FNCE 926
Empirical Methods in CF

Lecture 11 – Standard Errors & Misc.

Professor Todd Gormley

Announcements

- Exercise #4 is due
- Final exam will be in-class on April 26
  - After today, only two more classes of new material, structural & experiments
  - Practice exam available on Canvas

Background readings for today

- Readings for standard errors
  - Angrist-Pischke, Chapter 8
  - Bertrand, Duflo, Mullainathan (QJE 2004)
  - Petersen (RFS 2009)
- Readings for limited dependent variables
  - Angrist-Pischke, Sections 3.4.2 and 4.6.3
  - Greene, Section 17.3

Outline for Today

- Quick review of last lecture on matching
- Discuss standard errors and clustering
  - "Robust" or "Classical"?
  - Clustering: when to do it and how
- Discuss limited dependent variables
- Student presentations of "Matching" papers

Quick Review *[Part 1]*

- Matching is intuitive method
  - For each treated observation, find comparable untreated observations with similar covariates, X
    - They will act as estimate of unobserved counterfactual
    - Do the same thing for each untreated observation
  - Take average difference in outcome, y, of interest across all X to estimate ATE

Quick Review *[Part 2]*

- But, what are necessary assumptions for this approach to estimate ATE?
  - **Answer #1** = *Overlap*… Need both treated and control observations for X's
  - **Answer #2** = *Unconfoundedness*… Treatment is as good as random after controlling for X

Quick Review *[Part 3]*

- Matching is just a control strategy!
  - It does **NOT** control for unobserved variables that might pose identification problems
  - It is **NOT** useful in dealing with other problems like simultaneity and measurement error biases
- Typically used as robustness check on OLS or way to screen data before doing OLS

Quick Review *[Part 4]*

- Relative to OLS estimate of treatment effect…
  - Matching basically just weights differently
  - And, doesn't make functional form assumption
    - Angrist-Pischke argue you typically won't find large difference between two estimates if you have right X's and flexible controls for them in OLS

Quick Review *[Part 5]*

- **Many** choices to make when matching
  - Match on covariates or propensity score?
  - What distance metric to use?
  - What # of observations?
- Will want to show robustness of estimate to various different approaches

Standard Errors & LDVs – *Outline*

- Getting your standard errors correct
  - "Classical" *versus* "Robust" SE
  - Clustered SE
- Limited dependent variables

Getting our standard errors correct

- It is important to make sure we get our standard errors correct so as to avoid misleading or incorrect inferences
  - E.g. standard errors that are too small will cause us to reject the null hypothesis that our estimated β's are equal to zero too often
    - I.e. we might erroneously claim to found a "statistically significant" effect when none exists

Homoskedastic *or* Heteroskedastic?

- One question that typically comes up when trying figure out the appropriate SE is homoskedasticity *versus* heteroskedasiticity
  - Homoskedasticity assumes the variance of the residuals, u, around the CEF, does not depend on the covariates, X
  - Heteroskedasticity doesn't assume this

"Classical" versus "Robust" SEs *[Part 1]*

- What do the default standard errors reported by programs like Stata assume?
  - **Answer** = Homoskedasticity! This is what we refer to as "classical" standard errors
    - As we discussed in earlier lecture, this is typically **not** a reasonable assumption to make
    - "Robust" standard errors allow for heteroskedasticity and don't make this assumption

"Classical" versus "Robust" SEs *[Part 2]*

- Putting aside possible "clustering" (which we'll discuss shortly), should you always use robust standard errors?
  - **Answer** = Not necessarily! *Why?*
    - Asymptotically, "classical" and "robust" SE are correct, but both suffer from finite sample bias, that will tend to make them *too small* in small samples
    - "Robust" can sometimes be smaller than "classical" SE because of this bias or simple noise!

Finite sample bias in standard errors

- Finite sample bias is easily corrected in "classical" standard errors
  [Note: this is done automatically by Stata]
- This is not so easy with "robust" SEs…
  - Small sample bias can be <u>worse</u> with "robust" standard errors, and while finite sample corrections help, they typically don't fully remove the bias in small samples

Many different corrections are available

- Number of methods developed to try and correct for this finite-sample bias
  - By default, Stata automatically does one of these when use **vce(robust)** to calculate SE
  - But, there are other ways as well; e.g.,
    - regress y x, **vce(hc2)**
    - regress y x, **vce(hc3)** ←

Developed by Davidson and MacKinnon (1993); works better when heterogeneity is worse

Classical *vs.* Robust – *Practical Advice*

- Compare the robust SE to the classical SE and take **maximum** of the two
  - Angrist-Pischke argue that this will tend to be closer to the true SE in small samples that exhibit heteroskedasticity
    - If small sample bias is real concern, might want to use HC2 or HC3 instead of typical "robust" option
    - While SE using this approach might be too large if data is *actually* homoskedastic, this is less of concern

Standard Errors & LDVs – *Outline*

- Getting your standard errors correct
  - "Classical" *versus* "Robust" SE
  - Clustered SE
    - Violation of independence and implications
    - How big of a problem is it? And, when?
    - How do we correct for it with clustered SE?
    - When might clustering not be appropriate?
- Limited dependent variables

Clustered SE – *Motivation [Part 1]*

- "Classical" and "robust" SE depend on assumption of <u>independence</u>
  - i.e. our observations of *y* are random draws from some population and are hence uncorrelated with other draws
  - Can you give some examples where this is likely an unrealistic in CF? *[E.g. think of firm-level capital structure panel regression]*

Clustered SE – *Motivation [Part 2]*

- **Example Answers**
  - Firm's outcome (e.g. leverage) is likely correlated with other firms in same industry
  - Firm's outcome in year t is likely correlated to outcome in year t-1, t-2, etc.
- In practice, independence assumption is often unrealistic in corporate finance

Clustered SE – *Motivation [Part 3]*

- Moreover, this non-independence can cause **significant** <u>downward</u> biases in our estimated standard errors
  - E.g. standard errors can easily double, triple, etc. once we correct for this!
  - This is different than correcting for heterogeneity (i.e. "Classical" vs. "robust") tends to increase SE, at most, by about 30% according to Angrist-Pischke

Example violations of independence

- Violations tend to come in two forms

#1 – Cross-sectional "Clustering"

- E.g. outcome, y, *[e.g. ROA]* for a firm tends to be correlated with y of other firms in same industry because they are subject to same demand shocks

#2 – "Time series correlation"

- E.g. outcome, y, *[e.g. Ln(assets)]*for firm in year t tends to be correlated with the firm's y in other years because there serial correlation over time

Violation means non-*i.i.d.* errors

- Such violations basically mean that our errors, *u*, are not *i.i.d.* as assumed
  - Specifically, you can think of the errors as being correlated in groups, where

$$y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig} \longleftarrow$$

Error for observation *i*, which is group *g*

- $\text{var}(u_{ig}) = \sigma_u^2 > 0$
- $corr(u_{ig}, u_{jg}) = \rho_u \sigma_u^2 > 0$

ρ_u is called "intra-class correlation coefficient"

"Robust" and "classical" SEs assume this is zero

"Cluster" terminology

- **Key idea:** errors are correlated within groups (i.e. clusters), but <u>not</u> correlated across them
  - In cross-sectional setting with one time period, cluster might be industry; i.e. obs. within industry correlated but obs. in different industries are not
  - In time series correlation, you can think of the "cluster" as the multiple observations for each cross-section *[e.g. obs. on firm over time are the cluster]*

Why are classical SE too low?

- Intuition…
  - Broadly speaking, you don't have as much random variation as you really think you do when calculating your standard errors; hence, your standard errors are too small
    - E.g. if double # of observations by just replicating existing data, your classical SE will go down even though there is <u>no new information;</u> Stata does not realize the observations are <u>**not**</u> independent

Standard Errors & LDVs – *Outline*

- Getting your standard errors correct
  - "Classical" *versus* "Robust" SE
  - Clustered SE
    - Violation of independence and implications
    - How big of a problem is it? And, when?
    - How do we correct for it with clustered SE?
    - When might clustering not be appropriate?
- Limited dependent variables

How large, and what's important?

- By assuming a structure for the non-*i.i.d.* nature of the errors, we can derive a formula for are large the bias will be
- Can also see that two factors are key
  - Magnitude of intra-class correlation in *u*
  - Magnitude of intra-class correlation in *x*

Random effect version of violation

- To do this, we will assume the within-group correlation is driven by a random effect

$$y_{ig} = \beta_0 + \beta_1 x_{ig} + \underbrace{v_g + \eta_{ig}}_{u_{ig}}$$

All within-group correlation is captured by random effect v_g, and
$corr(\eta_{ig}, \eta_{jg}) = 0$

In this case, intra-class correlation coefficient is

$$\rho_u = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2}$$

Moulton Factor

- With this setting and a constant # of observations per group, *n*, we can show that

Correct SE of estimate

$$\frac{SE\left(\hat{\beta}_1\right)}{SE_c\left(\hat{\beta}_1\right)} = \left[1 + (n-1)\rho_u\right]^{\frac{1}{2}}$$

"Classical" SE you get when you don't account for correlation

This ratio is called the "Moulton Factor"; it tells you how much larger corrected SE will be

Moulton Factor – *Interpretation*

$$\frac{SE\left(\hat{\beta}_1\right)}{SE_c\left(\hat{\beta}_1\right)} = \left[1 + (n-1)\rho_u\right]^{\frac{1}{2}}$$

- **Interpretation** = If corrected for this non-*i.i.d.* structure within groups (i.e. clustering) classical SE will larger by factor equal to Moultan Factor
  - E.g. Moultan Factor = 3 implies your standard errors will triple in size once correctly account for correlation!

What affects the Moulton Factor?

$$\frac{SE\left(\hat{\beta}_1\right)}{SE_c\left(\hat{\beta}_1\right)} = \left[1+\left(n-1\right)\rho_u\right]^{\frac{1}{2}}$$

- Formula highlights importance of n and ρ_u
  - There is no bias if $\rho_u = 0$ or if $n = 1$ *[Why?]*
  - If ρ_u rises, the magnitude of bias rise *[Why?]*
  - If observations per group, n, rises bias is greater *[Why?]*

Answers about Moultan Factor

- **Answer #1:** ρ_u =0 implies each additional obs. provides new info. (as if they are *i.i.d.*), and (2) n=1 implies there aren't multiple obs. per cluster, so correlation is meaningless
- **Answer #2** = Higher intra-class correlation ρ_u means that new observations within groups provide even less new information, but classical standard errors don't realize this
- **Answer #3** = Classical SE thinks each additional obs. adds information, when in reality, it isn't adding that much. <u>So, bias is worse with more observations per group.</u>

Bottom line…

- Moultan Factor basically shows that downward bias is <u>greatest</u> when…
  - Dependent variable is highly correlated across observations within group *[e.g. high time series correlation in panel]*
  - And, we have a large # of observations per group *[e.g. large # of years in panel data]*

Expanding to uneven group sizes, we see that one other factor will be important as well…

Moulton Factor with *uneven* group sizes

$$\frac{SE(\hat{\beta}_1)}{SE_c(\hat{\beta}_1)} = \left(1 + \left[\frac{V(n_g)}{\bar{n}} + \bar{n} - 1\right]\rho_u \rho_x\right)^{\frac{1}{2}}$$

- n_g = size of group g
- $V(n_g)$ = variance of group sizes
- $\bar{n}$ = average group size
- ρ_u = intra-class correlation of errors, u
- ρ_x = intra-class correlation of covariate, x

Importance of non-*i.i.d.* *x*'s *[Part 1]*

$$\frac{SE\left(\hat{\beta}_1\right)}{SE_c\left(\hat{\beta}_1\right)} = \left(1 + \left[\frac{V\left(n_g\right)}{\bar{n}} + \bar{n} - 1\right]\rho_u \boxed{\rho_x}\right)^{\frac{1}{2}}$$

- Now we see that a non-zero correlation between *x*' s within groups is also important
  - **Question:** For what type of covariates will this correlation be high? *[i.e. when is clustering important?]*

Importance of non-*i.i.d.* *x*'s *[Part 2]*

- Prior formula shows that downward bias will also be bigger when…
  - Covariate only varies at group level; ρ_x will be exactly equal to 1 in those cases!
  - When covariate likely has a lot of time series dependence *[e.g. Ln(assets) of firm]*

Standard Errors & LDVs – *Outline*

- Getting your standard errors correct
  - "Classical" *versus* "Robust" SE
  - Clustered SE
    - Violation of independence and implications
    - How big of a problem is it? And, when?
    - How do we correct for it with clustered SE?
    - When might clustering not be appropriate?
- Limited dependent variables

How do we correct for this?

- There are many possible ways
  - *If* think error structure is random effects, as modeled earlier, then you could just multiply SEs by Moulton Factor…
  - But, more common way, which allows for any type of within-group correlation, is to "**cluster**" your standard errors
    - Implemented in Stata using **vce(cluster** *variable***)** option in estimation command

Clustered Standard Errors

- Basic idea is that it allows for **<u>any</u>** type of correlation of errors within group
  - E.g. if "cluster" was a firm's observations for years 1, 2, …, T, then it would allow corr(u_{i1}, u_{i2}) to be different than corr(u_{i1}, u_{i3})
    - Moultan factor approach would assume these are all the same which may be wrong
- Then, use independence across groups and asymptotics to estimate SEs

Clustering – *Cross-Sectional Example #1*

- Cross-sectional firm-level regression

$$y_{ij} = \beta_0 + \beta_1 x_j + \beta_2 z_{ij} + u_{ij}$$

  - y_{ij} is outcome for firm i in industry j
  - x_j only varies at industry level
  - z_{ij} varies within industry
  - **How should you cluster?**
    - **Answer =** Cluster at the industry level. Observations might be correlated within industries and one of the covariates, x, is <u>perfectly correlated</u> within industries

Clustering – *Cross-Sectional Example #2*

- Panel firm-level regression

$$y_{ijt} = \beta_0 + \beta_1 x_{jt} + \beta_2 z_{ijt} + u_{ijt}$$

  - y_{ijt} is outcome for firm i in industry j in year t
  - If you think firms are subject to similar industry shocks *over* time, how might you cluster?
    - **Answer =** Cluster at the industry-year level. Obs. might be correlated within industries in a given year
    - **But, what is probably even more appropriate?**

Clustering – *Time-series example*

- **Answer = cluster at industry level!**
  - This allows errors to be correlated over time within industries, which is *very* likely to the true nature of the data structure in CF
    - E.g. Shock to y (and error u) in industry j in year t is likely to be persistent and still partially present in year $t+1$ for many variables we analyze. So, corr(u_{ijt}, u_{ijt+1}) is <u>not</u> equal to zero. Clustering at industry level <u>would</u> account for this; clustering at industry-year level does **NOT** allow for any correlation across time

Time-series correlation

- **Such time-series correlation is very common in corporate finance**
  - E.g. leverage, size, etc. are all persistent over time
  - Clustering at industry, firm, or state level is a non-parametric and robust way to account for this!

Such serial correlation matters…

- When non-*i.i.d.* structure comes from serial correlation, the number of obs. per group, *n*, is the number of years for each panel
  - Thus, downward bias of classical or robust SE will be greater when have <u>more</u> years of data!
  - This can matter a lot in diff-in-diff… **[Why? Hint… there are three potential reasons]**

Serial correlation in diff-in-diff *[Part 1]*

- Serial correlation is particularly important in difference-in-differences because…

  #1 – Treatment indicator is highly correlated over time! *[E.g. for untreated firms is stays zero entire time, and for treated firms it stays equal to 1 after treatment]*

  #2 – We often have multiple pre- and post-treatment observations *[i.e. many observations per group]*

  #3 – And, dependent variables typically used often have a high time-series dependence to them

Serial correlation in diff-in-diff *[Part 2]*

- Bertrand, Duflo, and Mullainathan (QJE 2004) shows how bad this SE bias can be…
  - In standard type of diff-in-diff where <u>true</u> β=0, you'll find significant effect at 5% level in as much as 45 percent of the cases!
    - Remember… you should only reject null hypothesis 5% of time when the true effect is actually zero!

Firm FE *vs.* firm clusters

- Whether to use both FE and clustering often causes confusion for researchers
  - E.g. should you have both firm FE **<u>and</u>** clustering at firm level, and if so, what is it doing?

Easiest to understand why both might be appropriate with a few quick questions…

Firm FE *vs.* firm clusters *[Part 1]*

- Consider the following regression

$$y_{it} = \beta_0 + \beta_1 x_{it} + \underbrace{f_i + v_{it}}_{u_{it}}$$

  - y_{it} = outcome for firm i in year t
  - f_i = time-invariant unobserved heterogeneity
  - u_{it} is estimation error term if don't control for f_i
  - v_{it} is estimation error term if do control for f_i

Now answer the following questions…

Firm FE *vs.* firm clusters *[Part 2]*

- Why is it probably not a good idea to just use firm clusters with no firm FE?
  - **Answer =** Clustering only corrects standard errors; it doesn't deal with potential omitted variable bias if corr(x,f) $\neq$ 0!

Firm FE *vs.* firm clusters *[Part 3]*

- Why should we still cluster at firm level if even if we already have firm FE?
  - **Answer =** Firm FE removes <u>time-invariant</u> heterogeneity, f_i, from error term, but it doesn't account for possible *serial correlation*!
    - I.e. v_{it} might still be correlated with v_{it-1}, v_{it-2}, etc.
    - E.g. firm might get hit by shock in year t, and effect of that shock only *slowly* fades over time

Firm FE *vs.* firm clusters *[Part 4]*

- Will we get consistent estimates with both firm FE and firm clusters if serial dependence in error is driven by time-varying omitted variable that is correlated with x?
  - **Answer = No!**
    - Clustering only corrects SEs; it doesn't deal with potential bias in estimates because of an omitted variable problem!
    - And, Firm FE isn't sufficient in this case either because omitted variable isn't time-invariant

Clustering – *Practical Advice [Part 1]*

- Cluster at most aggregate level of variation in your covariates
  - E.g. if one of your covariates only varies at industry or state level, **<u>cluster at that level</u>**
- **<u>Always</u>** assume serial correlation
  - Don't cluster at state-year, industry-year, firm-year; cluster at state, industry, or firm *[this is particularly true in diff-in-diff]*

Clustering – *Practical Advice [Part 2]*

- Clustering is <u>not</u> a substitute for FE
  - Should use both FE to control for unobserved heterogeneity across groups and clustered SE to account for remaining serial correlation in *y*
- Be careful when # of clusters is small…

Standard Errors & LDVs – *Outline*

- Getting your standard errors correct
  - "Classical" *versus* "Robust" SE
  - Clustered SE
    - Violation of independence and implications
    - How big of a problem is it? And, when?
    - How do we correct for it with clustered SE?
    - When might clustering not be appropriate?
- Limited dependent variables

Need enough clusters…

- Asymptotic consistency of estimated clustered standard errors depends on # of clusters, **<u>not</u>** # of observations
  - I.e. only guaranteed to get precise estimate of correct SE if we have a lot of clusters
  - **If too few clusters, SE will be too low!**
    - This leads to practical questions like… "If I do firm-level panel regression with 50 states and cluster at state level, are there enough clusters?"

How important is this in practice?

- Unclear, but *probably* not a big problem
  - Simulations of Bertrand, et al (QJE 2004) suggest 50 clusters was plenty in their setting
    - In fact, bias wasn't that bad with 10 states
    - This is consistent with Hansen (JoE 2007), which finds that 10 clusters is enough when using clusters to account for serial correlation
  - But, can't guarantee this is always true, particularly in cross-sectional settings

If worried about # of clusters…

- You can try aggregating the data to remove time-series variation
  - E.g. in diff-in-diff, you would collapse data into one pre- and one post-treatment observation for each firm, state, or industry *[depending on what level you think is non-i.i.d]*, and then run the estimation
    - See Bertrand, Duflo, and Mullainathan (QJE 2004) for more details on how to do this

Cautionary Note on aggregating

- Can have very low power
  - Even if true $\beta \neq 0$, aggregating approach can often fail to reject the null hypothesis
- Not as straightforward (but still doable) when have multiple events at different times or additional covariates
  - See Bertrand, et al (QJE 2004) for details

Double-clustering

- Petersen (2009) emphasized idea of potentially clustering in second dimension
  - E.g. cluster for firm <u>and</u> cluster for year *[Note: this is not the same as a firm-year cluster!]*
  - Additional year cluster allows errors within year to be correlated in arbitrary ways
    - Year FE removes <u>common</u> error each year
    - Year clusters allows for things like when Firm A and B are highly correlated within years, but Firm A and C are not *[I.e. it isn't a common year error]*

But is double-clustering it necessary?

- In asset pricing, YES; in corporate finance… unclear, but **probably not**
  - In asset pricing, makes sense… some firms respond more to systematic shocks across years *[i.e. high equity beta firms!]*
  - But, harder to think why correlation or errors in a year would consistently differ across firms for CF variables
    - Petersen (2009) finds evidence consistent with this; adding year FE is probably sufficient in CF

Clustering in Panels – *More Advice*

- Within Stata, two commands can do the fixed effects estimation for you
  - **xtreg, fe**
  - **areg**
- They are identical, except when it comes to the cluster-robust standard errors
  - xtreg, fe cluster-robust SE are **smaller** because it doesn't adjust doF when clustering!

Clustering – **xtreg, fe** *versus* **areg**

- xtreg, fe are appropriate when FE are nested within clusters, which is commonly the case *[See Wooldridge 2010, Chapter 20]*
  - E.g. firm fixed effects are nested within firm, industry or state clusters. So, if you have firm FE and cluster at firm, industry, or state, use xtreg, fe
  - **Note:** xtreg, fe will give you an error if FE aren't nested in clusters; then you should use areg

Standard Errors & LDVs – *Outline*

- Getting your standard errors correct
  - "Classical" *versus* "Robust" SE
  - Clustered SE
- Limited dependent variables

Limited dependent variables (LDV)

- LDV occurs whenever outcome *y* is zero-one indicator *or* non-negative
  - If think about it, it is very common
    - Firm-level indicator for issuing equity, doing acquisition, paying dividend, etc.
    - Manager's salary *[b/c it is non-negative]*
  - Zero-one outcomes are also called discrete choice models

Common misperception about LDVs

- It is often thought that LDVs shouldn't be estimated with OLS
  - I.e. can't get causal effect with OLS
  - Instead, people argue you need to use estimators like Probit, Logit, or Tobit
- **But, this is wrong!**
  To see this, let's compare linear probability model to Probit & Logit

Linear probability model (LPM)

- LPM is when you use OLS to estimate model where outcome, *y*, is an indicator
  - Intuitive and very few assumptions
  - But admittedly, there are issues…
    - Predicted values can be outside [0,1]
    - Error will be heteroskedastic *[Does this cause bias?]*

      Answer = No! Just need to correct SEs

Logit & Probit *[Part 1]*

- Basically, they assume latent model

$$y^* = x'\beta + u$$

x ' is vector of controls, including constant

  - y^* is <u>unobserved</u> latent variable
  - And, we assume <u>observed</u> outcome, y, equals 1 if $y^*>0$, and zero otherwise
  - And, make assumption about error, u
    - Probit assumes u distributed normally
    - Logit assumes u is logistic distribution

What are Logit & Probit? *[Part 2]*

- With those assumptions, can show…
  - $Prob(y^* > 0 | x) = Prob(u < x'\beta | x) = F(x'\beta)$
  - And, thus $Prob(y = 1 | x) = F(x'\beta)$, where $F(x'\beta)$ is cumulative distribution function of u
- Because this is nonlinear, we use maximum likelihood estimator to estimate β
  - See Greene, Section 17.3 for details

What are Logit & Probit? *[Part 3]*

- **Note:** reported estimates in Stata are not marginal effects of interest!
  - I.e. you can't easily interpret them or compare them to what you'd get with LPM
  - Need to use post-estimation command "**margins**" to get marginal effects at average x

Logit, Probit *versus* LPM

- Benefits of Logit & Probit
  - Predicted probabilities from Logit & Probit will be between 0 and 1…

- **But, are they needed to estimate casual effect of some random treatment, *d*?**

NO! LPM is okay to use

- Just think back to natural experiments, where treatment, d, is exogenously assigned
  - Difference-in-differences estimators were shown to estimate average treatment effects
  - **Nothing** in those proofs required assumption that outcome y is continuous with full support!
- Same is true of non-negative y
  [I.e. Using Tobit isn't necessary either]

Instrumental variables and LDV

- Prior conclusions also hold in 2SLS estimations with exogenous instrument
  - 2SLS still estimates local average treatment effect with limited dependent variables

Caveat – Treatment with covariates

- There is, however, an issue when estimating treatment effects **when including other covariates**
  - CEF almost certainly won't be linear if there are additional covariates, x
    - It is linear if just have treatment, d, and no X's
- **But, Angrist-Pischke say not to worry…**

Angrist-Pischke view on OLS *[Part 1]*

- OLS still gives best <u>linear</u> approx. of CEF under less restrictive assumptions
  - If non-linear CEF has causal interpretation, then OLS estimate has causal interpretation as well
  - If assumptions about distribution of error are correct, non-linear models (e.g. Logit, Probit, and Tobit) basically just provide efficiency gain

Angrist-Pischke view on OLS *[Part 2]*

- But this efficiency gain (from using something like Probit or Logit) comes with cost…
  - Assumptions of Probit, Logit, and Tobit are <u>not</u> testable [can't observe u]
  - Theory gives little guidance on right assumption, and **if** assumption wrong, estimates <u>biased</u>!

Angrist-Pischke view on OLS *[Part 3]*

- Lastly, in practice, marginal effects from Probit, Logit, etc. will be similar to OLS
  - True *even* when average y is close to either 0 or 1 (i.e. there are a lot of zeros or lot of ones)

One other problem…

- Nonlinear estimators like Logit, Probit, and Tobit can't easily estimate interaction effects
  - E.g. can't have $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$
  - Marginal effects reported by statistical programs will be wrong; need to take additional steps to get correct interacted effects; See Ai and Norton (*Economic Letters* 2003)

One last thing to mention…

- With non-negative outcome y and random treatment indicator, d
  - OLS still correctly estimates ATE
  - But, **<u>don't</u>** condition on $y > 0$ when selecting your sample; that messes things up!
    - This is equivalent to "bad control" in that you're implicitly controlling for whether $y > 0$, which is also outcome of treatment!
    - See Angrist-Pischke, pages 99-100

Summary of Today *[Part 1]*

- Getting your SEs correct is important
  - If clustering isn't important, run both "classical" and "robust" SE; choose higher
  - But, use clustering when…
    - One of key independent variables only varies at aggregate level (e.g. industry, state, etc)
    - Or, dependent variable or independent variables likely exhibit time series dependence

Summary of Today *[Part 2]*

- Miscellaneous advice on clustering
  - Best to assume time series dependence; e.g. cluster at group level, not group-year
  - Firm FE and firm clusters are not substitutes
  - Use clustered SE produced by **xtreg** not **areg**

Summary of Today *[Part 3]*

- Can use OLS with LDVs
  - Still gives ATE when estimating treatment effect
  - In other settings (i.e. have more covariates), still gives best linear approx. of non-linear causal CEF
- Estimators like Probit, Logit, Tobit have their own problems

In First Half of Next Class

- Randomized experiments
  - Benefits…
  - Limitations
- Related readings; see syllabus

In Second Half of Next Class

- Papers are not necessarily connected to today's lecture on standard errors

Assign papers for next week…

- Heider and Ljungqvist (JFE Forthcoming)
  - Capital structure and taxes
- Iliev (JF 2010)
  - Effect of SOX on accounting costs
- Appel, Gormley, Keim (working paper, 2015)
  - Impact of passive investors on governance

Break Time

- Let’s take our 10 minute break
- We’ll do presentations when we get back