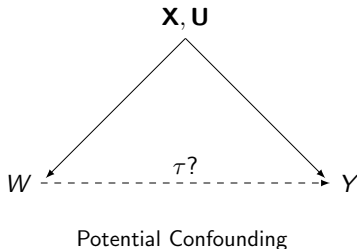


# PLSC 504 – Fall 2023

## Causal Inference with Observational Data

September 20, 2023

# What We're On About



Here:

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

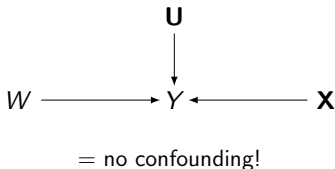
- **Randomize**

(or...)

- Instrumental Variables Approaches
- 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'll discuss these approaches in a couple weeks, as models for panel/time-series cross-sectional data.

# Under Randomization



## Note:

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

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

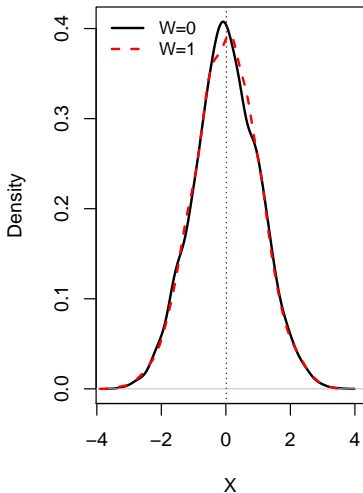
or, more demandinglly,

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

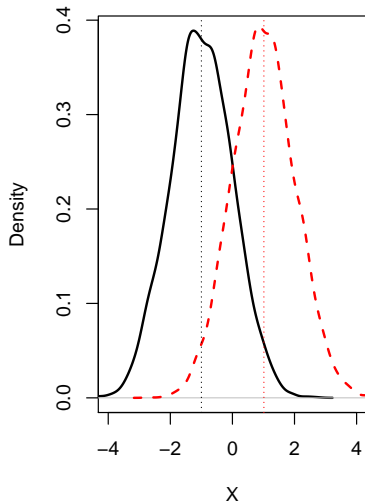
- Can yield imbalance by random chance...

# Covariate Balance / Imbalance

**Balanced X**



**Unbalanced X**



# 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 $W_i$

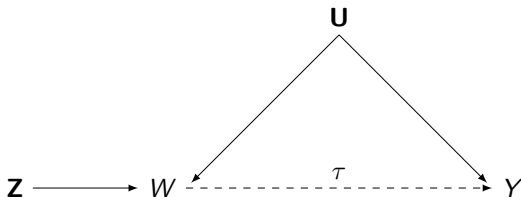
Valid causal inference requires  $Y_{0i}, Y_{1i} \perp W_i | \mathbf{X}_i, \mathbf{U}_i$

- 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



Instrumental Variables

As in the more general regression case where we have  $\text{Cov}(\mathbf{X}, u) \neq 0$ , 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  $W$  (“exclusion restriction”)
  3.  $\mathbf{Z}$  is exogenous [i.e.,  $\text{Cov}(\mathbf{Z}, \mathbf{U}) = 0$ ]
- 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 \(2021\)](#): 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)

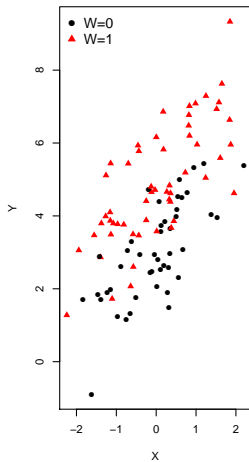
# Nonrandom Assignment of $W_i$ (continued)

So...

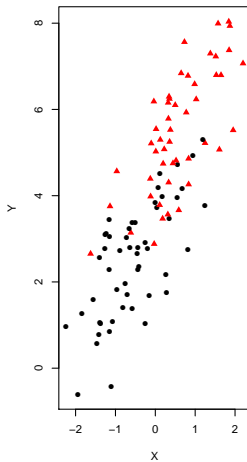
- Causal inference with observational data typically requires that  $\mathbf{U} = \emptyset$ ...
- This typically requires a strong theoretical motivation in order to assume that the observed  $\mathbf{X}$  exhausts the list of possible confounders.
- **Even if** this assumption is reasonable, there 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

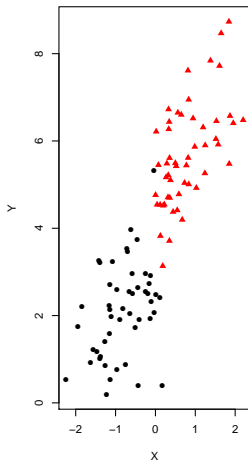
**Complete Overlap**



**Moderate Overlap**



**No 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 is a way of dealing with one or 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...

- 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: Genetic matching ([Diamond and Sekhon 2013](#)), etc.<sup>1</sup>

---

<sup>1</sup>[Shalizi \(2016\)](#) notes that “(A)pproximate matching is implicitly doing nonparametric regression by a nearest-neighbor method,” and that “(M)aybe it is easier to get doctors and economists to swallow “matching” than “nonparametric nearest neighbor regression”; this is not much of a reason to present the subject as though nonparametric smoothing did not exist, or had nothing to teach us about causal inference.”

Interestingly, quite a few of the good matching programs written for R have been written by political scientists...

- 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)



# Regression Discontinuity Designs

“RDD”:

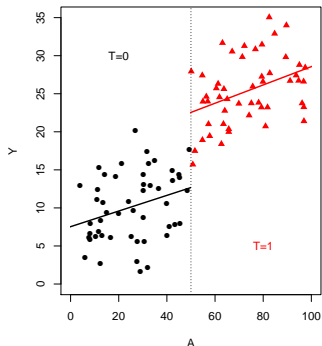
- Treatment changes abruptly [usually at some threshold(s)] according to the value(s) of some measured, continuous, pre-treatment variable(s)
  - This is known as the “assignment” or “forcing variable(s),” sometimes denoted **A**
  - Formally:

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

- 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`



- R
  - Packages for matching are listed above (Matching, MatchIt, etc.)
  - Similarly for RDD (rddtools, rdd, etc.)
  - IV regression: ivreg (in AER), ts1s (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

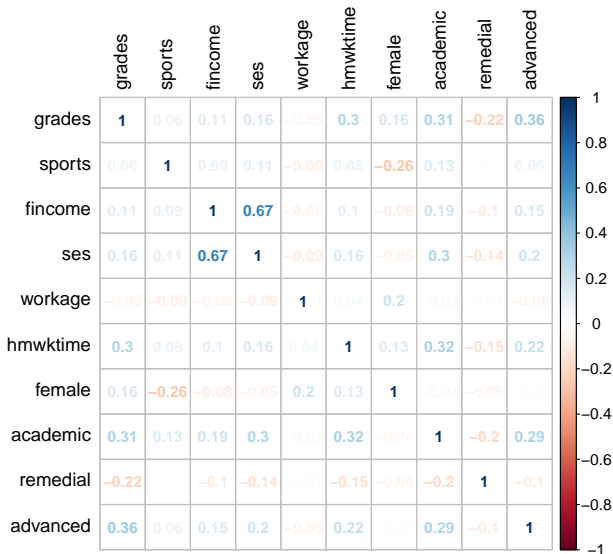
  

workage		hwmktime	female		academic
Min.	:11.0	Min. :1.0	Min.	:0.00	Min. :0.00
1st Qu.:	13.0	1st Qu.:4.0	1st Qu.:	0.00	1st Qu.:0.00
Median	:15.0	Median :4.0	Median	:1.00	Median :0.00
Mean	:14.6	Mean :4.5	Mean	:0.52	Mean :0.41
3rd Qu.:	16.0	3rd Qu.:6.0	3rd Qu.:	1.00	3rd Qu.:1.00
Max.	:21.0	Max. :7.0	Max.	: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 ***

---

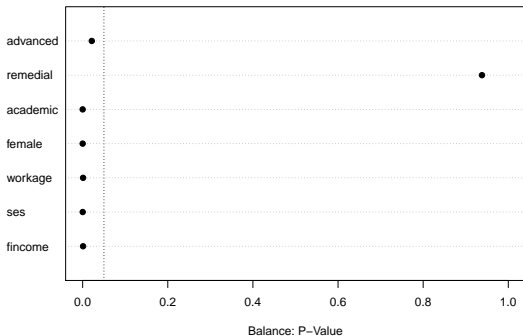
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

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 `sports = 0` and `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:

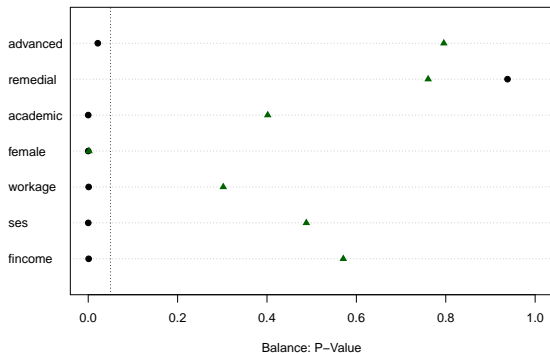
	Control	Treated
All	864	511
Matched	287	239
Unmatched	577	272

```
> # Output matched data:
```

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

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

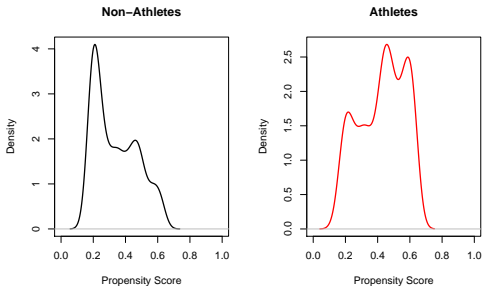
# Exact Matching: Balance



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

```
.
.
.
```

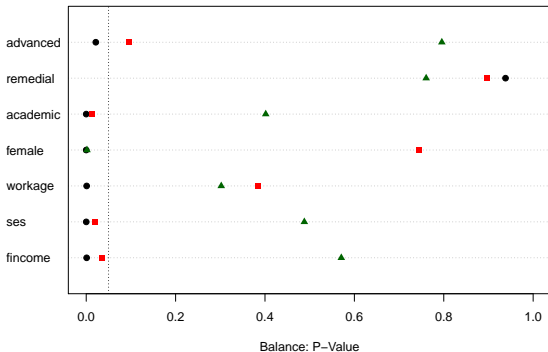
Percent Balance Improvement:

	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  $\text{sports} = 0$  and  $\text{sports} = 1$ . Black dots are pre-matching; green triangles are after exact matching; red squares are after propensity score matching.

# Differences in Means

```
> 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* (0.13)	3.05* (0.23)	2.84* (0.16)	2.75* (0.17)
sports	0.10* (0.04)	0.12* (0.06)	0.09* (0.04)	0.08 (0.05)
fincome	0.00 (0.01)	0.05 (0.03)	-0.00 (0.02)	0.01 (0.02)
ses	0.02 (0.03)	-0.14 (0.07)	0.05 (0.04)	0.03 (0.05)
workage	-0.02* (0.01)	-0.03* (0.01)	-0.03* (0.01)	-0.02* (0.01)
female	0.30* (0.04)	0.34* (0.06)	0.31* (0.05)	0.29* (0.05)
academic	0.29* (0.04)	0.24* (0.08)	0.31* (0.05)	0.31* (0.05)
remedial	-0.23* (0.04)	-0.28* (0.06)	-0.28* (0.05)	-0.21* (0.05)
advanced	0.44* (0.04)	0.51* (0.08)	0.43* (0.05)	0.40* (0.05)
R <sup>2</sup>	0.23	0.29	0.26	0.22
Adj. R <sup>2</sup>	0.23	0.28	0.25	0.21
N	1375	526	1022	939

\* $p < 0.05$

## 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?



- Good references:
  - Freedman (2012)\*
  - Shalizi (someday)\*
  - Morgan and Winship (2014)
  - Pearl et al. (2016)
  - Peters et al. (2017)
- Courses / syllabi (a sampling):
  - Eggers (2019)
  - Frey (2019)
  - Hidalgo (2020)
  - Imai (2022)
  - Simpson (2019)
  - Strezhnev (2023)
  - Visconti (2019)
  - Xu (2018)
  - Yamamoto (2018)
- Other useful things:
  - The Causal Inference Book
  - Some useful notes

\* I really like this one.