FNCE 926 Empirical Methods in CF

Lecture 9 – Common Limits & Errors

Professor Todd Gormley

Announcements

- Rough draft of research proposal due
  - Should have uploaded to Canvas
  - I'll try to e-mail feedback by next week

Background readings for today

- Ali, Klasa, and Yeung (RFS 2009)
- Gormley and Matsa (RFS 2014)

Outline for Today

- Quick review of last lecture on RDD
- Discuss common limitations and errors
  - Our data isn't perfect
  - Hypothesis testing mistakes
  - How not to control for unobserved heterogeneity
- Student presentations of "RDD" papers

Quick Review *[Part 1]*

- What is difference between "sharp" and "fuzzy" regression discontinuity design (RDD)?
  - **Answer =** With "sharp", change in treatment status only depends on *x* and cutoff *x'*; with "fuzzy", only *probability* of treatment varies at cutoff

Quick Review *[Part 2]*

- Estimation of **sharp** RDD

$$y_i = \alpha + \beta d_i + f(x_i - x') + d_i \times g(x_i - x') + u_i$$

  - d_i = indicator for $x \geq x'$
  - *f( )* and *g( )* are polynomial functions that control for effect of x on y
  - Can also do analysis in tighter window around threshold value x'

Quick Review *[Part 3]*

- Estimation of **fuzzy** RDD is similar…

$$y_i = \alpha + \beta d_i + f(x_i - x') + u_i$$

But use T_i and as instrument for d_i where

  - T_i is indicator for $x \geq x'$
  - And, d_i is indicator for treatment

Quick Review *[Part 4]*

- What are some standard internal validity tests you might want to run with RDD?
  - **Answers:**
    - Check robustness to different polynomial orders
    - Check robustness to bandwidth
    - Graphical analysis to show discontinuity in *y*
    - Compare other characteristics of firms around cutoff threshold to make sure no other discontinuities
    - And more…

Quick Review *[Part 5]*

- If effect of treatment is heterogeneous, how does this affect interpretation of RDD estimates?
  - **Answer =** They take on local average treatment effect interpretation, and fuzzy RDD captures only effect of compliers. Neither is problem for internal validity, but can sometimes limit external validity of finding

Common Limitations & Errors – *Outline*

- Data limitations
- Hypothesis testing mistakes
- How to control for unobserved heterogeneity

Data limitations

- The data we use is almost never perfect
  - Variables are often reported with error
  - Exit and entry into dataset typically not random
  - Datasets only cover certain types of firms

Measurement error – *Examples*

- Variables are often reported with error
  - Sometimes it is just innocent noise
    - E.g. Survey respondents self report past income with error *[because memory isn't perfect]*
  - Sometimes, it is more systematic
    - E.g. survey might ask teenagers # of times smoked marijuana, but teenagers that have smoked and have high GPA might say zero
- **How will these errors affect analysis?**

Measurement error – *Why it matters*

- **Answer** = Depends; but in general, hard to know exactly how this will matter
  - If *y* is mismeasured…
    - If only random noise, just makes SEs larger
    - But if <u>systematic</u> in some way *[as in second example]*, can cause bias if error is correlated with *x*'s
  - If *x* is mismeasured…
    - Even simple CEV causes attenuation bias on mismeasured *x* and biases on <u>all</u> other variables

Measurement error – *Solution*

- Standard measurement error solutions apply [see "Causality" lecture]
    - Though admittedly, measurement error is difficult to deal with unless know exactly source and nature of the error

Survivorship Issues – *Examples*

- In other cases, observations are included or missing for systematic reasons; e.g.
    - **Ex. #1** – Firms that do an IPO and are added to datasets that cover public firms may be different than firms that do not do an IPO
    - **Ex. #2** – Firms adversely affected by some event might subsequently drop out of data because of distress or outright bankruptcy
- **How can these issues be problematic?**

Survivorship Issues – *Why it matters*

- **Answer =** There is a selection bias, which can lead to incorrect inferences
  - **Ex. #1** – E.g. going public may not cause high growth; it's just that the firms going public were going to grow faster anyway
  - **Ex. #2** – Might not find adverse effect of event (or might understate it's effect) if some affected firms go bankrupt and are dropped

Survivorship Issues – *Solution*

- Again, no easy solutions; but, if worried about survivorship bias…
  - Check whether treatment (in diff-in-diff) is associated with observations being more or less likely to drop from data
  - In other analysis, check whether covariates of observations that drop are systematically different in a way that might be important

Sample is limited – *Examples*

- Observations in commonly used datasets are often limited to certain firms
  - **Ex. #1** – Compustat covers largest public firms
  - **Ex. #2** – Execucomp only provides incentives on CEOs of firms listed on S&P 1500
- **How this might affect our analysis?**

Sample is limited – *Why it matters*

- **Answer =** Need to be careful when making claims about external validity
  - **Ex. #1** – Might find no effect of treatment in Compustat because treatment effect is greatest for unobserved, smaller, private firms
  - **Ex. #2** – Observed correlations between incentives and risk-taking in Execucomp might not hold for smaller firms

Sample is limited – *Solution*

- Be careful with inferences to avoid making claims that lack external validity
- Argue that your sample is representative of economically important group
- Hand-collect your own data if theory your interested in testing requires it!
  - This can actually make for a great paper and is becoming increasingly important in finance

Interesting Example of Data Problem

- Ali, Klasa, and Yeung (RFS 2009) provide interesting example of data problem
  - They note the following…
    - Many theories argue that "industry concentration" is important factor in many finance settings
    - But researchers measure industry concentration (i.e. herfindahl index) using Compustat
  - **How might this be problematic?**

Ali, et al. – Example data problem *[P1]*

- **Answer =** Systematic measurement error!
  - Compustat doesn't include private firms; so using it causes you to mismeasure concentration
  - Ali, et al. find evidence of this by calculating concentration using U.S. Census data
    - Correlation between measures is just 13%
    - Moreover, error in Compustat measure is systematically related to some key variables, like turnover of firms in the industry

Ali, et al. – Example data problem *[P2]*

- Ali, et al. (RFS 2009) found it mattered; using Census measure overturns <u>four</u> previously published results
  - E.g. Concentration is positively related to R&D, not negatively related as previously argued
  - See paper for more details…

Common Limitations & Errors – *Outline*

- Data limitations
- Hypothesis testing mistakes
- How to control for unobserved heterogeneity

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

Sample =	Small Firms	Large Firms	Low D/E Firms	High D/E Firms
Treatment * Post	0.031	0.104**	0.056	0.081***
	(0.121)	(0.051)	(0.045)	(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

Sample =	**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"

Common Limitations & Errors – *Outline*

- Data limitations
- Hypothesis testing mistakes
- How to control for unobserved heterogeneity
  - How **not** to control for it
  - General implications
  - Estimating high-dimensional FE models

Unobserved Heterogeneity – *Motivation*

- Controlling for unobserved heterogeneity is a fundamental challenge in empirical finance
  - Unobservable factors affect corporate policies and prices
  - These factors may be correlated with variables of interest
- Important sources of unobserved heterogeneity are often common across groups of observations
  - Demand shocks across firms in an industry, differences in local economic environments, etc.

Many different strategies are used

- As we saw earlier, FE can control for unobserved heterogeneities and provide consistent estimates
- But, there are other strategies also used to control for unobserved group-level heterogeneity…
  - **"Adjusted-Y" (*Adj*Y)** – dependent variable is demeaned within groups *[e.g. 'industry-adjust']*
  - **"Average effects" (*Avg*E)** – uses group mean of dependent variable as control *[e.g. 'state-year' control]*

*Adj*Y and *Avg*E are widely used

- In *JF, JFE, and RFS…*
  - Used since at least the late 1980s
  - Still used, 60+ papers published in 2008-2010
  - Variety of subfields; asset pricing, banking, capital structure, governance, M&A, etc.
- Also been used in papers published in the *AER*, *JPE*, and *QJE* and top accounting journals, *JAR*, *JAE*, and *TAR*

But, *Adj*Y and *Avg*E are inconsistent

- As Gormley and Matsa (2014) shows…
  - Both can be **more** biased than OLS
  - Both can get **opposite** sign as true coefficient
  - In practice, bias is likely and trying to predict its sign or magnitude will typically be impractical
- **Now, let's see why they are wrong…**

The underlying model *[Part 1]*

- Recall model with unobserved heterogeneity

$$y_{i,j} = \beta X_{i,j} + f_i + \varepsilon_{i,j}$$

- i indexes groups of observations (e.g. industry); j indexes observations within each group (e.g. firm)
  - $y_{i,j}$ = dependent variable
  - $X_{i,j}$ = independent variable of interest
  - f_i = unobserved group heterogeneity
  - $\varepsilon_{i,j}$ = error term

The underlying model *[Part 2]*

- Make the standard assumptions:

N groups, *J* observations per group, where *J* is small and *N* is large

X and ε are *i.i.d.* across groups, but not necessarily *i.i.d.* within groups

$$\mathrm{var}(f) = \sigma_f^2, \mu_f = 0$$

$$\mathrm{var}(X) = \sigma_X^2, \mu_X = 0$$

$$\mathrm{var}(\varepsilon) = \sigma_\varepsilon^2, \mu_\varepsilon = 0$$

Simplifies some expressions, but doesn't change any results

The underlying model *[Part 3]*

- Finally, the following assumptions are made:

$$\text{cov}(f_i, \varepsilon_{i,j}) = 0$$

$$\text{co v}(X_{i,j}, \varepsilon_{i,j}) = \text{co v}(X_{i,j}, \varepsilon_{i,-j}) = 0$$

$$\text{cov}(X_{i,j}, f_i) = \sigma_{Xf} \neq 0$$

What do these imply?

Answer = Model is correct in that if we can control for *f*, we'll properly identify effect of *X;* but if we don't control for *f* there will be omitted variable bias

We already know that OLS is biased

True model is: $y_{i,j} = \beta X_{i,j} + f_i + \varepsilon_{i,j}$

But OLS estimates: $y_{i,j} = \beta^{OLS} X_{i,j} + u_{i,j}^{OLS}$

- By failing to control for group effect, f_i, OLS suffers from standard omitted variable bias

$$\hat{\beta}^{OLS} = \beta + \frac{\sigma_{Xf}}{\sigma_X^2}$$

Alternative estimation strategies are required…

Adjusted-Y (*AdjY*)

- Tries to remove unobserved group heterogeneity by demeaning the dependent variable within groups

***Adj*Y estimates:** $y_{i,j} - \bar{y}_i = \beta^{AdjY} X_{i,j} + u_{i,j}^{AdjY}$

where $\bar{y}_i = \frac{1}{J} \sum_{k \in group\ \mathrm{i}} \left(\beta X_{i,k} + f_i + \varepsilon_{i,k} \right)$

Note: Researchers often exclude observation at hand when calculating group mean or use a group median, but both modifications will yield similarly inconsistent estimates

Example *AdjY* estimation

- One example – firm value regression:

$$Q_{i,j,t} - \bar{Q}_{i,t} = \alpha + \boldsymbol{\beta}'\mathbf{X}_{i,j,t} + \varepsilon_{i,j,t}$$

  - $Q_{i,j,t}$ = Tobin's Q for firm *j*, industry *i*, year *t*
  - $\bar{Q}_{i,t}$ = mean of Tobin's Q for industry *i* in year *t*
  - X_{ijt} = vector of variables thought to affect value
  - Researchers might also include firm & year FE

Anyone know why *AdjY* is going to be inconsistent?

Here is why…

- Rewriting the group mean, we have:

$$\bar{y}_i = f_i + \beta \bar{X}_i + \bar{\varepsilon}_i ,$$

- Therefore, *Adj*Y transforms the true data to:

$$y_{i,j} - \bar{y}_i = \beta X_{i,j} - \beta \bar{X}_i + \varepsilon_{i,j} - \bar{\varepsilon}_i$$

What is the *Adj*Y estimation forgetting?

*Adj*Y can have omitted variable bias

- $\hat{\beta}^{adjY}$ can be inconsistent when $\beta \neq 0$

True model: $y_{i,j} - \bar{y}_i = \beta X_{i,j} - \boxed{\beta \bar{X}_i} + \varepsilon_{i,j} - \bar{\varepsilon}_i$

But, *Adj*Y estimates: $y_{i,j} - \bar{y}_i = \beta^{AdjY} X_{i,j} + u_{i,j}^{AdjY}$

- ❑ By failing to control for $\bar{X}_i$, *Adj*Y suffers from omitted variable bias when $\sigma_{X\bar{X}} \neq 0$

$$\hat{\beta}^{AdjY} = \beta - \beta \frac{\sigma_{X\bar{X}}}{\sigma_X^2}$$

In practice, a positive covariance between X and $\bar{X}$ will be common; e.g. industry shocks

Now, add a second variable, Z

- Suppose, there are instead **two** RHS variables

True model: $y_{i,j} = \beta X_{i,j} + \gamma Z_{i,j} + f_i + \varepsilon_{i,j}$

- Use same assumptions as before, but add:

$$\text{cov}(Z_{i,j}, \varepsilon_{i,j}) = \text{cov}(Z_{i,j}, \varepsilon_{i,-j}) = 0$$

$$\text{var}(Z) = \sigma_Z^2 , \mu_Z = 0$$

$$\text{cov}(X_{i,j}, Z_{i,j}) = \sigma_{XZ}$$

$$\text{cov}(Z_{i,j}, f_i) = \sigma_{Zf}$$

AdjY estimates with 2 variables

- With a bit of algebra, it is shown that:

$$\begin{bmatrix} \hat{\beta}^{AdjY} \\ \hat{\gamma}^{AdjY} \end{bmatrix} = \begin{bmatrix} \beta + \dfrac{\beta\left(\sigma_{xz}\sigma_{z\bar{x}} - \sigma_z^2\sigma_{x\bar{x}}\right) + \gamma\left(\sigma_{xz}\sigma_{z\bar{z}} - \sigma_z^2\sigma_{x\bar{z}}\right)}{\sigma_z^2\sigma_x^2 - \sigma_{xz}^2} \\ \gamma + \dfrac{\beta\left(\sigma_{xz}\sigma_{x\bar{x}} - \sigma_x^2\sigma_{z\bar{x}}\right) + \gamma\left(\sigma_{xz}\sigma_{x\bar{z}} - \sigma_x^2\sigma_{z\bar{z}}\right)}{\sigma_z^2\sigma_x^2 - \sigma_{xz}^2} \end{bmatrix}$$

Estimates of **both** β and γ can be inconsistent

Determining sign and magnitude of bias will typically be difficult

Average Effects (*Avg*E)

- *Avg*E uses group mean of dependent variable as control for unobserved heterogeneity

*Avg***E estimates:** $y_{i,j} = \beta^{AvgE} X_{i,j} + \gamma^{AvgE} \bar{y}_i + u_{i,j}^{AvgE}$

Average Effects (*Avg*E)

- Following profit regression is an *Avg*E example:

$$ROA_{i,s,t} = \alpha + \boldsymbol{\beta}'\mathbf{X}_{i,s,t} + \gamma\overline{ROA}_{s,t} + \varepsilon_{i,s,t}$$

  - $\overline{ROA}_{s,t}$ = mean of ROA for state *s* in year *t*
  - X_{ist} = vector of variables thought to profits
  - Researchers might also include firm & year FE

Anyone know why *Avg*E is going to be inconsistent?

*Avg*E has measurement error bias

- *Avg*E uses group mean of dependent variable as control for unobserved heterogeneity

*Avg***E estimates:** $y_{i,j} = \beta^{AvgE} X_{i,j} + \boxed{\gamma^{AvgE}\,\overline{y}_i} + u_{i,j}^{AvgE}$

Recall, true model: $y_{i,j} = \beta X_{i,j} + \boxed{f_i} + \varepsilon_{i,j}$

Problem is that $\overline{y}_i$ measures f_i with error

*Avg*E has measurement error bias

- Recall that group mean is given by $\bar{y}_i = f_i + \beta \bar{X}_i + \bar{\varepsilon}_i$,
  - Therefore, $\bar{y}_i$ measures f_i with error $-\beta \bar{X}_i - \bar{\varepsilon}_i$
  - As is well known, even classical measurement error causes all estimated coefficients to be inconsistent
- Bias here is complicated because error can be correlated with **both** mismeasured variable, f_i , and with $X_{i,j}$ when $\sigma_{X\bar{X}} \neq 0$

*Avg*E estimate of β with <u>one</u> variable

- With a bit of algebra, it is shown that:

$$\hat{\beta}^{AvgE} = \beta + \boxed{\frac{\sigma_{Xf}\left(\beta\sigma_{f\bar{X}} + \beta^2\sigma_{\bar{X}}^2 + \sigma_{\bar{\varepsilon}}^2 - \sigma_{\varepsilon\bar{\varepsilon}}\right) - \beta\sigma_{X\bar{X}}\left(\sigma_f^2 + \beta\sigma_{f\bar{X}} + \sigma_{\varepsilon\bar{\varepsilon}}\right)}{\sigma_X^2\left(\sigma_f^2 + 2\beta\sigma_{f\bar{X}} + \beta^2\sigma_{\bar{X}}^2 + \sigma_{\bar{\varepsilon}}^2\right) - \left(\sigma_{Xf} + \beta\sigma_{X\bar{X}}\right)^2}}$$

Determining magnitude and direction of bias is difficult

Covariance between X and $\bar{X}$ again problematic, but not needed for *Avg*E estimate to be inconsistent

Even non-*i.i.d.* nature of errors can affect bias!

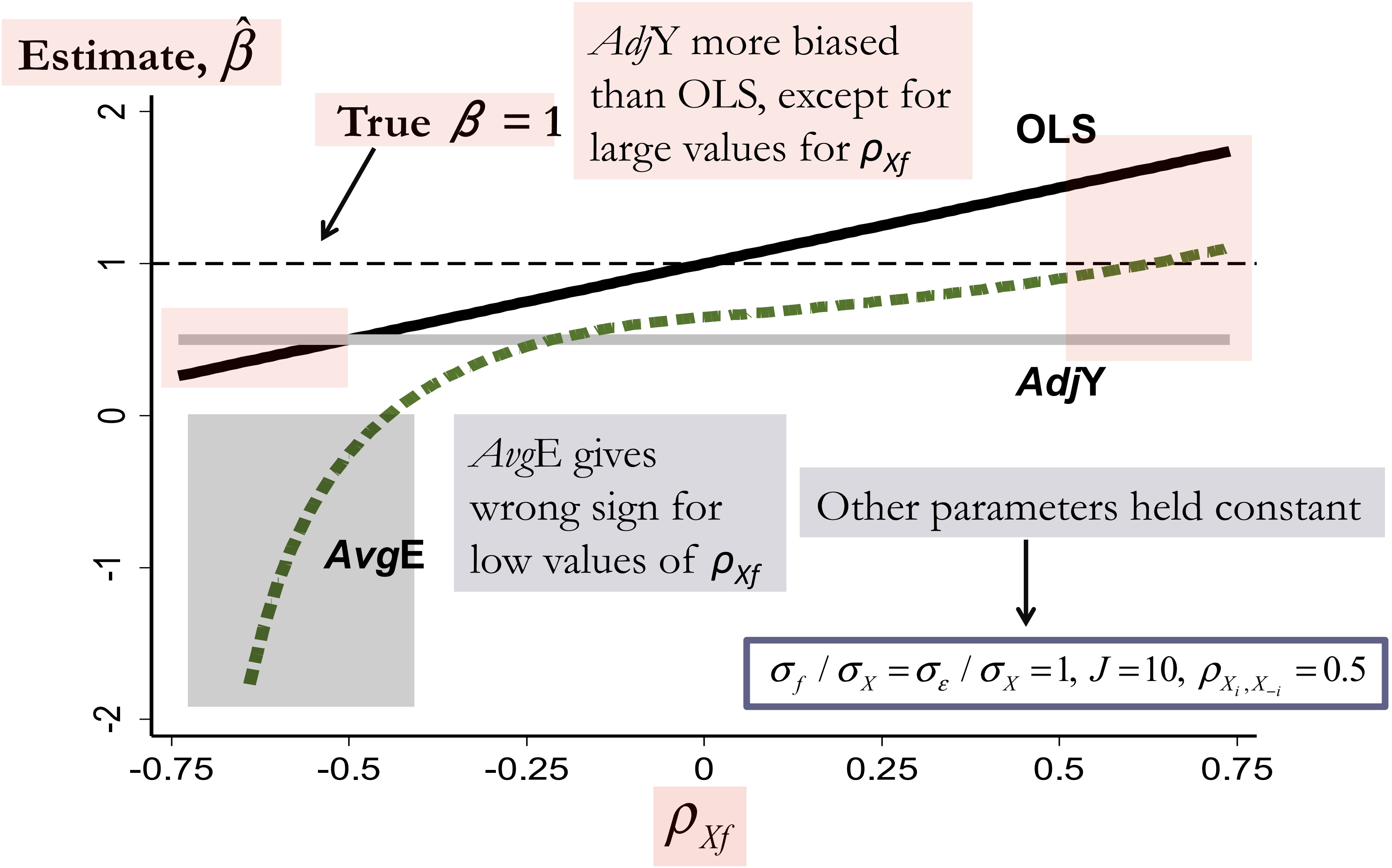
Comparing OLS, *Adj*Y, and *Avg*E

- Can use analytical solutions to compare relative performance of OLS, *Adj*Y, and *Avg*E
- To do this, we re-express solutions…
  - We use correlations (e.g. solve bias in terms of correlation between *X* and *f*, ρ_{Xf}, instead of σ_{Xf})
  - We also assume *i.i.d.* errors [just makes bias of *Avg*E less complicated]
  - And, we exclude the observation-at-hand when calculating the group mean, $\bar{X}_i$, …

Why excluding X_i doesn't help

- Quite common for researchers to exclude observation at hand when calculating group mean
  - It does remove mechanical correlation between X and omitted variable, $\bar{X}_i$, but it does **not** eliminate the bias
  - In general, correlation between X and omitted variable, $\bar{X}_i$, is non-zero whenever $\bar{X}_i$ is not the same for every group i
    - This variation in means across group is almost assuredly true in practice; *see paper for details*

ρ_{Xf} has large effect on performance



True $\beta = 1$

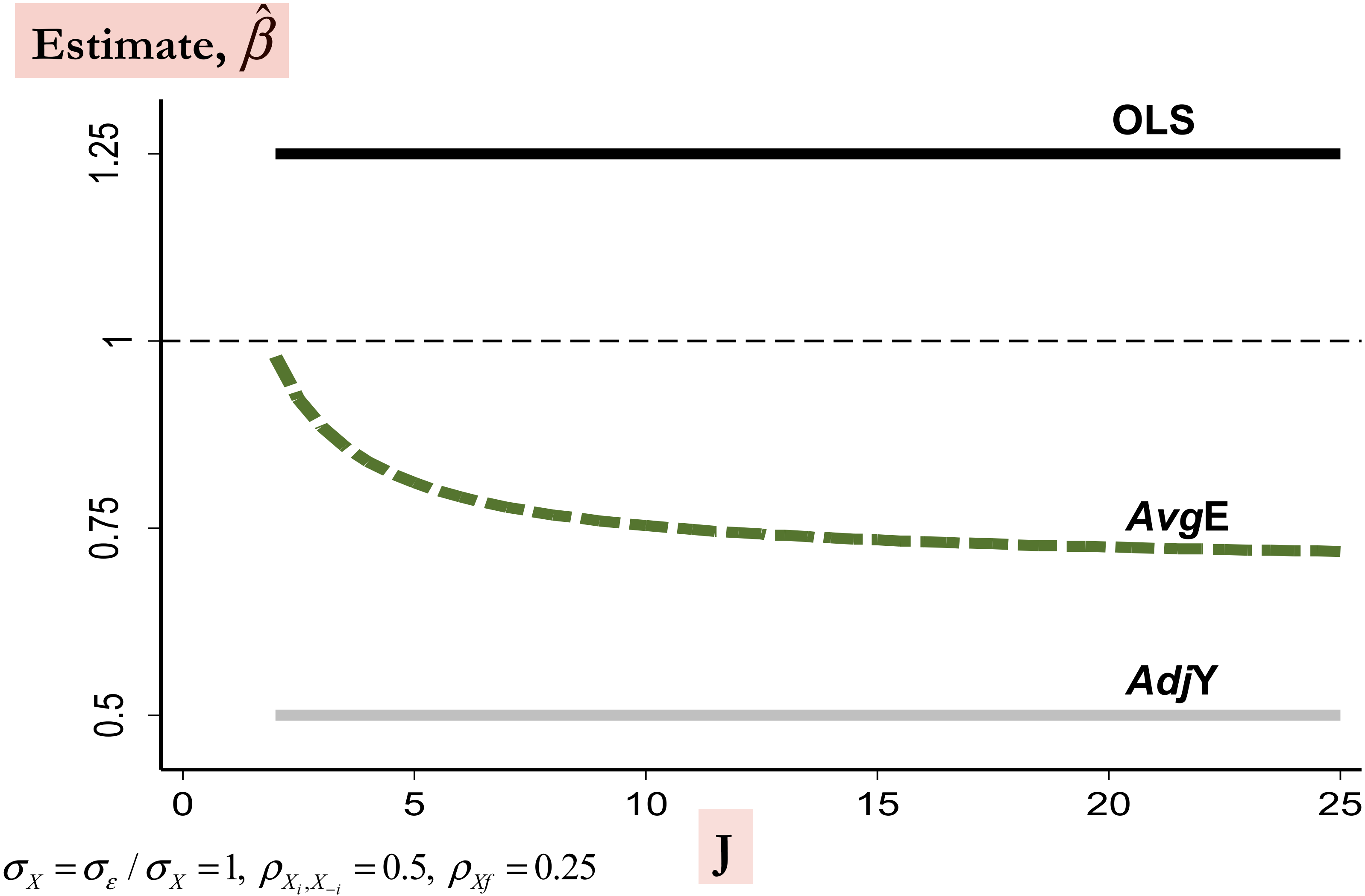
*Adj*Y more biased than OLS, except for large values for ρ_{Xf}

*Avg*E gives wrong sign for low values of ρ_{Xf}

Other parameters held constant

$\sigma_f / \sigma_X = \sigma_\varepsilon / \sigma_X = 1,\ J = 10,\ \rho_{X_i, X_{-i}} = 0.5$

More observations need not help!



$\sigma_f / \sigma_X = \sigma_\varepsilon / \sigma_X = 1,\ \rho_{X_i, X_{-i}} = 0.5,\ \rho_{Xf} = 0.25$

Summary of OLS, *Adj*Y, and *Avg*E

- In general, all three estimators are inconsistent in presence of unobserved group heterogeneity
- *Adj*Y and *Avg*E may not be an improvement over OLS; depends on various parameter values
- *Adj*Y and *Avg*E can yield estimates with opposite sign of the true coefficient

Fixed effects (FE) estimation

- **Recall:** FE adds dummies for each group to OLS estimation and is **consistent** because it directly controls for unobserved group-level heterogeneity

- Can also do FE by demeaning all variables with respect to group *[i.e. do 'within transformation']* and use OLS

FE estimates: $y_{i,j} - \bar{y}_i = \beta^{FE}\left(X_{i,j} - \bar{X}_i\right) + u_{i,j}^{FE}$

True model: $y_{i,j} - \bar{y}_i = \beta\left(X_{i,j} - \bar{X}_i\right) + \left(\varepsilon_{i,j} - \bar{\varepsilon}_i\right)$

Comparing FE to *Adj*Y and *Avg*E

- To estimate effect of X on Y controlling for Z
  - One could regress Y onto both X and Z…

    Add group FE
  - *Or*, regress residuals from regression of Y on Z onto residuals from regression of X on Z

    Within-group transformation!
- *Adj*Y and *Avg*E aren't the same as finding the effect of X on Y controlling for Z because...
  - *Adj*Y only partials Z out from Y
  - *Avg*E uses fitted values of Y on Z as control

The differences will matter! *Example #1*

- Consider the following capital structure regression:

$$(D/A)_{i,t} = \alpha + \boldsymbol{\beta}\mathbf{X}_{i,t} + f_i + \varepsilon_{i,t}$$

  - $(D/A)_{it}$ = book leverage for firm *i*, year *t*
  - X_{it} = vector of variables thought to affect leverage
  - f_i = firm fixed effect

- We now run this regression for each approach to deal with firm fixed effects, using 1950-2010 data, winsorizing at 1% tails…

Estimates vary considerably

Dependent variable = book leverage

	OLS	***Adj* Y**	***Avg* E**	**FE**
Fixed Assets/ Total Assets	0.270***	0.066***	0.103***	0.248***
	(0.008)	(0.004)	(0.004)	(0.014)
Ln(sales)	0.011***	0.011***	0.011***	0.017***
	(0.001)	0.000	0.000	(0.001)
Return on Assets	-0.015***	0.051***	0.039***	-0.028***
	(0.005)	(0.004)	(0.004)	(0.005)
Z-score	-0.017***	-0.010***	-0.011***	-0.017***
	0.000	(0.000)	(0.000)	(0.001)
Market-to-book Ratio	-0.006***	-0.004***	-0.004***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	166,974	166,974	166,974	166,974
R-squared	0.29	0.14	0.56	0.66

The differences will matter! *Example #2*

- Consider the following firm value regression:

$$Q_{i,j,t} = \alpha + \boldsymbol{\beta}'\mathbf{X}_{i,j,t} + f_{j,t} + \varepsilon_{i,j,t}$$

  - Q = Tobin's Q for firm *i*, industry *j*, year *t*
  - X_{ijt} = vector of variables thought to affect value
  - $f_{j,t}$ = industry-year fixed effect

- We now run this regression for each approach to deal with **industry-year** fixed effects…

Estimates vary considerably

Dependent Variable = Tobin's Q

	OLS	***Adj* Y**	***Avg* E**	**FE**
Delaware Incorporation	0.100***	0.019	0.040	0.086**
	(0.036)	(0.032)	(0.032)	(0.039)
Ln(sales)	-0.125***	-0.054***	-0.072***	-0.131***
	(0.009)	(0.008)	(0.008)	(0.011)
R&D Expenses / Assets	6.724***	3.022***	3.968***	5.541***
	(0.260)	(0.242)	(0.256)	(0.318)
Return on Assets	-0.559***	-0.526***	-0.535***	-0.436***
	(0.108)	(0.095)	(0.097)	(0.117)
Observations	55,792	55,792	55,792	55,792
R-squared	0.22	0.08	0.34	0.37

Common Limitations & Errors – *Outline*

- Data limitations
- Hypothesis testing mistakes
- How to control for unobserved heterogeneity
  - How **not** to control for it
  - General implications
  - Estimating high-dimensional FE models

General implications

- With this framework, easy to see that other commonly used estimators will be biased

Other *Adj*Y estimators are problematic

- Same problem arises with other *Adj*Y estimators
  - Subtracting off median or value-weighted mean
  - Subtracting off mean of matched control sample *[as is customary in studies if diversification "discount"]*
  - Comparing "adjusted" outcomes for treated firms pre-versus post-event *[as often done in M&A studies]*
  - Characteristically adjusted returns *[as used in asset pricing]*

AdjY-type estimators in asset pricing

- Common to sort and compare stock returns across portfolios based on a variable thought to affect returns
- But, returns are often first "characteristically adjusted"
  - I.e. researcher subtracts the average return of a benchmark portfolio containing stocks of similar characteristics
  - This is equivalent to *AdjY*, where "adjusted returns" are regressed onto indicators for each portfolio
- **Approach fails to control for how avg. independent variable varies across benchmark portfolios**

Asset Pricing *AdjY* – *Example*

- Asset pricing example; sorting returns based on R&D expenses / market value of equity

Characteristically adjusted returns by R&D Quintile (i.e., *Adj* Y)

Missing	Q1	Q2	Q3	Q4	Q5
-0.012***	-0.033***	-0.023***	-0.002	0.008	0.020***
(0.003)	(0.009)	(0.008)	(0.007)	(0.013)	(0.006)

We use industry-size benchmark portfolios and sorted using R&D/market value

Difference between Q5 and Q1 is 5.3 percentage points

Estimates vary considerably

Dependent Variable = Yearly Stock Return

	***Adj* Y**	**FE**
R&D Missing	0.021**	0.030***
	(0.009)	(0.010)
R&D Quintile 2	0.01	0.019
	(0.013)	(0.014)
R&D Quintile 3	0.032***	0.051***
	(0.012)	(0.018)
R&D Quintile 4	0.041***	0.068***
	(0.015)	(0.020)
R&D Quintile 5	0.053***	0.094***
	(0.011)	(0.019)
Observations	144,592	144,592
R^2	0.00	0.47

Same *Adj*Y result, but in regression format; quintile 1 is excluded

Use benchmark-period FE to transform both returns and R&D; this is equivalent to double sort

What if *Adj*Y or *Avg*E is true model?

- If data exhibited structure of *Avg*E estimator, this would be a peer effects model
  [i.e. group mean affects outcome of other members]
  - In this case, <u>none</u> of the estimators (OLS, *Adj*Y, *Avg*E, or FE) reveal the true β *[Manski 1993; Leary and Roberts 2010]*
- Even if interested in studying $y_{i,j} - \bar{y}_i$, *Adj*Y only consistent if $X_{i,j}$ does not affect $y_{i,j}$

Common Limitations & Errors – *Outline*

- Data limitations
- Hypothesis testing mistakes
- How to control for unobserved heterogeneity
  - How **not** to control for it
  - General implications
  - Estimating high-dimensional FE models

Multiple high-dimensional FE

- Researchers occasionally motivate using *Adj*Y and *Avg*E because FE estimator is computationally difficult to do when there are more than one FE of high-dimension

Now, let's see why this is (and isn't) a problem…

LSDV is usually needed with two FE

- Consider the below model with two FE

$$y_{i,j,k} = \beta X_{i,j,k} + f_i + \delta_k + \varepsilon_{i,j,k}$$

Two separate group effects

  - Unless panel is balanced, within transformation can only be used to remove one of the fixed effects
  - For other FE, you need to add dummy variables *[e.g. add time dummies and demean within firm]*

Why such models can be problematic

- Estimating FE model with many dummies can require a lot of computer memory
  - E.g., estimation with both firm and 4-digit industry-year FE requires ≈ 40 GB of memory

This is growing problem

- Multiple unobserved heterogeneities increasingly argued to be important
  - Manager and firm fixed effects in executive compensation and other CF applications *[Graham, Li, and Qui 2011, Coles and Li 2011]*
  - Firm and industry×year FE to control for industry-level shocks *[Matsa 2010]*

But, there are solutions!

- There exist two techniques that can be used to arrive at consistent FE estimates without requiring as much memory

  #1 – Interacted fixed effects

  #2 – Memory saving procedures

#1 – Interacted fixed effects

- Combine multiple fixed effects into one-dimensional set of fixed effect, and remove using within transformation
  - E.g. firm and industry-year FE could be replaced with firm-industry-year FE

But, there are limitations…

- Can severely limit parameters you can estimate
- Could have serious attenuation bias

#2 – Memory-saving procedures

- Use properties of sparse matrices to reduce required memory, *e.g. Cornelissen (2008)*
- Or, instead iterate to a solution, which eliminates memory issue entirely, *e.g. Guimaraes and Portugal (2010)*
  - See paper for details of how each works
  - Both can be done in Stata using user-written commands FELSDVREG and REGHDFE

These methods work…

- Estimated typical capital structure regression with firm and 4-digit industry×year dummies
  - Standard FE approach would not work; my computer did not have enough memory…
  - Sparse matrix procedure took 8 hours…
  - Iterative procedure took 5 minutes

Summary of Today *[Part 1]*

- Our data isn't perfect…
  - Watch for measurement error
  - Watch for survivorship bias
  - Be careful about external validity claims
- Make sure to test that estimates across subsamples are actually statistically different

Summary of Today *[Part 2]*

- Don't use *Adj*Y or *Avg*E!
- But, do use fixed effects
  - Should use benchmark portfolio-period FE in asset pricing rather than char-adjusted returns
  - Use iteration techniques to estimate models with multiple high-dimensional FE

In First Half of Next Class

- Matching
  - What it does…
  - And, what it doesn't do
- Related readings… see syllabus

Assign papers for next week…

- Gormley and Matsa (working paper, 2015)
  - Corporate governance & playing it safe preferences
- Ljungqvist, Malloy, Marston (JF 2009)
  - Data issues in I/B/E/S
- Bennedsen, et al. (working paper, 2012)
  - CEO hospitalization events

No comments needed from other groups

Break Time

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