Week 12: Repeated Observations and Panel Data

Brandon Stewart[1]

Princeton

November 16–20, 2020

[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
  - panel data
  - diff-in-diff
  - fixed effects
  - wrap-up
- The Following Week
  - ?
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causality

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
- NB: we won't be using T for treatment today because it is extremely consistently used for time. We will end up using D for treatment which is another common letter for treatment.

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:

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

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

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

- 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:

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

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

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* $E[u_{it}|\mathbf{X}] = 0$, then, $E[\Delta u_i | \Delta X] = 0$ 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
```

We Covered

- The basic panel data notation
- First Difference Models

Next Time: Difference-in-Differences

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

- Last Week
  - causal inference with unmeasured confounding
- This Week
  - panel data
  - diff-in-diff
  - fixed effects
  - wrap-up
- The Following Week
  - ?
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causality

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)

https://www.jstor.org/stable/2118030

Motivation: Studying the Minimum Wage

- Economics conventional wisdom: higher minimum wages decrease low-wage jobs.
- Card and Krueger (1994) study a 1992 NJ minimum wage increase (\$4.25 to \$5.05).
- Idea: compare employment rates in 410 fast-food restaurants in NJ and eastern PA (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
- "What would a skeptic consider convincing evidence?" David Card
- "There was a time when we thought econometric techniques would solve a lot of the data problems. Now I think the feeling is that there are a lot of problems for which it is easier to get better data." Alan Krueger

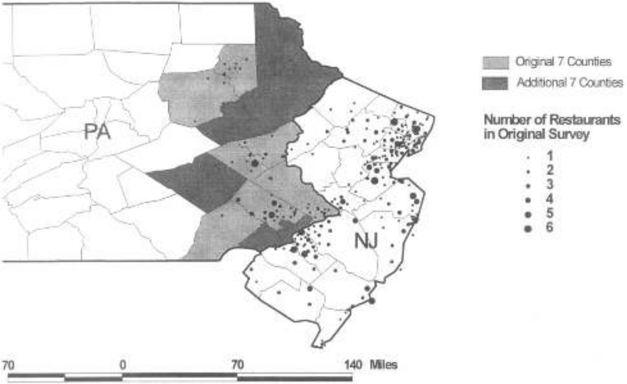


FIGURE 1. AREAS OF NEW JERSEY AND PENNSYLVANIA COVERED BY ORIGINAL SURVEY AND BLS DATA

Source: Card and Krueger 2000

Conversations / David Card and Alan Krueger

Two Economists Catch Clinton's Eye By Bucking the Common Wisdom

They use control groups in their research, and test minimum-wage theories by surveying the managers of fast-food restaurants.

'Let's Not Assume'

The New Jersey Test

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:

$$y_{it} = \beta_0 + \delta_0 I(t = 2) + \beta_1 x_{it} + a_i + u_{it}$$

- $I(t = 2)$ is a dummy variable for the second time period
- β_1 is the quantity of interest: it's the effect of being treated

Diff-in-Diff Mechanics

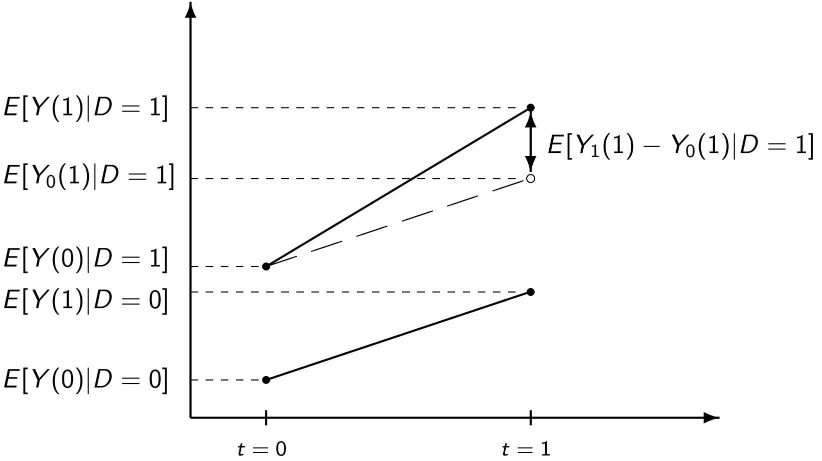
- Let's take differences:

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

- 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

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

$\square$

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
© 2009 SAGE Publications
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?

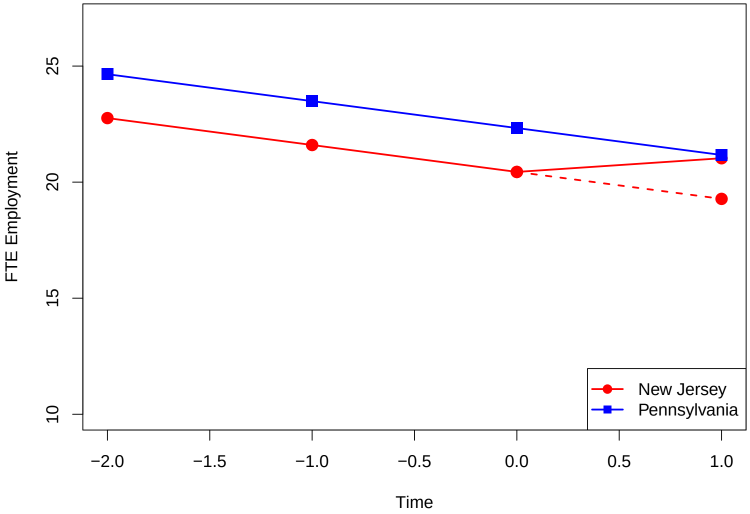
$$\text{employment}_{it} = \beta_0 + \beta_1 \text{minimum wage}_{it} + a_i + u_{it}$$

- 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

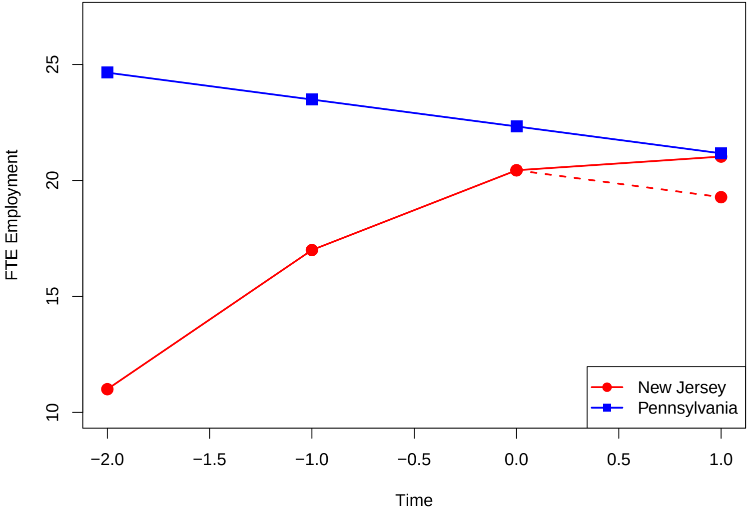
$$\Delta\text{employment}_i = \beta_0 + \beta_1 NJ_i + \Delta u_i$$

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

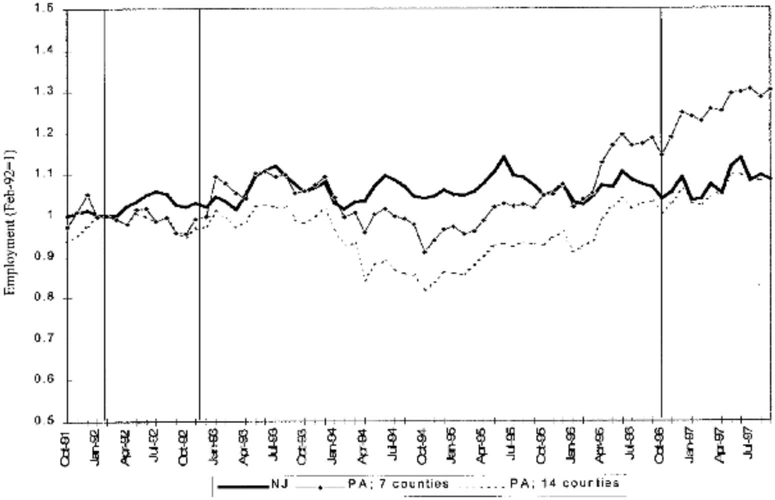
Parallel Trends?



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:

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

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?'

What to read next?

- Angrist and Pishke Chapter 5 Parallel Worlds: Fixed Effects, Differences-in-Differences and Panel Data
- Morgan and Winship Chapter 11 Repeated Observations and the Estimation of Causal Effects

We Covered

- Difference-in-Differences
- Parallel Trends Assumption

Next Time: Fixed Effects

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

- Last Week
  - causal inference with unmeasured confounding
- This Week
  - panel data
  - diff-in-diff
  - fixed effects
  - wrap-up
- The Following Week
  - ?
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causality

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 $E[\mathbf{u}|\mathbf{X}] = 0$: all errors are uncorrelated with covariates in every period:

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

- 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
```

Applying Fixed Effects

- We can use fixed effects for other data structures to restrict comparisons to within unit variation
  - Matched pairs
    - Twin fixed effects to control for unobserved effects of family background
  - Cluster fixed effects in hierarchical data
    - School fixed effects to control for unobserved effects of school

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 $\leadsto$ FE is more efficient
  - if $u_{it} = u_{i,t-1} + e_{it}$ with e_{it} iid (random walk) $\leadsto$ 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.

We Covered

- Fixed Effects!
- Computation for Fixed Effects!

Next Time: Non-parametric Identification and Fixed Effects

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

- Last Week
  - causal inference with unmeasured confounding
- This Week
  - panel data
  - diff-in-diff
  - fixed effects
  - wrap-up
- The Following Week
  - ?
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causality

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:

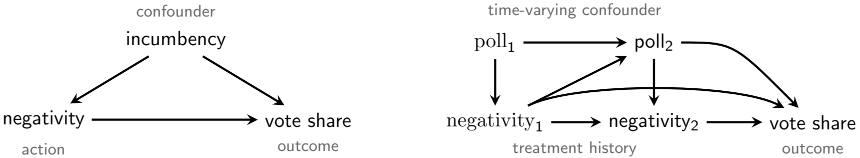
$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + a_i + u_{it}$$

- 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


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

MATTHEW BLACKWELL Harvard University

ADAM N. GLYNN Emory University

INTRODUCTION


AJPS

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

Kosuke Imai

In Song Kim

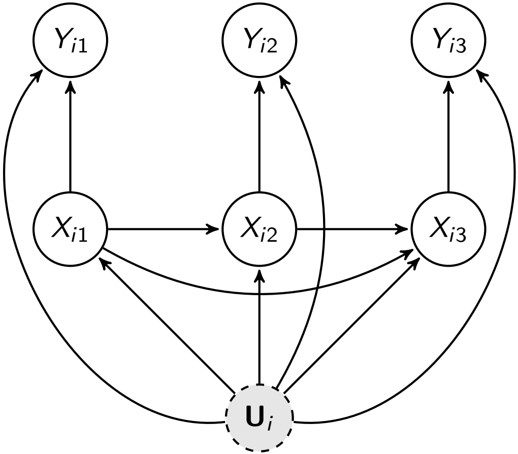
We are going to focus on addressing unobserved time-invariant confounders using the last paper.

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

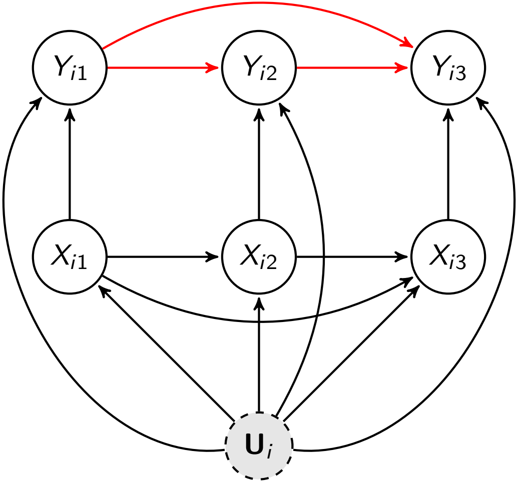
the result implies that the counterfactual outcome for a treated observation in a given time period is estimated using the observed outcomes of different time periods of the same unit. Since such a comparison is valid only when no causal dynamics exist, this finding underscores the important limitation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

- Imai and Kim (2019)

What Ideal Experiment Corresponds to the Fixed Effects Model?

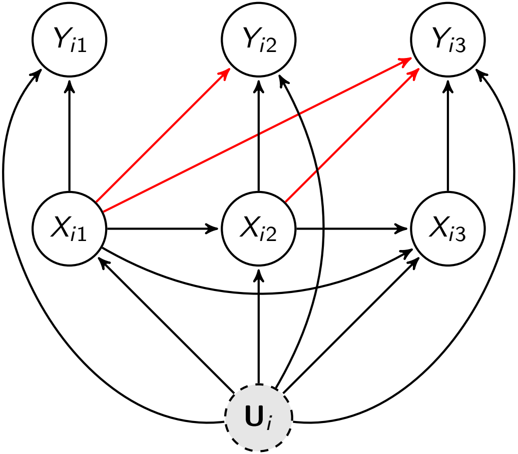
- Experiment that satisfies the model assumptions:
  1. randomize X_{i1} given $\mathbf{U}_i$
  2. randomize X_{i2} given X_{i1} and $\mathbf{U}_i$
  3. randomize X_{i3} given X_{i2}, X_{i1}, and $\mathbf{U}_i$
  4. and so on
- Experiment that does not satisfy the model assumptions:
  1. randomize X_{i1}
  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 (A2)



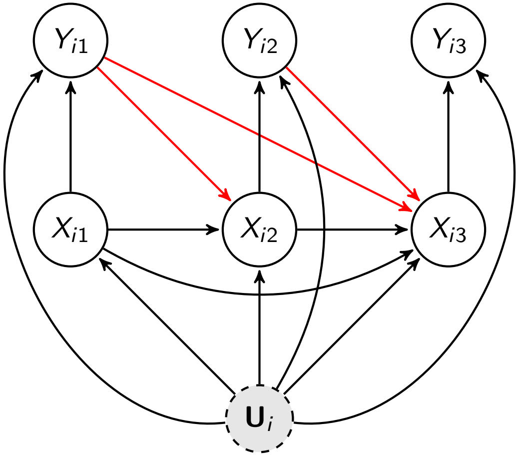
- Strict exogeneity still holds.
- Past outcomes do not confound $X_{it} \longrightarrow Y_{it}$ given $\mathbf{U}_i$.
- No need to adjust for past outcomes.
- Should use cluster robust standard errors for inference.
- Conclusion: The assumption can be relaxed

Past Treatments Don't Directly Affect Current Outcome (A4)



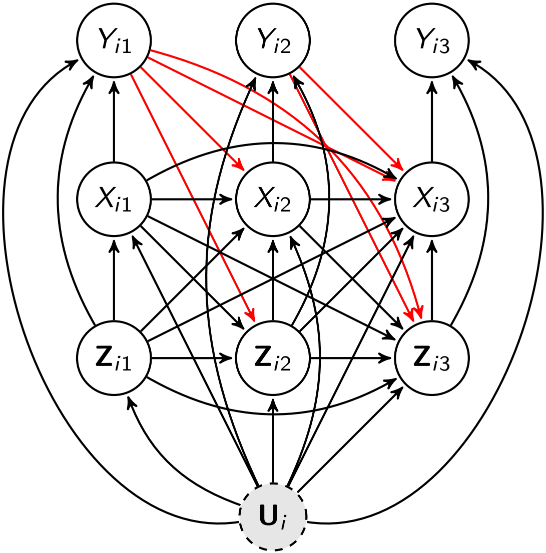
- Need to adjust for past treatments
- Strict exogeneity holds given past treatments and $\mathbf{U}_i$
- Impossible to adjust for an entire treatment history and $\mathbf{U}_i$ at the same time
- Adjust for a small number of past treatments $\rightsquigarrow$ often arbitrary
- Conclusion: The assumption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

Past Outcomes Don't Directly Affect Current Treatment (A3)



- 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

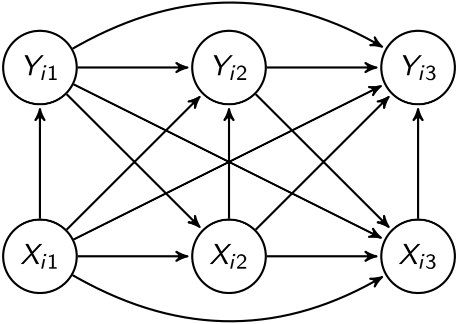
Can't We Just Adjust for Time-Varying Confounders?



- $Y_{it} = \alpha_i + \beta X_{it} + \gamma^\top \mathbf{Z}_{it} + \epsilon_{it}$
- 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 (2019) 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 Imai and Kim (2020) and 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

Summary Table (Imai and Kim 2019)

TABLE 1 Identification Assumptions of Various Estimators

	Linearity	Time-Invariant Unobservables	Past Outcomes Affect Current Treatment	Past Treatments Affect Current Outcome
$Y_{it} = \alpha_i + \beta X_{it} + \epsilon_{it}$	Yes	Allowed	Not allowed	Not allowed
$Y_{it} = \alpha_i + \beta X_{it} + \rho Y_{i,t-1} + \epsilon_{it}$	Yes	Allowed	Allowed	Not allowed
$Y_{it} = \alpha_i + \beta_1 X_{it} + \beta_2 X_{i,t-1} + \epsilon_{it}$	Yes	Allowed	Not allowed	Allowed
$Y_{it} = \alpha_i + \beta_1 X_{it} + \beta_2 X_{i,t-1} + \rho Y_{i,t-1} + \epsilon_{it}$	Yes	Allowed	Partially allowed	Partially allowed
Marginal structural models	No	Not allowed	Allowed	Allowed

What to read next?

- Morgan and Winship Chapter 11 Repeated Observations and the Estimation of Causal Effects
- Imai and Kim (2019) "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science*, `http://dx.doi.org/10.1111/ajps.12417`

We Covered

- Non-parametric identification for fixed effects.
- A glimpse at dynamic causal inference.

Next Time: Review

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

- Last Week
  - causal inference with unmeasured confounding
- This Week
  - panel data
  - diff-in-diff
  - fixed effects
  - wrap-up
- The Following Week
  - ?
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causality

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)

Almost 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 (lightly covered)
- Weighting (not covered)

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 can make standard errors larger or smaller?

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 Freedman's 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`

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

Q: In the real world of research, when you have your data, how do you know which method to use? For example, how do you know that you need to use matching/regression discontinuity?

A: Let's chat.

Q: What would you have covered with more time?

1. Missing Data
2. Sensitivity Analysis
3. Mediation (and Front Door Adjustment)
4. Bounds and Partial Identification
5. Weighting Approaches
6. Quantile Regression
7. On the Ground Decision Making

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 in matrix notation
  - Regression assumptions (classical and agnostic)
  - Properties: Bias, Efficiency, Consistency
  - Standard errors, testing, p-values, and confidence intervals
  - Polynomials, Interactions, Dummy Variables
  - F-tests

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

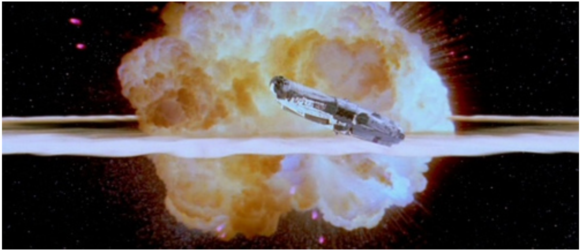
There is so much more to learn! Take classes, read books!

- Take SOC 504 or the POL sequence!
- Pursue the SML certificate!
- Go to workshops and seminars etc.!
- Read the suggested books in the syllabus!

I've tried to offer a perspective on technique in addition to just a list of methods. See Lundberg, Johnson and Stewart "What's Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory"

Thanks!

Thanks so much for an amazing semester.



I will see you in the final synchronous session!