
Empirical Methods in CF

Lecture 8 – Standard Errors & Misc.

Professor Todd Gormley

Background readings for today

- Readings for matching
 - Roberts-Whited, *Section 6*
 - Angrist-Pischke, *Sections 3.3.1-3.3.3*
 - Wooldridge, *Section 21.3.5*

- Readings for standard errors
 - Angrist-Pischke, Chapter 8
 - Bertrand, Duflo, Mullainathan (QJE 2004)
 - Petersen (RFS 2009)

Miscellaneous & Errors – *Outline*

- Matching
- Hypothesis testing mistakes
- Standard errors & clustering

Matching Methods – *Basic Idea [Part 1]*

- Matching approach to estimate treatment effect is very intuitive and simple
 - For each treated observation, you find a “matching” untreated observation that serves as the de facto counterfactual
 - Then, compare outcome, y , of treated observations to outcome of matched obs.

Matching Methods – *Basic Idea [Part 2]*

- A bit more formally...
 - For each value of X , where there is both a treated and untreated observation...
 - Match treated observations with $X=X'$ to untreated observations with same $X=X'$
 - Take difference in their outcomes, y
 - Then, use average difference across all the X 's as estimate of treatment effect

Matching Methods – *Intuition*

- What two things is matching approach basically assuming about the treatment?
 - **Answer #1** = Treatment isn't random; if it were, would not need to match on X before taking average difference in outcomes
 - **Answer #2** = Treatment is random *conditional* on X ; i.e. controlling for X , untreated outcome captures the unobserved treated counterfactual

Matching is a “Control Strategy”

- Can think of matching as just a way to control for necessary X 's to ensure CMI strategy necessary for causality holds

What is another control strategy we could use to estimate treatment effect?

Matching and OLS; not that different

- **Answer = Regression!**
 - I.e. could just regress y onto indicator for treatment with necessary controls for X to ensure CMI assumption holds
 - E.g. to mirror matching estimator, you could just put in indicators for each value of X as the set of controls in the regression

So, how are matching & regression different?

Matching *versus* Regression

- Basically, can think of OLS estimate as particular weighted matching estimator
 - Demonstrating this difference in weighting can be a bit technical...
 - See Angrist-Pischke Section 3.3.1 for more details on this issue, but following example will help illustrate this...

Matching *vs* Regression – Example [P1]

- Example of difference in weighting...
 - First, do simple matching estimate
 - Then, do OLS where regress y on treatment indicator and you control for X 's by adding indicators for each value of X
 - This is very nonparametric and general way to control for covariates X
 - If think about it, this is very similar to matching; OLS will be comparing outcomes for treated and untreated with same X 's

Matching *vs* Regression – Example [P2]

- But, *even in this example*, you'll get different estimates from OLS and matching
 - Matching gives more weight to obs. with $X=X'$ when there are more treated with that X'
 - OLS gives more weight to obs. with $X=X'$ when there is more variation in treatment [*i.e. we observe a more equal ratio of treated & untreated*]

Matching *vs* Regression – Bottom Line

- Angrist-Pischke argue that, in general, differences between matching and OLS are not of much empirical importance
- **Moreover, similar to OLS, matching has a serious limitation...**

Matching – *Key Limitation* [Part 1]

- What sets matching estimator apart from other estimators like IV, natural experiments, and regression discontinuity?
 - **Answer** = It does not rely on any clear source of exogenous variation!
 - I.e. If OLS estimate of treatment effect is biased, so is a matching estimator of treatment effect!

Matching – *Key Limitation [Part 2]*

- And, we abandoned OLS for a reason...
 - If original treatment isn't random (i.e. exogenous), it is often difficult to believe that controlling for some X 's will somehow restore randomness
 - E.g. there could be problematic, unobserved heterogeneity
 - **Note:** regression discontinuity design is exception
 - Matching estimator suffers same problem!

Matching – *Key Limitation [Part 3]*

- Please remember this!
- Matching does **NOT** and **cannot** be used...
 - To fix simultaneity bias problem
 - To eliminate measurement error bias...
 - To fix omitted variable bias from unobservable variables [*can't match on what you can't observe!*]

Matching – *Summary*

- See my online “Matching” slides for details on how to do matching

Miscellaneous & Errors – *Outline*

- Matching
- Hypothesis testing mistakes
- Standard errors & clustering

Hypothesis testing mistakes

- As noted in lecture on natural experiments, triple-difference can be done by running double-diff in two separate subsamples
 - E.g. estimate effect of treatment on small firms; then estimate effect of treatment on large firms

Example inference from such analysis

<i>Sample =</i>	Small Firms	Large Firms	Low D/E Firms	High D/E Firms
Treatment * Post	0.031 (0.121)	0.104** (0.051)	0.056 (0.045)	0.081*** (0.032)
N	2,334	3,098	2,989	2,876
R-squared	0.11	0.15	0.08	0.21
Firm dummies	X	X	X	X
Year dummies	X	X	X	X

□ From above results, researcher often concludes...

- “Treatment effect is larger for bigger firms”
- “High D/E firms respond more to treatment”

Do you see any problem with either claim?

Be careful making such claims!

- **Answer = Yes!** The difference across subsamples may not actually be statistically significant!
 - Hard to know if different just eyeballing it because whether difference is significant depends on covariance of the two separate estimates
- **How can you properly test these claims?**

Example triple interaction result

	<i>Sample =</i>	All Firms
Treatment * Post		0.031 (0.121)
Treatment * Post * Large		0.073 (0.065)
N		5,432
R-squared		0.12
Firm dummies		X
Year dummies		X
Year * Large dummies		X

Difference is not actually statistically significant

Remember to interact year dummies & triple difference; otherwise, estimates won't match earlier subsamples

Practical Advice

- **Don't make claims you haven't tested; they could easily be wrong!**
 - Best to show relevant p -values in text or tables for any statistical significance claim you make
 - If difference isn't statistically significant [*e.g. p -value = 0.15*], can just say so; triple-diffs are noisy, so this isn't uncommon
 - Or, be more careful in your wording...
 - I.e. you could instead say, “we found an effect for large firms, but didn't find much evidence for small firms”

Miscellaneous & Errors – *Outline*

- Matching
- Hypothesis testing mistakes
- Standard errors & clustering

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* heteroskedasticity
 - 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 worse 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 independence
 - 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 downward** 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. $\ln(\text{assets})$] 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

$$y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig} \quad \leftarrow \text{Error for observation } i, \text{ which is group } g$$

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

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 not 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 no new information; Stata does not realize the observations are **not** independent

Standard Errors & LDVs – *Outline*

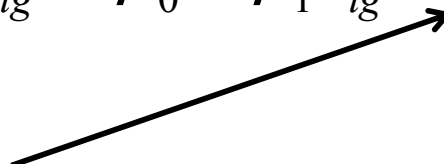
- Getting your standard errors correct
 - “Classical” *versus* “Robust” SE
 - Clustered SE
 - Violation of independence and implications
 - How big of a problem is it? And, when?
 - How do we correct for it with clustered SE?
 - When might clustering not be appropriate?
- Limited dependent variables

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 how 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 $\text{corr}(\eta_{ig}, \eta_{jg}) = 0$

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

“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*

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

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

What affects the Moulton Factor?

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

- 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 Moulton Factor

- **Answer #1:** $\rho_u = 0$ 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. So, bias is worse with more observations per group.

Bottom line...

- Moulton Factor basically shows that downward bias is greatest 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]

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


- 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; p_x will be exactly equal to 1 in those cases!
 - When covariate likely has a lot of time series dependence [e.g. $\text{Ln}(\text{assets})$ 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 any 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 $\text{corr}(u_{i1}, u_{i2})$ to be different than $\text{corr}(u_{i1}, u_{i3})$
 - Multivariate 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_j , is perfectly correlated 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 be 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, $\text{corr}(u_{ijt}, u_{ijt+1})$ is not equal to zero. Clustering at industry level would 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 more 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 it 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 true $\beta=0$, 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 **and** 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 $\text{corr}(x, f) \neq 0!$

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 time-invariant 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, **cluster at that level**
- **Always** 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 not 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, **not** # 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 and 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 common 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

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**