Week 12: Repeated Observations and Panel Data

Brandon Stewart[1]

Princeton

December 10 and 12, 2018

[1] These slides are heavily influenced by Matt Blackwell, Adam Glynn, Jens Hainmueller and Erin Hartman.

Where We've Been and Where We're Going...

- Last Week
  - causal inference with unmeasured confounding
- This Week
  - Monday:
    - panel data
    - diff-in-diff
    - fixed effects
  - Wednesday:
    - spillover of material
    - Q&A
    - wrap-up
- The Following Week
  - break!
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causality

Questions?

1 Set Up

2 Differencing Models

3 Difference-in-Differences

4 Fixed Effects

5 Non-parametric Identification and Fixed Effects

6 (Almost) Twenty Questions
- Review
- Topics Beyond the Course
- Research Practice
- Opinions and Musings

7 Concluding Thoughts for the Course

8 Appendix: Why Does Weighting Work?

Motivation

Is Democracy Good for the Poor?

Michael Ross University of California, Los Angeles

- Relationship between democracy and infant mortality?
- Compare levels of democracy with levels of infant mortality, but...
- Democratic countries are different from non-democracies in ways that we can't measure?
  - they are richer or developed earlier
  - provide benefits more efficiently
  - possess some cultural trait correlated with better health outcomes
- If we have data on countries over time, can we make any progress in spite of these problems?

Ross Data

```
##      cty_name year democracy infmort_unicef
## 1 Afghanistan 1965         0            230
## 2 Afghanistan 1966         0             NA
## 3 Afghanistan 1967         0             NA
## 4 Afghanistan 1968         0             NA
## 5 Afghanistan 1969         0             NA
## 6 Afghanistan 1970         0            215
```

Notation for Panel Data

- Units, $i = 1, \dots, n$
- Time, $t = 1, \dots, T$
- Slightly different focus than clustered data we covered earlier
  - Panel: we have repeated measurements of the same units
  - Clustering: units are clustered within some grouping.
  - The main difference is what level of analysis we care about (individual, city, county, state, country, etc).
- Time is a typical application, but applies to other groupings:
  - counties within states
  - states within countries
  - people within professions

Nomenclature

Names are used in different ways across fields but generally:

- Panel data: large n, relatively short T
- Time series, cross-sectional (TSCS) data: smaller n, large T
- We are primarily going to focus on similarities today but there are some differences.

A Baseline Linear Model

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + a_i + u_{it}$$

- $\mathbf{x}_{it}$ is a vector of (possibly time-varying) covariates
- a_i is an unobserved time-constant unit effect ("fixed effect")
- u_{it} are the unobserved time-varying "idiosyncratic" errors
- $v_{it} = a_i + u_{it}$ is the combined unobserved error:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + v_{it}$$

- Covers the case of separable, linear unmeasured confounding.

We will start by considering performance of estimators assuming this model is true.

Naive Strategy: Pooled OLS

- Pooled OLS: pool all observations into one regression
- Treats all unit-periods (each *it*) as an iid unit.
- Has two problems:
  1. Heteroskedasticity (see clustering from diagnostics week)
  2. Possible violation of zero conditional mean errors
- Both problems arise out of ignoring the unmeasured heterogeneity inherent in a_i

Pooled OLS with Ross data

```
pooled.mod <- lm(log(kidmort_unicef) ~ democracy + log(GDPcur),
                 data = ross)
summary(pooled.mod)
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.76405    0.34491   28.31   <2e-16 ***
## democracy    -0.95525    0.06978  -13.69   <2e-16 ***
## log(GDPcur) -0.22828    0.01548  -14.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7948 on 646 degrees of freedom
##   (5773 observations deleted due to missingness)
## Multiple R-squared:  0.5044, Adjusted R-squared:  0.5029
## F-statistic: 328.7 on 2 and 646 DF,  p-value: < 2.2e-16
```

Unmeasured Heterogeneity

- Assume that zero conditional mean error holds for the idiosyncratic error:

$$\mathbb{E}[u_{it}|\mathbf{X}] = 0$$

- But time-constant effect, a_i, is correlated with the $\mathbf{X}$:

$$\mathbb{E}[a_i|\mathbf{X}] \neq 0$$

- Example: democratic institutions correlated with time-invariant unmeasured aspects of health outcomes, like quality of health system or a lack of ethnic conflict.
- Ignore the heterogeneity $\rightsquigarrow$ correlation between the combined error and the independent variables:

$$\mathbb{E}[v_{it}|\mathbf{X}] = \mathbb{E}[a_i + u_{it}|\mathbf{X}] \neq 0$$

- Pooled OLS will be biased and inconsistent because zero conditional mean error fails for the combined error.

1. Set Up
2. Differencing Models
3. Difference-in-Differences
4. Fixed Effects
5. Non-parametric Identification and Fixed Effects
6. (Almost) Twenty Questions
   - Review
   - Topics Beyond the Course
   - Research Practice
   - Opinions and Musings
7. Concluding Thoughts for the Course
8. Appendix: Why Does Weighting Work?

First Differencing

- First approach: compare changes over time as opposed to levels
- Intuitively, the levels include the unobserved heterogeneity, but changes over time should be free of time-invariant heterogeneity
- Two time periods:

$$
\begin{aligned}
y_{i1} &= \mathbf{x}'_{i1}\boldsymbol{\beta} + a_i + u_{i1} \\
y_{i2} &= \mathbf{x}'_{i2}\boldsymbol{\beta} + a_i + u_{i2}
\end{aligned}
$$

- Look at the change in y over time:

$$
\begin{aligned}
\Delta y_i &= y_{i2} - y_{i1} \\
&= (\mathbf{x}'_{i2}\boldsymbol{\beta} + a_i + u_{i2}) - (\mathbf{x}'_{i1}\boldsymbol{\beta} + a_i + u_{i1}) \\
&= (\mathbf{x}'_{i2} - \mathbf{x}'_{i1})\boldsymbol{\beta} + (a_i - a_i) + (u_{i2} - u_{i1}) \\
&= \Delta\mathbf{x}'_i\boldsymbol{\beta} + \Delta u_i
\end{aligned}
$$

First Differences Model

$$\Delta y_i = \Delta \mathbf{x}_i' \boldsymbol{\beta} + \Delta u_i$$

- Coefficient on the levels $\mathbf{x}_{it}$ is the same as the coefficient on the changes $\Delta \mathbf{x}_i$!
- fixed effect/unobserved heterogeneity, a_i drops out (relies on unobserved component being constant over time!)
- *If* $\mathbb{E}[u_{it}|\mathbf{X}] = 0$, then, $\mathbb{E}[\Delta u_i | \Delta X] = 0$ and zero conditional mean error holds.
- Due to 'no perfect collinearity': $\mathbf{x}_{it}$ has to change over time for some units. High variance if its slow moving.
- Differencing will reduce the variation in the independent variables and thus increase standard errors.

First Differences in R (via plm package)

```
library(plm)

fd.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur), data = ross,
                      index = c("id", "year"), model = "fd")
summary(fd.mod)

## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = log(kidmort_unicef) ~ democracy + log(GDPcur),
##     data = ross, model = "fd", index = c("id", "year"))
##
## Unbalanced Panel: n=166, T=1-7, N=649
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -0.9060 -0.0956  0.0468  0.1410  0.3950
##
## Coefficients :
##              Estimate Std. Error  t-value Pr(>|t|)
## (intercept) -0.149469   0.011275 -13.2567  < 2e-16 ***
## democracy   -0.044887   0.024206  -1.8544  0.06429 .
## log(GDPcur) -0.171796   0.013756 -12.4886  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    23.545
## Residual Sum of Squares: 17.762
## R-Squared      :  0.24561
##       Adj. R-Squared :  0.24408
## F-statistic: 78.1367 on 2 and 480 DF, p-value: < 2.22e-16
```

1 Set Up

2 Differencing Models

3 Difference-in-Differences

4 Fixed Effects

5 Non-parametric Identification and Fixed Effects

6 (Almost) Twenty Questions
- Review
- Topics Beyond the Course
- Research Practice
- Opinions and Musings

7 Concluding Thoughts for the Course

8 Appendix: Why Does Weighting Work?

Motivation: Studying the Minimum Wage

Minimum Wages and Employment:
A Case Study of the Fast-Food Industry
in New Jersey and Pennsylvania

By DAVID CARD AND ALAN B. KRUEGER*

On April 1, 1992, New Jersey's minimum wage rose from $4.25 to $5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise. Comparisons of employment growth at stores in New Jersey and Pennsylvania (where the minimum wage was constant) provide simple estimates of the effect of the higher minimum wage. We also compare employment changes at stores in New Jersey that were initially paying high wages (above $5) to the changes at lower-wage stores. We find no indication that the rise in the minimum wage reduced employment. (*JEL* J30, J23)

- Economics conventional wisdom: higher minimum wages decrease low-wage jobs.
- Card and Krueger (1994) study a 1992 New Jersey minimum wage increase ($4.25 to $5.05).
- Idea: compare employment rates in 410 fast-food restauarants in New Jersey and eastern Pennsylvania (where there wasn't a wage increase) both before and after the change.
- Based on survey data:
  - Wave 1: March 1992, one month before the minimum wage increased
  - Wave 2: December 1992, eight months after increase

Difference-in-Differences

- Often called "diff-in-diff" (DiD), it is a special kind of FD model
- Let x_{it} be an indicator of a unit being "treated" at time t.
- Focus on two-periods where:
  - $x_{i1} = 0$ for all i
  - $x_{i2} = 1$ for the "treated group"
- Assume the model:

$$y_{it} = \beta_0 + \delta_0 d_t + \beta_1 x_{it} + a_i + u_{it}$$

- d_t is a dummy variable for the second time period
  - $d_2 = 1$ and $d_1 = 0$
- β_1 is the quantity of interest: it's the effect of being treated

Diff-in-Diff Mechanics

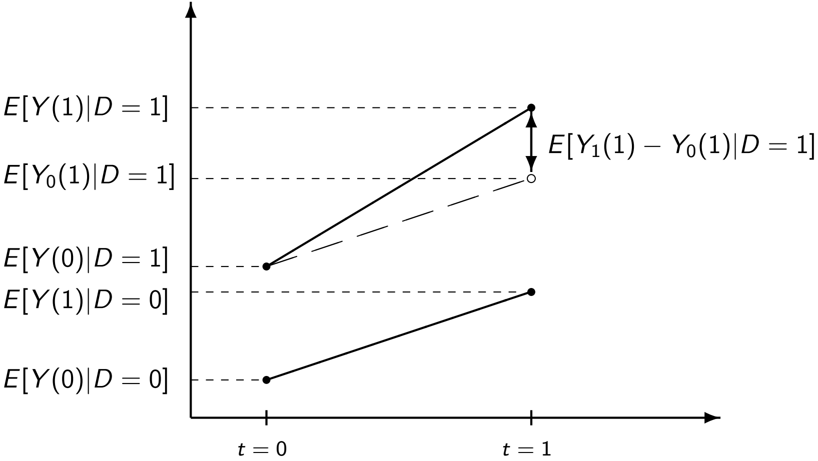
- Let's take differences:

$$\begin{aligned}(y_{i2} - y_{i1}) &= \delta_0(1 - 0) + \beta_1(x_{i2} - x_{i1}) + (a_i - a_i) + (u_{i2} - u_{i1}) \\ (y_{i2} - y_{i1}) &= \delta_0 + \beta_1(x_{i2} - x_{i1}) + (u_{i2} - u_{i1})\end{aligned}$$

- This represents
  - δ_0: the difference in the average outcome from period 1 to period 2 in the untreated group
  - $(x_{i2} - x_{i1}) = 1$ for the treated group and 0 for the control group
  - β_1 represents the additional change in y over time (on top of δ_0) associated with being in the treatment group.

Graphical Representation: Difference-in-Differences

Define $D = 1$ when $x_{i2} - x_{i1} = 1$ and 0 otherwise



Identification with Difference-in-Differences

Identification Assumption (parallel trends)

$E[Y_0(1) - Y_0(0)|D = 1] = E[Y_0(1) - Y_0(0)|D = 0]$

Identification Result

Given parallel trends the ATT is identified as:

$$\begin{aligned} E[Y_1(1) - Y_0(1)|D = 1] &= \Big\{E[Y(1)|D = 1] - E[Y(1)|D = 0]\Big\} \\ &- \Big\{E[Y(0)|D = 1] - E[Y(0)|D = 0]\Big\} \end{aligned}$$

Identification with Difference-in-Differences

Identification Assumption (parallel trends)

$$E[Y_0(1) - Y_0(0)|D = 1] = E[Y_0(1) - Y_0(0)|D = 0]$$

Proof.

Note that the identification assumption implies
$E[Y_0(1)|D = 0] = E[Y_0(1)|D = 1] - E[Y_0(0)|D = 1] + E[Y_0(0)|D = 0]$
plugging in we get

$$\begin{aligned}
& \{E[Y(1)|D = 1] - E[Y(1)|D = 0]\} - \{E[Y(0)|D = 1] - E[Y(0)|D = 0]\} \\
= & \{E[Y_1(1)|D = 1] - E[Y_0(1)|D = 0]\} - \{E[Y_0(0)|D = 1] - E[Y_0(0)|D = 0]\} \\
= & \{E[Y_1(1)|D = 1] - (E[Y_0(1)|D = 1] - E[Y_0(0)|D = 1] + E[Y_0(0)|D = 0])\} \\
- & \{E[Y_0(0)|D = 1] - E[Y_0(0)|D = 0]\} \\
= & E[Y_1(1) - Y_0(1)|D = 1] + \{E[Y_0(0)|D = 1] - E[Y_0(0)|D = 0]\} \\
- & \{E[Y_0(0)|D = 1] - E[Y_0(0)|D = 0]\} \\
= & E[Y_1(1) - Y_0(1)|D = 1]
\end{aligned}$$

□

Difference-in-Differences Interpretation

- Key idea: comparing the changes over time in the control group to the changes over time in the treated group.
- The differences between these differences is our estimate of the causal effect:

$$\beta_1 = \overline{\Delta y}_{\text{treated}} - \overline{\Delta y}_{\text{control}}$$

- Why more credible than simply looking at the treatment/control differences in period 2?
  - Unmeasured reasons why the treated group has higher or lower outcomes than the control group
  - $\rightsquigarrow$ bias due to violation of zero conditional mean error
  - DiD estimates the bias using period 1 and corrects for it.
- DiD works for additive and time-invariant confounding (i.e. satisfies parallel trends)

Example: Lyall (2009)

Journal of Conflict Resolution
Volume 53 Number 3
June 2009 331-362

10.1177/0022002708330881
http://jcr.sagepub.com
hosted at
http://online.sagepub.com

Does Indiscriminate Violence Incite Insurgent Attacks?

Evidence from Chechnya

Jason Lyall
Department of Politics and the Woodrow Wilson School
Princeton University, New Jersey

Example: Lyall (2009)

- Does Russian shelling of villages cause insurgent attacks?

$$\text{attacks}_{it} = \beta_0 + \beta_1 \text{shelling}_{it} + a_i + u_{it}$$

- We might think that artillery shelling by Russians is targeted to places where the insurgency is the strongest
- That is, part of the village fixed effect, a_i might be correlated with whether or not shelling occurs, x_{it}
- This would cause our pooled estimates to be biased
- Instead Lyall takes a diff-in-diff approach: compare attacks over time for shelled and non-shelled villages:

$$\Delta\text{attacks}_i = \beta_0 + \beta_1 \Delta\text{shelling}_i + \Delta u_i$$

- Counterintuitive findings: shelled villages experience a 24% reduction in insurgent attacks relative to controls.

Example: Card and Krueger (2000)

- Do increases to the minimum wage depress employment at fast-food restaurants?

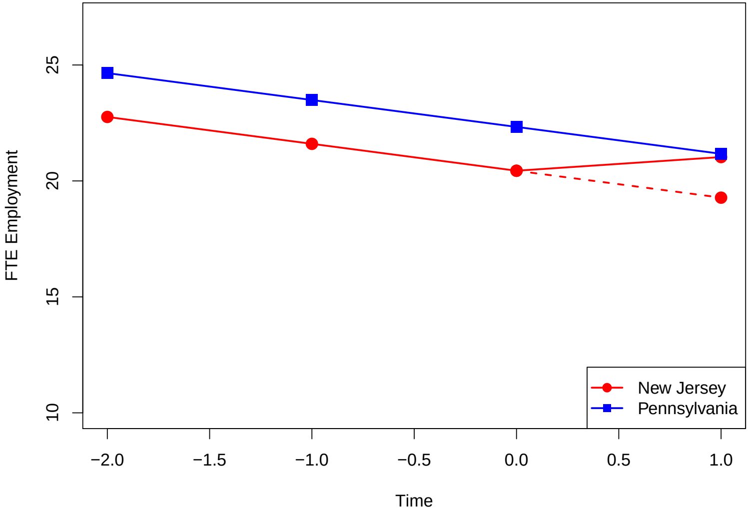
$$\text{employment}_{it} = \beta_0 + \beta_1 \text{minimum wage}_{it} + a_i + u_{it}$$

- Each i here is a different fast food restaurant in either New Jersey or Pennsylvania
- Between $t = 1$ and $t = 2$ NJ raised its minimum wage
- Employment in fast food might be driven by other state-level policies correlated with minimum wage
- Diff-in-diff approach: regress changes in employment on store being in NJ

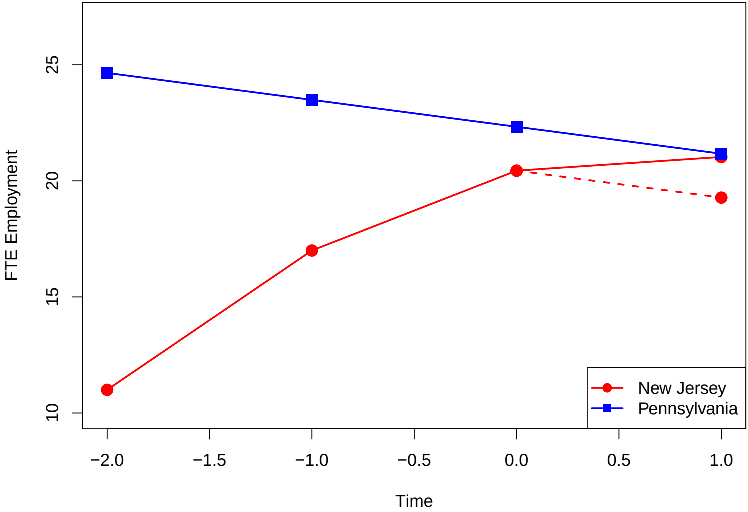
$$\Delta \text{employment}_i = \beta_0 + \beta_1 NJ_i + \Delta u_i$$

- NJ_i indicates which stores received the treatment of a higher minimum wage at time period $t = 2$

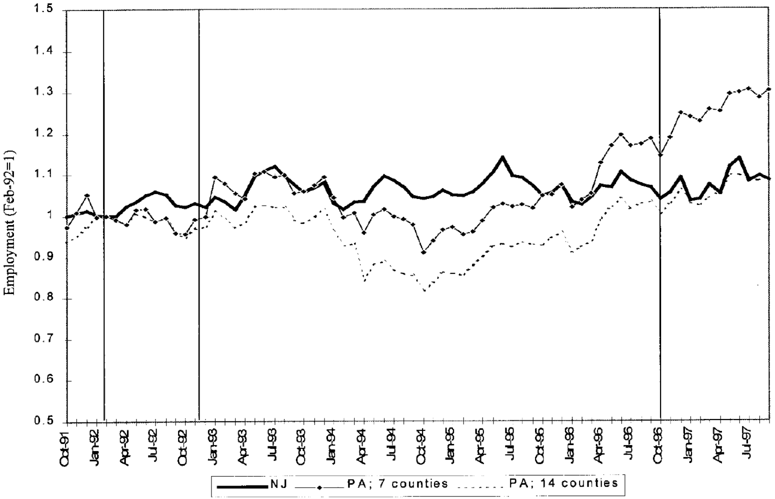
Parallel Trends?



Parallel Trends?



Longer Trends in Employment (Card and Krueger 2000)



First two vertical lines indicate the dates of the Card-Krueger survey. October 1996 line is the federal minimum wage hike which was binding in PA but not NJ

Threats to Identification

1) Failure of Exogeneity
Treatment needs to be independent of the idiosyncratic shocks:

$$\mathbb{E}[(u_{i2} - u_{i1})|x_{i2}] = 0$$

2) Non-parallel dynamics
variation in the outcome over time is the same for the treated and control groups (i.e. no omitted time-varying confounders). e.g. Ashenfelter's dip: people who enroll in job training programs see their earnings decline prior to that training (presumably why they are entering)

3) Changes in Composition of Treatment/Control Groups
we don't want composition of sample to change between periods. what if workers move from eastern PA to NJ in search of higher paying jobs?

4) Long-term vs. Short-term Effects
parallel trends are less credible over a long time horizon, but many policies need time to take effect.

Threats to Identification

5) Functional Form Dependence
   difference in levels and difference in logs can be quite different (example via Justin Grimmer)
   - imagine a training program for the young
   - employment for the young increases from 20% to 30%
   - employment for the old increases from 5% to 10%
   - positive DiD effect: $(30 - 20) - (10 - 5) = 5\%$
   - but if you consider log changes:
     $[log(30) - log(20)] - [log(10) - log(5)] = log(1.5) - log(2) < 0$
   - how do we tell which (if either) yields parallel trends?

6) Endogenous Control Variables
   can add (time-varying) covariates to help with some of above concerns $\rightsquigarrow$ "regression diff-in-diff"

$$y_{i2} - y_{i1} = \delta_0 + \mathbf{z}_i'\tau + \beta(x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

but need to be careful that they aren't affected by the treatment.

Concluding Thoughts on Panel Differencing Models

- Useful toolkit for leveraging panel data, often quite straightforward to explain to people
- Be cautious of assumptions required
  - parallel trends assumptions are most likely to hold over a shorter time-window. Impossible to test.
  - can conduct placebo tests which can build confidence, but hard to provide definitive evidence.
  - some approaches use placebos to correct bias (DDD), but this is just a difference assumption.
- Two questions to ask:
  1. 'what is the counterfactual?' or
  2. 'what variation is used to identify this effect?'
- Personal Gripe: 'Two-way Fixed Effects' models often called a DiD or Generalized-DiD design but the parallel trend assumptions are different in important ways.

1 Set Up

2 Differencing Models

3 Difference-in-Differences

4 Fixed Effects

5 Non-parametric Identification and Fixed Effects

6 (Almost) Twenty Questions
- Review
- Topics Beyond the Course
- Research Practice
- Opinions and Musings

7 Concluding Thoughts for the Course

8 Appendix: Why Does Weighting Work?

Basic Model Review

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + a_i + u_{it}$$

- Recall our standard linear model with unobserved time-invariant confounding
- We discussed a differencing approach to this model
- The Fixed effects model is an alternative way to remove time-invariant unmeasured confounding
- We will start by assuming the model and discussing properties and in the next section, we will consider non-parametric identification.

Fixed Effects Models

- Core idea is to focus on within-unit comparisons: changes in y_{it} and x_{it} relative to their within-group means
- First note that taking the average of the y's over time for a given unit leaves us with a very similar model:

$$
\begin{aligned}
\overline{y}_i &= \frac{1}{T}\sum_{t=1}^{T}\left[\mathbf{x}_{it}'\boldsymbol{\beta} + a_i + u_{it}\right] \\
&= \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{it}'\right)\boldsymbol{\beta} + \frac{1}{T}\sum_{t=1}^{T}a_i + \frac{1}{T}\sum_{t=1}^{T}u_{it} \\
&= \overline{\mathbf{x}}_i'\boldsymbol{\beta} + a_i + \overline{u}_i
\end{aligned}
$$

- Key fact: because it is time-constant the mean of a_i is just a_i
- This regression is sometimes called the "between regression"

Within Transformation

- The "fixed effects," "within," or "time-demeaning" transformation is when we subtract off the over-time means from the original data:

$$(y_{it} - \overline{y}_i) = (\mathbf{x}'_{it} - \overline{\mathbf{x}}'_i)\boldsymbol{\beta} + (u_{it} - \overline{u}_i)$$

- If we write $\ddot{y}_{it} = y_{it} - \overline{y}_i$, then we can write this more compactly as:

$$\ddot{y}_{it} = \ddot{\mathbf{x}}'_{it}\boldsymbol{\beta} + \ddot{u}_{it}$$

- Degrees of freedom: $nT - n - k - 1$, which accounts for within transformation (i.e. either use a package like `plm` or adjust the degrees of freedom manually).
- We are now modeling observations as deviation from their group mean.
- NB: you must demean the X variables not just the Y variables.

Fixed Effects with Ross data

```
fe.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur), data = ross, index = c("id", "year"),
 model = "within")
summary(fe.mod)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(kidmort_unicef) ~ democracy + log(GDPcur),
##     data = ross, model = "within", index = c("id", "year"))
##
## Unbalanced Panel: n=166, T=1-7, N=649
##
## Residuals :
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -0.70500 -0.11700  0.00628  0.12200  0.75700
##
## Coefficients :
##              Estimate Std. Error  t-value  Pr(>|t|)
## democracy   -0.143233   0.033500  -4.2756 2.299e-05 ***
## log(GDPcur) -0.375203   0.011328 -33.1226 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    81.711
## Residual Sum of Squares: 23.012
## R-Squared      :  0.71838
##       Adj. R-Squared :  0.53242
## F-statistic: 613.481 on 2 and 481 DF, p-value: < 2.22e-16
```

Strict Exogeneity

- FE models are valid if $\mathbb{E}[\mathbf{u}|\mathbf{X}] = 0$: all errors are uncorrelated with covariates in every period:

$$\mathbb{E}[\ddot{u}_{it}|\ddot{\mathbf{X}}] = \mathbb{E}[u_{it}|\ddot{\mathbf{X}}] - \mathbb{E}[\overline{u}_i|\ddot{\mathbf{X}}] = 0 - 0 = 0$$

- This is because the composite errors, $\ddot{u}_{it}$ are function of the errors in every time period through the average, $\overline{u}_i$
- This rules out, for instance, lagged dependent variables, since $y_{i,t-1}$ has to be correlated with $u_{i,t-1}$. Thus it can't be a covariate for y_{it}.

Fixed Effects and Time-Invariant Covariates

- What if there is a covariate that doesn't vary over time?
- Then $x_{it} = \overline{x}_i$ and $\ddot{x}_{it} = 0$ for all periods t.
- If the time-demeaned covariate is always 0, then it will be perfectly collinear with the intercept and will violate full rank. R/Stata and the like will drop it from the regression.
- Basic message: any time-constant variable gets "absorbed" by the fixed effect. It has nothing to contribute because the comparison is within the units.
- Can include interactions between time-constant and time-varying variables, but lower order term of the time-constant variables get absorbed by fixed effects too

Time-constant variables

- Pooled model with a time-constant variable, proportion Islamic:

```
library(lmtest)
p.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur) + islam,
             data = ross, index = c("id", "year"), model = "pooling")
coeftest(p.mod)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept) 10.30607817  0.35951939  28.6663 < 2.2e-16 ***
## democracy   -0.80233845  0.07766814 -10.3303 < 2.2e-16 ***
## log(GDPcur) -0.25497406  0.01607061 -15.8659 < 2.2e-16 ***
## islam        0.00343325  0.00091045   3.7709 0.0001794 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Time-constant variables

- FE model, where the islam variable drops out, along with the intercept:

```
fe.mod2 <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur) + islam,
               data = ross, index = c("id", "year"), model = "within")
coeftest(fe.mod2)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## democracy   -0.129693   0.035865  -3.6162 0.0003332 ***
## log(GDPcur) -0.379997   0.011849 -32.0707 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Alternate Computation: Least Squares Dummy Variable

- As an alternative to the within transformation, we can also include a series of $n-1$ dummy variables for each unit:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + d_i^{(1)}\alpha_1 + d_i^{(2)}\alpha_2 + \cdots + d_i^{(n)}\alpha_n + u_{it}$$

- Here, $d_i^{(1)}$ is a binary variable which is 1 if $i = 1$ and 0 otherwise—just a unit dummy.
- Gives the exact same estimates/standard errors as with time-demeaning
  - Advantage: easy to implement in base `R` (so is the de-meaning but you have to recompute standard errors by changing the degrees of freedom manually).
  - Disadvantage: computationally difficult with large data sets, since we have to run a regression with $n+k$ variables.
- Why are these equivalent? (remember partialing out strategy and Frisch-Waugh-Lovell theorem)

Example with Ross data

```
library(lmtest)
lsdv.mod <- lm(log(kidmort_unicef) ~ democracy + log(GDPcur) +
               as.factor(id), data = ross)
coeftest(lsdv.mod)[1:6,]
coeftest(fe.mod)[1:2,]
```

```
##                      Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)        13.7644887 0.26597312  51.751427 1.008329e-198
## democracy          -0.1432331 0.03349977  -4.275644  2.299393e-05
## log(GDPcur)        -0.3752030 0.01132772 -33.122568 3.494887e-126
## as.factor(id)AGO    0.2997206 0.16767730   1.787485  7.448861e-02
## as.factor(id)ALB   -1.9309618 0.19013955 -10.155498  4.392512e-22
## as.factor(id)ARE   -1.8762909 0.17020738 -11.023558  2.386557e-25
```

```
##               Estimate Std. Error    t value      Pr(>|t|)
## democracy   -0.1432331 0.03349977  -4.275644  2.299393e-05
## log(GDPcur) -0.3752030 0.01132772 -33.122568 3.494887e-126
```

Fixed Effects Versus First Differences

- Key assumptions:
  - Strict exogeneity: $E[u_{it}|\mathbf{X}, a_i] = 0$
  - Time-constant unmeasured heterogeneity, a_i
- Together $\implies$ fixed effects and first differences are unbiased and consistent
- With $T = 2$ the estimators produce identical estimates, but not more generally although they have the same target estimand.
- So which one is better when $T > 2$? Which one is more efficient?
  - if u_{it} uncorrelated $\rightsquigarrow$ FE is more efficient
  - if $u_{it} = u_{i,t-1} + e_{it}$ with e_{it} iid (random walk) $\rightsquigarrow$ FD is more efficient.
- In between, not clear which is better (although if using FD, the errors are serially correlated and need correction).
- Large differences between FE and FD should make us worry about assumptions.
- Note that when the second dimension isn't time, fixed effects will be relevant more often.

1 Set Up

2 Differencing Models

3 Difference-in-Differences

4 Fixed Effects

5 Non-parametric Identification and Fixed Effects

6 (Almost) Twenty Questions
- Review
- Topics Beyond the Course
- Research Practice
- Opinions and Musings

7 Concluding Thoughts for the Course

8 Appendix: Why Does Weighting Work?

Moving Beyond Linear Separable Confounding

- One reason we like DAGs is that the identification results don't have to start with a statement like, assume the following linear model:

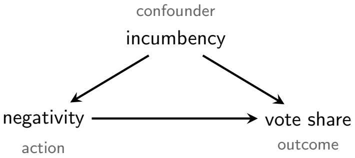
$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + a_i + u_{it}$$

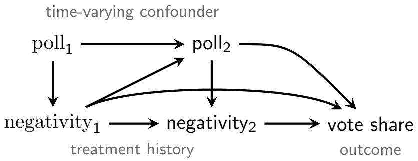
- What assumptions have we made so far?
  - constant effects
  - linearity
  - strict exogeneity
- We've seen the trouble with constant effects before, it goes back to Lecture 10 and results on regression with heterogenous treatment effects more generally.

Contemporaneous, Cumulative and Dynamic Effects

- Another assumption we have been making is that our interest is in a single contemporaneous effect: $\mathbf{x}'_{it}\boldsymbol{\beta}$
- What if we want to consider the history of a treatment or the effect of a treatment regime (i.e. a treatment that varies over time)?
- Opens up new estimands, but have to think about how time-varying confounders affect treatment assignment.

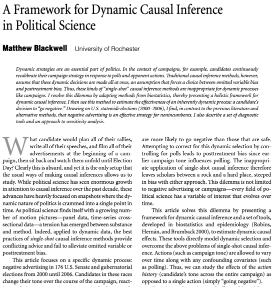
Examples of static and dynamic causal inference problems:





Core Conundrum

There is a (possibly irresolvable) tension: modeling causal dynamics between treatment and outcomes OR addressing unobserved time-invariant confounders. Three great recent papers:


A Framework for Dynamic Causal Inference in Political Science

Matthew Blackwell University of Rochester

Dynamic strategies are an essential part of politics. In the context of campaigns, for example, candidates continuously recalibrate their campaign strategy in response to polls and opponent actions. Traditional causal inference methods, however, assume that these dynamic decisions are made all at once, an assumption that forces a choice between omitted variable bias and posttreatment bias. Thus, these kinds of "single-shot" causal inference methods are inappropriate for dynamic processes like campaigns. I resolve this dilemma by adapting methods from biostatistics, thereby presenting a holistic framework for dynamic causal inference. I then use this method to estimate the effectiveness of an inherently dynamic process: a candidate's decision to "go negative." Drawing on U.S. statewide elections (2000–2006), I find, in contrast to the previous literature and alternative methods, that negative advertising is an effective strategy for nonincumbents. I also describe a set of diagnostic tools and an approach to sensitivity analysis.

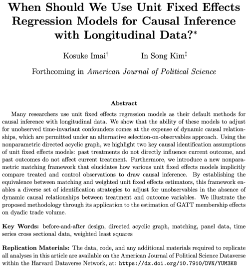

American Political Science Review (2018) 112, 4, 1067–1082
doi:10.1017/S0003055418000357 © American Political Science Association 2018

How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables

MATTHEW BLACKWELL *Harvard University*
ADAM N. GLYNN *Emory University*

Repeated measurements of the same countries, people, or groups over time are vital to many fields of political science. These measurements, sometimes called time-series cross-sectional (TSCS) data, allow researchers to estimate a broad set of causal quantities, including contemporaneous effects and direct effects of lagged treatments. Unfortunately, popular methods for TSCS data can only produce valid inferences for lagged effects under some strong assumptions. In this paper, we use potential outcomes to define causal quantities of interest in these settings and clarify how standard models like the autoregressive distributed lag model can produce biased estimates of these quantities due to post-treatment conditioning. We then describe two estimation strategies that avoid these post-treatment biases—inverse probability weighting and structural nested mean models—and show via simulations that they can outperform standard approaches in small sample settings. We illustrate these methods in a study of how welfare spending affects terrorism.

INTRODUCTION


When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?*

Kosuke Imai† In Song Kim‡

Forthcoming in *American Journal of Political Science*

Abstract

Many researchers use unit fixed effects regression models as their default methods for causal inference with longitudinal data. We show that the ability of these models to adjust for unobserved time-invariant confounders comes at the expense of dynamic causal relationships, which are permitted under an alternative selection-on-observables approach. Using the nonparametric directed acyclic graph, we highlight two key causal identification assumptions of unit fixed effects models: past treatments do not directly influence current outcome, and past outcomes do not affect current treatment. Furthermore, we introduce a new nonparametric matching framework that elucidates how various unit fixed effects models implicitly compare treated and control observations to draw causal inference. By establishing the equivalence between matching and weighted unit fixed effects estimators, this framework enables a diverse set of identification strategies to adjust for unobservables in the absence of dynamic causal relationships between treatment and outcome variables. We illustrate the proposed methodology through its application to the estimation of GATT membership effects on dyadic trade volume.

Key Words: before-and-after design, directed acyclic graph, matching, panel data, time series cross sectional data, weighted least squares

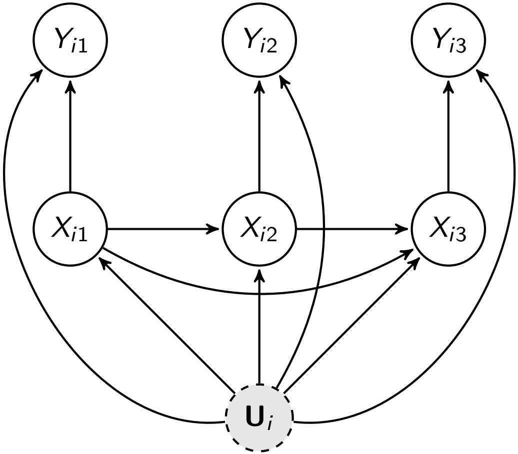
Replication Materials: The data, code, and any additional materials required to replicate all analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: https://dx.doi.org/10.7910/DVN/YUM3K8

We are going to focus on addressing unobserved time-invariant confounders using the last paper.
Next several slides are based on slides graciously provided by In Song Kim and Kosuke Imai.

Directed Acyclic Graph (DAG)

Non-parametric identification assumptions for fixed effects:

$$Y_{it} \quad = \quad g(X_{it}, \mathbf{U}_i, \epsilon_{it}) \quad \text{and} \quad \epsilon_{it} \perp\!\!\!\perp \{\mathbf{X}_i, \mathbf{U}_i\}$$



Assumptions:

1. No unobserved time-varying confounders
2. Past outcomes do not directly affect current outcome
3. Past outcomes do not directly affect current treatment
4. Past treatments do not directly affect current outcome

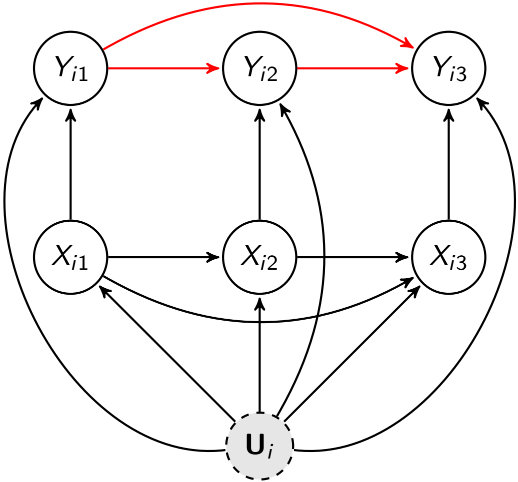
the result implies that the counterfactual outcome for a treated observation in a given time period is estimated using the observed outcomes of different time periods of the same unit. Since such a comparison is valid only when no causal dynamics exist, this finding underscores the important limitation of linear regression models with unit fixed effects.

- Imai and Kim (Forthcoming)

What Ideal Experiment Corresponds to the Fixed Effects Model?

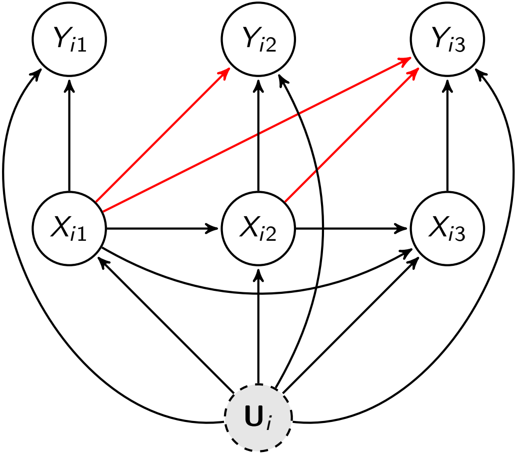
- Experiment that satisfies the model assumptions:
  1. randomize X_{i1} given $\mathbf{U}_i$
  2. randomize X_{i2} given X_{i1} and $\mathbf{U}_i$
  3. randomize X_{i3} given X_{i2}, X_{i1}, and $\mathbf{U}_i$
  4. and so on
- Experiment that does not satisfy the model assumptions:
  1. randomize X_{i1}
  2. randomize X_{i2} given X_{i1} and Y_{i1}
  3. randomize X_{i3} given X_{i2}, X_{i1}, Y_{i1}, and Y_{i2}
  4. and so on
- Now let's consider each assumption in turn.

Past Outcomes Don't Directly Affect Current Outcome



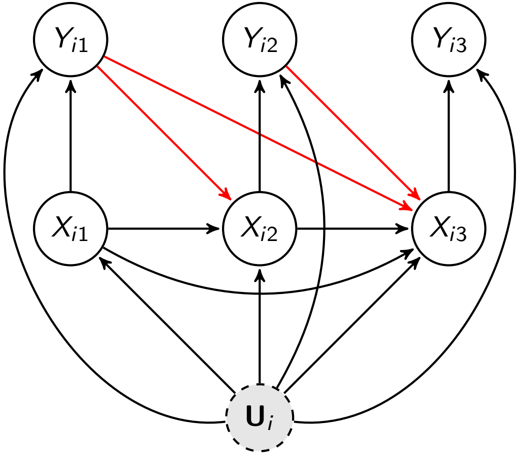
- Strict exogeneity still holds.
- Past outcomes do not confound $X_{it} \longrightarrow Y_{it}$ given $\mathbf{U}_i$.
- No need to adjust for past outcomes.
- Should use cluster robust standard errors for inference.
- Conclusion: The assumption can be relaxed

Past Treatments Don't Directly Affect Current Outcome



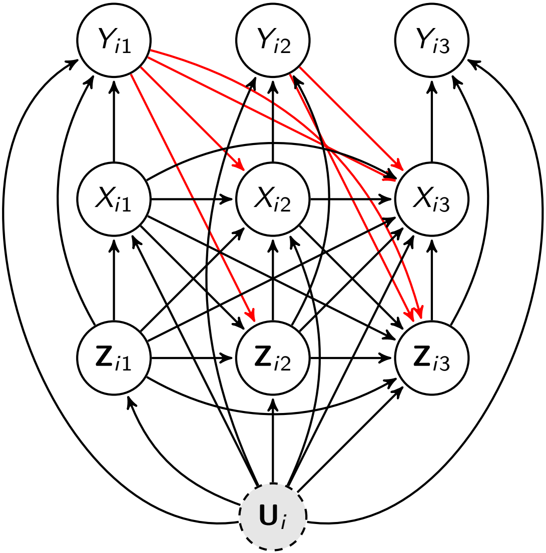
- Need to adjust for past treatments
- Strict exogeneity holds given past treatments and $\mathbf{U}_i$
- Impossible to adjust for an entire treatment history and $\mathbf{U}_i$ at the same time
- Adjust for a small number of past treatments $\rightsquigarrow$ often arbitrary
- Conclusion: The assumption can be partially relaxed

Past Outcomes Don't Directly Affect Current Treatment



- Correlation between error term and future treatments
- Violation of strict exogeneity
- No adjustment is sufficient
- Implication: No dynamic causal relationships between treatment and outcome variables
- Conclusion: The assumption cannot be relaxed

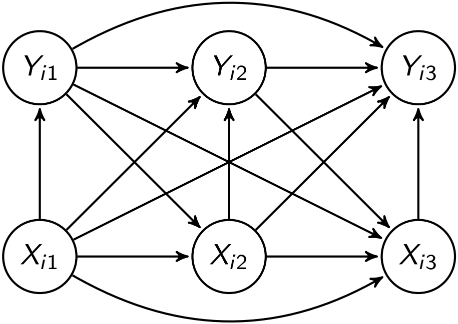
Can't We Just Adjust for Time-Varying Confounders?



- $Y_{it} = \alpha_i + \beta X_{it} + \gamma^\top \mathbf{Z}_{it} + \epsilon_{it}$
- past outcomes cannot directly affect current treatment
- past outcomes cannot *indirectly* affect current treatment through $\mathbf{Z}_{it}$

But What If I Have Causal Dynamics?

Alternative: Marginal Structural Models (Robins, Hernán and Brumback, 2000) — see Blackwell 2013 and Blackwell and Glynn 2018 for accessible introductions.



- Absence of unobserved time-invariant confounders $\mathbf{U}_i$
- past treatments can directly affect current outcome
- past outcomes can directly affect current treatment

- Comparison across units within the same time rather than across different time periods within the same unit
- Can identify the average effect of an entire treatment sequence
- Trade-off $\rightsquigarrow$ no free lunch

Conclusions and Nonparametric Estimation

- Imai and Kim (Forthcoming) offer a matching framework for fixed effects models which exploits an equivalence to weighted unit fixed effects estimators (see `wfe` package in `R` as well).
- The paper clarifies assumptions for fixed effects and first difference estimators.
- Follow-up working paper by Imai, Kim and Wang extends to two-way fixed effects estimator.
- Tradeoff:
  1) unobserved time-invariant confounders $\rightsquigarrow$ fixed effects
  2) causal dynamics between treatment and outcome $\rightsquigarrow$ selection-on-observables

1. Set Up
2. Differencing Models
3. Difference-in-Differences
4. Fixed Effects
5. Non-parametric Identification and Fixed Effects
6. (Almost) Twenty Questions
   - Review
   - Topics Beyond the Course
   - Research Practice
   - Opinions and Musings
7. Concluding Thoughts for the Course
8. Appendix: Why Does Weighting Work?

Q: What conditions do we need to infer causality?

A: A clear estimand, an identification strategy and an estimation strategy.

Identification Strategies in This Class

- Experiments (ignorability via randomization)
- Selection on Observables (conditional ignorability)
- Natural Experiments (ignorability via quasi-randomization)
- Instrumental Variables (instrument + exclusion restriction)
- Regression Discontinuity (continuity assumption)
- Difference-in-Differences (parallel trends)
- Fixed Effects (time-invariant unobserved heterogeneity, strict ignorability)

Essentially everything assumes: consistency/SUTVA (no interference between units, variation in the treatment is irrelevant) and positivity (there is some chance of all getting treatment)

Some Estimation Strategies

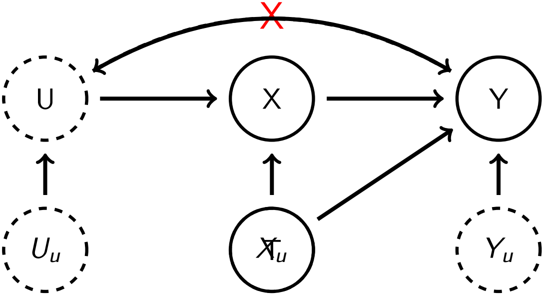
- Stratification
- Regression (and relatives)
- Matching (not covered)
- Weighting (not covered)

Q: Can you review how to read DAGs?

A: Sure[2]

[2] Courtesy of Erin Hartman's slides for this.

Notation

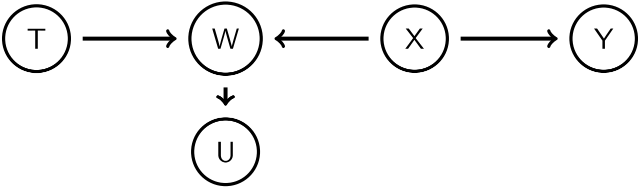


DAGs encode non-parametric structural models.

$$X = f_X(U)$$

$$Y = f_Y(X, U)$$

d-separation



A path p is blocked by a set of nodes Z if and only if:

(1) p contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node B is in Z or

(2) p contains a collider $A \rightarrow B \leftarrow C$ such that the collision node B is not in Z and no descendant of B is in Z

If Z blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on Z, and thus are conditionally independent given Z.

Q: Can you review how instrumental variables deal with issues of confounding?

A: We use only the units whose treatment status was effectively randomized by the instrument (because they are compliers).

Q: What are degrees of freedom and how do they play into standard errors?

A: Let's consider the anatomy of a standard error.

Anatomy of the Standard Error

Imagine we have a regression of Y on a variable of interest X and a vector of other variables $\mathbf{Z}$.

$$\widehat{\text{Var}}(\widehat{\beta}_X) = \frac{\frac{1}{(n-k-1)}\sum_{i=1}^{n}\hat{u}_i^2}{(1-R^2_{X\sim\mathbf{Z}})\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

- the numerator is our estimator for σ_u^2 the unknown error variance. It is formed by the degrees of freedom correction times the sum of the squared residuals.
- the denominator includes one minus the R^2 from the regression of X_i on $\mathbf{Z}_i$
- we complete the denominator by multiplying a measure of the variation in X_i, the sum of squared deviations from the mean.

$$\widehat{\text{SE}}(\widehat{\beta_X}) = \sqrt{\widehat{\text{Var}}(\widehat{\beta}_X)}$$

Q: When conducting an experiment, should we avoid OLS and always go for difference in means?

A: Regression adjustment of experiments can be helpful for improving precision. We don't need it for confounding, but it can improve our standard errors to adjust for pre-treatment covariates that are highly predictive of the output. If done correctly and in moderate-to-large samples, this can dramatically improve your standard errors. Even better though is blocking which is adjustment by design.

Further Reading:

- Lin, W., 2013. 'Agnostic notes on regression adjustments to experimental data: Reexamining Freedmans critique.' *The Annals of Applied Statistics*
- Athey, S. and Imbens, G.W., 2017. 'The Econometrics of Randomized Experiments.' In *Handbook of Economic Field Experiments* (Vol. 1, pp. 73-140).
- Egap Methods Guide: 10 things you need to know about covariate adjustment. `https://egap.org/methods-guides/10-things-know-about-covariate-adjustment`

1. Set Up
2. Differencing Models
3. Difference-in-Differences
4. Fixed Effects
5. Non-parametric Identification and Fixed Effects
6. (Almost) Twenty Questions
   - Review
   - Topics Beyond the Course
   - Research Practice
   - Opinions and Musings
7. Concluding Thoughts for the Course
8. Appendix: Why Does Weighting Work?

Q: Can you discuss the difference between having an instrument and having a mediator?

A: If we think of the treatment as the mediator of the instrument, it is by the exclusion restriction a total mediator (the direct effect is 0).

Q: How do propensity scores and matching fit into all of this?

A: They are different ways of conditioning on variables in a selection on observables strategy. Importantly: they are tools for estimation not tools for identification.

Propensity Score as a Low-Dimensional Summary

- Summary: The propensity score is the probability of treatment given some covariates X.
- Stratification is hard when X has has many dimensions
- Curse of dimensionality: there will be very few, if any, units in a given stratum of X_i.
- We can instead stratify on a low-dimensional summary, the propensity score:

$$e(x) = \mathbb{P}[D_i = 1 | X_i = x]$$

- Rosenbaum and Rubin (1983) showed that:

$$D_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \mid X_i \implies D_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \mid e(X_i)$$

- $\rightsquigarrow$ stratifying on e_i is the same in expectation as stratifying on the full X_i.
- The true propensity score is actually a balancing score, which means that $D_i \perp\!\!\!\perp X_i \mid e(X_i)$

Propensity score specifics

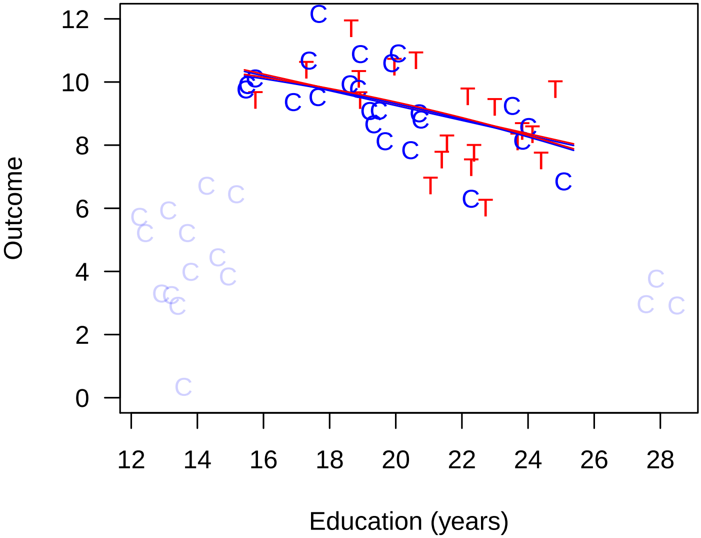
- What variables do we include in the propensity score model?
  - Any set of variables that blocks all the backdoor paths from D_i to Y_i.
- Check balance within strata of $\hat{e}_i$. Covariates should be balanced:

$$f(X_i|D_i = 1, \hat{e}_i) = f(X_i|D_i = 0, \hat{e}_i)$$

- Can also use automated/nonparametric tools for estimating $\hat{e}_i$.
- How do we use propensity scores?
  - Propensity score can be used in many contexts: weighting, matching, regression or even just stratification
  - It also shows up in a number of more advanced methods for heterogeneous treatment effects, causal inference in longitudinal data etc.
  - Typically it is a tool to achieve balance.
  - NB: propensity scores only achieve balance in expectation

Matching as Non-Parametric Preprocessing

(Ho, Imai, King, Stuart, 2007: fig.1, Political Analysis)



Three Approaches to Matching

- There are many approaches to matching. We will cover just three for the sake of time.
- This isn't a statement that these are the best three, just a set which are straightforward to learn.
- Which is the best method? The one that produces the best balance!

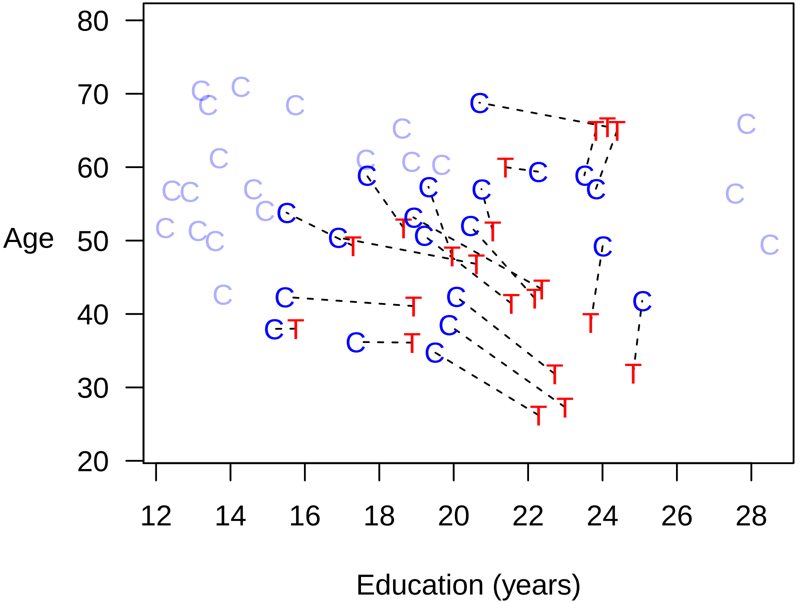
Next few slides based on slides by Gary King and Rich Nielsen

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance$(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance>caliper
2. Checking Measure imbalance, tweak, repeat, ...
3. Estimation Difference in means or a model

Mahalanobis Distance Matching

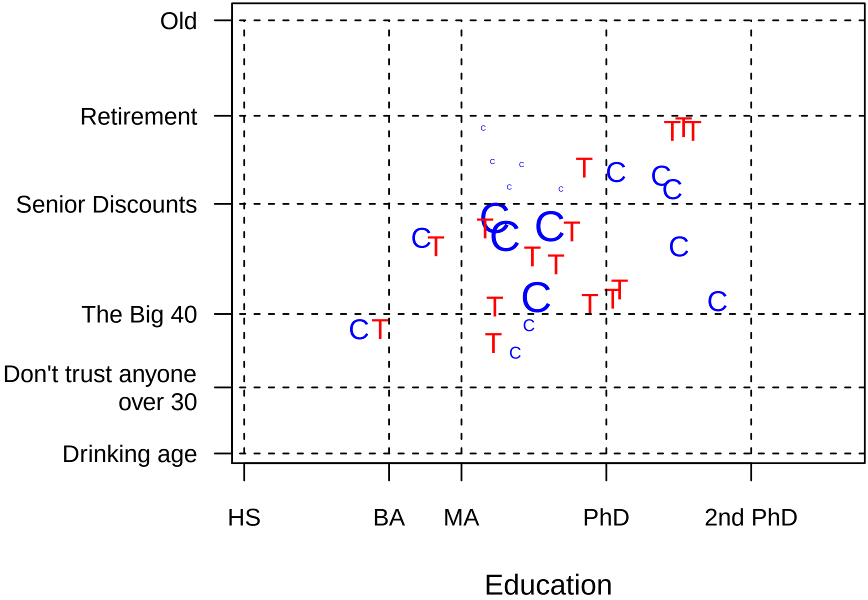


Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
   - Pass on original (uncoarsened) units except those pruned
2. Checking Determine matched sample size, tweak, repeat, ...
   - Easier, but still iterative
3. Estimation Difference in means or a model
   - Need to weight controls in each stratum to equal treateds

Coarsened Exact Matching

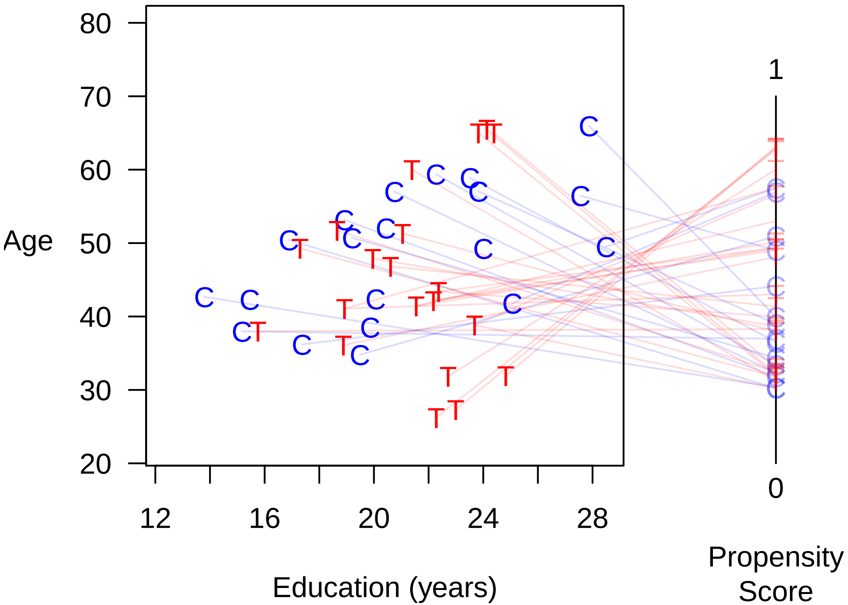


Method 3: Propensity Score Matching

(Approximates Completely Randomized Experiment)

1. Preprocess (Matching)
   - Reduce k elements of X to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - $\text{Distance}(X_i, X_j) = |\pi_i - \pi_j|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance>caliper
2. Checking Measure imbalance, tweak, repeat, ...
3. Estimation Difference in means or a model

Propensity Score Matching



Q: Could you discuss hierarchical models?

A: Sure. Generally speaking, they are a way of borrowing information.

Eight Schools Data

School	Est. Effect	SE
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Policy Question: What is the effect size in School A?

Eight Schools Background

- ETS analyzes special coaching program on test scores
- 8 separate parallel experiments in different high schools
- Outcome was the score on a special administration of SAT-V with scores varying between 200 and 800 ($\mu = 500, \sigma = 100$)
- SAT is designed to be resistant to short-term efforts intended to boost performance, but each school thought it was a success.
- No prior reason to believe that one program would be more effective than the others
- Treatment effects estimated controlling for PSAT-M and PSAT-V scores
- A bit over the 30 students in each school
- For the sake of scale: an 8-point increase in the score indicates about 1 more test item was answered correctly.
- (Analysis is from Rubin 1981, treatment via Gelman et al 2015)

What do we know?

- Unbiased estimate: 28 points
- Hypothesis test fails to reject hypothesis that true effect is the same for all of them
- Should we analyze them all together? All separately?
- It is the "same course" in every school, but they are different schools.

Options for Analysis

There are two clear options:

1. an unpooled analysis in which we use separate estimates for every school- in this case directly from the table
   - 2 moderate effects, 4 small effects and 2 small negative effects
   - standard errors are large, 95% intervals overlap substantially
2. a pooled analysis that generates a single estimate for all schools
   - assume that all effects are exactly the same
   - we get the single effect size and standard error with inverse variance weighting of the unpooled estimates.

$$\bar{y}_{\cdot} = \frac{\sum_{j=1}^{8} \frac{1}{\sigma_j^2} \bar{y}_j}{\sum_{j=1}^{8} \frac{1}{\sigma_j^2}}$$

$$\sigma_{\cdot}^2 = \left(\sum_{j=1}^{8} \frac{1}{\sigma_j^2} \right)^{-1}$$

   - the pooled estimate is 7.7 with standard error of 4.1. Thus the confidence interval is $[-.5, 15.9]$

Problems with Separate and Pooled Analysis

- The two approaches radically different results for school A: 28.4 (s.e. 14.9) vs. 7.7 (s.e. 4.1)
- Under a Bayesian framework, the separate analysis implies the probability statement "the probability is $\frac{1}{2}$ that the true effect in A is more than 28.4"
- This seems ... dubious given the other results (remember we had no reason to believe one school would perform stronger than the others)
- The pooled analysis implies the statement "the probability is $\frac{1}{2}$ that the true effect in A is less than 7.7", it also implies that "the probability is $\frac{1}{2}$ that the true effect in A is less than the true effect in C"
- Again these seem unlikely given the data

Borrowing Information

- We want an estimate that combines information from the 8 experiments without assuming that all the effects are equal
- Rubin suggests a middle path: a hierarchical model in which we
  1. assume that each school's true effect is drawn a Normal distribution with some unknown mean and standard deviation
  2. assume that the observed effect in each school is sampled from a normal distribution with a mean equal to its true effect and standard deviation given in the table
- This model contains both the separate and pooled estimates as limiting special cases. If we force the standard deviation of the true effects to be zero, then all school get the same estimate, if we let it go to infinity we get the separate estimates

The Model

$$\bar{y}_j|\theta_j \sim \text{Normal}(\theta_j, \sigma_j^2)$$
$$\theta_j|\mu, \tau \sim \text{Normal}(\mu, \tau^2)$$
$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$$

Known: $\bar{y}_j, \sigma_j^2$
Unknown: τ, μ, θ

Some Mechanics

How do the calculations work conditional on some values of the hyperparameters?

The θs are latent variables which have a distribution. In Bayesian statistics we call this the posterior distribution.

$$\theta_j|\mu, \tau, y \sim \mathrm{N}(\hat{\theta}_j, V_j)$$

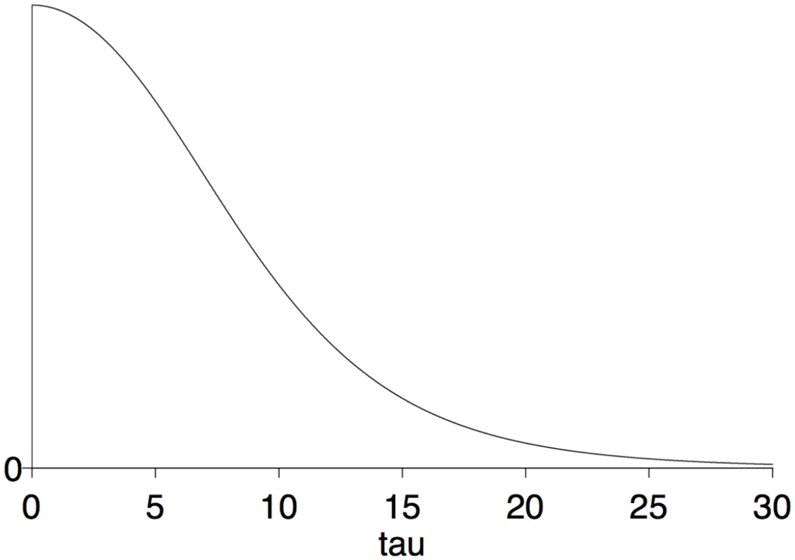
$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{y}_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

$$V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

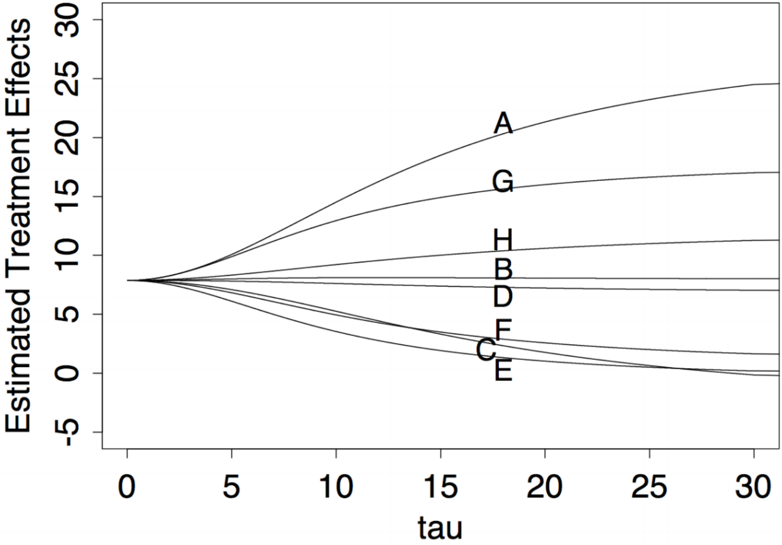
What is Happening?

- We are borrowing information between the schools
- Alternatively- we are regularizing estimates of the individual effects towards their grand mean
- This captures our intuition that while School A might have a larger effect, it is perhaps an overestimate
- The form show us that the amount of shrinkage is relative to our certainty about the estimate and how much we believe the individual effects matter
- Our final guess is that the median effect for school A is about 10 points with 50% probability between 7 and 16

Results

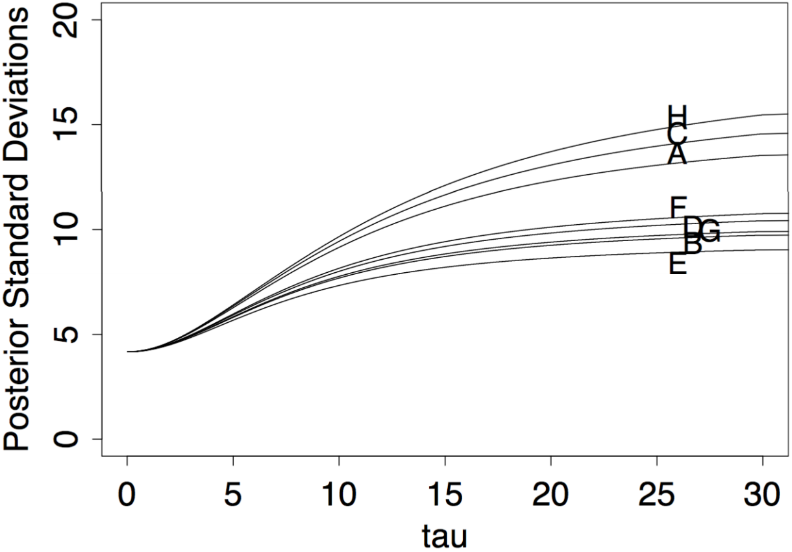


Results

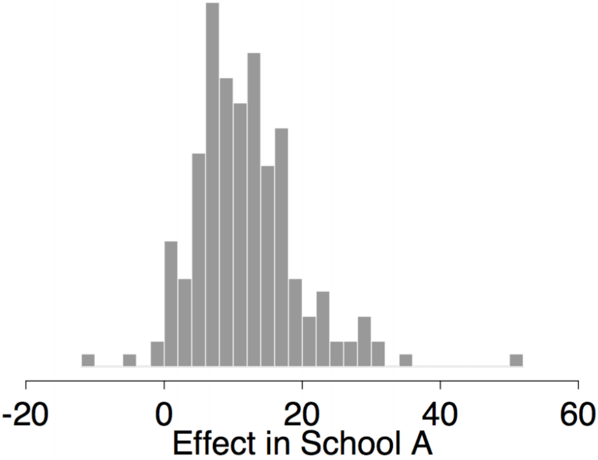


Estimated Treatment Effects
-5
0
5
10
15
20
25
30
A
G
H
B
D
F
C
E
0
5
10
15
20
25
30
tau

Results


Posterior Standard Deviations
0
5
10
15
20
0
5
10
15
20
25
30
tau
H
C
A
F
D
G
B
E

Results



The Great Thing About Eight Schools

- This is a microcosm of hierarchical modeling
- Works well when we have a decent number of groups and the individual group sample sizes are lowish
- Allows us to capture variability in our treatment effects, variances etc.
- Allows us to model dependence in our error terms

Q: How do we determine power?

A: A combination of the effect size, the variance and the sample size. Unfortunately, only one of which we know. See the `DeclareDesign` suite of packages for this and so much more!

1. Set Up
2. Differencing Models
3. Difference-in-Differences
4. Fixed Effects
5. Non-parametric Identification and Fixed Effects
6. (Almost) Twenty Questions
   - Review
   - Topics Beyond the Course
   - Research Practice
   - Opinions and Musings
7. Concluding Thoughts for the Course
8. Appendix: Why Does Weighting Work?

Q: Could we discuss more examples of missteps/misuses of certain statistical techniques/methods in papers published in prominent journals? I think seeing how other researchers have made mistakes and why mistakes arise could be helpful for diagnosing similar mistakes in our own work?

A: I think the biggest and most frequent mistakes I see are:

- not being clear about the estimand
- mistaking not significant results for a finding of zero effect (need equivalence tests)
- lack of clarity about the counterfactual and common support

Q: When should you pick your statistical strategy? How do you balance pre-planning research / literature reviews with potential problems with data/causal assumptions? How much data exploration should you do up front compared to exploration throughout the question? If you have a causal question or idea but arent sure of data, how should you go about searching for potential data and making sure assumptions are reasonable?

A: Let's chat.

Q: What do you believe will be the biggest applications for social statistics in the future?

A: Let's chat.

Q: What are your favorite resources for learning tricky concepts?

I've used the following procedure many times:

1. Identify approx. the best textbook (often can do this via syllabi hunting)
2. Read the relevant textbook material
3. Derive the equations/math
4. Try to explain it to someone else

1. Set Up
2. Differencing Models
3. Difference-in-Differences
4. Fixed Effects
5. Non-parametric Identification and Fixed Effects
6. (Almost) Twenty Questions
   - Review
   - Topics Beyond the Course
   - Research Practice
   - Opinions and Musings
7. Concluding Thoughts for the Course
8. Appendix: Why Does Weighting Work?

Where are you?

You've been given a powerful set of tools

Your New Weapons

- Basic probability theory
  - Probability axioms, random variables, marginal and conditional probability, building a probability model
  - Expected value, variances, independence
  - CDF and PDF (discrete and continuous)
- Properties of Estimators
  - Bias, Efficiency, Consistency
  - Central limit theorem
- Univariate Inference
  - Interval estimation (normal and non-normal Population)
  - Confidence intervals, hypothesis tests, p-values
  - Practical versus statistical significance

Your New Weapons

- Simple Regression
  - regression to approximate the conditional expectation function
  - idea of conditioning
  - kernel and loess regressions
  - OLS estimator for bivariate regression
  - Variance decomposition, goodness of fit, interpretation of estimates, transformations
- Multiple Regression
  - OLS estimator for multiple regression
  - Regression assumptions
  - Properties: Bias, Efficiency, Consistency
  - Standard errors, testing, p-values, and confidence intervals
  - Polynomials, Interactions, Dummy Variables
  - F-tests
  - Matrix notation

Your New Weapons

- Diagnosing and Fixing Regression Problems
  - Non-normality
  - Outliers, leverage, and influence points, Robust Regression
  - Non-linearities and GAMs
  - Heteroscedasticity and Clustering
- Causal Inference
  - Frameworks: potential outcomes and DAGs
  - Measured Confounding
  - Unmeasured Confounding
  - Methods for repeated data
- And you learned how to use `R`: you're not afraid of trying something new!

Using these Tools

So, Admiral Ackbar, now that you've learned how to run these regressions we can just use them blindly, right?


IT'S A TRAP!

Beyond Linear Regressions

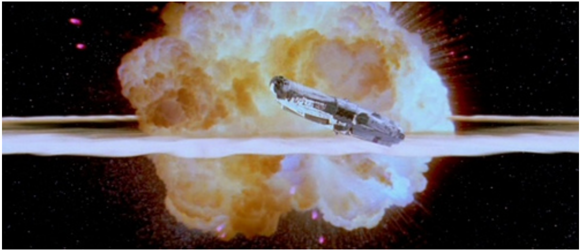
You need more training

Beyond Linear Regressions

There is so much more to learn! Take classes, read books!

Thanks!

Thanks so much for an amazing semester.



Fill out your evaluations!

1 Set Up

2 Differencing Models

3 Difference-in-Differences

4 Fixed Effects

5 Non-parametric Identification and Fixed Effects

6 (Almost) Twenty Questions
- Review
- Topics Beyond the Course
- Research Practice
- Opinions and Musings

7 Concluding Thoughts for the Course

8 Appendix: Why Does Weighting Work?

Weighting with the Propensity Score

Intuition

- Treated and control samples are unrepresentative of the overall population.
- Leads to imbalance in the covariates.
- Reweight them to be more representative.

Survey samples

- Useful to review survey samples to understand the logic
- Finite population: $\{1, \dots, N\}$
- Suppose that we wanted estimate the population mean of Y_i:

$$\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

- We have a sample of size n, where $Z_i = 1$ indicates that i is included in the sample.
- Unequal sampling probability: $\mathbb{P}(Z_i = 1) = \pi_i$
  - $\rightsquigarrow$ sample is not representative.
  - $\sum_{i=1}^{N} \pi_i = n$

Survey weights

- Sample mean is biased:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{N} Z_i Y_i\right] = \frac{1}{n}\sum_{i=1} \pi_i Y_i$$

- Inverse probability weighting: To correct, weight each unit by the reciprocal of the probability of being included in the sample: Y_i/π_i.
- Horvitz-Thompson estimator is unbiased:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{Z_i Y_i}{\pi_i}\right] = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{E}[Z_i] Y_i}{\pi_i} = \frac{1}{N}\sum_{i=1}^{N} \frac{\pi_i Y_i}{\pi_i} = \bar{Y}_N$$

- Reweights the sample to be representative of the population.

Back to causal effects

- With a completely randomized experiment, we can just use the simple differences in means:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$

- With no unmeasured confounders, we need to adjust for X_i.

$$\begin{aligned}
\mathbb{E}[Y_i(d)] &= \mathbb{E}\left[\mathbb{E}[Y_i(d)|X_i]\right] \\
&= \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i(d)|X_i = x]\mathbb{P}(X_i = x) \\
&= \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i(d)|D_i = d, X_i = x]\mathbb{P}(X_i = x) \\
&= \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i|D_i = d, X_i = x]\mathbb{P}(X_i = x)
\end{aligned}$$

- With subclassification, we binned X_i, calclulated within-bin differences and then averaged across the bins, just like this.

Searching for the weights

$$\mathbb{E}[Y_i(d)] = \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i | D_i = d, X_i = x]\mathbb{P}(X_i = x)$$

- Compare this to the the within treatment group average:

$$\begin{aligned}
\mathbb{E}[Y_i | D_i = d] &= \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i | D_i = d, X_i = x]\mathbb{P}(X_i = x | D_i = d) \\
&= \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i | D_i = d, X_i = x] \frac{\mathbb{P}(D_i = d | X_i = x)\mathbb{P}(X_i = x)}{\mathbb{P}(D_i = d)}
\end{aligned}$$

- How should we reweight the data from an observational study?
- If we were to reweight the data by $W_i = 1/\mathbb{P}(D_i = d | X_i)$, then we would break the relationship between D_i and X_i.

Weights

- Single binary covariate. Define the weight function:

$$w(d, x) = \frac{1}{e(x)^d (1 - e(x))^{1-d}}$$

- To get the weight for i, plug in observed treatment, covariate: $W_i = w(D_i, X_i)$
- If $(D_i, X_i) = (1, 1)$,

$$W_i = \frac{1}{e(1)} = \frac{1}{\mathbb{P}(D_i = 1 | X_i = 1)}$$

- If $(D_i, X_i) = (0, 0)$:

$$W_i = \frac{1}{1 - e(0)} = \frac{1}{\mathbb{P}(D_i = 0 | X_i = 0)}$$

Example

	$X_i = 0$	$X_i = 1$
$D_i = 0$	4	3
$D_i = 1$	4	9

- $\mathbb{P}(D_i = 1 | X_i = 0) = 0.5$
- $\mathbb{P}(D_i = 1 | X_i = 1) = 0.75$
- Weights:

	$X_i = 0$	$X_i = 1$
$D_i = 0$	$1/0.5$	$1/0.25$
$D_i = 1$	$1/0.5$	$1/0.75$

- Weighted data (the pseudo-population):

	$X_i = 0$	$X_i = 1$
$D_i = 0$	8	12
$D_i = 1$	8	12

- $\mathbb{P}_W(D_i = 1 | X_i = x) = 0.5$ for all x

Properties of reweighted data

- Let's calculate the weighted probability that $D_i = 1$.

$$\begin{aligned}
&\mathbb{P}_W[D_i = 1 | X_i = x] \\
&= \frac{w(1, x) \cdot \mathbb{P}[D_i = 1 | X_i = x]}{\omega^*} \\
&= \frac{\frac{1}{\mathbb{P}[D_i = 1 | X_i = x]} \cdot \mathbb{P}[D_i = 1 | X_i = x]}{\omega^*} \\
&= \frac{1}{\omega^*}.
\end{aligned}$$

- ω^* is a normalization factor to make sure probabilities sum to 1.
- Important point: $\mathbb{P}_W(D_i = 1 | X_i = 1) = \mathbb{P}_W(D_i = 1 | X_i = 0) = \frac{1}{\omega^*}$
- $\rightsquigarrow$ D_i independent of X_i in the reweighted data.

Overall mean

- What is the weighted mean for the treated group?
- Use a similar approach to survey weights, where D_i is the "sampling indicator":

$$\bar{Y}_i^w = \frac{1}{N} \sum_{i=1}^{N} D_i W_i Y_i$$

- $W_i Y_i$ is the weighted outcome, D_i is there to select out the treated observations.
- We want to see what the conditional weighted mean identifies:

$$\mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} W_i D_i Y_i\right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[W_i D_i Y_i] = \mathbb{E}[W_i D_i Y_i]$$

Proving unbiasedness

- Weighted mean of treated units is mean of potential outcome:

$$
\begin{aligned}
\mathbb{E}[W_i D_i Y_i] &= \mathbb{E}\left[\frac{D_i Y_i}{e(X_i)}\right] && \text{(Weight Def.)} \\
&= E\left[\frac{D_i Y_i(1)}{e(X_i)}\right] && \text{(Consistency)} \\
&= E\left[E\left[\frac{D_i Y_i(1)}{e(X_i)}\Big| X_i\right]\right] && \text{(Iterated Expectations)} \\
&= E\left[\frac{E[D_i|X_i]E[Y_i(1)|X_i]}{e(X_i)}\right] && \text{(n.u.c.)} \\
&= E\left[\frac{e(X_i)E[Y_i(1)|X_i]}{e(X_i)}\right] && \text{(Propensity Score Definition)} \\
&= E[Y_i(1)] && \text{(Iterated Expectations)}
\end{aligned}
$$

Putting it all together

- The same logic would give us the mean potential outcomes under control:
$$E\left[\frac{(1-D_i)Y_i}{1-e(X_i)}\right] = E[Y_i(0)]$$
- These two facts provide an estimator for the average treatment effect:
$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{e(X_i)} - \frac{(1-D_i)Y_i}{1-e(X_i)}\right)$$
- The above two results give us that this esimator is unbiased.
- This is sometimes called the Horvitz-Thompson estimator due to the close connection to the survey sampling estimator.