Specification Bias

Dr. Seema Gupta
Department of Statistics
Ram Lal Anand College

How does one go about finding the "correct" model?

1. what are the criteria in choosing a model for empirical analysis?
2. What types of model specification errors is one likely to encounter in practice?
3. What are the consequences of specification errors?
4. How does one detect specification errors? In other words, what are some of the diagnostic tools that one can use?
5. Having detected specification errors, what remedies can one adopt and with what benefits?
6. How does one evaluate the performance of competing models?

Model chosen for empirical analysis should satisfy the following criteria:

- *Be data admissible*- that is, predictions made from the model must be logically possible.
- *Be consistent with theory*- that is, it must make good economic sense. For example,
- *Have weakly exogenous regressors*- that is, the explanatory variables, or regressors, must be uncorrelated with the error term.
- *Exhibit parameter constancy*- the values of the parameters should be stable. Otherwise, forecasting will be difficulty.
- *Exhibit data coherency*- the residuals estimated from themodel must be purely random.
- *Be encompassing* - the model should 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.

TYPES OF SPECIFICATION ERRORS

- let this model be

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

where Y = total cost of production and X = output.

- But suppose for some reason a researcher decides to use the following model
- $$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + u_{2i} \quad (2)$$
- This would constitute a specificationerror, the error consisting in **omitting a relevant variable** (X_i^3). Therefore, the error term u_{2i} is infact

$$u_{2i} = \beta_4 X_i^3 + u_{1i} \quad (3)$$

- Now suppose that another researcher uses the following model

$$Y_i = \gamma_1 + \gamma_2 X_i + \gamma_3 X_i^2 + \gamma_4 X_i^3 + \gamma_5 X_i^4 + u_{3i} \quad (4)$$

- If (1) is Correct (4) also constitutes a specification error, the error here consisting in **including an unnecessary or irrelevant variable.** The new error term is in fact

$$u_{3i} = u_{1i} - \gamma_5 X_i^4 \quad (5)$$

= u_{1i} since $\gamma_5 = 0$

- Now assume that yet another researcher postulates the following model
- $lnY_i = \tau_1 + \tau_2 X_i + \tau_3 X_i^2 + \tau_4 X_i^3 + u_{4i}$ (6)
- In relation to the true model would also constitute a specification bias, the bias here being the use of the **wrong functional form.** In (1) *Y* appears linearly, whereas in (6) it appears log-linearly.
- Finally, consider the researcher who uses the following model:

$$Y_i^* = \beta_1^* + \beta_2^* X_i^* + \beta_3^* X_i^{*2} + \beta_4^* X_i^{*3} + u_i^* \quad (7)$$

- where Y_i^*= Y_i+ $\varepsilon_i$and $X_i^* = X_i$+ ω_i, $\varepsilon_i$and $\omega_i$being the errors of measurement.
- What (7) states is that instead of using the true Y_i and $X_i$we use their proxies, $Y_i^*$and X_i^*, which may contain errors of measurement. Therefore, in (7) we commit the **errors of measurement bias.**
- Another type of specification error relates to the way the stochastic error
- u_i(or u_t) enters the regression model. Consider for instance, the following bivariate regression model without the intercept term
- $Y_i = \beta X_i u_i$ (8)

- where the stochastic error term enters multiplicatively with the property that $\ln u_i$ satisfies the assumptions of the CLRM, against the following model $Y_i = \alpha X_i + u_i$ (9)
- where the error term enters additively. Although the variables are the same in the two models, we have denoted the slope coefficient in (8) by β and the slope coefficient in (9) by α. Now if (8) is the "correct" or "true" model, would the estimated α provide an unbiased estimate of the true β?

In developing an empirical model, one is likely to commit one or more of the following specification errors:

- **1.** Omission of a relevant variable(s)
- **2.** Inclusion of an unnecessary variable(s)
- **3.** Adopting the wrong functional form
- **4.** Errors of measurement
- **5.** Incorrect specification of the stochastic error term

1-4 are essentially in the nature of model specification errors in that we have in mind a “true” model but somehow we do not estimate the correct model. In model mis-specification errors, we do not know what the true model is to begin with.

CONSEQUENCES OF MODEL SPECIFICATION ERRORS

Underfitting a Model (Omitting a Relevant Variable)

- Suppose the true model is:

$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ **(10)**

- but for some reason we fit the following model:

$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i$ **(11)**

- The consequences of omitting variable X_3 are as follows
- 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,* $\widehat{\alpha_1}$ and $\widehat{\alpha_2}$ are *biased as well as inconsistent.* That is, $E(\widehat{\alpha_1}) = \beta_1$ and $E(\widehat{\alpha_2}) = \beta_2$, and the bias does not disappear as the sample size gets larger.

- Even if X_2 and X_3 are not correlated, $\widehat{\alpha_1}$ is biased, although$\widehat{\alpha_2}$ is now unbiased.
- The disturbance variance σ2 is incorrectly estimated.
- The conventionally measured variance of $\widehat{\alpha_2}$ $(= \sigma^2 / \sum x_{2i}^2)$ is a *biased* estimator of the variance of the true estimator $\widehat{\beta_2}$.
- In consequence, the usual confidence interval and hypothesis-testing procedures are likely to give misleading conclusions about the statistical significance of the estimated parameters.
- The forecasts based on the incorrect model and the forecast (confidence) intervals will be unreliable.

$$E(\widehat{\alpha_2}) = \beta_2 + \beta_3 b_{23}$$

where b_{32} is the slope in the regression of the excluded variable X_3 on the included variable X_2

$$b_{23} = \Sigma x_{3i} x_{2i} / \Sigma x_{2i}^2$$

- We can see that $\widehat{\alpha_2}$ is biased unless β_3 or b_{32} or both are zero. β_3 can not be zero otherwise there will be no specification error to begin with. B_{32} will be zero when X_2 and X_3 are uncorrelated which is unlikely.

- Generally, however, the extent of the bias will depend on the *bias term* $\beta_3 b_{32}$. If, for instance, β_3 is positive (i.e., X_3 has a positive effect on Y) and $b_{32}$is positive (i.e., X_2 and X_3 are positively correlated),$\widehat{\alpha_2}$, on average, will overestimate the true β_2. But this result should not be surprising, for X_2 represents not only its *direct effect* on Y but also its *indirect effect* (via X_3) on Y. In short, X_2 gets credit for the influence that is rightly attributable to X_3.$X_3$is prevented from showing its effect explicitly because it is not “allowed” to enter the model.

Variances of $\widehat{\alpha_2}$ and $\widehat{\beta_2}$ are

- $var(\widehat{\alpha_2}) = \frac{\sigma^2}{\Sigma x_{2i}^2}$
- $var(\widehat{\beta_2}) = \frac{\sigma^2}{\Sigma x_{2i}^2(1-r_{23}^2)} = \frac{\sigma^2}{\Sigma x_{2i}^2} VIF$
- In general the two variance will be different.
- Since 0 <r_{23}^2<1, therefore$var(\widehat{\alpha_2})$<$var(\widehat{\beta_2})$
- Although $\widehat{\alpha_2}$ is biased, its variance is smaller than the variance of the unbiased estimator $\widehat{\beta_2}$(we rule out the case where r_{23} = 0, since in practice there is some correlation between regressors).Therefore a dilemma.

- σ^2estimated from model (11) and that estimated from the true model (10) are not the same because the RSS of the two models as well as their degrees of freedom (df) are different. You may recall that we obtain an estimate of σ^2 as $\widehat{\sigma^2}$= RSS/df, which depends on the number of regressors included in the model as well as the df . Now if we add variables to the model, the RSS generally decreases (as more variables are added to the model, the *R*2 increases), but the degrees of freedom also decrease because more parameters are estimated. The net outcome depends on whether the RSS decreases sufficiently to offset the loss of degrees of freedom due to the addition of regressors. It is quite possible that if a regressorhas a strong impact on the regressand—for example, it may reduce RSS more than the loss in degrees of freedom as a result of its addition to the model—inclusion of such variables will not only reduce the bias but will also increase precision (i.e., reduce standard errors) of the estimators.

- On the other hand, if the relevant variables have only a marginal impact on the regressand, and if they are highly correlated (i.e., VIF is larger), we may reduce the bias in the coefficients of the variables already included in the model, but increase their standard errors (i.e., make them less efficient).
- The tradeoff in this situation between bias and precision can be substantial and will depend on the relative importance of the various regressors.

- When X_2 and X_3 are uncorrelated $var(\widehat{\alpha_2})\ and\ var(\widehat{\beta_2})$
- are same then there is no harm in dropping X_3.
- **The point is clear: Once a model is formulated on the basis of the relevant theory, one is illadvisedto drop a variable from such a model.**

Inclusion of an Irrelevant Variable (Overfitting a Model)

- let us assume that $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$ (12)

is the true model but we fit the model

$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + v_i$ (13)

We commit a specification bias of including an un-necessary variable.

Consequences

- The OLS estimators of the parameters of the "incorrect" model are all *unbiased and consistent,* that is, $E(\widehat{\alpha_1}) = \beta_1$, $E(\widehat{\alpha_2}) = \beta_2, E(\widehat{\alpha_3}) = \beta_3 = 0$.
- 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 $\hat{\beta}$ of the true model.

- $var(\widehat{\beta_2}) = \frac{\sigma^2}{\Sigma x_{2i}^2}$
- $var(\widehat{\alpha_2}) = \frac{\sigma^2}{\Sigma x_{2i}^2(1-r_{23}^2)}$
- Therefore $\frac{var(\widehat{\alpha_2})}{var(\widehat{\beta_2})} = \frac{1}{(1-r_{23}^2)}$
- Since $0 \leq r_{23}^2 \leq 1$, therefore $var(\widehat{\alpha_2}) \geq var(\widehat{\beta_2})$

- The implication of this finding is that the inclusion of the unnecessary variable *X*3 makes the variance of $\widehat{\alpha_2}$larger than necessary, thereby making$\widehat{\alpha_2}$ less precise. This is also true of $\widehat{\alpha_1}$.

The **asymmetry** in the two types of specification biases

- 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.
- On the other hand, including 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 only penalty we pay for the inclusion of the superfluous variable is that the estimated variances of the coefficients are larger, and as a result our probability inferences about the parameters are less precise.

- An unwanted conclusion here would be that it is better to include irrelevant variables than to omit the relevant ones. But this philosophy is not to be espoused because addition of unnecessary variables will lead to loss in efficiency of the estimators and may also lead to the problem of multicollinearity, not to mention the loss of degrees of freedom.

Therefore,

- In general, the best approach is to include only explanatory variables that, on theoretical grounds, *directly* influence the dependent variable and that are not accounted for by other included variables.