Specification and Data Issues

Tyler Ransom

Univ of Oklahoma

Mar 12, 2019

Today's plan

1. Review reading topics on when $E(u|\mathbf{x}) \neq 0$
   1.1 Using Proxy Variables for Unobserved Explanatory Variables
   1.2 Measurement Error
   1.3 Nonrandom Sampling and Missing Data
   1.4 Outlying Observations
   1.5 Alternatives to OLS: LAD, Quantile Regression
2. In-class activity: start working on your project!

Proxy Variables

Using Proxy Variables

- **Omitted Variable Bias:** When something in u is correlated with x and y

- A solution: Collect information on proxy variables for the omitted one

- Example: ability in a wage equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$$

- Can't observe (or measure!) *abil*, but can observe its proxy: IQ test score

- Can put *IQ* in the wage equation instead of *abil*, under certain conditions

Conditions for proxy variables to be valid

- Need *abil* only changes with *IQ* and not the other *x*'s

- In math terms, we need:

$$E(abil|educ, exper, IQ) = E(abil|IQ)$$

- Otherwise we would still have omitted variable bias in ability

- Key requirement: *IQ* is such a good predictor of *abil* ...

  ... that none of the other *x*'s add information

Lagged dependent variables as proxies

- y_{-1} can be a good proxy for many omitted factors
- Example: school spending and test scores

$$score = \beta_0 + \beta_1 poverty + \beta_2 spending + \alpha_1 score_{-1} + u$$

- Want to compare obtain effect of spending on test scores
- Districts with high achievement may be better funded
- Including $score_{-1}$ holds fixed prior achievement, so β_2 is closer to causal
- More on this when we get to panel data (next month)

Measurement Error

Measurement error

- Recall: if x_j is correlated with u, we say x_j is **endogenous**
- Reasons x_j can be endogenous:
    - it's correlated with an omitted variable (in u)
    - x_j is measured with error
- Measurement error is sometimes fine, sometimes a serious problem

When y is measured with error

- We might have measurement error in y
- e.g. hourly wage in survey data
- Might be misreported by worker (rounding, bad memory, etc.)
- If measurement error is uncorrelated with x's, we don't have any problem
- OLS on the (mismeasured) y will give unbiased, consistent estimates
- If not, then all β's will be biased, inconsistent

When x is measured with error

- Typically, measurement error in an x is more problematic

- But depends on assumptions about the measurement error

- Suppose we have one x, and x_1^* is what we *would like to observe*

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1^* + u \\
x_1 &= x_1^* + e_1
\end{aligned}
$$

- Assume u uncorrelated with x_1^* and e_1

Measurement error in x

- We can plug in for x_1^* in the previous equation:

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

- Is u uncorrelated with e_1?

- The **classical errors-in-variables (CEV)** assumption is

$$Cov(x_1^*, e_1) = 0$$

  meaning the measurement error is uncorrelated with the true x_1

Implications of the CEV assumption

- The CEV assumption implies

$$\begin{aligned} Cov(x_1, e_1) = Cov(x_1^* + e_1, e_1) &= Var(e_1) &&= \sigma_{e_1}^2 \\ Var(x_1) = Var(x_1^* + e_1) &= Var(x_1^*) + Var(e_1) &&= \sigma_{x_1^*}^2 + \sigma_{e_1}^2 \end{aligned}$$

- Key Question: What is the cov. between x_1 and the error term, $u - \beta_1 e_1$?

$$\begin{aligned} Cov(x_1, u - \beta_1 e_1) &= Cov(x_1, u) - \beta_1 Cov(x_1, e_1) \\ &= -\beta_1 \sigma_{e_1}^2 < 0 \end{aligned}$$

because $Cov(x_1, u) = 0$ by assumption

Attenuation bias

- If you take the previous formulas to the limit ($N \to \infty$), can show that

$$\left|plim(\hat{\beta}_1)\right| < |\beta_1|$$

- This is called **attenuation bias**
- The estimator is systematically too close to zero when compared with β_1
- This is an important part of empirical work in economics
- But one should understand that it depends critically on the CEV assumption

Other assumption about measurement error in x

- The CEV assumption assumes

$$Cov(x_1^*, e_1) = 0$$

- If we instead assume that

$$Cov(x_1, e_1) = 0$$

  then there is no attenuation bias

Example: Measurement error in years of schooling

- Some people: standard $\log(wage)$ eq. underestimates returns to schooling

- Because *educ* is measured with error (i.e. attenuation bias), i.e.

$$\begin{aligned} educ &= educ^* + e_1 \\ Cov(educ^*, e_1) &= 0 \end{aligned}$$

  where $educ^*$ is actually schooling and *educ* is reported in a survey

- If this assumption is true, then $\hat{\beta}_{educ}$ will be biased towards 0

Example: Measurement error in years of schooling

- But is CEV assumption reasonable in this context?
- If *educ* is a discretized version of $educ^*$ (e.g.$educ^* = 12.5$)

  but *educ* is the highest grade completed (e.g. $educ = 12$),

  then $educ = educ^* + e_1$ with $Cov(educ^*, e_1) = 0$ is not plausible
- An implication of the CEV assumption is

$$Var(educ) > Var(educ^*)$$

  so that reported schooling is more variable than actual schooling
- This seems unlikely if *educ* is a truncated version of $educ^*$

Nonrandom Sampling and Missing Data

Nonrandom sampling

- Earlier, we discussed when it might be good to do weighted OLS

- One case: when we're interested in population-level descriptive stats

- There are two types of nonrandom sampling:

  1. Exogenous sampling

     - Sampling based on x's; poses no problems for OLS

  2. Endogenous sampling

     - Sampling based on y's or u's (in addition to x's); serious problems for OLS

Endogenous sampling: Examples

- Students at OU who have lower GPAs less likely to report GPA in a survey

- People with high incomes less likely to report income in a survey

- ... etc.

- This introduces a **sample-selection problem**

- The reason a unit has missing value is systematically related to u

- Need advanced econometrics to resolve sample-selection problem

Outlying Observations

Outliers

- **Outlier:** a data point that seems fundamentally different from the rest

- Data sets are messy and mistakes happen

- e.g. one row has proportion instead of percent

- Sometimes outliers are valid

- How do we tell which outliers are valid and which aren't?

- Often, just have to use best judgment

Influential observations

- We can also decide if the outlier is **influential**
- If we remove the observation, how much does it affect our $\hat{\beta}$'s?
- OLS is extremely sensitive to outlying observations
- This is because it is like an “average” as opposed to a “median”
- Can compute **studentized residuals**, but this doesn’t solve everything

Median/Quantile Regression

Least Absolute Deviations (LAD)

- Recall the Population Regression Function from the beginning of class:

$$E\left(y|x\right) = \beta_0 + \beta_1 x$$

- Instead of the conditional mean of y, how about the conditional median?

$$Med\left(y|x\right) = \alpha_0 + \alpha_1 x$$

- OLS gives us the conditional mean

- **Least Absolute Deviations (LAD)** gives the conditional median

Least Absolute Deviations (LAD)

- OLS minimizes:

$$\sum_i (y_i - \beta_0 - \beta_1 x_1)^2$$

- and LAD minimizes:

$$\sum_i |y_i - \alpha_0 - \alpha_1 x_1|$$

- What's the formula for the $\hat{\alpha}$'s? We can't write it down (no closed form)

- LAD is much less sensitive to outliers

- Similar to how Median is less sensitive than mean

Quantile Regression

- We might also be interested in other points of the conditional dist'n of y

- e.g. 10th, 25th, 75th, or 90th percentiles, etc.

- **Quantile regression** allows us to choose any percentile

- The idea is the same as LAD. To estimate in R:

```
library(quantreg)
est <- rq(y ~ x, data=df, tau = 0.5)
```

- Can change `tau` to other values between 0 and 1