Financial Econometrics II – Cross Section and Panel Data

Endogeneity, Panel Data, and Standard Errors

Andreas Fuster
Swiss Finance Institute @ EPFL

SFI Léman PhD program – 2024, Lecture 1

Who I am

- Associate Professor of Finance at SFI@EPFL since 2021
- Previously 10 years in central banking (7 years at NY Fed, 3 years at Swiss National Bank)
  - before that: BA (‘licence’) at HEC Lausanne, MPhil at Oxford, PhD at Harvard, all in Economics
- Research:
  - Household finance
  - Real estate finance
  - Banking / financial intermediation / fintech
  - Macro
  - Behavioral/experimental economics

This (half-)course

s:fi

- Overview of cross section and panel data (“micro-econometric”) methods that are commonly used in finance
  - ”theory”, applications in literature, own practice
- 4 lectures
  - Today, Dec 5, 12, and 19; 14:00-17:00
- 2 problem sets:
  - first one due December 12 (before lecture 3), already on Moodle
  - second one due in January (exact date tbd)
- The problem sets jointly account for 50% of grade for this part (so 25% for course overall); exam (Jan 20) for the other 50%

Course readings (cf. Syllabus)

- The course is most closely based on the handbook chapter by Roberts and Whited (2012)
  - would recommend reading (SSRN version), though a bit dated now
- More recent useful book with finance applications: Verbeek (2021)
- "Chatty" overviews of much of the material are Angrist and Pischke (2009), Cunningham (2021) and Huntington-Klein (2022)
  - Cunningham & Huntington-Klein are freely available online
- For technical background, the Wooldridge book is the classic
- Great overviews of recent developments on certain topics are the NBER Summer Institute "Methods Lectures"

Paper discussions

- In each of the remaining three lectures after today, we will spend 30-45 minutes discussing one empirical paper in detail

- You are expected to read this paper ahead of time & come prepared for the discussion – I will "cold call"

- I suggest that you write notes for yourself that cover the following
  - What is the empirical approach? Potential endogeneity issues & how does the paper address them?
  - Data used & main results? Economic interpretation?
  - What do you like about the paper?
  - What could be improved / wasn't clear to you?

  Try to link in particular to things we discussed in the lectures. Also think about the way results are communicated (tables/figures/writing).

Paper discussions

- Reading for next time (on Moodle):

The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXXV, NO. 4 • AUGUST 2020

Do CEOs Matter? Evidence from Hospitalization Events

MORTEN BENNEDSEN, FRANCISCO PÉREZ-GONZÁLEZ, and DANIEL WOLFENZON*

ABSTRACT

Using variation in firms' exposure to their CEOs resulting from hospitalization, we estimate the effect of chief executive officers (CEOs) on firm policies, holding firm-CEO matches constant. We document three main findings. First, CEOs have a significant effect on profitability and investment. Second, CEO effects are larger for younger CEOs, in growing and family-controlled firms, and in human-capital-intensive industries. Third, CEOs are unique: the hospitalization of other senior executives does not have similar effects on the performance. Overall, our findings demonstrate that CEOs are a key driver of firm performance, which suggests that CEO contingency plans are valuable.

Approach / acknowledgment

- My goal in this course is not to cover the technical details of different methods in detail, but to
  1. show you the basics of the most common methods (so you can understand and assess papers using them), and
  2. direct you towards the (practical) "research frontier" – an important goal is for you to "know where to look" if you want to apply these methods in your own research

- The course syllabus and slides incorporates material from several other lecturers (see syllabus) – most directly Prof. Philip Valta from the University of Bern, who taught this course in previous years
  — all errors are my own

Overview

- Motivation
- OLS and endogeneity
- Biases from endogeneity
- Panel data
- Standard errors

Motivation

- In empirical corporate finance research (broadly defined), the most important and pervasive issue confronting researchers is **endogeneity**.
- Endogeneity may be loosely defined as the **correlation between the explanatory variables and the error term** in a regression.
- Endogeneity leads to **biased** and **inconsistent** parameter estimates that make reliable inference virtually impossible.
- Endogeneity concerns are present in almost every study.
  - exception: randomized controlled trials (experiment)
- How can we address endogeneity concerns in finance and economics research?

Causality

- Ideally, when doing empirical research, we want to make **causal** statements
  - Example 1: What is the **effect** of a change in the bankruptcy law on firms' investment choices?
  - Example 2: How does competition **affect** the firms' financial leverage?
  - Example 3: How do stock and option ownership **affect** bankers' incentives to take on risk?

- In other words, we would like to go beyond saying that variables x and y are **associated** or **correlated** with each other.

Identification

- Empirical economists often discuss the ability to make causal statements in the context of “identification”
- The term identification has many different meanings in econometrics
  - Lewbel (2019, JEL): “Identification zoo” – over two dozen different identification terms in the literature
- “Technical” definition: an econometric model is identified if the parameters of interest (e.g., the coefficients in a regression model) can be uniquely determined or estimated from the observed data
- But when discussing empirical work, we typically mean by ”identification” the process of figuring out **what part of the variation in your data answers your (causal) research question**
  - e.g. how do you know you have a causal effect and not a correlation

How prevalent are the methods we will see?

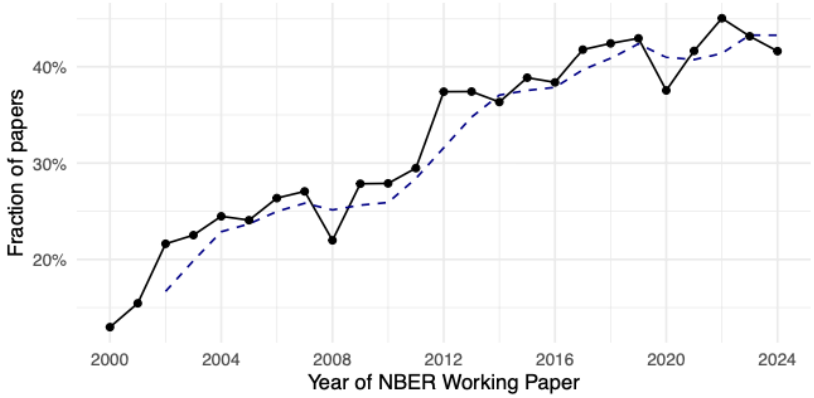
- Currie, Kleven, and Zwiers (2020) tracked the use of different "identification technologies" in academic papers (top econ journals and NBER working papers) over time, focusing on "applied micro".

- Goldsmith-Pinkham (2024) extends this study through 2024 and including macro and finance papers (but only using NBER WPs).

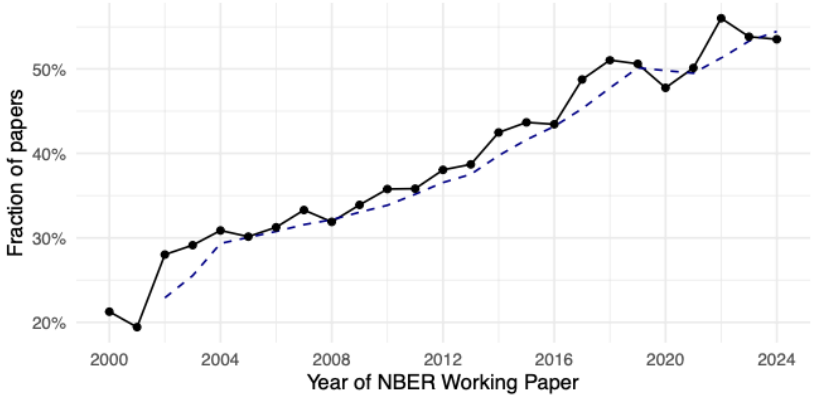
- First result (next slide): strong upward trend in
  - mentions of "identification"
  - use of experimental and quasi-experimental (our focus) methods
  - administrative data (large datasets that are often not publicly accessible)
  - use of figures vs. tables ("graphical revolution")

- Often referred to as "the **credibility revolution**"

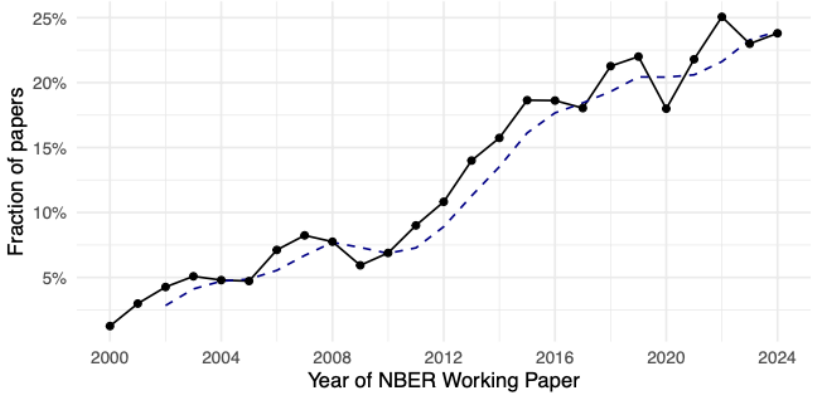
How prevalent are the methods we will see?



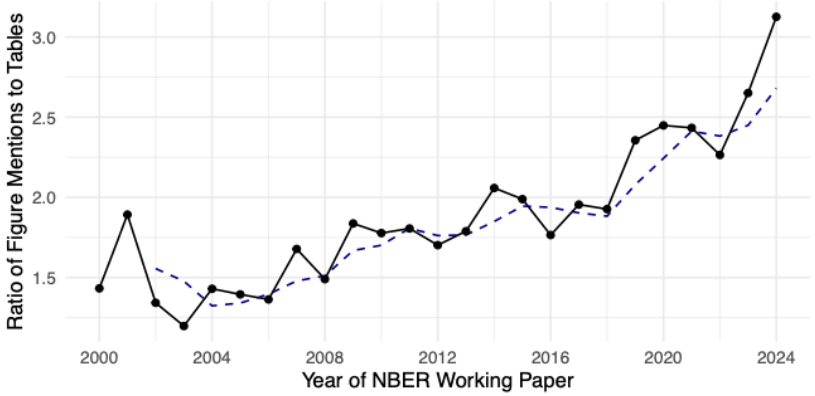
(a) Identification



(b) All experimental and quasi-experimental methods



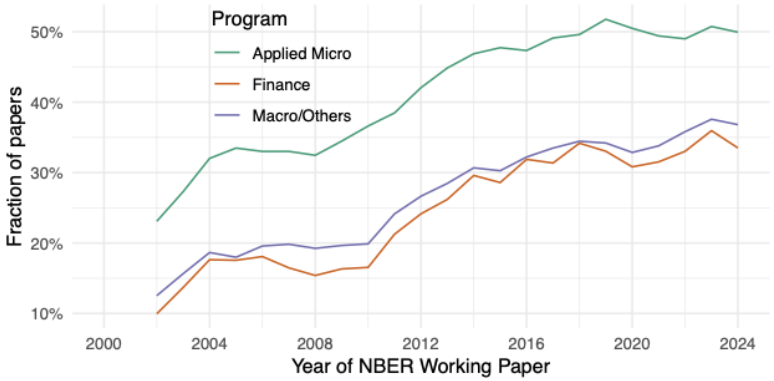
(c) Administrative data



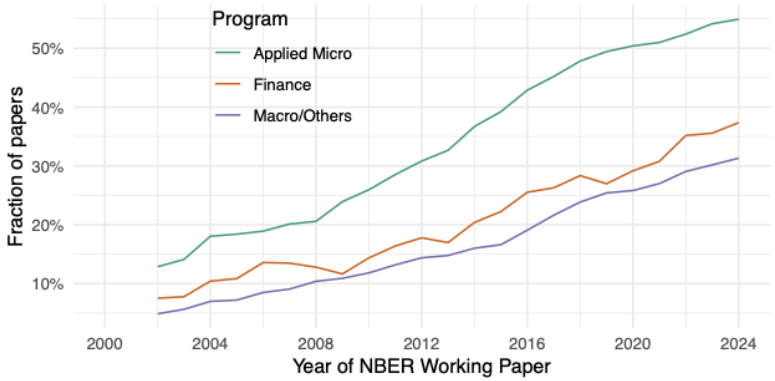
(d) Graphical revolution

How prevalent are the methods we will see?

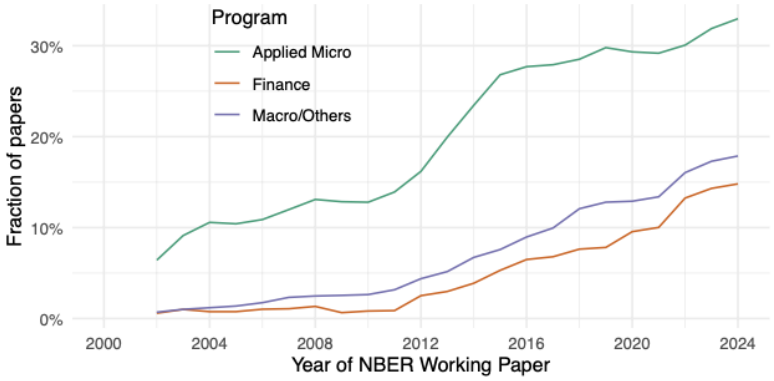
- Finance is below applied micro but upward trend similar



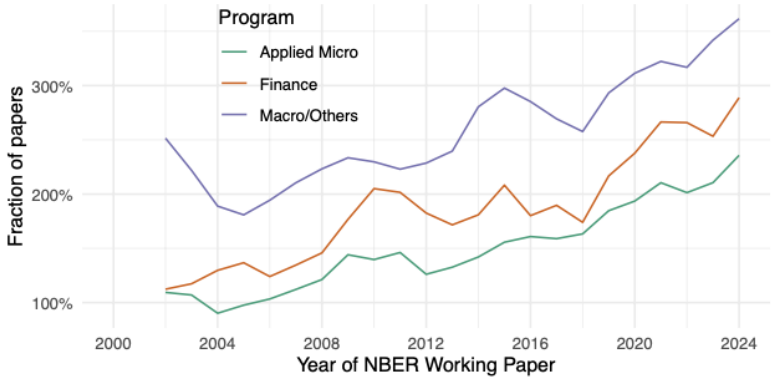
(a) Identification



(b) All experimental and quasi-experimental methods



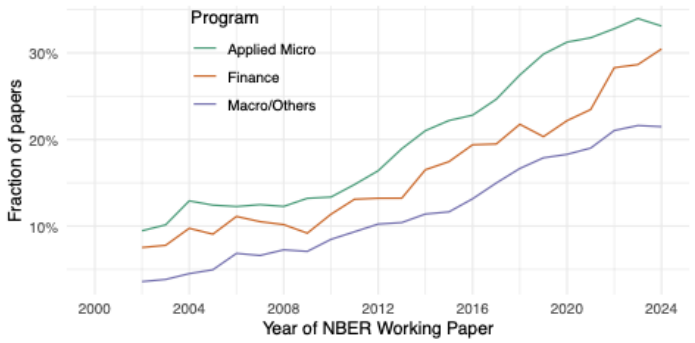
(c) Administrative data



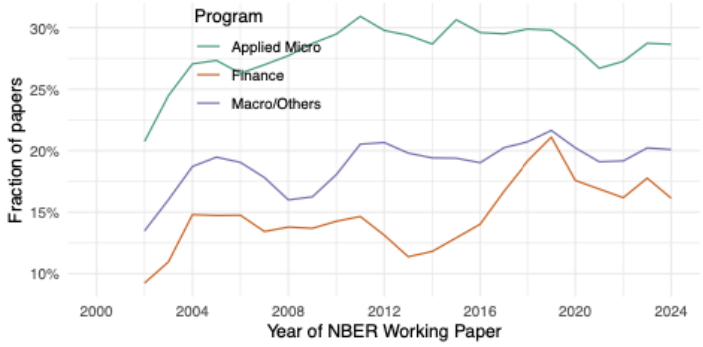
(d) Graphical revolution

How prevalent are the methods we will see?

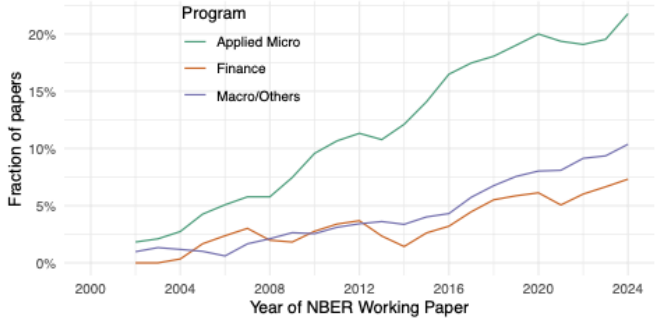
- Specific methods – difference-in-differences dominates (but partly because “event study” also included):



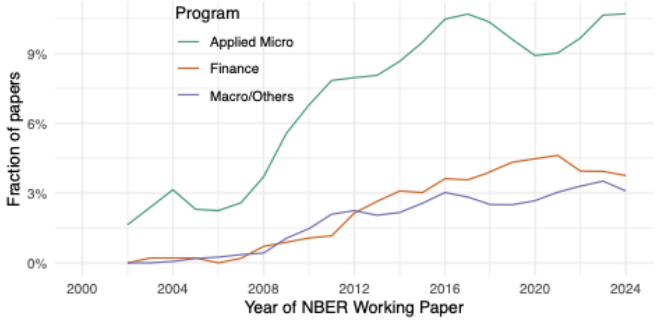
(a) Difference-in-differences



(b) Instrumental variables



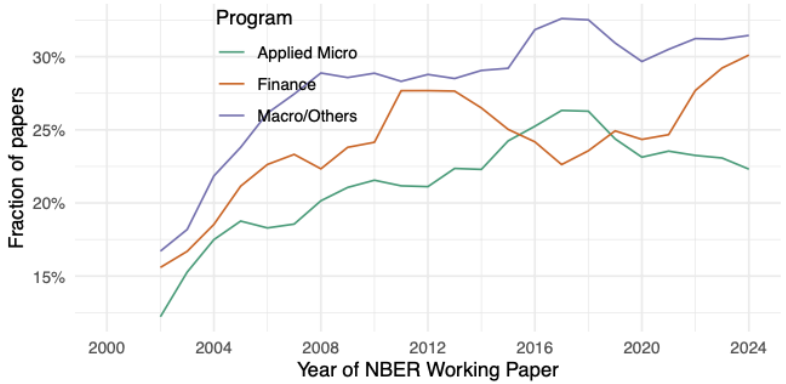
(a) Experiments



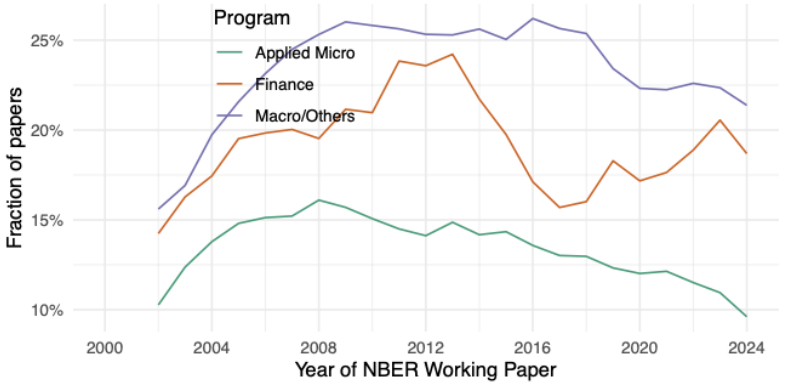
(b) Regression discontinuity

How prevalent are the methods we will see?

- Structural methods (2nd part of the course) also prevalent in finance and macro – more so than in other fields



(a) Structural Models



(b) Structural Models without mention of experimental or quasi-experimental methods

Figure 8: Panel (a) reports the share of papers that mention structural model estimation. Figure (b) reports the share of papers that mention structural model estimation and do not mention any form of experimental or non-experimental methods. See Table 2 for the breakdown of fields, and the Appendix for definitions on keywords.

Microeconometrics & finance research

s:fi

- Most of the methods we will see were first applied in non-finance contexts – e.g. labor economics
- These are also the fields where methodological innovations are still ongoing (or new insights from econometric theorists tend to be incorporated first)
- Adoption in finance has usually been lagging a bit, but the lag seems to be shrinking
- Nevertheless, there are often potential "arbitrage opportunities" in research technology
- Also note that the innovation is sometimes "destructive"
  - "standard" methods shown to have bad properties (at least under some circumstances) – we will discuss some examples


David Clingingsmith
@dclingi
There are two kinds of metrics papers.
1. We made you new toys.
2. We are taking your toys away.
22:25 · 19.07.20 · Twitter for iPhone
100 Retweets 14 Quote Tweets 1'002 Likes

Overview

- Motivation
- OLS and endogeneity
- Biases from endogeneity
- Panel data
- Standard errors

Linear model / OLS

- Suppose that we have the following linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- The key assumptions needed for OLS to produce **consistent** estimates of the parameters are the following:

OLS assumptions for consistency

1. A random sample of observations on y and $(x_1,\ldots, x_k)$

2. Linearity (in parameters)

3. No linear relationships among the explanatory variables (i.e. no perfect collinearity, or full rank)

4. An error term that is **uncorrelated with each explanatory variable** $(Cov(x_j, u) = 0 \text{ for } j = 1, \ldots, k)$

- Consistency is a **large sample property**: if N is large enough, the estimate is likely to be close to the true value.

OLS assumptions for unbiasedness

- For **unbiased** estimates, assumption 4 needs to be replaced with:

5. An error term with zero mean conditional on the explanatory variables ($E(u \mid X) = 0$).
   - The average of u (i.e., the unexplained portion of y) does not depend on the values of x.
   - This is the **strict exogeneity assumption**, sometimes also referred to as «conditional mean independence»

- Unbiasedness is a **finite sample property**

Unobservable error term

- Assumptions 4 and 5 are the **primary focus** of most research designs.
- **Problem:** we **cannot test** these assumptions because we cannot observe u.
- In other words, the error term is unobservable, and we cannot empirically test whether a variable is correlated with the regression error term.
- However, we need assumptions 4 (or 5) to hold to make **causal** inferences.

Overview

- Motivation
- OLS and endogeneity
- Biases from endogeneity
- Panel data
- Standard errors

Biases

- When assumption 4 (and 5) is **violated**, we typically observe one (or multiple) of three different **biases**:

1. **Omitted variable bias**
2. Simultaneity bias
3. Measurement error bias

Omitted variable bias

- **Omitted variables** refer to those variables that should be included in the vector of explanatory variables, but for various reasons are not.
- Most corporate (and individual) decisions are determined at least in part by factors that are **unobservable** to the econometrician.
- The inability to observe these determinants means that instead of appearing among the explanatory variables, these **omitted variables appear in the error term**.
- If these omitted variables are **correlated** with the included explanatory variables, we have an **endogeneity problem**, and **our estimates will be inconsistent.**

Omitted variable bias

- The **true** model is:

$$y = \beta_0 + \beta_1 x + \gamma w + u$$

where w is an **unobservable** variable, and γ its coefficient.

- The researcher estimates: $y = \beta_0 + \beta_1 x + v$

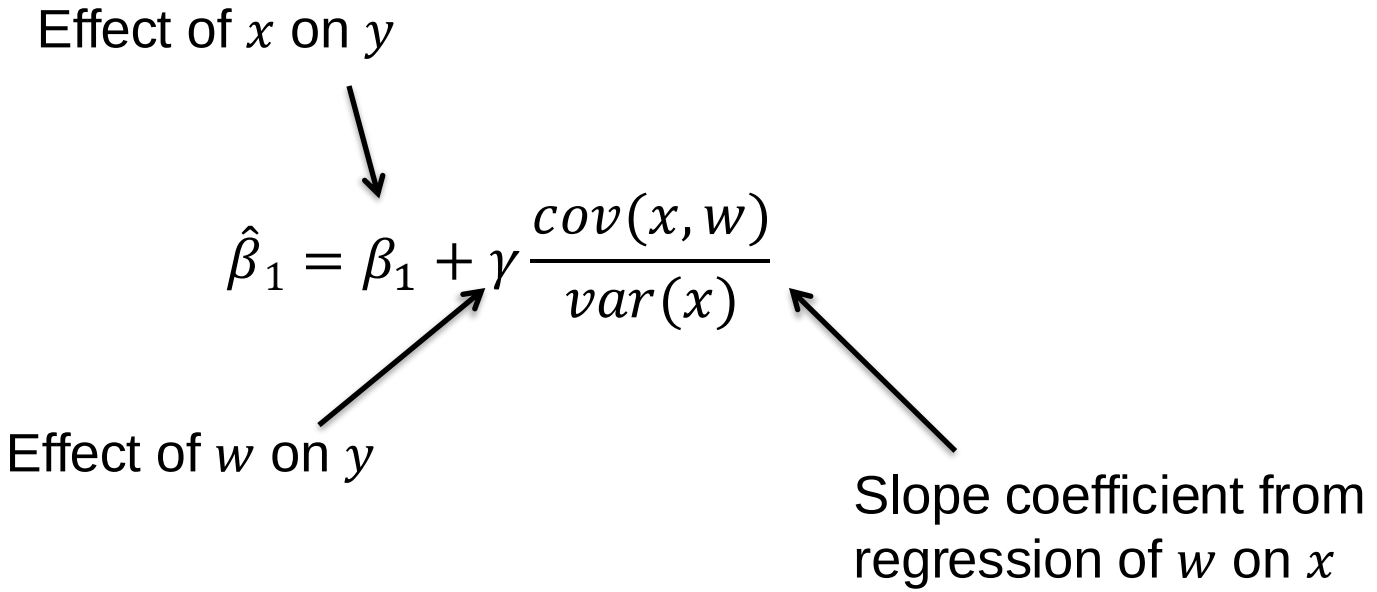
  where $v = \gamma w + u$ is the composite error term

- If the omitted variable w is **correlated** with the explanatory variable x, then the composite error term v is correlated with the explanatory variable.

- OLS will produce inconsistent (and biased) estimates **for <u>all</u> explanatory variables** in the regression model.

Omitted variable bias

- To determine the sign and magnitude of the bias (in case of a single omitted variable), we can compute



- Estimated coefficient only unbiased if $cov(x,w) = 0$ or w has no direct effect on y
- Things get quite messy with more than one omitted var.

Example – omitted variable bias

- Suppose we estimate:

$$\text{CEO Compensation} = \beta_0 + \beta_1 size + v$$

- But the true model is:

$$\text{CEO Compensation} = \beta_0 + \beta_1 size + \gamma ability + u$$

- Partial correlation between ability and compensation likely positive ($\gamma > 0$).
- Partial correlation between ability and firm size likely positive ($cov(x, w) > 0$).
- Bias is likely to be positive.

Omitted variable bias

- How do we **deal** with the omitted variable bias?
- If the **omitted variable is observable**, we just add it as a control variable.
- If the **omitted variable is unobservable**, it is much harder to deal with it.
- One possibility is to find a so called **proxy variable**.

Proxy variables

- Proxy variables require rather stringent (implicit) assumptions in order to “work”.
- Consider the following model (where we are interested in β_1, β_2):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 q + u$$

- q is unobserved, but we have a proxy z
- Further suppose: $q = \delta_0 + \delta_1 z + v$

  — v is an error associated with the proxy’s imperfect representation of the unobservable variable q.

Proxy variables – assumptions

- **Assumption 1:** The proxy variable is redundant in the structural equation. The proxy variable z is irrelevant if we could control for q.

$$E(y|x,q,z) = E(y|x,q)$$

- **Assumption 2:** z is a good proxy for q such that after controlling for z, q does not depend on x_1 and x_2.

$$E(v|x_1,x_2,z) = 0$$

- i.e. $E(q|x_1,x_2,z) = E(q|z)$

Proxy variables

- Recall the true model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 q + u$
- Plug-in for q, using $q = \delta_0 + \delta_1 z + v$

$$y = \underbrace{(\beta_0 + \beta_3 \delta_0)}_{\alpha_0} + \beta_1 x_1 + \beta_2 x_2 + \underbrace{(\beta_3 \delta_1)}_{\alpha_1} z + \underbrace{(u + \beta_3 v)}_{e}$$

- Prior assumptions ensure that $E(e|x_1, x_2, z) = 0$, such that the estimates $(\alpha_0, \alpha_1, \beta_1, \beta_2)$ are consistent.
- Note that β_0 and β_3 are **not** identified.

Proxy variables – example

- In the previous example, we do not observe managerial ability.
- We can try to find a proxy variable that is correlated with the unobserved variable.
- For example, a proxy for CEO's ability could be the CEO's IQ.
- Consider the previous compensation estimation:

$$\text{CEO Compensation} = \beta_0 + \beta_1 size + \gamma ability + u$$

Proxy variables – example

- If we use **IQ as a proxy for unobserved ability**, we need to assume that the proxy variable is redundant in the structural equation. That's OK.

- In addition, we need to assume that

$$E(ability|size, IQ) = E(ability|IQ)$$

- The average ability does not change with firm size after accounting for IQ.

- Reasonable assumption?

Bounding the bias from unobservables

- In many cases, available proxies only incompletely capture the omitted variable(s) one may be worried about.
- An intuitive approach is to explore the sensitivity of the coefficient of interest to the inclusion of observed controls (which are plausibly correlated with the omitted variable). If a coefficient is stable after inclusion of the observed controls, this is taken as a sign that omitted variable bias is limited.
- However, this requires the observed controls to actually have explanatory power for *y* ➡ as more controls are added, the R^2 needs to increase, otherwise the intuitive argument is not meaningful.

Bounding the bias from unobservables

- This point is formalized in Oster (2019), and her method is increasingly used in finance papers as well.
- E.g. Gargano et al. (JF 2023), https://ssrn.com/abstract=3519635

$$\beta^* - \hat{\beta} \approx \delta \left(\hat{\beta} - \beta^\circ\right) \frac{R_{max} - \hat{R}}{\hat{R} - R^\circ}$$

β^* is an unbiased estimator of the population value of β, β° and R° are the coefficient estimate and R-square from the short regression, and $\hat{\beta}$ and $\hat{R}$ are the estimate and R-square from the long regression. R_{max} is the maximum feasible R-square for the regression. This exact relationship holds under restrictive assumptions, but Oster (2019) shows that it can be generalized. We use her framework and code for our calculations.

- "Short" regression: limited set of controls
- "Long" regression: full set of controls (all possible proxies)
- If the estimated β doesn't change much, while R-sq increases a lot, can claim that OV bias limited. (R_{max} usually set = 1)
  - bound δ, the ratio of the sensitivity of the outcome to unobservable characteristics over the sensitivity to observable characteristics

Biases

- When assumption 4 (and 5) is **violated**, we typically observe one (or multiple) of three different **biases** :

1. Omitted variable bias
2. **Simultaneity bias**
3. Measurement error bias

Simultaneity bias

- **Simultaneity bias** (aka reverse causality bias) occurs when y and one or more x's are determined in equilibrium.
- We can then plausibly argue that x causes y, or that y causes x.

- **Example:** regress a valuation multiple (market-to-book ratio) on an index of anti-takeover provisions. The coefficient on the index typically has a negative sign.
  - It does not necessarily mean that the presence of anti-takeover provisions leads to a loss in firm value.
  - It is also possible that managers of low-value firms adopt anti-takeover provisions to entrench themselves.

Simultaneity bias

- Suppose that both y and x have zero mean, and that y and x are determined jointly in equilibrium:

$$y = \beta x + u$$
$$x = \alpha y + v$$

and u and v are uncorrelated.

- Think of y as the market-to-book ratio, and x as a measure of anti-takeover provisions.

- Other examples: (i) firm size and efficiency/profitability; (ii) loan interest rate and default risk.

Simultaneity bias

- The population estimate of the slope coefficient of the first equation is

$$\hat{\beta} = \frac{cov(x,y)}{var(x)} = \frac{cov(x,\beta x+u)}{var(x)} = \beta + \frac{cov(x,u)}{var(x)}$$

$$= \beta + \frac{\alpha(1-\alpha\beta)var(u)}{\alpha^2 var(u)+var(v)}$$

- Because x is correlated with u, we have a bias.

- The simultaneity bias is difficult to sign (more so than omitted variable bias).

Simultaneity bias

- If the x is affected by y (reverse causality), we will not be able to make causal inferences using OLS.
- Sometimes using **lagged x's** helps to address this concern.
- But we will typically need other tools, such as **instrumental variables** or **quasi-natural experiments** to deal with this problem.

Biases

- When assumption 4 (and 5) is **violated**, we typically observe one (or multiple) of three different **biases**:

1. Omitted variable bias

2. Simultaneity bias

3. **Measurement error bias**

Measurement error bias

- Estimation will have measurement error whenever a variable is **measured imprecisely**.
  - Firm's book value of debt is a noisy measure of firm's market value of debt
  - Average tax rate is a noisy measure of the marginal tax rate
  - Estimated loan-to-value ratio on an outstanding mortgage
- Such measurement error can cause bias, and the bias can be quite complicated.

Measurement error in y

- Measurement error in the **dependent variable** is typically not a problem; it causes the standard errors to be larger.

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

  - But, we measure y^* with error $e = y - y^*$
  - We only observe y, and estimate
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + (u + e)$
  - As long as $E(e|x) = 0$, the OLS estimates are consistent and unbiased.

Measurement error bias

- Measurement error in the **independent variable** is more subtle.

- Assume the model is:

$$y = \beta_0 + \beta_1 x^* + u$$

- x^* is observed with an error, $e = x - x^*$

  - We assume that $E(y|x^*, x) = E(y|x^*) \rightarrow x$ does not affect y after controlling for x^*.

Measurement error bias

- The bias depends on the **assumptions about the measurement error** e.
- Case 1: Measurement error is uncorrelated with the **observed** measure, x.
- Case 2: Measurement error is uncorrelated with the **unobserved** measure x^*

Measurement error bias

- **Case 1:** $Cov(x, e) = 0$
- Substituting $x^* = x - e$ into the true model yields

$$y = \beta_0 + \beta_1 x + u - \beta_1 e$$

- There is no bias.
- The standard errors are larger.

Measurement error bias

- **Case 2:** $Cov(x^*, e) = 0$
- We still have the model: $y = \beta_0 + \beta_1 x + u - \beta_1 e$, but now x is correlated with e:

$$cov(x, e) = E(xe) = E(x^* e) + E(e^2) = \sigma_e^2$$

- Because an independent variable is correlated with the error, the **estimates will be biased**.

- Assuming Case 2 is referred to as the **classical error-in-variables (CEV)** assumption.
- The OLS estimate will be biased towards zero **(attenuation bias)**.

Measurement error bias

- In the CEV case, the OLS regression with more than one independent variables generally gives **inconsistent estimates of all the coefficients**.
- And **bias of other coefficients not necessarily toward zero**.

- Example: Fazzari, Hubbard, and Petersen (1988).
  - They regress investment on Tobin's *q* and cash
  - They find a positive coefficient on cash and argue that there must be financial constraints.
  - But *q* as a measure of investment opportunities is noisy, and all coefficients are biased.
    - See the work by Erickson and Whited (2000)

Summary

- Whenever an independent variable is correlated with the error term, OLS estimates are no longer consistent.
  - Omitted variables bias
  - Simultaneity bias
  - Measurement error bias
- In every empirical setting, it is important to try understanding what type of biases could occur.
  - What is the endogenous variable? Why are they endogenous? What are the implications for inference?

Overview

- Motivation
- OLS and endogeneity
- Biases from endogeneity
- Panel data
- Standard errors

Motivation

- A big concern in corporate finance settings are **omitted variables**.
- If the omitted variable is captured in the error term of a regression, and the omitted variable is correlated with other explanatory variables in that regression, OLS will produce inconsistent estimates.
- **Example:**

$$leverage_{i,j,t} = \beta_0 + \beta_1 profit_{i,j,t-1} + u_{i,j,t}$$

- *leverage* is debt/assets for firm i, in industry j, in year t. *profit* is net income/assets.
- What are examples of omitted variables?

Motivation

- Sometimes we can find valid **proxy variables** for the omitted variables. But often that is difficult.
- Panel data can be useful in helping to address the omitted variable bias.
- Panel data helps to address omitted variable bias if the omitted variable is **time-invariant**.
- Specifically, it **helps with any unobserved variable that does not vary within groups of observations.**

Panel data

- A panel data set is a data set for which we have repeated observations over T time periods (e.g. years) on a cross section of N individuals, families, firms, cities etc.

- When the panel is balanced, we have the same time periods for each cross-section observation.

  — You observe 500 CEOs over a ten year period ($N = 500$ and $T = 10$)

  — You observe 2'000 firms in Compustat over a twenty year period ($N = 2'000$ and $T = 20$)

Panel data

- Take the following model:

$$y_{i,t} = \alpha + \beta x_{i,t} + \delta c_i + u_{i,t}$$

- c_i is the **unobserved, time-invariant** variable. We make the following assumptions:

$E(u_{i,t}) = 0$

$Cov(x_{i,t}, c_i) \neq 0$

$Cov(u_{i,t}, c_i) = 0$

$Cov(x_{i,t}, u_{i,s}) = 0$ for all s,t ⟵ **Strict exogeneity assumption**

Panel data – omitted variable bias

- If we estimate the model:

$$y_{i,t} = \alpha + \beta x_{i,t} + \underbrace{\varepsilon_{i,t}}_{\delta c_i + u_{i,t}}$$

- x is correlated with the disturbance ε (through its correlation with the unobserved variable, c, which is now part of the disturbance

- We have an omitted variable bias.

Transforming the data

- If we take the population mean of the dependent variable for each unit of observation, i, we get

$$\bar{y}_i = \alpha + \beta \bar{x}_i + \delta c_i + \bar{u}_i$$

- Where

$$\bar{y}_i = \frac{1}{T}\sum_t y_{i,t},\ \bar{x}_i = \frac{1}{T}\sum_t x_{i,t},\ \bar{u}_i = \frac{1}{T}\sum_t u_{i,t}$$

- (assuming there are T observations per unit i).

Transforming the data

- Now we can subtract $\bar{y}_i$ from $y_{i,t}$ and we get

$$y_{i,t} - \bar{y}_i = \beta\left(x_{i,t} - \bar{x}_i\right) + \left(u_{i,t} - \bar{u}_i\right)$$

- The unobserved variable $\boldsymbol{c_i}$ **is gone** (as is the constant) because it is time-invariant!
- With the strict exogeneity assumption, $\left(x_{i,t} - \bar{x}_i\right)$ and $\left(u_{i,t} - \bar{u}_i\right)$ are uncorrelated, and we get a **consistent estimate of $\boldsymbol{\beta}$**
  - Note: strict exogeneity rules out lagged y among the x

- The transformation is called the within transformation (because it demeans all variables within their group).
- It is also called the **fixed effect (FE) estimator**.

Fixed effect estimator

- The FE estimator allows us to capture "**unobserved heterogeneity**" – any type of unobserved variable that does **not vary** within the group.

- Statistical programs easily implement this estimator for you:
  - Stata: "xtreg, fe"; areg; **reghdfe**
  - R: fixest (often much faster – https://stata2r.github.io/fixest)

- Do not do the demeaning yourself. It is often tedious, and the standard errors need special care. Let the software do it for you.

Fixed effect estimator

- An alternative way to do the FE estimation is by adding **dummy variables**.
  - For each group i, create a dummy variable and add it to the regression.
- This is called the **least squares dummy variable model**.
- We get consistent estimates and standard errors that are identical to what we would get with the within estimator.
- It can be easily implemented in Stata:
  - regress y x i.c
  - but: with large data sets, much slower than reghdfe

Fixed effect estimator

- **Pros:**
  - Allows for arbitrary correlation between each fixed effect c and each x within group i
  - Identification from within-group variation. Intuitive interpretation.
  - Can use fixed effects to control for many types of unobserved heterogeneity:
    - Common shocks to all firms across time: time (year) fixed effect
    - Unobserved firm, CEO, country, industry etc. effects
    - Unobserved industry demand or supply shocks etc. (industry × year fixed effects, location x year fixed effects, …)
    - [Note: using multiple types of FEs used to be tedious, but reghdfe (Stata) and fixest (R) have solved that problem]

Fixed effect estimator

- **Cons / limitations:**
  - If no (not enough) within-group variation in the independent variables, we cannot disentangle it from group FE.
  - May remove the interesting variation across (e.g.) firms and inflate standard errors (→ higher chance of Type II error)
  - Measurement error of independent variable can be amplified – **"understand your variation"!**
  - FE estimator not well suited for nonlinear models (e.g. Probit, Logit, Tobit). For dummy outcomes, nowadays LPM standard (i.e., just estimate with OLS).

How do FEs impact standard errors?
(based on deHaan, 2021 – recommended reading)

- Will discuss standard errors more below, but a useful way to think about them is provided by the formula under homosked.:

$$\hat{\sigma}_{\beta_x} = \sqrt{\frac{SSR/(N-K)}{SST(x) \cdot (1-R_x^2)}}$$

  - SSR, the sum of squared residuals, will typically decrease with FE (but so will $(N-K)$, where K is the number of parameters)
  - $SST(x)$ is total amount of variation in x. In principle unaffected by FEs, but may decrease if many 'singletons' (= only one observation per FE group) are dropped from the sample
  - R_x^2: R-sq. from regressing x on all other explanatory variables, incl. FE. **Mechanically increases as FEs are added** (unless orthogonal to x) **– tends to dominate & increase st. errors**

FE "diagnostics" (from deHaan, 2021)

s:fi

Appendix A: Author Checklist – Good Practices when Using Fixed Effects

Below are some of the major points to consider when using FE.

*Indicates that my SUMHDFE Stata package produces the information needed to complete these evaluations. A beta version of SUMHDFE is available at https://github.com/ed-dehaan/sumhdfe. See Appendix B for an example of how to report the * items in your own paper.

1. Keep in mind that FE restrict analyses to only variation in *X* that exists within the FE groupings.

2. Carefully consider whether you need FE, and why:
   i. If an unobservable *Z* is thought to be correlated with **both** *X* and *Y* and is constant within FE groups, then including FE are likely necessary to reduce type 1 errors
   ii. If the unobserved *Z* is not correlated with *Y*, then FE are not necessary and will increase the risk of type 2 errors, especially if *Z* is correlated with *X*
   iii. If the unobserved *Z* is **only** correlated with *Y*, then including FE can improve model fit and reduce the risk of type 2 errors.

(see Section 3.2 of deHaan's paper for details)

FE "diagnostics" (from deHaan, 2021)

3. *Report the number of singletons that exist within your FE structure.
   i. Lots of singletons indicate that your FE structure is likely too narrow for your data.
   ii. Consider dropping singletons during the sample construction.
   iii. Always drop singletons when running each regression (reghdfe will do automatically)

4. *Report the number of observations that have no within-FE variation in *X*
   i. Lots of observations with no within-FE variation raises concerns about whether the observations with variation in *X* are similar to those without variation
   ii. Evaluate the similarity of observations with and without within-FE variation in *X*. If dissimilar, try to improve similarity through matching or refining the sample
   iii. If possible, evaluate regression results with and without no-variation observations

5. *For non-binary variables, report the pooled standard deviations of *X* and other key variables, the within-FE standard deviations, and the reduction in standard deviations caused by the FE
   i. If little variation remains, re-consider whether the FE structure is appropriate
   (particular worry: variation due to "measurement error" or special events – e.g. mergers etc.)
6. *When using the standard deviation of *X* (or any other distributional statistic) to interpret the economic magnitude of a regression coefficient, use the within-FE standard deviation

Random effect estimator

- The model is very similar as FE

$$y_{i,t} = \alpha + \beta x_{i,t} + c_i + u_{i,t}$$

- But here we need to assume that $\boldsymbol{c_i}$ **and the observed** $\boldsymbol{x_{i,t}}$ **are uncorrelated.**
- In most corporate finance settings, this assumption is unrealistic.
- If this assumption holds, OLS would give consistent estimate of β. Then why bother?
- In (corporate finance) practice, the random effect estimator is not very useful.

First difference model

- Taking the **first difference** of a model is another way to remove unobserved heterogeneity.
- Rather than subtracting off the group mean of the variable from each variable, you instead subtract the **lagged observation**.
- Take two cross sections

$$y_{i,t} = \alpha + \beta x_{i,t} + c_i + u_{i,t}$$
$$y_{i,t-1} = \alpha + \beta x_{i,t-1} + c_i + u_{i,t-1}$$

- In **first differences** we get

$$y_{i,t} - y_{i,t-1} = \beta(x_{i,t} - x_{i,t-1}) + (u_{i,t} - u_{i,t-1})$$

First difference model

- With the **strict exogeneity assumption** ($Cov(x_{i,t}, u_{i,s}) = 0$ for all *s,t*), OLS will produce a consistent estimate of β.
- Note that we lose one observation per cross section
- Produces identical results as FE estimator with just two observations per group.
- In other cases, both FE and first difference models are consistent; difference generally lies in the efficiency
  - FE is more efficient if disturbances are serially uncorrelated
  - First differences is more efficient if error terms follow a random walk. Intermediate cases (more typical): difficult to say.

Summary on panel data

- Panel data is useful to control for **unobserved** variables
  - FE estimator useful to control for unobserved heterogeneity in flexible ways
  - Reduces the scope for potential omitted variable biases
  - But: use FEs with care; understand what they do & how much variation they absorb
  - RE estimator not very useful in corporate finance settings

- Workhorse model in corporate finance: **Firm and year FE model (also known as two-way FE , or TWFE, model)**
  - typical for difference-in-differences studies (focus in lecture 3)

Selected topics we did not cover

- Panel data techniques specific to asset pricing – namely the "Fama-MacBeth" approach
  - you will surely cover this in empirical asset pricing
  - Verbeek Ch. 2.12 provides an overview

- Dynamic panel models with lagged dependent variables, e.g.

$$y_{i,t} = \alpha + \beta x_{i,t} + \gamma y_{i,t-1} + u_{i,t}$$

  - strict exogeneity is violated & coefficients are inconsistent – important to keep in mind
  - can be solved via GMM ("Arellano-Bond" etc.)
  - see e.g. Verbeek Ch. 5

Overview

- Motivation
- OLS and endogeneity
- Biases from endogeneity
- Panel data
- Standard errors

Standard errors

- To make correct inferences, we must make sure that the **standard errors are correct**

- If standard errors are too small, we would reject the null hypothesis in favor of the alternative “too often”. We would then claim that an effect is **statistically significant**, even though it is not.

- What do the default (classical) standard errors reported in a program like Stata assume?

  — **Homoskedasticity**: $Var(u|x) = \sigma^2$, i.e. the conditional variance of u (or y) is constant.

Standard errors

- The assumption of homoskedasticity is typically not reasonable in corporate finance settings.

- "**Robust**" (White) standard errors allow for heteroskedasticity and do not make this assumption.
  - Easily implemented in Stata (but: see next slide!)
  - Note that we generally **prefer robust standard errors over classical standard errors**, but even robust standard errors can be too small in small samples and biased downward if heteroskedasticity is modest (though some versions correct for this)
  - **Common approach:** Estimate with both classical and robust standard error, and use the specification with the larger standard errors (Angrist-Pischke)

Which robust standard errors?

- http://datacolada.org/99 (on Young, QJE 2019):

> It turns out that there are **five** main ways to compute robust standard errors. STATA has a default way, *and it is **not** the best way.*
>
> In other words, when in STATA you run: ***reg y x, robust*** you are not actually getting the most robust results available. Instead, you are doing something we have known to be not quite good enough since the first George W Bush administration (the in-hindsight good old days).
>
> The five approaches for computing robust standard errors are unhelpfully referred to as HC0, HC1, HC2, HC3, and HC4. STATA's default is HC1. That is the procedure the QJE article used to conclude regression results are inferior to randomization tests, but HC1 is known to perform poorly with 'small' samples. Long & Ervin 2000 unambiguously write "*When N<250 . . . HC3 should be used*". A third of the simulations in the QJE paper have N=20, another third N=200.

- (HC1 are the original White s.e. with DF adjustment; HC3 are based on Davidson-MacKinnon 1993.)
- R uses HC3 as default.

Clustered standard errors

- Both classical and robust standard errors assume that the **observations of y are random draws** from some population and are **uncorrelated** with other draws.
  - Firm's investment choice at time t is uncorrelated with the firm's investment choice at time $t-1$ (**no time-series correlation**)
  - Firm's leverage decision in industry j is uncorrelated with another firm's leverage decision in the same industry (**no cross sectional correlation**)

- In corporate finance, this independence assumption is often unrealistic.
- As a result, standard errors are **significantly biased downward**. This can make much larger difference than classical vs. robust standard errors.

Clustered standard errors

- When we talk about **clustered standard errors**, we assume that errors are correlated within the group (cluster), but **not** correlated across them. For example,
  - Correlation between firms' financing choices within an industry, but not across industries.
  - Correlation of firm's investment choices over time, etc.

- The **bias** in s.e. if we **do not cluster** can be **very large**.
- Rule of thumb: with single x, clustered SE are about $\tau = \sqrt{1 + \rho_x \rho_u (M - 1)}$ times the incorrect default SE, with:
  - ρ_x (ρ_u) = within-cluster correlation in x (… in error term u)
  - M = average cluster size (= N / # of clusters) → in panel, the more observations per group (e.g. firm), the larger the bias

Clustered standard errors

- Take the following panel firm-level regression

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

- y_{ijt} is outcome for firm *i* in country *j* in year *t*
- x_{jt} only varies at country-year level (e.g. some gvmt policy)

- If firms are subject to similar country shocks over time, how should we cluster the standard errors? What will likely give larger standard errors?
  - Country-year?
  - Country?

Clustering vs. firm FE

- Consider the following regression

$$y_{i,t} = \alpha + \beta x_{i,t} + \underbrace{c_i + u_{i,t}}_{e_{i,t}}$$

- $y_{i,t}$ is outcome for firm i in year t
- c_i is time-invariant unobserved heterogeneity
- $e_{i,t}$ is estimation error if we do not control for c_i
- $u_{i,t}$ is estimation error if we do control for c_i

- Which problems do firm FE and clustering help addressing?

Clustering vs. firm FE

- **Clustering corrects standard errors** for cross sectional or serial dependence; it does **not help** dealing with potential omitted variable bias.
- The **firm FE** removes **time-invariant** heterogeneity from the error term; it does **not help** dealing with possible serial correlation.
- **Therefore:** clustering is not a substitute for FE. You **should use both firm FE and clustered standard errors** in your regressions.

- Advice: cluster at **most aggregate level** of variation in the covariates.
  - e.g. if law change affects firms depending on industry, cluster at industry level

How many clusters do you need?

s:fi

- Theory relies on #clusters going to infinity – but typically we have many fewer clusters

- How many do you need? One rule of thumb is that 50 is fine. Often fewer is ok. But also depends on balance of cluster sizes (unbalanced = lower # of "effective" clusters).
  - clustered SE most likely much too small (i.e. over-reject null) even with large number of clusters **if**: (i) one or a few clusters are unusually large, or (ii) only a few clusters are treated.

- If you're worried (or somebody else is), can use "wild bootstrap" methods – see e.g. Roodman et al. (2019)

Double-clustering (aka two-way clustering)

- Should we **cluster** standard errors **along two dimensions**, e.g. firm and year (which is not the same as firm-year)?
  - e.g. in reghdfe, vce(cluster *variable1 variable2*)
  - not the same as vce(cluster *variable1#variable2*)

- Additional cluster at the year level allows errors within a year to be correlated in arbitrary ways.

- Probably more important in **asset pricing** than in **corporate finance** settings**.**
  - see Petersen paper, cited on next slide

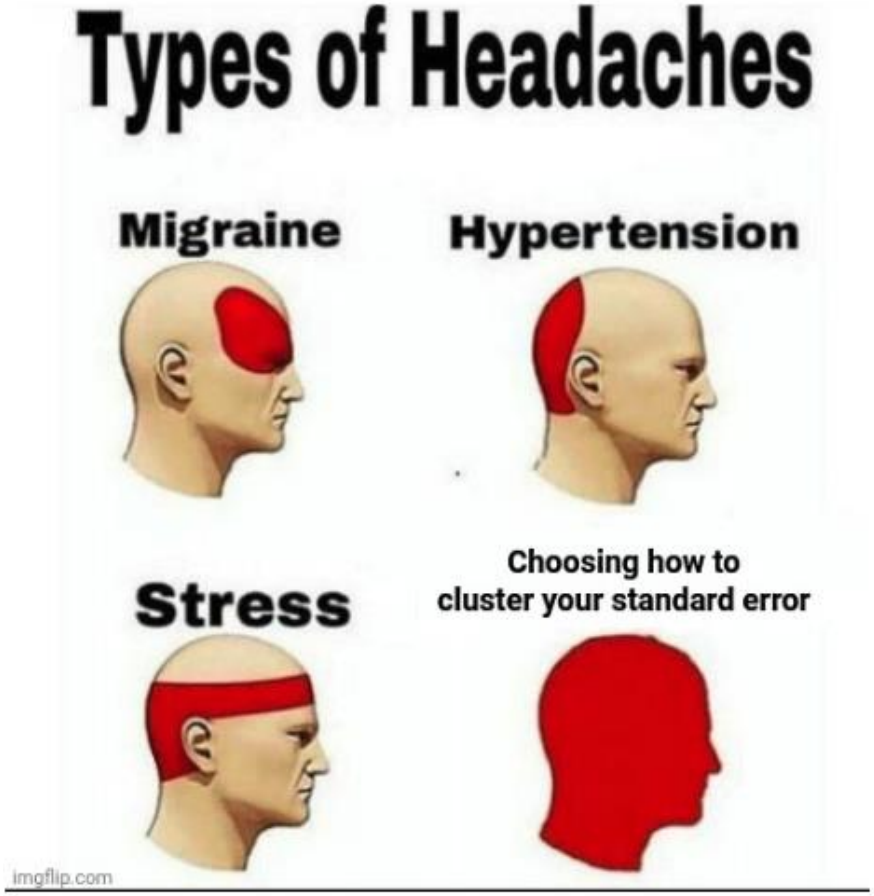
Clustering – final words

- The practice of when/how to cluster is not set in stone, and theory is still evolving
  - esp. for two-way clustering and clustering in models other than OLS – e.g. IV, nonlinear models, etc.

- The classic reference in finance is Petersen (RFS 2008)

- Other good references: Cameron and Miller (JHR 2015); **MacKinnon et al. (J of Econometrics 2023)**, who provide recommendations on what researchers should do:
  - report summary stats on #clusters, median size, max size; leverage and influence of individual clusters
  - use *p*-values/c.i. from wild cluster bootstrap as "verification"

Very final word: statistical inference is hard!



Journal of Econometrics Session: What is a Standard Error?

Panel Session

Saturday, Jan. 7, 2023 · 2:30 PM - 4:30 PM (CST)

Hilton Riverside, Grand Salon A Sec 3

Hosted By: ECONOMETRIC SOCIETY

Moderators: Serena Ng, Columbia University

Elie Tamer, Harvard University

Panelist(s)

Andrew Gelman, Columbia University

Patrick Kline, University of California-Berkeley

James Powell, University of California-Berkeley

Jeffrey M. Wooldridge, Michigan State University

Bin Yu, University of California-Berkeley

Source: Khoa Vu (X), via Peter Hull