Model Specification Continued

EC 320: Introduction to Econometrics

David Hall
University of Oregon

Review and Today's Goal

Last time, we looked at:

- Omitted Variables

What happens when we omit a variable that should be there? Bias!

Today we cover variable non-inclusion – what happens when we include a variable that does not actually do any explaining?

But First, Reminders

- This is your last non-review, non-project class. So basically, the last full lecture from me.
- I may be teaching Health Economics in the fall if there is enough interest.
- Part 1 of your project is due tonight.
- A4 is out.

Warm Up i

I am going to now give you 15 minutes to consult with each other regarding the first set of problems in your final assignment, Assignment 4. The first problems are back from Chapter 4 and have to do more with non-linearity.

- I will allow 15 minutes to consult each other, then we will discuss as a class what you got after.
- I will not tell you any answers, only confirm or deny answers given to me.

Do Not Actually Forget All That

Inclusion of a Redundant Variable

Comparison

Table 6.1 Consequences of variable specification

		True model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted model	$\hat{Y} = b_1 + b_2 X_2$	Correct specification, no problems	Coefficients are biased (in general). Standard errors are invalid
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$	Coefficients are unbiased (in general) but inefficient. Standard errors are valid (in general)	Correct specification, no problems

Super Formally, The Issue Is...

True Model: $Y = \beta_0 + \beta_1 X_1 + u$
You Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$
and so obviously, $\hat{\beta}_1$ will not be our usual estimator.

Instead, it will be...

Super Formally, the Issue Is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_{1i}-\bar{X}_1\right)\left(Y_i-\bar{Y}\right)\sum_{i=1}^{n}\left(X_{2i}-\bar{X}_2\right)^2-\sum_{i=1}^{n}\left(X_{2i}-\bar{X}_2\right)\left(Y_i-\bar{Y}\right)\sum_{i=1}^{n}\left(X_{1i}-\bar{X}_1\right)\left(X_{2i}-\bar{X}_2\right)}{\sum_{i=1}^{n}\left(X_{1i}-\bar{X}_1\right)^2\sum_{i=1}^{n}\left(X_{2i}-\bar{X}_2\right)^2-\left(\sum_{i=1}^{n}\left(X_{1i}-\bar{X}_1\right)\left(X_{2i}-\bar{X}_2\right)\right)^2}$$

... but I never really discussed the multi-variate estimator because it is long and complicated.

A Few Notes on This

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_{1i}-\bar{X}_1\right)\left(Y_i-\bar{Y}\right)\sum_{i=1}^{n}\left(X_{2i}-\bar{X}_2\right)^2-\sum_{i=1}^{n}\left(X_{2i}-\bar{X}_2\right)\left(Y_i-\bar{Y}\right)\sum_{i=1}^{n}\left(X_{1i}-\bar{X}_1\right)\left(X_{2i}-\bar{X}_2\right)}{\sum_{i=1}^{n}\left(X_{1i}-\bar{X}_1\right)^2\sum_{i=1}^{n}\left(X_{2i}-\bar{X}_2\right)^2-\left(\sum_{i=1}^{n}\left(X_{1i}-\bar{X}_1\right)\left(X_{2i}-\bar{X}_2\right)\right)^2}$$

In principle, the only issue with this estimator compared to our typical OLS estimator is **inefficiency.**

Q: What is inefficiency in this case? Why is it bad?

But what is a redundant variable?

In this case, I am discussing a variable that has no explanatory power. In other words, **the coefficient on this variable when included as a regressor is 0.**

$$Y = \beta_0 + \beta_1 X_1 + 0X_2 + u \tag{1}$$

and yet, X_1 and X_2 might have some correlation $r^2_{X_1,X_2}$.

Redundant Correlation Bad

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{\sum_{i=1}^{n}\left(X_{2i} - \bar{X}_1\right)^2} \quad \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{\sum_{i=1}^{n}\left(X_{2i} - \bar{X}_1\right)^2} \frac{1}{\left(1 - r^2_{X_1 X_2}\right)}$$

Questions?

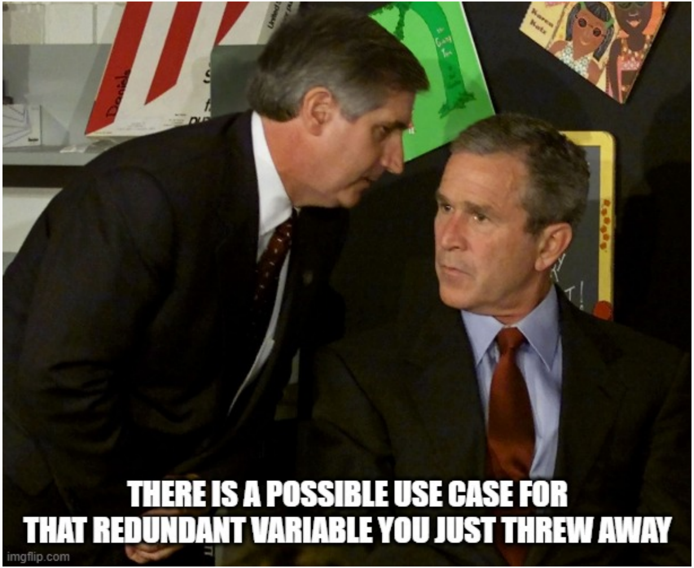
- Adding some non-intuitive/non-theory-backed/redundant variable will just hurt your ability to do inference.
- I won't expect you to mathematically write that, I do expect you to understand that (a) redundancy = inefficient and (b) it depends on the correlation between your regressors and the redundant variable

Poor Redundant Variable

Redundant Slide of the Poor Redundant Variable

But Wait! There's More!

Instrumental Variables



Two Characteristics

There are two characteristics about this redundant variable that makes it bad for a basic regression.

Let the redundant variable be denoted Z.

- $Cov(Z, X) \neq 0$
- $Cov(Z, Y) = 0$

These are the exact properties we want in an *instrument.*

Situation and Setup

An instrumental variable is used for situations when a regressor and outcome are plausibly endogenous (jointly determined, reverse causal, etc.):

- A shift in demand will have an effect on price, that price will have an effect on quantity demanded.
- Hospitals might determine the health of an area, but the health (and therefore, need for health services) of an area determines whether a hospital enters.
- People from poorer areas enlist in the military more. Their leaving might worsen local labor market conditions.

I often think about this as non-random assignment of a treatment. An instrumental variable helps to restore the randomness.

Some Examples i

Demand and Price

- Goal: Estimate the causal effect of price on quantity demanded.
- Problem: Supply and demand shocks both move price.
  $\Rightarrow Cov(Price, \epsilon) \neq 0$.
- Instrument: Weather shocks, cost shifters, transportation disruptions.

Hospitals and Health

- Goal: Estimate the effect of hospital access on health outcomes.

Some Examples ii

- Problem: Hospitals open where people are sick. Selection.
  $\Rightarrow Cov(HospitalAccess, \epsilon) \neq 0$.
- Instrument: Distance to the nearest hospital, certificate-of-need rules, historical facility placement.

Military Enlistment

- Goal: Estimate the effect of military service on earnings or health.
- Problem: Enlistment is non-random (ability, preferences, local economy).
  $\Rightarrow Cov(Service, \epsilon) \neq 0$.
- Instrument: Vietnam draft lottery numbers, ASVAB eligibility cutoffs.

The Setup: Why OLS is Biased

Suppose we want to estimate the causal effect of X on Y:

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

But X is endogenous:

$$Cov(X_i, u_i) \neq 0.$$

This makes the OLS estimator biased:

$$\hat{\beta}_1^{OLS} = \beta_1 + \frac{Cov(X, u)}{Var(X)}.$$

Setup Continued

Idea: We have a variable Z_i such that:

$$Cov(Z_i, X_i) \neq 0 \quad \text{(relevance)}$$

$$Cov(Z_i, u_i) = 0 \quad \text{(exogeneity)}$$

Use Z to isolate the variation in X that is uncorrelated with u.

We do this in 2 Stages

2 Chainz says to use 2 Stage Least Squares (2SLS)

Two-Stage Least Squares: Stage 1

Goal: Extract the part of X that is driven by the instrument Z.

- Z is correlated with X (relevance).
- Z is not correlated with u (exogeneity).

First Stage Regression:

$$X_i = \pi_0 + \pi_1 Z_i + v_i.$$

This gives us predicted values:

$$\hat{X}_i = \pi_0 + \pi_1 Z_i.$$

Two-Stage Least Squares: Stage 2

Interpretation: $\hat{X}_i$ is the "clean," exogenous variation in X coming only from Z.

Stage 2: Use the exogenous part of X (the fitted values) to estimate the causal effect on Y.

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + e_i.$$

Stage 2 Continued

Interpretation:

- We no longer use X_i (which was correlated with u_i).
- We use $\hat{X}_i$, which is constructed only from Z_i.
- Because Z is unrelated to u, so is $\hat{X}$.

$$Cov(\hat{X}_i, u_i) = 0 \quad \Rightarrow \quad \hat{\beta}_1^{2SLS} \text{ is causal.}$$

Takeaways and Intuition

- Z can be thought of as creating random assignment.
- We use Z to isolate the part of X that is exogenous to u.
- In reality, finding an instrument is **extremely hard**.
- *You have to make the logical case that your instrument is exogenous to your outcome!*
- On A4, the extra credit asks you to run your own instrumental variable estimation.

Questions?

Next Class

- Bring your computers if you have them to work in-class on the project.
- If you do not have one, go to wherever you usually do your work and I will be reachable by email next Tuesday.