# GSERM 2022

## Regression for Publishing

June 17, 2022

# Potential Outcomes and Counterfactual Inference

The goal: **Making causal inferences from observational data.**

- Establish and measure the *causal* relationship between variables in a non-experimental setting.
- The "fundamental problem of causal inference":

  *It is impossible to observe the causal effect of a treatment / predictor on a single unit.*

- Specific challenges:
  - *Confounding*
  - *Selection bias*
  - *Heterogenous treatment effects*

# Causation and Counterfactuals

**Causal statements imply <u>counterfactual</u> reasoning.**

- "If the cause(s) had been different, the outcome(s) would be different, too."
- Conditioning, probabilistic and causal:

| Probabilistic conditioning | Causal conditioning |
|---|---|
| $\Pr(Y \mid X = x)$ | $\Pr[Y \mid do(X = x)]$ |
| Factual | Counterfactual |
| Select a sub-population | Generate a new population |
| Predicts passive observation | Predicts active manipulation |
| Calculate from full DAG* | Calculate from surgically-altered DAG* |
| Always identifiable when $X$ and $Y$ are observable | Not always identifiable even when $X$ and $Y$ are observable |

*See below. Source: Swiped from Shalizi, "Advanced Data Analysis from an Elementary Point of View", Table 23.1.

- Causality (typically) implies / requires:
  - *Temporal ordering*
  - *Mechanism*
  - *Correlation*

**Notation**

- $N$ observations indexed by $i$, $i \in \{1, 2, ...N\}$
- Outcome variable $Y$
- Interest: the effect on $Y$ of a treatment variable $W$:
  - $W_i = 1 \leftrightarrow$ observation $i$ is "treated"
  - $W_i = 0 \leftrightarrow$ observation $i$ is "control"

**Potential Outcomes**

- $Y_{0i}$ = the value of $Y_i$ if $W_i = 0$
- $Y_{1i}$ = the value of $Y_i$ if $W_i = 1$
- $\delta_i = (Y_{1i} - Y_{0i})$ = the treatment effect of $W$

The average treatment effect (ATE) is just:

$$\begin{aligned} \text{ATE} \equiv \bar{\delta} &= \text{E}(Y_{1i} - Y_{0i}) \\ &= \frac{1}{N}\sum_{i=1}^{N} Y_{1i} - Y_{0i}. \end{aligned}$$

BUT we observe only $Y_i$:

$$Y_i = \begin{cases} Y_{0i} \text{ if } W_i = 0, \\ Y_{1i} \text{ if } W_i = 1. \end{cases}$$

or (equivalently)

$$Y_i = W_i Y_{1i} + (1 - W_i) Y_{0i}.$$

# Estimating Treatment Effects

Key to estimating treatment effects: **Assignment mechanism for** $W$.

Neyman/Rubin/Holland: Treat inability to observed $Y_{0i}$ / $Y_{1i}$ as a missing data problem.

Q: How do we think about missing data?

A: Time for a digression...

Notation:

$$\underset{N \times k}{\mathbf{X}_i} \cup \{\mathbf{W}_i, \mathbf{Z}_i\}$$

$\mathbf{W}_i$ have some missing values,
$\mathbf{Z}_i$ are "complete"

$$R_{ik} = \begin{cases} 1 & \text{if } W_{ik} \text{ is missing,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\pi_{ik} = \Pr(R_{ik} = 1)$$

Rubin's flavors of missingness:

- Missing completely at random ("MCAR") (= "ignorable"):

$$\mathbf{R} \perp \{\mathbf{Z}, \mathbf{W}\}$$

- Missing at random ("MAR") (conditionally "ignorable"):

$$\mathbf{R} \perp \mathbf{W} | \mathbf{Z}$$

- Anything else is "informatively" (or "non-ignorably") missing.

# Estimating Treatment Effects

Key to estimating treatment effects: **Assignment mechanism for $W$.**

Neyman/Rubin/Holland: Treat inability to observed $Y_{0i}$ / $Y_{1i}$ as a missing data problem.

- If the "missingness" due to the value of $W_i$ is orthogonal to the values of $Y$, then it is ignorable. Formally:

$$\Pr(W_i|\mathbf{X}_i, Y_{0i}, Y_{1i}) = \Pr(W_i|\mathbf{X}_i)$$

- If that "missingness" is non-orthogonal, then it is not ignorable, and can lead to bias in estimation
- Non-ignorable assignment of $W$ requires understanding the mechanism by which that assignment occurs

# SUTVA

One more thing: the stable unit-treatment value assumption ("SUTVA")

- Requires that there be two and only two possible values of $Y$ for each observation $i$.

- That is: "the observation (of $Y_i$) on one unit should be unaffected by the particular assignment of treatments to the other units."

- $\equiv$ the "assumption of no interference between units," meaning:
  - Values of $Y$ for any two $i, j$ ($i \neq j$) observations do not depend on each other
  - Treatment effects are homogenous within categories defined by $W$

# Treatment Effects Under Randomization of $W$

If $W_i$ is assigned randomly, then:

$$\Pr(W_i) \perp Y_{0i}, Y_{1i}$$

and so:

$$\Pr(W_i | Y_{0i}, Y_{1i}) = \Pr(W_i) \, \forall \, Y_{0i}, Y_{1i}.$$

This means that the "missing" data on $Y_0 / Y_1$ are ignorable (here, in the special case where the $\mathbf{X}_i$ on which $W_i$ depends is null).

This in turn means that:

$$f(Y_{0i} | W_i = 0) = f(Y_{0i} | W_i = 1) = f(Y_i | W_i = 0) = f(Y_i | W_i = 1)$$

and

$$f(Y_{1i} | W_i = 0) = f(Y_{1i} | W_i = 1) = f(Y_i | W_i = 0) = f(Y_i | W_i = 1)$$

# Randomized $W$ (continued)

Implication: $Y_{0i}$ and $Y_{1i}$ are (not identical but) *exchangeable*...

This in turn means that:

$$E(Y_{0i}|W_i) = E(Y_{1i}|W_i)$$

and so

$$\begin{aligned}\widehat{\text{ATE}} &= \text{E}(Y_i|W_i = 1) - \text{E}(Y_i|W_i = 0) \\ &= \bar{Y}_{W=1} - \bar{Y}_{W=0}.\end{aligned}$$

will be an unbiased estimate of the ATE.

# Observational Data: $W$ Depends on $\mathbf{X}$

Formally,

$$Y_{0i}, Y_{1i} \perp W_i | \mathbf{X}_i.$$

Here,

- $\mathbf{X}$ are *known confounders* that (stochastically) determine the value of $W_i$,
- Conditioning on $\mathbf{X}$ is necessary to achieve exchangeability.

So long as $W$ is entirely due to $\mathbf{X}$, we can condition:

$$f(Y_{1i}|\mathbf{X}_i, W_i = 1) = f(Y_{1i}|\mathbf{X}_i, W_i = 0) = f(Y_i|\mathbf{X}_i, W_i)$$

and similarly for $Y_{0i}$.

# $W$ Depends on $\mathbf{X}$ (continued)

Estimands:

- the *average treatment effect for the treated* (ATT):

$$\mathrm{ATT} = \mathrm{E}(Y_{1i}|W_i = 1) - \mathrm{E}(Y_{0i}|W_i = 1).$$

- the *average treatment effect for the controls* (ATC):

$$\mathrm{ATC} = \mathrm{E}(Y_{1i}|W_i = 0) - \mathrm{E}(Y_{0i}|W_i = 0).$$

Corresponding estimates:

$$\widehat{\mathrm{ATT}} = \mathrm{E}\{[\mathrm{E}(Y_i|\mathbf{X}_i, W_i = 1) - \mathrm{E}(Y_i|\mathbf{X}_i, W_i = 0)]|W_i = 1\}.$$

and

$$\widehat{\mathrm{ATC}} = \mathrm{E}\{[\mathrm{E}(Y_i|\mathbf{X}_i, W_i = 1) - \mathrm{E}(Y_i|\mathbf{X}_i, W_i = 0)]|W_i = 0\}.$$

Note that in both cases **the expectation of the whole term is conditioned on** $W_i$.

# Confounding

Confounding occurs when one or more observed or unobserved factors $\mathbf{X}$ affect the causal relationship between $W$ and $Y$.

Formally, confounding requires that:

- $\text{Cov}(\mathbf{X}, W) \neq 0$ (the confounder is associated with the "treatment")
- $\text{Cov}(\mathbf{X}, Y) \neq 0$ (the confounder is associated with the outcome)
- $\mathbf{X}$ does not "lie on the path" between $W$ and $Z$ (that is, $\mathbf{X}$ is not affected by either $W$ or $Y$).

# DAGs

Directed acyclic graphs (DAGs) are a tool for visualizing and interpreting structural/causal phenomena.

- DAGs comprise:
  - Nodes (typically, variables / phenomena) and
  - Edges (or lines; typically, relationships/causal paths).
- Directed means each edge is *unidirectional*.
- Acyclical means exactly what it suggests: If a graph has a "feedback loop," it is not a DAG.
- Read more at the Wikipedia page, or at this useful page.

# Know your DAG

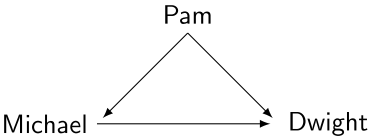



Figure: A DAG

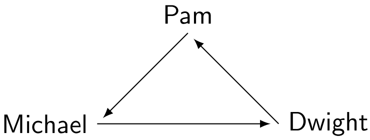



Figure: Not a DAG

# DAGs and Confounding


Figure: No Confounding

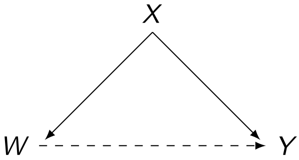



Figure: Confounding

# Confounding Bias: Some Toy Examples

Example One: $\text{Cov}(W, Y) = 0$ (ATE=2)

| $i$ | $W_i$ | $Y_{0i}$ | $Y_{1i}$ | $Y_{1i} - Y_{01}$ | $Y_i$ | $(\bar{Y}\|W = 1) - (\bar{Y}\|W = 0)$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 8 | (10) | (2) | 8 | - |
| 2 | 0 | 10 | (12) | (2) | 10 | - |
| 3 | 0 | 12 | (14) | (2) | 12 | - |
| 4 | 1 | (8) | 10 | (2) | 10 | - |
| 5 | 1 | (10) | 12 | (2) | 12 | - |
| 6 | 1 | (12) | 14 | (2) | 14 | - |
| $\text{Mean}_{\text{obs}}$ | - | 10 | 12 | - | 11 | 2 |
| $\text{Mean}_{\text{all}}$ | - | (10) | (12) | (2) | - | - |

$t = -1.22, \; p = 0.14$

# Confounding Bias: Some Toy Examples

Example Two: Cov$(W, Y) > 0$ (ATE=2)

| $i$ | $W_i$ | $Y_{0i}$ | $Y_{1i}$ | $Y_{1i} - Y_{01}$ | $Y_i$ | $(\bar{Y}\|W = 1) - (\bar{Y}\|W = 0)$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 8 | (10) | (2) | 8 | - |
| 2 | 0 | 8 | (10) | (2) | 8 | - |
| 3 | 0 | 10 | (12) | (2) | 10 | - |
| 4 | 1 | (10) | 12 | (2) | 12 | - |
| 5 | 1 | (12) | 14 | (2) | 14 | - |
| 6 | 1 | (12) | 14 | (2) | 14 | - |
| $\text{Mean}_{\text{obs}}$ | - | 8.67 | 13.33 | - | 11 | 4.67 |
| $\text{Mean}_{\text{all}}$ | - | (10) | (12) | (2) | - | - |

$t = -4.95,\ p < 0.001$

# Confounding Bias: Some Toy Examples

Example Three: $\text{Cov}(W, Y) < 0$ (ATE=2)

| $i$ | $W_i$ | $Y_{0i}$ | $Y_{1i}$ | $Y_{1i} - Y_{01}$ | $Y_i$ | $(\bar{Y}\|W = 1) - (\bar{Y}\|W = 0)$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 12 | (14) | (2) | 12 | - |
| 2 | 0 | 12 | (14) | (2) | 12 | - |
| 3 | 0 | 10 | (12) | (2) | 10 | - |
| 4 | 1 | (10) | 12 | (2) | 12 | - |
| 5 | 1 | (8) | 10 | (2) | 10 | - |
| 6 | 1 | (8) | 10 | (2) | 10 | - |
| $\text{Mean}_{\text{obs}}$ | - | 11.33 | 10.67 | - | 11 | -0.67 |
| $\text{Mean}_{\text{all}}$ | - | (10) | (12) | (2) | - | - |

$t = 0.71, \ p = 0.74$

# Confounding Illustrated

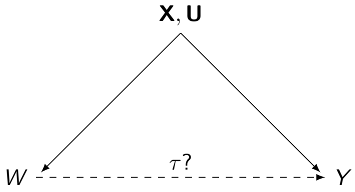



Figure: Potential Confounding

Here:

- $Y$ is the outcome of interest,
- $W$ is the primary predictor / covariate ("treatment") of interest,
- $T_i$ is the "treatment indicator" for observation $i$, where $T_i \equiv W_i$ (most of the time...)
- We're interested in estimating $\tau$, the "treatment effect" of $W$ on $Y$,
- **X** are observed confounders,
- **U** are unobserved confounders.

# Things We Can Do

- **Randomize**

  (or...)

- Instrumental Variables (various)
- Selection on Observables:
  - Regression / Weighting
  - Matching (propensity scores, multivariate/minimum-distance, genetic, etc.)
- Regression Discontinuity Designs ("RDD")
- Differences-In-Differences ("DiD")*
- Synthetic Controls*
- Others...

* We discuss these approaches a bit more in the *Analyzing Panel Data* course.

# Under Randomization

$$\begin{tikzcd} & \mathbf{U} \arrow[d] & \\ W \arrow[r] & Y & \mathbf{X} \arrow[l] \end{tikzcd}$$

Figure: = no confounding!

**Note:**

- Randomized assignment of $W$ "balances" covariate values – both observed and unobserved – *on average*...
- That is, under randomization of $W$:

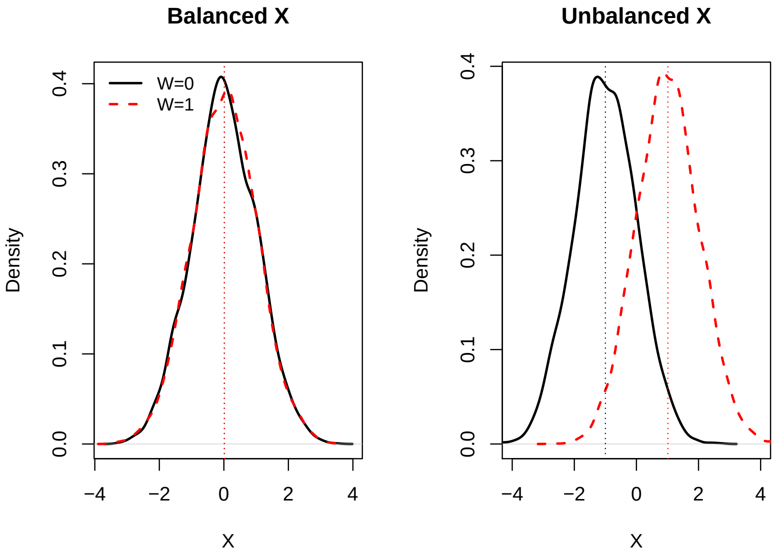
$$\mathrm{E}(\mathbf{X}_i, \mathbf{U}_i \mid W_i = 0) = \mathrm{E}(\mathbf{X}_i, \mathbf{U}_i \mid W_i = 1)$$

  or, more demandingly,

$$\mathrm{E}[f(\mathbf{X}, \mathbf{U}) \mid W_i = 0] = \mathrm{E}[f(\mathbf{X}, \mathbf{U}) \mid W_i = 1]$$

- <u>Can</u> yield imbalance by random chance...

# Covariate Balance / Imbalance



# Covariate Imbalance Under Randomization

Why seek balance when randomizing?

- More accurate estimates of treatment effects
- Higher statistical power

Possible Approaches:

1. Force balance by design:
   - Stratification / blocking
   - Matching / paired randomization (see below)
   - Rerandomization approaches (e.g., Morgan and Rubin 2012)
2. Post-randomization analysis:
   - Pre- vs. post-treatment $Y$ values / "gain scores"
   - (Post-treatment) stratification by $\mathbf{X}$
   - (Pre-treatment) covariate adjustment via weighting / regression

# Nonrandom Assignment of $W_i$

Valid causal inference requires $Y_{0i}, Y_{1i} \perp W_i | \mathbf{X}_i, \mathbf{U}_i$

- That is, treatment assignment $W_i$ is *conditionally ignorable*

**"What if I have unmeasured confounders?"**

- In general, that's a bad thing.
- One approach: obtain *bounds* on possible values of $\tau$
  - Assume you have one or more unmeasured confounders
  - Undertake one of the methods described below to get $\hat{\tau}$
  - Calculate the range of values for $\hat{\tau}$ that could occur, depending on the degree and direction of confounding bias
  - Or ask: How strong would the effect of the $\mathbf{U}$s have to be to make $\hat{\tau} \rightarrow 0$?
- Some useful cites:
  - Rosenbaum and Rubin (1983)
  - Rosenbaum (2002)
  - DiPrete and Gangl (2004)
  - Liu et al. (2013)
  - Ding and VanderWeele (2016)

# Digression: Instrumental Variables

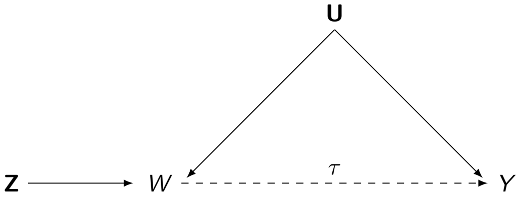



Figure: Instrumental Variables

As in the more general regression case where we have $\text{Cov}(\mathbf{X}, \epsilon) \neq 0$, instrumental variables can be used to address confounding in causal analyses.

# Instrumental Variables (continued)

Considerations:

- Requires:
    1. $\text{Cov}(\mathbf{Z}, W) \neq 0$)
    2. $\mathbf{Z}$ has no independent effect on $Y$, except through $W$
    3. $\mathbf{Z}$ is exogenous [i.e., $\text{Cov}(\mathbf{Z}, \epsilon) = 0$]
- Arguably most useful when treatment compliance is uncertain / driven by unmeasured factors ("intent to treat" analyses)
- Mostly, they're not that useful at all...
    - Bound et al. (1995): Weak instruments are worse than endogeneity bias
    - Young (2020): Inferences in published IV work (in economics) are wrong and terrible
    - Shalizi (2020, chapters 20-21): Gathers all the issues together, sometimes hilariously
- Other useful references:
    - Imbens et al. (1996) (the overly-cited one)
    - Hernan and Robins (2006) (making sense of things)
    - Lousdal (2018) (a good intuitive introduction)

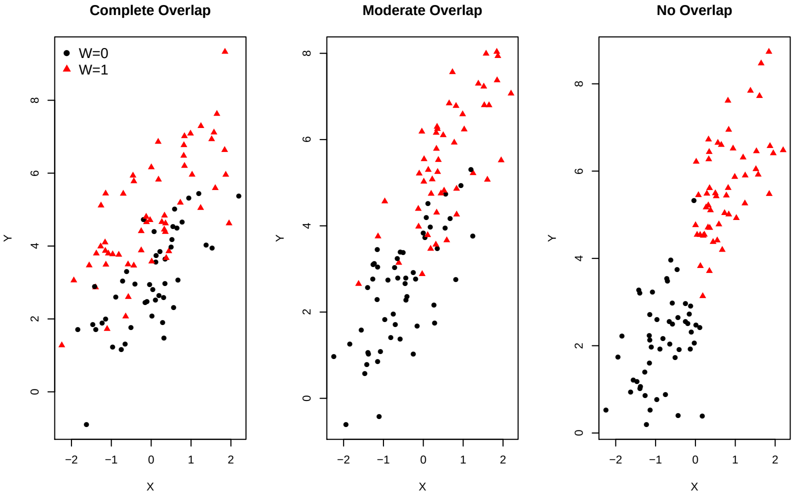
# Nonrandom Assignment of $W_i$ (continued)

So...

- Causal inference with observational data typically requires that $\mathbf{U} = \varnothing$...
- This typically requires a strong theoretical motivation in order to assume that the observed $\mathbf{X}$ exhausts the list of possible confounders.

- **Even if** this assumption is reasonable, there are two (related) important concerns:
  - Lack of *covariate balance* (as above)
  - Lack of *overlap* among observations with $W_i = 0$ vs. $W_i = 1$
  - The latter is related to *positivity*, the requirement that each observation's probability of receiving (or not receiving) the treatment is greater than zero

# Overlap



**In general**:

- Ensuring overlap allows us to make counterfactual statements from observational data
  - Requires that we have comparable $W_i = 0$ and $W_i = 1$ units
  - It's *necessary* – no overlap means any counterfactual statements are based on assumption
  - Think of this as an aspect of *model identification* (Crump et al. 2009)
  - Most often handled via matching

- Ensuring covariate balance corrects potential bias in $\hat{\tau}$ due to (observed) confounding
  - This can be done a number of different ways: stratification, weighting, regression...
  - Key: Adjusting for (observable) differences across groups defined by values of $W$
- In general, we usually address overlap first, then balance...

# Matching

Matching is a way of dealing with one of both of covariate overlap and (im)balance.

The process, generally:

1. Choose the $\mathbf{X}$ on which the observations will be matched, and the matching procedure;
2. Match the observations with $W_i = 0$ and $W_i = 1$;
3. Check for balance in $\mathbf{X}_i$; and
4. Estimate $\hat{\tau}$ using the matched pairs.

Variants / considerations:

- 1:1 vs. 1:$k$ matching
- "Greedy" vs. "Optimal" matching (see Gu and Rosenbaum 1993)
- Distances, calipers, and "common support"
- Post-matching: Balance checking...

# Flavors of Matching

- Simplest: Exact Matching
  - For each of the $n$ observations $i$ with $W = 1$, find a corresponding observation $j$ with $W = 0$ that has identical values of $\mathbf{X}$
  - Calculate $\hat{\tau} = \frac{1}{n}\sum(Y_i - Y_j)$
  - Generally not practical, especially for high-dimensional $\mathbf{X}$
  - Variants: "coarsened" exact matching (e.g., Iacus et al. 2011)

- Multivariate Matching
  - Match each observation $i$ which has $W = 1$ with a corresponding observation $j$ with $W = 0$, and whose values on $\mathbf{X}_j$ are the most similar to $\mathbf{X}_i$
  - One example: Mahalanobis distance matching, based on the distance:

$$d_M(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{S}^{-1}(\mathbf{X}_i - \mathbf{X}_j)}.$$

# Flavors of Matching (continued)

- Propensity Score Matching
  - Match observation $i$ which has $W = 1$ with observation $j$ having $W = 0$ based on the closeness of their *propensity score*
  - The propensity score is, $\Pr(W_i = 1|\mathbf{X}_i)$, typically calculated as the predicted value of $T_i$ (the treatment indicator) from a logistic (or other) regression of $T$ on $\mathbf{X}$.
  - The assumptions about matching [that $Y$ is orthogonal to $W|\mathbf{X}$ and that $\Pr(W_i = 1|\mathbf{X}_i) \in (0, 1)$] mean that $Y \perp W| \Pr(T|\mathbf{X})$.
  - In practice: read this...

- Other variants [e.g., genetic matching (Diamond and Sekhon 2013), etc.]

# Matching Software

In R :

- the `Match` package (does propensity score, *M*-distance, and genetic matching, plus balance checking and other diagnostics)

- the `MatchIt` package (for pre-analysis matching; also has nice options for checking balance)

- the `optmatch` package (suite for 1:1 and 1:$k$ matching via propensity scores, *M*-distance, and optimum balancing)

- `matching` (in the `arm` package)

"RDD":

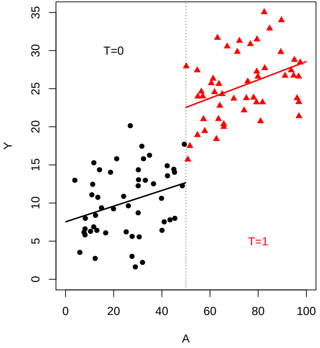
- Treatment changes abruptly [usually at some threshold(s)] according to the value(s) of some measured, continuous, pre-treatment variable(s)
  - This is known as the "assignment" or "forcing variable(s)," sometimes denoted **A**
  - Formally:

$$T_i = \begin{cases} 0 \text{ if } A_i \leq c \\ 1 \text{ if } A_i > c \end{cases}$$

- Intuition: Observations near but on either side of the threshold(s) are highly comparable, and can be used to (locally) identify $\tau$
- This is because variation in $T_i$ near the threshold is effectively random (a "local randomized experiment")
- E.g. Carpenter and Dobkin (2011) (on the relationship between the legal drinking age and public health outcomes like accidental deaths)

# RDD (continued)

- Pluses:
  - Can be estimated straightforwardly, as: $$Y_i = \beta_0 + \beta_1 A_i + \tau T_i + \gamma A_i T_i + \epsilon_i$$
  - Generally requires fewer assumptions than IV or DiD (and those assumptions are easier to observe and test)
- Minuses:
  - Provides only an estimate of a local treatment effect
  - Fails if (say) subjects can manipulate $A$ in the vicinity of $c$
- Lee and Lemieux (2010) is an excellent (if fanboi-ish) review
- R packages: `rddtools`, `rdd`, `rdrobust`, `rdpower`, `rdmulti`



# Software Matters

- R
  - Packages for matching are listed above (`Matching`, `MatchIt`, etc.)
  - Similarly for RDD (`rddtools`, `rdd`, etc.)
  - IV regression: `ivreg` (in `AER`), `tsls` (in `sem`), others
  - See generally the *Econometrics* and *SocialSciences* CRAN Task Views
- Stata also has a large suite of routines for attempting causal inference with observational data
- And there's a pretty good NumPy/SciPy-dependent package for Python, called (creatively) Causalinference

# Example: Sports and Grades in High School

**Question: Does participation in high school varsity sports help or hinder academic achievement (i.e., grades)?**

Data: "High School And Beyond" survey (1983 wave) ($N = 1375$)

Variables:

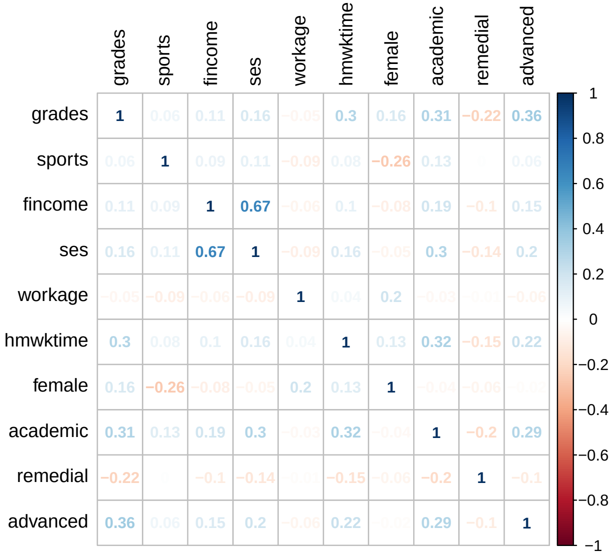
- `grades`: As=4, As & Bs=3.5, etc.
- `sports`: 1 if participated in varsity sports, 0 otherwise
- `fincome`: Family income (7-point scale)
- `ses`: Socioeconomic Status: 1=low, 2-middle, 3=high
- `workage`: Age at which started working
- `hmwktime`: Time spent on homework (7-point scale)*
- `female`: 1 = female student, 0 = male student
- `academic`: 1 if the student is on an academic track, 0 else
- `remedial`: 1 if the student took $\geq$1 remedial course
- `advanced`: 1 if the student took $\geq$1 advanced course

* Likely post-treatment, so we'll omit in the examples below.

# Summary Statistics

```
> summary(sports)
     grades           sports          fincome          ses       
 Min.   :0.0   Min.   :0.00   Min.   :1.0   Min.   :1.00  
 1st Qu.:2.5   1st Qu.:0.00   1st Qu.:3.0   1st Qu.:1.00  
 Median :3.0   Median :0.00   Median :5.0   Median :2.00  
 Mean   :2.9   Mean   :0.37   Mean   :4.4   Mean   :1.96  
 3rd Qu.:3.5   3rd Qu.:1.00   3rd Qu.:6.0   3rd Qu.:2.00  
 Max.   :4.0   Max.   :1.00   Max.   :7.0   Max.   :3.00  
    workage        hmwktime        female        academic   
 Min.   :11.0   Min.   :1.0   Min.   :0.00   Min.   :0.00  
 1st Qu.:13.0   1st Qu.:4.0   1st Qu.:0.00   1st Qu.:0.00  
 Median :15.0   Median :4.0   Median :1.00   Median :0.00  
 Mean   :14.6   Mean   :4.5   Mean   :0.52   Mean   :0.41  
 3rd Qu.:16.0   3rd Qu.:6.0   3rd Qu.:1.00   3rd Qu.:1.00  
 Max.   :21.0   Max.   :7.0   Max.   :1.00   Max.   :1.00  
    remedial        advanced   
 Min.   :0.00   Min.   :0.00  
 1st Qu.:0.00   1st Qu.:0.00  
 Median :0.00   Median :0.00  
 Mean   :0.36   Mean   :0.37  
 3rd Qu.:1.00   3rd Qu.:1.00  
 Max.   :1.00   Max.   :1.00  
```

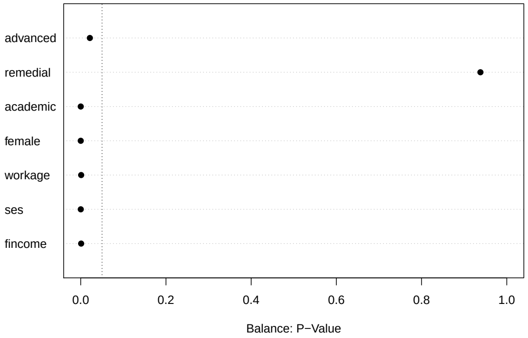
# Correlation Plot



# Simple $t$-test & Regression

```
> with(sports, t.test(grades~sports))

Welch Two Sample t-test

data:  grades by sports
t = -2, df = 1064, p-value = 0.02
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.183 -0.014
sample estimates:
mean in group 0 mean in group 1
            2.9             3.0


> summary(lm(Model,data=sports))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.71145    0.13397   20.24  < 2e-16 ***
sports       0.10119    0.03969    2.55    0.011 *
fincome      0.00435    0.01378    0.32    0.753
ses          0.02216    0.03487    0.64    0.525
workage     -0.01879    0.00794   -2.37    0.018 *
female       0.30062    0.03881    7.75  1.8e-14 ***
academic     0.29063    0.04099    7.09  2.1e-12 ***
remedial    -0.23215    0.03919   -5.92  4.0e-09 ***
advanced     0.44435    0.04004   11.10  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.68 on 1366 degrees of freedom
Multiple R-squared:  0.231,Adjusted R-squared:  0.226
F-statistic: 51.2 on 8 and 1366 DF,  p-value: <2e-16
```

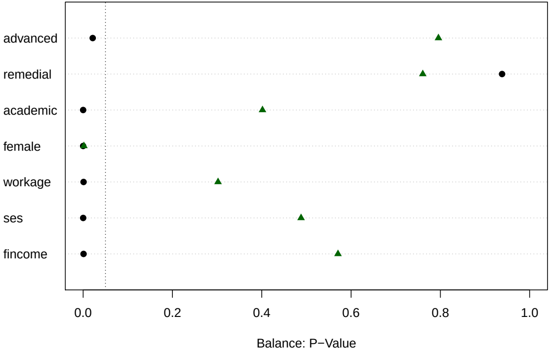
# Balance Tests (Pre-Matching)




These are $P$-values associated with $t$-tests (for binary predictors) or Kolmogorov-Smirnov tests (for continuous predictors) for balance between $\texttt{sports} = 0$ and $\texttt{sports} = 1$.

```
> M.exact <- matchit(sports~fincome+ses+workage+female+academic+
+                    remedial+advanced,data=sports,method="exact")
> M.exact

Call:
matchit(formula = sports ~ fincome + ses + workage + female +
    academic + remedial + advanced, data = sports, method = "exact")

Exact Subclasses: 166

Sample sizes:
          Control Treated
All           864     511
Matched       287     239
Unmatched     577     272

> # Output matched data:

> sports.exact <- match.data(M.exact,group="all")

> dim(sports.exact)
[1] 526  12
```

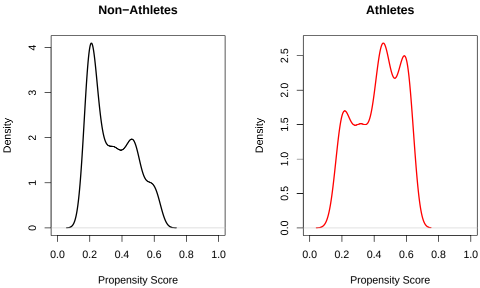
# Exact Matching: Balance




These are $P$-values associated with $t$-tests (for binary predictors) or Kolmogorov-Smirnov tests (for continuous predictors) for balance between $\texttt{sports} = 0$ and $\texttt{sports} = 1$. Black dots are pre-matching; green triangles are after exact matching.

# Propensity Score Matching

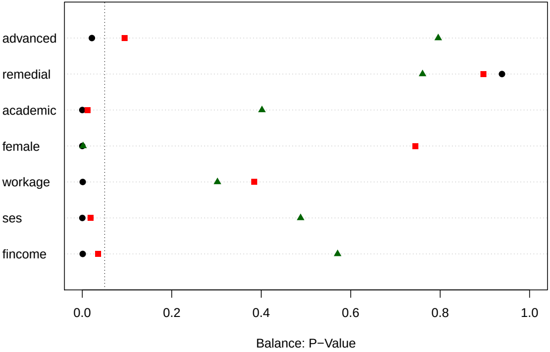
```
> PSfit <- glm(sports~fincome+ses+workage+female+academic+remedial+
+                  advanced,data=sports,family=binomial(link="logit"))

> # Generate scores & check common support:

> PS.df <- data.frame(PS = predict(PSfit,type="response"),
+                       sports = PSfit$model$sports)
```



# Propensity Score Matching

```
> M.prop<-matchit(sports~fincome+ses+workage+female+academic+
+                      remedial+advanced,data=sports,
+                      method="nearest")
> summary(M.prop)
.
.
.
Percent Balance Improvement:
         Mean Diff. eQQ Med eQQ Mean eQQ Max
distance         80      83       80      63
fincome          29       0       30       0
ses              34       0       35       0
workage          71       0       68      25
female           96       0       96       0
academic         41       0       41       0
remedial        -88       0     -100       0
advanced         19       0       19       0

Sample sizes:
          Control Treated
All           864     511
Matched       511     511
Unmatched     353       0
Discarded       0       0
```

# Propensity Score Matching: Balance




These are $P$-values associated with $t$-tests (for binary predictors) or Kolmogorov-Smirnov tests (for continuous predictors) for balance between $\texttt{sports} = 0$ and $\texttt{sports} = 1$. Black dots are pre-matching; green triangles are after exact matching; red squares are after propensity score matching.

```
> with(sports, t.test(grades~sports))$statistic # No matching
     t
-2.286


> with(sports.exact, t.test(grades~sports))$statistic # Exact
     t
-1.395


> with(sports.prop, t.test(grades~sports,paired=TRUE))$statistic # PS
    t
-2.98


> with(sports.genetic, t.test(grades~sports))$statistic # Genetic
     t
-1.367
```

# Regression Results

| | No Matching | Exact | Propensity Score | Genetic |
|---|---|---|---|---|
| (Intercept) | $2.71^*$ | $3.05^*$ | $2.86^*$ | $2.78^*$ |
| | (0.13) | (0.23) | (0.16) | (0.17) |
| Plays Sports | $0.10^*$ | $0.12^*$ | $0.10^*$ | $0.09^*$ |
| | (0.04) | (0.06) | (0.04) | (0.05) |
| Family Income | 0.00 | 0.05 | 0.00 | 0.01 |
| | (0.01) | (0.03) | (0.02) | (0.02) |
| Socioeconomic Status | 0.02 | $-0.14$ | 0.04 | 0.02 |
| | (0.03) | (0.07) | (0.04) | (0.05) |
| Working Age | $-0.02^*$ | $-0.03^*$ | $-0.03^*$ | $-0.02^*$ |
| | (0.01) | (0.01) | (0.01) | (0.01) |
| Female | $0.30^*$ | $0.34^*$ | $0.30^*$ | $0.31^*$ |
| | (0.04) | (0.06) | (0.05) | (0.05) |
| Academic Track | $0.29^*$ | $0.24^*$ | $0.32^*$ | $0.32^*$ |
| | (0.04) | (0.08) | (0.05) | (0.05) |
| Remedial Track | $-0.23^*$ | $-0.28^*$ | $-0.29^*$ | $-0.20^*$ |
| | (0.04) | (0.06) | (0.05) | (0.05) |
| Advanced Course | $0.44^*$ | $0.51^*$ | $0.44^*$ | $0.39^*$ |
| | (0.04) | (0.08) | (0.05) | (0.05) |
| $R^2$ | 0.23 | 0.29 | 0.26 | 0.22 |
| Adj. $R^2$ | 0.23 | 0.28 | 0.26 | 0.22 |
| Num. obs. | 1375 | 526 | 1022 | 939 |

$^*p < 0.05$

Table:

# Some Questions...

- What – if anything – can the general robustness of our results tell us about the relationship between varsity athletics and grades?
- What can they tell us about our model?
- What mechanism(s) / circumstances might allow us to investigate the relationship between varsity athletic participation and grades using an RDD?
- What circumstances – if any – might allow us to investigate this relationship using instrumental variables?
- What sort(s) of experiments – natural or otherwise – might allow us to investigate this same relationship?

# Resources

- Good references:
  - Freedman (2012)*
  - Morgan and Winship (2014)
  - Pearl et al. (2016)
  - Peters et al. (2017)
- Courses / syllabi (a sampling):
  - Frey (2019)
  - Imai (2019)
  - Sekhon (2015)
  - Simpson (2019)
  - Xu (2018)
  - Yamamoto (2018)
- Other useful things:
  - The Causal Inference Book
  - Some useful notes

* I really like this one.

# Models for Panel / Time-Series Cross-Sectional Data

# Terminology + Data Structure

- "Longitudinal" $\neq$ "Time Series"
- Terminology:
  - "Unit" / "Units" / "Units of observation" / "Panels" = Things we observe repeatedly
  - "Observations" = Each (one) measurement of a unit
  - "Time points" = When each observation on a unit is made
  - $i \in \{1....N\}$ indexes units
  - $t \in \{1...T\}$ or $\{1...T_i\}$ indexes observations / time points
  - If $T_i = T \; \forall i$ then we have "balanced" panels / units
  - $NT$ = Total number of observations (if balanced)

- $N >> T \;\rightarrow$ "panel" data

- $T >> N$ or $T \approx N \;\rightarrow$ "time-series cross-sectional" ("TSCS") data

# Within- and Between-Unit Variation

Define:

$$\bar{Y}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} Y_{it}$$

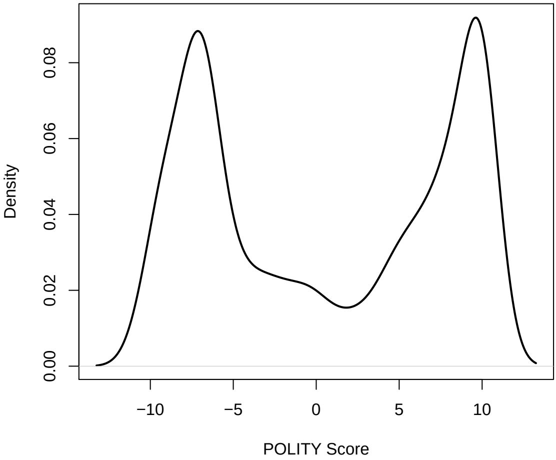
Then:

$$Y_{it} = \bar{Y}_i + (Y_{it} - \bar{Y}_i).$$

- The <u>total</u> variation in $Y_{it}$ can be decomposed into
- The <u>between-unit</u> variation in the $\bar{Y}_i$s, and
- The <u>within-unit</u> variation around $\bar{Y}_i$ (that is, $Y_{it} - \bar{Y}_i$).

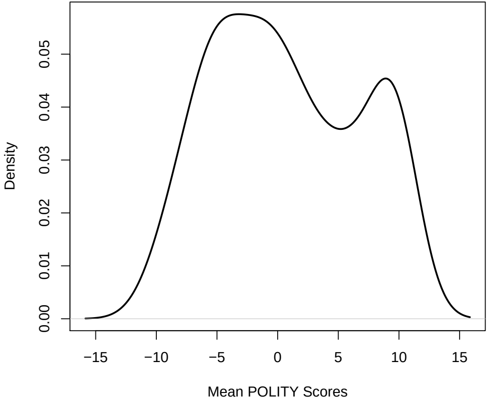
# POLITY: Total Variation

```
> with(Demos, describe(POLITY)) # all variation
   vars    n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 9260 0.63 7.47      0    0.71 10.4 -10  10    20    0    -1.68 0.08
```



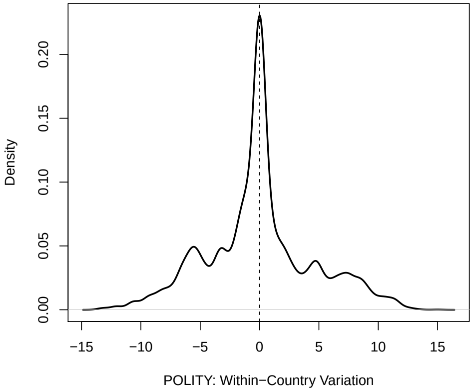
# POLITY: "Between" Variation

```
> POLITYmeans <- ddply(Demos,.(ccode),summarise,
+                  POLITYmean = mean(POLITY))
>
> with(POLITYmeans, describe(POLITYmean)) # "between" variation
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 162 0.73 5.99   -0.1     0.7 7.36 -10  10    20 0.16    -1.21 0.47
```



# POLITY: "Within" Variation

```
> Demos <- ddply(Demos, .(ccode), mutate,
+                 POLITYmean = mean(POLITY))
> Demos$POLITYwithin <- with(Demos, POLITY-POLITYmean)
>
> with(Demos, describe(POLITYwithin))
   vars    n mean   sd median trimmed  mad   min max range skew kurtosis   se
X1    1 8404    0 4.48      0   -0.11 3.16 -13.5  15  28.5 0.19     0.27 0.05
```



# One- and Two-Way Unit Effects

Two-way variation:

$$Y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \gamma V_i + \delta W_t + u_{it}$$

$\longrightarrow$ two-way effects:

$$Y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \alpha_i + \eta_t + u_{it}$$

One-way effects:

$$Y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \eta_t + u_{it} \quad \text{(time)}$$

$$Y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \alpha_i + u_{it} \quad \text{(units)}$$

# "Fixed" Effects

"Brute force" model:

$$\begin{aligned} Y_{it} &= \mathbf{X}_{it}\boldsymbol{\beta} + \alpha_i + u_{it} \\ &= \mathbf{X}_{it}\boldsymbol{\beta} + \alpha_1 I(i=1)_i + \alpha_2 I(i=2)_i + ... + u_{it} \end{aligned}$$

Alternatively:

$$\bar{X}_i = \frac{\sum_{N_i} X_{it}}{N_i}$$

and

$$\tilde{X}_{it} = X_{it} - \bar{X}_i.$$

Yields:

$$Y_{it} = \bar{\mathbf{X}}_i\boldsymbol{\beta}_B + (\mathbf{X}_{it} - \bar{X}_i)\boldsymbol{\beta}_W + \alpha_i + u_{it}$$

So, "fixed effects" are:

$$\begin{aligned} Y_{it}^{*} &= Y_{it} - \bar{Y}_i \\ \mathbf{X}_{it}^{*} &= \mathbf{X}_{it} - \bar{\mathbf{X}}_i \end{aligned}$$

$$Y_{it}^{*} = \boldsymbol{\beta}_{FE} \mathbf{X}_{it}^{*} + u_{it}.$$

$\equiv$ a "Within-Effects" Model.

Additionally, a standard $F$-test for

$$H_0 : \alpha_i = \alpha_j \forall i \neq j$$

versus

$$H_A : \alpha_i \neq \alpha_j \text{ for some } i \neq j$$

is $\sim F_{N-1, NT-(N-1)}$.

# An Example: Demonstrations, 1945-2014

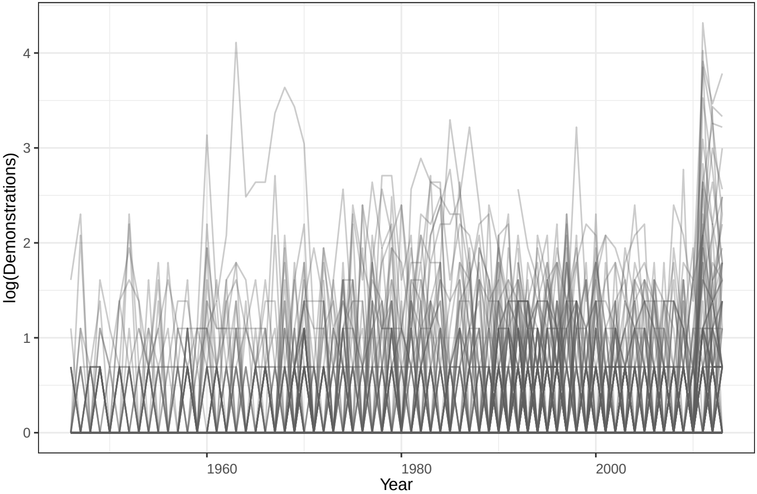
Data:

- 180 countries
- 70 years
- $i$ indexes countries, $t$ indexes years

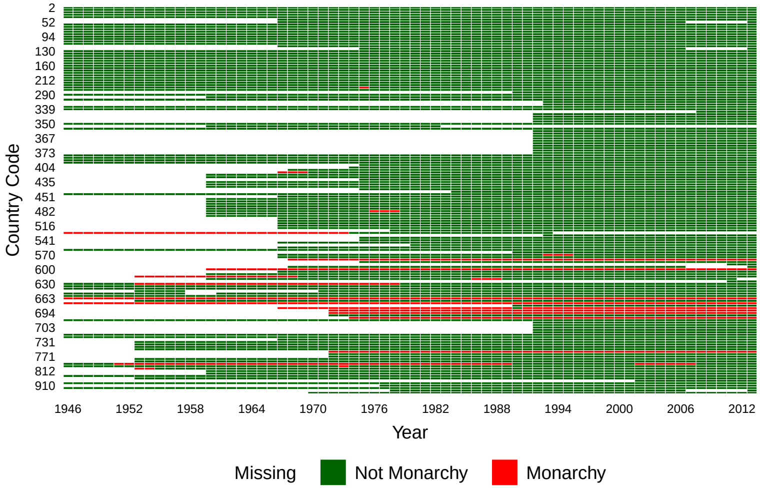
Model:

$$
\begin{aligned}
ln(\text{Demonstrations} + 1)_{it} \quad = \quad & \beta_0 + \beta_1 \text{POLITY}_{it} + \beta_2 \text{POLITY}^2_{it} + \beta_3 \ln(\text{GDP})_{it} + \\
& \beta_4 \text{Monarch}_{it} + \beta_5 \text{Cold War}_{it} + u_{it}
\end{aligned}
$$

# Visualizing Panel Data: Continuous $X$



# Visualizing Panel Data: Discrete $X$




(Created using `panelView`.)

```
> OLS<-lm(lnDemons~POLITY+I(POLITY^2)+lnGDP+Monarch+ColdWar,
+           data=PDF)
>
> summary(OLS)

Call:
lm(formula = lnDemons ~ POLITY + I(POLITY^2) + lnGDP + Monarch +
    ColdWar, data = PDF)

Residuals:
   Min     1Q Median     3Q    Max
-0.450 -0.293 -0.218 -0.075  4.107

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.124639   0.058208   -2.14    0.032 *
POLITY       0.006296   0.001179    5.34  9.5e-08 ***
I(POLITY^2) -0.002267   0.000255   -8.90  < 2e-16 ***
lnGDP        0.057679   0.007513    7.68  1.9e-14 ***
Monarch     -0.046393   0.028572   -1.62    0.104
ColdWar      0.027883   0.013961    2.00    0.046 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.526 on 6499 degrees of freedom
  (2863 observations deleted due to missingness)
Multiple R-squared:  0.0261,Adjusted R-squared:  0.0253
F-statistic: 34.8 on 5 and 6499 DF,  p-value: <2e-16
```

# "Fixed" (Within) Effects

```
> FE<-plm(lnDemons~POLITY+I(POLITY^2)+lnGDP+Monarch+ColdWar,
+            data=PDF, effect="individual",model="within")
>
> summary(FE)
Oneway (individual) effect Within Model

Call:
plm(formula = lnDemons ~ POLITY + I(POLITY^2) + lnGDP + Monarch +
    ColdWar, data = PDF, effect = "individual", model = "within")

Unbalanced Panel: n = 145, T = 1-62, N = 6505

Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
-1.3556 -0.2120 -0.0768  0.0193  4.0496

Coefficients:
             Estimate Std. Error t-value Pr(>|t|)
POLITY       0.001526   0.001553    0.98  0.32604
I(POLITY^2) -0.001942   0.000296   -6.55  6.1e-11 ***
lnGDP        0.054586   0.015200    3.59  0.00033 ***
Monarch      0.047976   0.068071    0.70  0.48097
ColdWar     -0.035487   0.016235   -2.19  0.02887 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Total Sum of Squares:    1410
Residual Sum of Squares: 1400
R-Squared:      0.013
Adj. R-Squared: -0.0102
F-statistic: 16.7177 on 5 and 6355 DF, p-value: <2e-16
```

# A Nicer Table

Table: Models of Demonstrations

| | OLS | FE |
|---|---|---|
| POLITY | 0.006*** | 0.002 |
| | (0.001) | (0.002) |
| POLITY Squared | −0.002*** | −0.002*** |
| | (0.0003) | (0.0003) |
| ln(GDP) | 0.058*** | 0.055*** |
| | (0.008) | (0.015) |
| Monarch | −0.046 | 0.048 |
| | (0.029) | (0.068) |
| Cold War | 0.028** | −0.035** |
| | (0.014) | (0.016) |
| Constant | −0.125** | |
| | (0.058) | |
| Observations | 6,505 | 6,505 |
| $R^2$ | 0.026 | 0.013 |
| Adjusted $R^2$ | 0.025 | -0.010 |
| Residual Std. Error | 0.526 (df = 6499) | |
| F Statistic | 34.820*** (df = 5; 6499) | 16.720*** (df = 5; 6355) |

*p<0.1; **p<0.05; ***p<0.01

# Issues (?) with "Fixed" Effects

Pros:

- Specification Bias
- Intuitive
- Widely Used/Understood

Cons:

- Can't Estimate $\boldsymbol{\beta}_B$
- Slowly-Changing **X**s
- (In)Efficiency / Inconsistency (Incidental Parameters)

# “Between” Effects

From:

$$Y_{it} = \bar{\mathbf{X}}_i \boldsymbol{\beta}_B + (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)\beta_W + \alpha_i + u_{it}.$$

“Between” effects:

$$\bar{Y}_i = \bar{\mathbf{X}}_i \boldsymbol{\beta}_B + u_{it}$$

- Essentially cross-sectional
- Based on $N$ observations

# A Nicer Table (Again)

Table: Models of Demonstrations

| | OLS | FE | BE |
|---|---|---|---|
| POLITY | $0.006^{***}$ | 0.002 | 0.006 |
| | (0.001) | (0.002) | (0.005) |
| POLITY Squared | $-0.002^{***}$ | $-0.002^{***}$ | $-0.003^{***}$ |
| | (0.0003) | (0.0003) | (0.001) |
| ln(GDP) | $0.058^{***}$ | $0.055^{***}$ | $0.069^{**}$ |
| | (0.008) | (0.015) | (0.027) |
| Monarch | $-0.046$ | 0.048 | $-0.050$ |
| | (0.029) | (0.068) | (0.103) |
| Cold War | $0.028^{**}$ | $-0.035^{**}$ | $0.259^{***}$ |
| | (0.014) | (0.016) | (0.085) |
| Constant | $-0.125^{**}$ | | $-0.306$ |
| | (0.058) | | (0.208) |
| Observations | 6,505 | 6,505 | 145 |
| $R^2$ | 0.026 | 0.013 | 0.127 |
| Adjusted $R^2$ | 0.025 | -0.010 | 0.096 |
| Residual Std. Error | 0.526 (df = 6499) | | |
| F Statistic | $34.800^{***}$ (df = 5; 6499) | $16.700^{***}$ (df = 5; 6355) | $4.060^{***}$ (df = 5; 139) |

$^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

# "Random" Effects

Model:

$$Y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + u_{it}$$

with:

$$u_{it} = \alpha_i + \lambda_t + \eta_{it}$$

and

$$\begin{aligned}
E(\alpha_i) = E(\lambda_t) = E(\eta_{it}) &= 0, \\
E(\alpha_i\lambda_t) = E(\alpha_i\eta_{it}) = E(\lambda_t\eta_{it}) &= 0, \\
E(\alpha_i\alpha_j) &= \sigma_\alpha^2 \text{ if } i = j, \ 0 \text{ otherwise}, \\
E(\lambda_t\lambda_s) &= \sigma_\lambda^2 \text{ if } t = s, \ 0 \text{ otherwise}, \\
E(\eta_{it}\eta_{js}) &= \sigma_\eta^2 \text{ if } i = j, \ t = s, \ 0 \text{ otherwise}, \\
E(\alpha_i\mathbf{X}_{it}) = E(\lambda_t\mathbf{X}_{it}) = E(\eta_{it}\mathbf{X}_{it}) &= 0.
\end{aligned}$$

"Variance Components":

$$Var(Y_{it}|\mathbf{X}_{it}) = \sigma_\alpha^2 + \sigma_\lambda^2 + \sigma_\eta^2$$

If we assume $\lambda_t = 0$, then we get a model like:

$$Y_{it} = \mathbf{X}_{it}\beta + \alpha_i + \eta_{it}$$

with total error variance:

$$\sigma_u^2 = \sigma_\alpha^2 + \sigma_\eta^2.$$

# “Random” Effects, Conceptually

# A Nicer Table (Yet Again)

Table: Models of Demonstrations

| | OLS | FE | BE | RE |
|---|---|---|---|---|
| POLITY | 0.006*** | 0.002 | 0.006 | 0.003* |
| | (0.001) | (0.002) | (0.005) | (0.001) |
| POLITY Squared | −0.002*** | −0.002*** | −0.003*** | −0.002*** |
| | (0.0003) | (0.0003) | (0.001) | (0.0003) |
| ln(GDP) | 0.058*** | 0.055*** | 0.069** | 0.057*** |
| | (0.008) | (0.015) | (0.027) | (0.012) |
| Monarch | −0.046 | 0.048 | −0.050 | −0.007 |
| | (0.029) | (0.068) | (0.103) | (0.053) |
| Cold War | 0.028** | −0.035** | 0.259*** | −0.024 |
| | (0.014) | (0.016) | (0.085) | (0.015) |
| Constant | −0.125** | | −0.306 | −0.132 |
| | (0.058) | | (0.208) | (0.105) |
| Observations | 6,505 | 6,505 | 145 | 6,505 |
| $R^2$ | 0.026 | 0.013 | 0.127 | 0.012 |
| Adjusted $R^2$ | 0.025 | -0.010 | 0.096 | 0.012 |
| Residual Std. Error | 0.526 (df = 6499) | | | |
| F Statistic | 34.800*** (df = 5; 6499) | 16.700*** (df = 5; 6355) | 4.060*** (df = 5; 139) | 81.100*** |

*p<0.1; **p<0.05; ***p<0.01

# Practical “Fixed” vs. “Random” Effects

- Hausman tests (just don’t...)
- “Panel” vs. “TSCS” Data
- Data-Generating Process
- Covariate Effects

# Unit Effects Models: Software

R :

- The `plm` package; `plm` command
- The `fixest` package; `feols` command
- The `lme4` package; command is `lmer`
- The `nlme` package; command `lme`
- Various commands in the `Paneldata` package (dated)

Stata : `xtreg`

- the `re` (the default) = random effects
- the `fe` = fixed (within) effects
- the `be` = between-effects

One approach: Lagged $Y$

$$Y_{it} = \phi Y_{it-1} + \mathbf{X}_{it}\boldsymbol{\beta}_{LDV} + u_{it}$$

Here,

- If $u$s are uncorrelated, $\hat{\beta}_{LDV}$ is biased (but consistent),
- If $u$s are correlated, $\hat{\beta}_{LDV}$ is biased and inconsistent

Key: In LDV, long-run impact of a unit change in $X$ is:

$$\hat{\beta}_{LR} = \frac{\hat{\beta}_{LDV}}{1 - \hat{\phi}}$$

In addition: Lagged dependent variables + unit effects (generally) = inconsistency ("Nickell bias")...

# Panel Data: Dynamics (continued)

Time-series dynamics in panel data...

- Unit roots, integration / cointegration, error-correction models...
- Also: Models of first-differences (changes-on-changes)
- Short series + asymptotic tests → "borrow strength"
- Typically require uniform unit roots across cross-sectional units

Things to think about:

- $N$ vs. $T$...
- Are dynamics nuisance or substance?
- What problem(s) do you <u>really</u> care about?

# Panel Data GLMs: Binary Responses

One-way unit effects (logit):

$$\Pr(Y_{it} = 1) = \frac{\exp(\mathbf{X}_{it}\beta + \alpha_i)}{1 + \exp(\mathbf{X}_{it}\beta + \alpha_i)} \equiv \Lambda(\mathbf{X}_{it}\beta + \alpha_i)$$

<u>Incidental Parameters</u>

- Nonlinearity $\rightarrow$ inconsistency in both $\hat{\alpha}$s <u>and</u> $\hat{\boldsymbol{\beta}}$.
- Chamberlain:

$$L^C = \prod_{i=1}^{N} \Pr\left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, ...Y_{iT} = y_{iT} \mid \sum_{t=1}^{T} Y_{it}\right)$$

  - Fixed effects = no estimates for $\boldsymbol{\beta}_b$
  - Interpretation: per logit, but $\mid \hat{\alpha}_i$.

Model is:

$$
\begin{aligned}
Y_{it}^{*} &= \mathbf{X}_{it}\beta + u_{it} \\
Y_{it} &= 0 \text{ if } Y_{it}^{*} \leq 0\ ; \\
&= 1 \text{ if } Y_{it}^{*} > 0
\end{aligned}
$$

with:

$$u_{it} = \alpha_i + \eta_{it}$$

with $\eta_{it} \sim$ i.i.d. N(0,1), and $\alpha_i \sim N(0, \sigma_\alpha^2)$. This implies:

$$\text{Var}(u_{it}) = 1 + \sigma_\alpha^2$$

and so:

$$\text{Corr}(u_{it}, u_{is},\ t \neq s) \equiv \rho = \frac{\sigma_\alpha^2}{1 + \sigma_\alpha^2}.$$

# Panel Data GLMs: Count Responses

Fixed effects:

- No "incidental parameters" problem (see e.g. Cameron and Trivedi, pp. 281-2)
- Means "brute force" approach also works
- Can be fit via:
  - `pglm` (in `pglm`)
  - `feglm` (in `fixest`)
  - `glmmML`

Random effects:

$$
\begin{aligned}
\Pr(Y_{i1} = y_{i1}, \ldots Y_{iT} = y_{iT}) &= \int_0^\infty \Pr(Y_{i1} = y_{i1}, \ldots Y_{iT} = y_{iT}) f(\alpha_i) d\alpha_i \\
&= \int_0^\infty \left[ \prod_{t=1}^{T} \Pr(Y_{it} \mid \alpha_i) \right] f(\alpha_i) d\alpha_i
\end{aligned}
$$

- Simplest to assume $\alpha_i \sim \Gamma(\theta)$
- Yields a model with $\mathrm{E}(Y_{it}) = \lambda_{it}$ and $\mathrm{Var}(Y_{it}) = \lambda_{it} + \frac{\lambda_{it}^2}{\theta}$
- Fit via `glmmML` or `glmer` (or others)
- $\exists$ random effects negative binomial too...

# Some Useful Packages

- `pglm`
  - Workhorse package for panel (FE, RE, BE) GLMs
  - Binary + ordered logit/probit, Poisson / negative binomial
  - Discussed + used extensively in Croissant and Millo (2018) *Panel Data Econometrics with R*
- `fixest`
  - Fast / efficient fitting of FE models
  - Fits linear models, logit, Poisson, and negative binomial
  - Includes easy coefficient plots & tables; simple multi-threading; built-in "robust" S.E.s
- `alpaca`
  - Fast / efficient fitting of GLMs with high-dimensional fixed effects
  - *Includes bias correction for incidental parameters after binary-response models*
  - Also includes useful panel data simulation routines + average partial effects

# Generalized Estimating Equations (GEEs)

Start with:

$$Y_{it} = \mu_{it} + u_{it}$$

where

- $i \in \{1, ...N\}$ are i.i.d. "units,"
- $t \in \{1, ... T\}$, $T > 1$ are "time points,"
- we want $g(\mu_{it}) = \mathbf{X}_{it}\boldsymbol{\beta}$.

**Key issue**: Accounting for (conditional) dependence in $Y$ over time.

Define:

$$\underset{T \times T}{\mathbf{R}_i(\boldsymbol{\alpha})} = \begin{pmatrix} 1.0 & \alpha_{12} & \cdots & \alpha_{1,T} \\ \alpha_{21} & 1.0 & \cdots & \alpha_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{T,1} & \cdots & \alpha_{T,T-1} & 1.0 \end{pmatrix},$$

$\rightarrow$ "working correlation" matrix.

- Completely defined by $\boldsymbol{\alpha}$,
- Structure specified by the analyst.

Two alternatives:

- $\hat{\boldsymbol{\Sigma}}_{\text{Model}}$ ("Naive" variance estimator)
  - Requires that $\mathbf{R}_i(\boldsymbol{\alpha})$ be "correct" for consistency.
  - Is slightly more efficient than $\hat{\boldsymbol{\Sigma}}_{\text{Robust}}$ if so.

- $\hat{\boldsymbol{\Sigma}}_{\text{Robust}}$ (Clustered Huber/White estimator)
  - Is consistent *even if* $\mathbf{R}_i(\boldsymbol{\alpha})$ *is misspecified*.
  - Is slightly less efficient than $\hat{\boldsymbol{\Sigma}}_{\text{Model}}$ if $\mathbf{R}_i(\boldsymbol{\alpha})$ is correct.

Means that GEEs:

- Are a straightforward variation on GLMs, and so
- Can be applied to a range of data types (continuous, binary, count, proportions, etc.),
- Yield robustly consistent point estimates of $\boldsymbol{\beta}$s,
- Account for within-unit correlation in an informed way, but also
- Yield consistent inferences even if that correlation is misspecified.

# Panel Data: Resources

Good references (various topics):


- Beck, Nathaniel, and Jonathan N. Katz. 1995. "What To Do (And Not To Do) With Time- Series Cross-Section Data." American Political Science Review 89(September): 634-647.
- Clark, Tom S. and Drew A. Linzer. 2015. "Should I Use Fixed Or Random Effects?" Political Science Research and Methods 3(2):399-408.
- Kropko ,Jonathan, and Robert Kubinec. 2020. "Interpretation and Identification of Within-Unit and Cross-Sectional Variation in Panel Data Models." PLoS ONE 15(4): e0231349.
- Liu, Licheng, Ye Wang, Yiqing Xu. 2020. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." Working paper: Stanford University.
- Mizik, Natalie, and Eugene Pavlov. 2018. "Panel Data Methods in Marketing Research." In Natalie Mizik and Dominique M. Hanssens, Eds. Handbook of Marketing Analytics. Northampton, MA: Edward Elgar.
- Zorn, Christopher. 2001. "Generalized Estimating Equation Models for Correlated Data: A Review with Applications." American Journal of Political Science 45(April):470-90.


Good texts:


- Croissant, Yves, and Giovanni Millo. 2018. Panel Data Econometrics with R. New York: Wiley.
- Hsiao, Cheng. 2014. *Analysis of Panel Data*, 3rd Ed. New York: Cambridge University Press.
- Wooldridge, Jeffrey. 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd Ed. Cambridge: MIT Press.


Also:


- Zorn, Christopher. 2022. *Analyzing Panel Data*. GSERM course, summer 2022.