# Model Specifications

## EC 320: Introduction to Econometrics

Emmett Saulnier
Spring 2022

# Prologue

# Rest of the term

## Last 4 lectures

- Today: Model specification
- Thursday: Differences in Differences part 1
- Next Tuesday: Differences in Differences part 2
- Next Thursday: Final exam review

## Last two problem sets

- Analytical Problem Set 7 Due Friday
- Computational Problem Set 7 Due Monday

# Model Specification

## Today we'll be diving deeper on a few concepts

What are the consequences of excluding a variable that should be in the model?

**Omitted Variable Bias**

How do we test restrictions to model specifications?

**t and F tests**

## Plus one new topic:

What happens if you have difficulty finding data on a variable and use a proxy instead?

**Today's lesson**

# Omitted Variable Bias

# Omitted Variable Bias

**Omitted variable bias** (OVB) arises when we omit a variable, $X_k$ that

1. Affects the outcome variable $Y$, $\beta_k \neq 0$
2. Correlates with an explanatory variable $X_j$, $Cov(X_j, X_k) \neq 0$,

Biases OLS estimator of $\beta_j$.

**What is our formula for OVB?**

If we omit $X_k$, then the formula for the bias it creates in $\hat{\beta}_j$ is...

$$Bias = \beta_k \frac{Cov(X_j, X_k)}{Var(X_j)}$$

**or equivalently** $Bias = \hat{\beta}_j^{short} - \hat{\beta}_j^{long}$

# Omitted Variable Bias

Suppose the true model is $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ and all of our assumptions hold for this model.

**What are the 6 OLS assumptions?**

Thus, $\hat{\beta}_1^{long}$ will be an unbiased estimate of $\beta_1$.

Suppose we do not have data on $x_{2i}$ and so we estimate
$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$

**What is OLS formula for $\hat{\beta}_1^{short}$ ?**

$$\hat{\beta}_1^{short} = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{1i} - \bar{x})^2}$$

# Omitted Variable Bias

$$
\begin{aligned}
\hat{\beta}_1^{short} &= \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{1i} - \bar{x})^2} \\
&= \frac{Cov(x_{1i}, y_i)}{Var(x_{1i})} \\
&= \frac{Cov(x_{1i}, \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i)}{Var(x_{1i})} \\
&= \frac{Cov(x_{1i}, \beta_0) + \beta_1 Cov(x_{1i}, x_{1i}) + \beta_2 Cov(x_{1i}, x_{2i}) + Cov(x_{1i}, u_i)}{Var(x_{1i})} \\
&= 0 + \beta_1 \frac{Cov(x_{1i}, x_{1i})}{Var(x_{1i})} + \beta_2 \frac{Cov(x_{1i}, x_{2i})}{Var(x_{1i})} + \frac{Cov(x_{1i}, u_i)}{Var(x_{1i})} \\
&= \beta_1 + \beta_2 \frac{Cov(x_{1i}, x_{2i})}{Var(x_{1i})} + \frac{Cov(x_{1i}, u_i)}{Var(x_{1i})}
\end{aligned}
$$

# Omitted Variable Bias

$$\begin{aligned}\hat{\beta}_j^{short} - \hat{\beta}_j^{long} &= \left(\beta_1 + \beta_2 \frac{Cov(x_{1i}, x_{2i})}{Var(x_{1i})} + \frac{Cov(x_{1i}, u_i)}{Var(x_{1i})}\right) \\ &\quad - \left(\beta_1 + \frac{Cov(x_{1i}, u_i)}{Var(x_{1i})}\right) \\ &= \beta_2 \frac{Cov(x_{1i}, x_{2i})}{Var(x_{1i})}.\end{aligned}$$

## In Summary

- Omitted bias happens when we do not include a variable that affects $y$ (and thus is included in the error term) and also is correlated with $x$.
- It is an example of a violation of exogeneity. $E[u_i|x] \neq 0$ if we have an omitted variable

**It also invaldiates our standard errors**

# Proxies in Model Specifications

# Proxies

Suppose you are considering the following model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

While $x_1$ is observed, suppose $x_2$ **not able to be observed**. Cases of unobservable data could include:

- Vaguely defined with no explicit measure (e.g. quality of education)
- Intangible and cannot be quantified (e.g. utility, ability)
- Requires so much time/energy that measurement is infeasible
- Confidentiality, privacy concerns may limit observed data availability

# Proxies

Rather than exclude the unobservable, you may wish to use an adequate **proxy** variable for $x_2$.

A **proxy** variable is a substitute variable that may reasonably be supposed to maintain similar properties to our missing variable.

**Example:** For quality of education, we could use the student-teacher ratio to have a measure of how many resources the educational institution makes available to students. Where quality is high, student-teacher ratios are low.

# Proxies

Returning to the model, our true data generating process for $y$:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

In the case where we have no data on $x_2$, suppose we have an **ideal proxy** for it such that there exists an *exact linear relationship* $x_2$ *and* $Z$:

$$x_2 = \lambda + \mu Z,$$

where $\lambda$ and $\mu$ are fixed, unknown constants. We cannot estimate them by running a regression of the above relationship, since we have no data on $x_2$ . **Let's sub in our expression and see what using Z achieves**.

# Inference using Proxies

Using $X_2 = \lambda + \mu Z$,

$$\begin{aligned} y =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\ =& \beta_0 + \beta_1 x_1 + \beta_2 (\lambda + \mu Z) + u \\ =& (\beta_0 + \beta_2 \lambda) + \beta_1 x_1 + \beta_2 \mu Z + u \\ =& \quad \alpha \quad + \beta_1 x_1 + \quad \gamma_2 Z + u \end{aligned}$$

1. $\beta_1$, its standard error and t-stat will be same as if $X_2$ used
2. $R^2$ will be the same as if $X_2$ was used instead of Z
3. The coefficient of Z, $\gamma_2$, will be estimate of $\beta_2 \mu$, and not possible to estimate $\beta_2$ directly
4. t-stat for $\gamma_2$ same as $\beta_2$, so able to assess significance of $X_2$
5. Not possible to estimate $\beta_0$ since we now only see $\alpha$

# Risks of using Proxies

Validity of all the subsequent takeaways rely on the condition that $Z$ is an ideal proxy for $x_2$

- In practice, unusual to find a proxy that is exactly linearly related to our missing variable
- If the relationship is close to linear, the results will hold **approximately**
- The biggest problem faced is that there is never any manner by which to test this condition of whether the approximated difference is sufficiently small
- Critical thinking in explaining your logical link between $X_2$ & $Z$ and accepting that you are relying on a strong assumptions are often required to convince an audience that a proxy is indeed valid

# Testing multiple restrictions

# F-test review

Suppose we have the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

and want to test $H_0 : \beta_1 = 0$ against the alternative $H_a : \beta_1 \neq 0$.

**What are our options?**

1. Run a t-test using the standard error of $\hat{\beta}_1$
2. Run an F-test with the coefficient restriction $\beta_1 = 0$.

$$F_{q,n-k-1} = \frac{RSS_r - RSS_u/q}{RSS_u/(n-k-1)}$$

# Multiple Coefficient Restrictions

Suppose now we want to test $H_0 : \beta_1 = 10,\ \beta_2 = \beta_3/5$. The alternative hypothesis is that at least one of these restrictions is wrong.

**What are the two regressions we have to run?**

1. **Unrestricted** $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$, gives us $RSS_u$
2. **Restricted** $y_i = \beta_0 + 10 x_{1i} + (\beta_3/5) x_{2i} + \beta_3 x_{3i} + u_i$, gives us $RSS_r$

But we can't actually "run" the restricted model as-is, rearranging...

$$(y_i - 10x_{1i}) = \beta_0 + \beta_3(x_{2i}/5 + x_{3i}) + u_i$$

- Any variables without a parameter to be estimated go on the LHS with your outcome
- Any variables with the same parameter are combined together

# Multiple Coefficient Restrictions

We now have everything we need to calculate the F-statistic. Suppose we have $n = 500$ and

1. Unrestricted model: $RSS_u = 1000$
2. Restricted model: $RSS_r = 1010$

$$F_{2,496} = \frac{1010 - 1000/2}{1000/500 - 3 - 1} = 4.96$$

and critical value $F_{crit} = 3.01$

**What is the conclusion of this test?**

We reject the null hypothesis in favor of the alternative. At least one of the restrictions we imposed is wrong.