Econometrics Discussion Section 2

John Green

Spring 2024

Omitted Variable Bias

- In economics, we often want to determine the impact of X on Y
  - But X is not the only thing going on! Usually many variables are effecting Y
  - If we don't include these other variables in our model, we will get biased estimates of the effect of X on Y, a problem referred to as *omitted variable bias*
- Let's think about what this does to our estimate with a silly example
  - We are an AC company, and we want to know what causes people to buy more AC units
  - Our statistician tells us that the number of swimming pool accidents is a good predictor of AC units sold
  - We run a regression of AC units sold on swimming pool accidents and find a positive relationship
  - Should we conclude that swimming pool accidents cause people to buy more AC units?

Omitted Variable Bias

- Should we conclude that swimming pool accidents cause people to buy more AC units?
- Of course not! The omitted variable here is temperature. When it's hot, people buy more AC units and more people go swimming
- Intuitively, what direction do we expect the bias in the effect of swimming pool accidents to be?

Omitted Variable Bias

- Should we conclude that swimming pool accidents cause people to buy more AC units?
- Of course not! The omitted variable here is temperature. When it's hot, people buy more AC units and more people go swimming
- Intuitively, what direction do we expect the bias in the effect of swimming pool accidents to be?
- It should bias the estimate *upwards*: we are attributing the effect of temperature to swimming pool accidents and thus making swimming pool accidents look more important than they are

Omitted Variable Bias

- We have omitted variable bias when we have a variable Z that is correlated with X and is a determinant of Y
  - Note: if Z is a determinant of Y but is not correlated with X, then it is not a problem!
  - These conditions mean that our OLS assumption of $E(u|X) = 0$ is violated
- Formula for OVB:

$$\hat{\beta} \rightarrow_p \beta + \frac{\sigma_u}{\sigma_X} \rho_{Xu}$$

  - The intuition from our example shows up in the correlation term

Causal effects

- We can use OLS to summarize a relationship without attaching any directionality
  - In this case we need to be careful in our language: "a change in X is associated with a change in Y"
- Usually economists want to be able to say something *causal*
- Ideal is a randomized controlled trial (RCT): some group gets the treatment, some group doesn't, and we compare outcomes between the two
  - Observational data usually differs from this in important ways

Solution to OVB

- An RCT eliminates OVB because we randomly assign the treatment, which will then not be correlated with any other variables!
  - This is usually not possible in economics — this is why economics is hard!
- Cross-tabulation eg run the regression on a subsample of your population where there is no OVB problem (but other issues emerge)
- Try and include omitted variables (obviously) in a multivariable regression

Multivariate regression

- Same logic as before:
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$
- Estimators derived in same way as before (just using matrices)
- Before we looked at R^2 as a measure of fit, but now we have to be careful: adding in more variables on the RHS will always increase R^2
- Adjusted R^2 is a better measure of fit which includes a degrees-of-freedom correction to penalize for adding in more variables
- Add one more assumption our previous 3 from the single-variable case: no perfect multicolinearity

Perfect multicolinearity

- This is when one of your RHS variables is a perfect linear combination of others
- Why is this a problem?

Perfect multicolinearity

- This is when one of your RHS variables is a perfect linear combination of others
- Why is this a problem?
- Think about basic algebra:
  - $y = a * x$
  - What is a?

Perfect multicolinearity

- $y = a * x$
- What is a?
- $a = \frac{y}{x}$
- But now what if I give you:
  - $y = b * x + c * x$
  - What are a and b?

Perfect multicolinearity

- But now what if I give you:
  - $y = b * x + c * x$
  - What are b and c?
- This question has no unique answer! Any combination of b and c such that $a = b + c$ will work (an infinite number)
- This is what's going on when their is perfect multicolinarity: our coefficient estimates β are not identified/unique

Binary variables

- For this reason we need to be careful when we have binary (0/1) variables ("indicator" variables) in our regression:
  - Treatment status
  - Gender
  - Party affiliation
- We either need to leave one out (the omitted group) or we need to drop our intercept
  - Either is fine, just changes interpretation of the coefficient estimates
  - Failing to do so is sometimes called the the "dummy variable trap"

Binary variables

- For this reason we need to be careful when we have binary (0/1) variables ("indicator" variables) in our regression:
  - Treatment status
  - Gender
  - Party affiliation
- We either need to leave one out (the omitted group) or we need to drop our intercept
  - Either is fine, just changes interpretation of the coefficient estimates
  - Failing to do so is sometimes called the the "dummy variable trap"
- Exact same math, intuition, problem and solution extends to *categorical* variables:
  - Seasonal effects
  - Race or ethnicity
- In practice, we take our categories and create a bunch of dummy variables, so we need to leave one category out as the "base" to avoid multicolinearity

Imperfect multicolinearity

- Perfect multicolinearity is straightforward to deal with; generally speaking, the regression simply will not work
- Slightly trickier is *imperfect* multicolinearity: X_1 and X_2 are highly correlated
  - For example, controlling for both age and years of work experience in a wage equation
- Imperfect multicolinearity will lead to imprecise estimation (large standard errors)
- Intuition: what does the data tell us about the effect of X_1 when X_2 is held constant?

Imperfect multicolinearity

- Perfect multicolinearity is straightforward to deal with; generally speaking, the regression simply will not work
- Slightly trickier is *imperfect* multicolinearity: X_1 and X_2 are highly correlated
  - For example, controlling for both age and years of work experience in a wage equation
- Imperfect multicolinearity will lead to imprecise estimation (large standard errors)
- Intuition: what does the data tell us about the effect of X_1 when X_2 is held constant?
- What can we do?

Imperfect multicolinearity

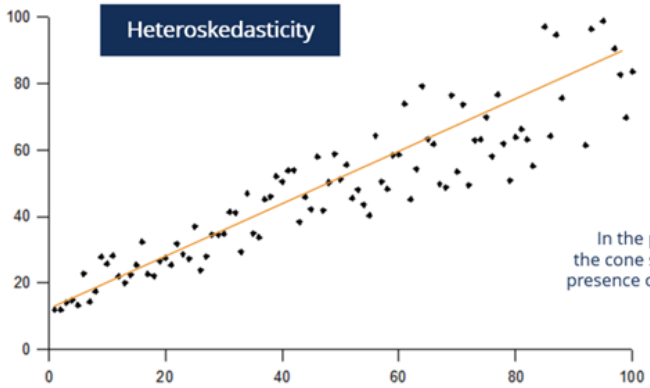
- Perfect multicolinearity is straightforward to deal with; generally speaking, the regression simply will not work
- Slightly trickier is *imperfect* multicolinearity: X_1 and X_2 are highly correlated
  - For example, controlling for both age and years of work experience in a wage equation
- Imperfect multicolinearity will lead to imprecise estimation (large standard errors)
- Intuition: what does the data tell us about the effect of X_1 when X_2 is held constant?
- What can we do?
- Can check scatterplots: if two variables are essentially linear together, probably only include one of them
- Guided by theory: it's usually obvious if two variables are highly linear

Control variables

- Usually, we cannot just include the omitted variables
- Instead, look for control variable(s) W which are correlated with the omitted variables
- Helps us address the OVB problem and get the causal impact of a variable X ($\rightarrow$ conditional mean independence holds)
- Example: impact of years of higher education on wages
  - Problem: ability is correlated with both! (In which direction is OVB?)
  - Solution: control for ability, eg use parental income, quality of local high schools, etc.

Residuals

- Usually, we cannot just include the omitted variables
- Instead, look for control variable(s) W which are correlated with the omitted variables
- Helps us address the OVB problem and get the causal impact of a variable X ($\rightarrow$ conditional mean independence holds)
- Example: relationship between GPA and job performance
  - Problem: ability is correlated with both! (In which direction is OVB?)
  - Solution: control for ability, eg use parental income, quality of local high schools, etc.


Heteroskedasticity
100
80
60
40
20
0
0
20
40
60
80
100
In the plot of residuals,
the cone shape indicates the
presence of heteroskedasticity

Hypothesis testing

- We already know about the hypothesis test on a single coefficient: simple t-test with a known distribution
- What about a joint hypothesis test? For example:
  - $H_0: \ \beta_1 = \beta_2 = 0$
  - $H_0: \ \beta_1 = \alpha_1$ & $\beta_2 = \alpha_2$
- What is the difference between testing this hypothesis one piece at a time, and testing them together (a *joint* test)?
- What is the alternate hypothesis?

Hypothesis testing

- We already know about the hypothesis test on a single coefficient: simple t-test with a known distribution
- What about a joint hypothesis test? For example:
  - $H_0: \ \beta_1 = \beta_2 = 0$
  - $H_0: \ \beta_1 = \alpha_1$ & $\beta_2 = \alpha_2$
- What is the difference between testing this hypothesis one piece at a time, and testing them together (a *joint* test)?
- What is the alternate hypothesis?
- $H_A: \beta_1 \neq \alpha_1$ or $\beta_2 \neq \alpha_2$)
- These two events will not be independent: so $P(\beta_1 \neq \alpha_1$ or $\beta_2 \neq \alpha_2)$ cannot be derived just by considering the probabilities each on their own with a t-test

Joint hypothesis test

- To test a joint hypothesis, we can use an F-statistic which follows an $F-$distribution
  - Depends on two parameters, both "degrees of freedom" (d_1 and d_2)
  - $d_1 = q$ is number of restrictions we are testing
  - $d_2 = n - (k + 1)$ is number of observations minus number of parameters estimated (careful to include the intercept)
  - In large samples, we often assume $n \to \infty$
- Intuition: look at the fit of the regression under the null and alternate hypotheses; if fit under the null is much worse than under the alternate, reject the null
  - What do we use to check for fit?

Joint hypothesis test

- To test a joint hypothesis, we can use an F-statistic which follows an $F-$distribution
  - Depends on two parameters, both "degrees of freedom" (d_1 and d_2)
  - $d_1 = q$ is number of restrictions we are testing
  - $d_2 = n - (k + 1)$ is number of observations minus number of parameters estimated (careful to include the intercept)
  - In large samples, we often assume $n \to \infty$
- Intuition: look at the fit of the regression under the null and alternate hypotheses; if fit under the null is much worse than under the alternate, reject the null
  - What do we use to check for fit?
  - Test statistic based on R^2
  - If R^2 much better for unrestricted regression, then our restriction is likely not true

F-test

- This F-test is very powerful:
- Can test complex hypotheses over multiple coefficients
- Confidence *set* for coefficients: all combos not rejected under a given test
- In practice: most common F-test is that "our data matters"
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
  - F-test for $\beta_1 = \beta_2 = 0$
  - E.g., do we do meaningfully better at explaining variation in Y with our X variables than we would by just taking the mean?
  - This F-stat automatically reported in stata

Hypotheses on coefficient relationships

- Sometimes we may have a hypothesis on how β_1 and β_2 relate to each other, but not on their specific values
- Eg $\beta_1 = \beta_2$ in $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- Often, we can use simple algebra to rewrite our model in a way that is easy to test:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ &= \beta_0 + \beta_1 X_1 + \beta_1 X_2 \\ &= \beta_0 + \beta_1 (X_1 + X_2) \\ &= \beta_0 + \beta_1 (X_{12}) \end{aligned}$$

- So, just make a new variable X_{12} and do an F-test for the restricted regression against the unrestricted regression

Hypotheses on coefficient relationships

- Might get more complicated, with more coefficients and different relationships
- Math (and code) becomes easier if we can write our restrictions in the form:

$$a_1\beta_1 + a_2\beta_2 + \cdots + a_k\beta_k = 0$$

- Our example is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ and $H_0 : \ \beta_1 = \beta_2$
  - Want to write in the form $a_1\beta_1 + a_2\beta_2 = 0$
  - What are a_1 and a_2?

Hypotheses on coefficient relationships

- Might get more complicated, with more coefficients and different relationships
- Math (and code) becomes easier if we can write our restrictions in the form:

$$a_1\beta_1 + a_2\beta_2 + \cdots + a_k\beta_k = 0$$

- Our example is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ and $H_0 : \ \beta_1 = \beta_2$
  - Want to write in the form $a_1\beta_1 + a_2\beta_2 = 0$
  - What are a_1 and a_2?
  - $\beta_1 = \beta_2 \rightarrow \beta_1 - \beta_2 = 0$ so $a_1 = 1$ and $a_2 = -1$