EC203: Applied Econometrics
Model misspecification

Dr. Tom Martin

University of Warwick

Illustrative reading:

- Wooldridge: Chapter 3, 5, 9
- Dougherty: Chapter 6, 8
- Gujarati: Chapter 7

Introduction: Endogeneity

Consider the usual CLRM:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

and associated assumptions:

- $E[\epsilon] = 0$
- $E[\epsilon|\mathbf{X}] = 0$
- No perfect collinearity and all $X's$ must exhibit variation
- $Cov(\epsilon_i, \epsilon_j|\mathbf{X}) = 0$ forall $i \neq j$
- $V(\epsilon|\mathbf{X}) = \sigma^2$
- $\epsilon|\mathbf{X} \sim N(0, \sigma^2)$

Introduction: endogeneity

Given the model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

- Our primary aim is to answer the question: What is the causal impact of X_j on Y?
- If $E[\epsilon|X_j] = 0$ we say that X_j is an **exogenous** variable.
- If all our variables are exogenous then all our estimators are **unbiased**: Each β_j picks up the causal effect of X_j on Y.
- If, however, $E[\epsilon|X_j] \neq 0$ we then say that X_j is an **endogenous** variable.
- If one of our variables is endogenous then all our estimators are **biased**.

Introduction: endogeneity

What are the primary causes of endogeneity?

- One can think of the error term as being everything we cannot observe and have not modeled correctly. Thus, there are many reasons to believe our variables will not be exogenous.
- We cover a few prominent examples of misspecification:
  1. omitting a relevant variable (under-fitting a model)
  2. including an irrelevant variable (over-fitting a model)
  3. simultaneity
  4. functional form misspecification (model misspecification)
- In general, when the model is thought to contain an endogenous variable we say the model is misspecified.

Case 1: omission of a relevant variable

Suppose we run the FALSE following regression:

$$Y_i = \alpha + \beta_1 X_i + \epsilon_i \ (F)$$

However, the TRUE regression is:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + v_i \ (T)$$

- In the true model all CLRM assumptions hold. That is: $E[v] = 0$, $E[v|X, Z] = 0$, $Cov(v_i, v_j|X, Z) = 0$, $V(v|X, Z) = \sigma^2$ and $v|X, Z \sim N(0, \sigma^2)$

Omission of a relevant variable

In the FALSE regression we have omitted the variable Z:

- Thus, biases that arise from this situation are termed omitted variable bias (OVB).[1]
- The main consequences of omitting a relevant variable are:
  1. all estimators are biased.
  2. all estimator variances are incorrectly estimated: usual t and F-statistics are wrong.
- An important variable may be omitted if:
  1. it is not identified as an important control: insufficient research has been undertaken in specifying the model.
  2. it is identified as an important control: but not having the data.

[1] It should be referred to as omitted relevant variable bias.

Omission of a relevant variable: consequences

Usefully, we can gain some insight into the sign of any possible bias. The FALSE model is:

$$Y_i = \alpha + \beta_1 X_i + \epsilon_i \ (F)$$

The TRUE regression is:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + v_i \ (T)$$

If we estimate F our estimate of β_1 equals:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Will this give us an unbiased estimate of β_1? To answer this question we need to consider the expected value of b_1.

Omission of a relevant variable: consequences

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\alpha + \beta_1 X_i + \epsilon_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_1 = \alpha\frac{\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \beta_1\frac{\sum_{i=1}^{n}(X_i - \bar{X})X_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})\epsilon_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})\epsilon_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Omission of a relevant variable: consequences

If $E[\epsilon|X] = 0$, then we could show the usual result that

$$b_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})\epsilon_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_1 = \beta_1 + \sum_{i=1}^{n} \lambda_i \epsilon_i$$

$$E[b_1] = E[\beta_1 + \sum_{i=1}^{n} \lambda_i \epsilon_i]$$

$$E[b_1] = \beta_1 + \lambda_1 E[\epsilon_1|X] + \lambda_2 E[\epsilon_2|X] + ... + \lambda_n E[\epsilon_n|X]$$

$$E[b_1] = \beta_1$$

Omission of a relevant variable: consequences

However, in in the current situation we know $\epsilon_i = \beta_2 Z_i + v_i$, therefore:

$$b_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})\epsilon_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\beta_2 Z_i + v_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\beta_2 Z_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(v_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_1 = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_1 = \beta_1 + \beta_2 \frac{cov(X, Z)}{V(X)}$$

Omission of a relevant variable: consequences

This is a crucial equation for understanding OVB:

$$b_1 = \beta_1 + \beta_2 \frac{cov(X, Z)}{var(X)}$$

- The bias is due to two main components:
  1. β_2: the regression coefficient of the excluded variable. That is, the partial effect Z on Y.
  2. $\frac{cov(X,Z)}{var(X)}$: the regression coefficient of the excluded variable on the included variable. That is, the partial effect of Z on X.
- Thus, an omitted variable will bias your coefficient unless either the partial effect of Z on Y, or Z on X is zero.

Omission of a relevant variable: consequences

Direction of bias:

	$corr(X, Z) > 0$	$corr(X, Z) < 0$
$\beta_2 > 0$	+ve bias	-ve bias
$\beta_2 < 0$	-ve bias	+ve bias

Let us consider two example: i) upward bias and ii) downward bias.

Omission of a relevant variable: consequences

Consider the following example of schooling on wage, a relationship we expect to be positive:

- Suppose the true population model is: $ln(wage_i) = \alpha + \beta_1 school_i + \beta_2 ability_i + \epsilon_i$
- While we estimate: $ln(wage_i) = a + b_1 school_i + e_i$
- Further, suppose we estimate $b_1 > 0$: schooling has a positive impact on wage. Are we correct?
- How do you expect our estimate b_1 to be biased?
- $b_1 = \beta_1 + \beta_2 \frac{cov(school,ability)}{var(school)}$
- β_2: suppose we expect that ability has a positive effect on wages.
- $\frac{cov(school,ability)}{var(school)} > 0$: we expect ability to have a positive effect on school duration.
- Therefore, we have an upward bias: $b_1 > \beta_1$.
- That is, we expect b_1 in our model to be greater than that in the true model.

Omission of a relevant variable: consequences

Intuition:

- $b_1 > 0$: suggests that more years of education increases your wages.
- However, people who have more years of education have more ability.
- Thus, our education variable is picking up the effect of ability also.
- Overall: given we expect $b_1 > \beta_1$, if we find that $b_1 > 0$ it does not necessarily imply that $\beta_1 > 0$. If the bias is so big, the true parameter could actually be negative: school lowers expected wages.

Omission of a relevant variable: consequences

Consider the following example of schooling on crime, a relationship we expect to be negative:

- Suppose the true population model is: $crime_i = \alpha + \beta_1 school_i + \beta_2 drugs_i + \epsilon_i$
- While instead we estimate: $crime_i = a + b_1 school_i + e_i$
- Further, suppose we find that $b_1 < 0$: schooling reduces the chance you commit a crime. Are we correct?
- How do you expect our estimate b_1 to be biased?
- $b_1 = \beta_1 + \beta_2 \frac{cov(school,drugs)}{var(school)}$
- β_2: suppose we expect a positive relationship between drugs and crime.
- $\frac{cov(school,drugs)}{var(school)}$: suppose we expect a negative relationship between school and drugs.
- Therefore, we expect a downward bias: $b_1 < \beta_1$.
- That is, in our model we expect b_1 to be more negative than that in the true model.

Omission of a relevant variable: consequences

Intuition:

- $b_1 < 0$: suggests that more years of education reduces the likelihood you commit a crime.
- However, people who have more years of education are less likely to use drugs.
- Thus, our education variable is also accounting for the fact that more educated people are less likely to use drugs.
- Overall: given we expect $b_1 < \beta_1$, if we find that $b_1 < 0$ it does not necessarily imply that $\beta_1 < 0$. If the bias is so big, the true parameter could actually be positive: school increases expected crime.

Omission of a relevant variable: testing

Is it possible to test for the omission of a relevant variable?

- Case 1: Data on a known omitted variable: if you have data on a possible omitted variable simply put it in the model and observe the consequence. That is observe whether its inclusion affects the coefficient on your variable of interest.
- Case 2: No data on a known omitted variable: if you think you have an omitted variable (a variable which is suggested by theory/empirical literature/common sense) yet you do not have data for it, then there is no test available. But you should consider possible bias.
- Case 3: No data on an unknown omitted variable: a possible omitted variable you have not even considered. Again there is no test available.[2]

[2] In a way this is worse than case 2, since it shows a lack of awareness on the researchers part.

Omission of a relevant variable: solutions

Possible solutions include:

- 1. If you have identified the relevant variable and can find data for it - include it!
- 2. Find a proxy variable. Note however, it is only a proxy and therefore is not a perfect solution. For example, when we regress wages on education we often assume ability is an important omitted variable. We could use IQ as a proxy. However, when people think of ability it is not necessarily IQ they think of.
- 3. Use an instrumental variable - this is a very common method in econometrics. We will discuss it shortly.
- 4. Use fixed effects estimation to control for unobserved time constant unobservables.

If you cannot find a solution, it is important to consider the direction of any possible bias that omitted variables may be causing in your analysis.

Omission of a relevant variable: note I

- A quick note on proxy variables.
- You frequently encounter the problem that you cannot obtain the precise data you would like. Indeed, often concepts are so vague it is impossible in principle to measure them: i.e. ability, socioeconomic status, quality of education ... etc.
- Others may be measurable but the cost of taking the measure make it infeasible (household surveys used in empirical research often omit important variables some researchers think to be very important).
- Thus, it is a better idea to search for a proxy variable rather than omit the variable completely: i.e. use IQ, income, staff-student ratio, ... etc.
- A good **proxy variable** will be one that is **strongly associated** with the omitted relevant variable.
- However, you cannot test this condition as you do not have data on your omitted relevant variable.

Omission of a relevant variable: note II

- We have only considered the bias formula when we omit one explanatory variable from a true model which has two explanatory variables.
- This is one reason why it is important to have a focused question: the effect of an main explanatory variable X_j on Y, and consider possible sources of endogeneity closely. In this way, the formula can provide a good starting point for considering bias.
- However, the concept can be generalised to many explanatory variables. In general omitting relevant variables will cause bias in all coefficients. Unless all omitted variables are uncorrelated with the explanatory variables (in which case they are not relevant).
- Hence our next question: what happens when we include an irrelevant variable?

Case 2: Inclusion of a irrelevant variables

Suppose the true population model is the following:

$$Y_i = \alpha + \beta_1 X_i + \epsilon_i \ (T)$$

However, instead the FALSE regression is specified:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + v_i \ (F)$$

Where $\beta_2 = 0$. The main consequences are:

1. The estimator's are unbiased: $E[\beta_1] = b_1$.
2. The estimator variances are also correctly estimated; however, including an irrelevant variable will make variances needlessly large.
3. Hence t-statistics will be needlessly small, and you will fail to reject your null hypothesis more often. Alternatively stated, our estimators lose precision/efficiency.

Inclusion of a irrelevant variables: consequences

To see this result, recall that in the SLR case:

$$V(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

While in the MLR case:

$$V(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2(1 - R_j^2)}$$

In the current case, R_j^2 is the r-squared from the regression: $X_i = \delta_0 + \delta_1 Z_i + \nu_i$. Thus, we can see that the greater the correlation between X and Z, the greater the loss of precision. The two models will be exactly identical when there is identically zero correlation between X and Z (i.e. when $R_j^2 = 1$).

Inclusion of a irrelevant variables: testing and solution

Suppose the true population model is the following:

$$Y_i = \alpha + \beta_1 X_i + \epsilon_i \ (T)$$

However, instead the FALSE regression is specified:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + v_i \ (F)$$

Where $\beta_2 = 0$. The test is simple:

1. Test the null $H_0 : \beta_2 = 0$.
2. If you fail to reject the hypothesis you could consider dropping the variable from the model.
3. As usual there is a trade-off: you may not want to drop it because you want to explicitly show the effect is non-significant - this may well be the case if it is a variable which theory/literature dictates is in the model.

Case 3: Simultaneity

Simultaneity, or reverse causality, can be represented in the following fashion. Suppose our regression of interest is:

$$Y_{1i} = \beta_0 + \beta_1 Y_{2i} + \beta_2 Z_{1i} + u_i \ (A)$$

Where Z_{1i} is an exogenous variable and u_i satisfies all CLRM assumptions, except now we know that Y_{2i} is itself a function:

$$Y_{2i} = \delta_0 + \delta_1 Y_{1i} + \delta_2 Z_{2i} + v_i \ (B)$$

That is, not only does Y_{1i} depend on Y_{2i}, but also, Y_{2i} depends on Y_{1i}. Z_{2i} is another exogenous variable.

An important question is: What are the consequences for OLS estimators if one uses them to estimate the equation (A)?

Simultaneity

To illustrate the consequences Y_{1i} into Y_{2i} and then solve for Y_{2i}:

$$\begin{aligned}
Y_{2i} &= \delta_0 + \delta_1(\beta_0 + \beta_1 Y_{2i} + \beta_2 Z_{1i} + u_i) + \delta_2 Z_{2i} + v_i \\
Y_{2i} &= \delta_0 + \delta_1\beta_0 + \delta_1\beta_1 Y_{2i} + \delta_1\beta_2 Z_{1i} + \delta_1 u_i + \delta_2 Z_{2i} + v_i \\
Y_{2i}(1 - \delta_1\beta_1) &= \delta_0 + \delta_1\beta_0 + \delta_1\beta_2 Z_{1i} + \delta_2 Z_{2i} + (v_i + \delta_1 u_i) \\
Y_{2i} &= \frac{\delta_0 + \delta_1\beta_0}{1 - \delta_1\beta_1} + \frac{\delta_1\beta_2}{1 - \delta_1\beta_1} Z_{1i} + \frac{\delta_2}{1 - \delta_1\beta_1} Z_{2i} + \frac{(v_i + \delta_1 u_i)}{1 - \delta_1\beta_1}
\end{aligned}$$

Simultaneity

$$Y_{2i} = \frac{\delta_0 + \delta_1\beta_0}{1 - \delta_1\beta_1} + \frac{\delta_1\beta_2}{1 - \delta_1\beta_1} Z_{1i} + \frac{\delta_2}{1 - \delta_1\beta_1} Z_{2i} + \frac{(v_i + \delta_1 u_i)}{1 - \delta_1\beta_1}$$

Therefore, Y_{2i} will depend on the u_i, implying $cov(Y_2, u) \neq 0$, thus $E[Y_{2i}|u_i] \neq 0$. Therefore, OLS estimation of equation (A) will result in biased coefficients. This will be true unless $\delta_1 = 0$, in which case there would be no reverse causality in the first instance.

To deal with this situation we can utilise, for example:

- panel data which observed individuals at multiple points in time and therefore may allow cause to be separated from effect
- An alternative is to use instrumental variable estimation, where in the above example, the instrument needs to be correlated with Y_2 and not with Y_1 - other than through its effect on Y_2.

Case 4: Incorrect functional form

A regression is said to suffer from **functional form misspecification** when it does not properly account for the relationship between the dependent and the observed explanatory variables.

Example 1:

- suppose the true model is $ln(wage_i) = \alpha + \beta_1 educ_i + \beta_2 exp_i + \beta_3 exp_i^2 + \epsilon_i$
- while you estimate $ln(wage_i) = \alpha + \beta_1 educ_i + \beta_2 exp_i + \epsilon_i$.
- That is, we haven't properly accounted for the relationship between wage and experience.

Incorrect functional form: consequences

Example 2:

- suppose the true model is:

$$\begin{aligned} ln(wage_i) =& \alpha + \beta_1 educ_i + \beta_2 exp_i + \beta_3 exp_i^2 + \\ & \beta_4 female_i + \beta_5 female_i * educ_i + \epsilon_i \end{aligned}$$

- while we estimate:

$$ln(wage_i) = \alpha + \beta_1 educ_i + \beta_2 exp_i + \beta_3 exp_i^2 + \epsilon_i$$

- That is, we haven't properly accounted for the heterogeneity in the returns to education.

Incorrect functional form: consequences

Consequences:

- It is essentially an omitted relevant variable problem. Thus, OLS estimators are biased and t and F-statistics are wrong.
- Therefore, it is a serious problem.
- Note, however, it is special kind of omitted variable: we have the data but we use it incorrectly.
- Thus, although the consequences are serious, you should not suffer from incorrect functional form.
- It is important to consider how theory stipulate models should be specified.
- It is important to test your regression to changes in specification.
- If your model is well thought out, there is a limited number of specifications you want to test: most relationships studied can be described by polynomials, interactions or logs.

Incorrect functional form: testing

Testing for incorrect functional form:

- You can use t and F-tests. For example, in example 1 above, you could simply use an F-test on the experience and the experience squared term to test if it is significant.
- Further, you can use chow tests to test for structural breaks - which are another form of misspecification bias.
- There are also a number of other formal tests for misspecification - including the RESET.

Incorrect functional form: testing

Suppose you run the model:

$$Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i \ (A)$$

To implement the RESET test:

- Calculate the fitted values - $\hat{Y}_i$.
- Then regress:

$$Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \gamma_1 \hat{Y}_i^2 + \gamma_2 \hat{Y}_i^3 + \epsilon_i (B)$$

Incorrect functional form: testing

- Then test $H_0 : \gamma_1 = \gamma_2 = 0$ against H_0 : at least one $\neq 0$. A null of no incorrect functional form.
- Use an F-test: where the unrestricted model is (B) and the restricted model is (A).
- Intuitively, $\hat{Y}_i^2$ and $\hat{Y}_i^3$ are non-linear functions of the X variables: the polynomials of the fitted values.
- Thus, if you reject H_0 there is some evidence that some of the X's should be entered in their quadratic form.

Incorrect functional form: testing

- The above test specification is the most common; however, you may only include a squared term or you may include higher order polynomials.
- Note this is a very low power test, since the polynomials of the fitted values are only rough proxies for the correct functional form.
- Recall low power implies you have difficultly in rejecting or failing to reject any hypothesis.
- That is, if you reject H_0 it doesn't necessarily mean you have a misspecification problem and if you fail to reject it doesn't mean you don't.
- The main point from this discussion is that it is very important to consider the robustness of your model to model specification.