X_{(i)}}{k} \\ \bar{X}_{(3)} &= \frac{\sum_{2K+1}^{3K} X_{(i)}}{k} \\ \bar{Y}_{1}^{*} &= \frac{\sum_{i=1}^{K} Y_{i}^{*}}{k}, \ \bar{Y}_{2}^{*} &= \frac{\sum_{K+1}^{2K} Y_{i}^{*}}{k} \\ \bar{Y}_{3}^{*} &= \frac{\sum_{2K+1}^{3K} Y_{i}^{*}}{k}, \ \bar{X}_{0} &= \frac{\sum_{i=1}^{n} X_{(i)}}{n}, \\ \bar{Y}^{*} &= \frac{\sum_{i=1}^{K} \bar{Y}_{i}^{*}}{n} \end{split}$$

Then α and β are estimated on the basis of first and third group means

$$\hat{\beta} = \frac{\bar{Y}_{3}^{*} - \bar{Y}_{1}^{*}}{\bar{X}_{(3)} - \bar{X}_{(1)}} , \ \hat{\alpha} = \bar{Y}^{*} - \hat{\beta}\bar{X}_{0}$$

As shown in Wald's test, if n = 3k + 1

We can find median first and delete it from the data and proceed further as above.

Method of Instrumental Variables (IVM)

To illustrate this method let us consider the model as

 $Y = \alpha + \beta X + e$ Where $X = \xi + u$ $Y = \eta + \vartheta$ $\eta = \alpha + \beta \xi$

So that $e = \vartheta - \beta u$

If we apply here OLS method, we do not get BLUE for α and β due to measurement errors.

Here the remedy is to devise the instrumental variable (IV) denoted by z such that its observations are correlated with x and uncorrelated with u and ϑ . Then we write

$$\hat{\beta} = \frac{\sum_{i}^{n} (Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum (X_i - \bar{X})(Z_i - \bar{Z})}$$

Now $X_i = \xi_i + U_i$, $Y_i = \eta_i + \vartheta_i$, $\eta_i = \alpha + \beta \xi_i$

and $Y_i = \alpha + \beta X_i + e_i$ where $e_i = \vartheta_i - \beta u_i$

 $\therefore Y_i - \overline{Y} = \beta(X_i - \overline{X}) + (e_i - \overline{e})$

Thus $\hat{\beta} = \frac{\sum_{i=1}^{n} (z_i - \overline{z}) \{\beta(x_i - \overline{x}) + (e_i - \overline{e})\}}{\sum (x_i - \overline{x})(z_i - \overline{z})}$

$$= \beta + \frac{\sum (z_i - \bar{z})(e_i - \bar{e})}{\sum (z_i - \bar{z})(x_i - \bar{x})}$$

Since $Corr(z,u) = Corr(z,\vartheta) = 0 \Rightarrow Corr(z,e) = 0$

Hence we get $plim\{\sum (Z_i - \bar{Z})(e_i - \bar{e})\} = 0$

so that $plim(\hat{\beta}) = \beta$

Thus if $Corr(Z, X) \neq 0$, We get consistent estimator of β by using IV method.

If Corr(Z,X) = 0, this method is not applicable.

In practice, this method is not more suitable due to the difficulty of obtaining the instrumental variable x.

(Note that IV method as shown above can be extended further in the case of classical general linear model).

