GSERM 2023
Regression for Publishing

June 23, 2023

Sample Selection In Theory

- Challenge: Inference to a Population from a Non-Random Sample
- Widespread Problem...
  - Heckman's wage equations...
  - Self-selection (e.g., into groups)
  - Surveys: "Screening" questions (sometimes...)
- Parallels in Missing Data, Causal/Counterfactual Inference

Sample Selection Basics

Consider latent Y^*s:

$$\begin{aligned} Y_{1i}^* &= \mathbf{X}_i\boldsymbol{\beta} + u_{1i} \\ Y_{2i}^* &= \mathbf{Z}_i\gamma + u_{2i} \end{aligned}$$

Observe:

$$Y_{1i} = \begin{cases} Y_{1i}^* \text{ if } Y_{2i}^* > 0 \\ \text{missing if } Y_{2i}^* \leq 0 \end{cases}$$

- Y_{2i}^* unobserved (except for sign);
- $\mathbf{X}_i$ observed iff Y_{1i} is observed;
- $\mathbf{Z}_i$ observed in every case.

Sample Selection Basics

When do we observe Y_1?

$$\begin{aligned}
\Pr(Y_{2i}^* \leq 0|\mathbf{X},\mathbf{Z}) &= \Pr(u_{2i} \leq -\mathbf{Z}_i\gamma) \\
&= 1 - \Pr(u_{2i} \geq -\mathbf{Z}_i\gamma) \\
&= 1 - \Pr(-u_{2i} \leq \mathbf{Z}_i\gamma) \\
&= 1 - \int_{-\infty}^{\mathbf{Z}_i\gamma} f(u_2)du_2 \\
&= 1 - F_{u_2}(\mathbf{Z}_i\gamma)
\end{aligned}$$

Define:

$$D_i = \begin{cases} 1 & \text{if } Y_{1i} \text{ is observed.} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\Pr(D_i = 1) = F_{u_2}(\mathbf{Z}_i\gamma).$$

An Assumption

Assume:

$$\{u_1, u_2\} \sim \mathcal{BVN}(0, 0, \sigma_1^2, 1, \sigma_{12})$$

Means

$$\Pr(D_i = 1 | \mathbf{Z}_i, \mathbf{X}_i) = \Phi(\mathbf{Z}_i \gamma).$$

Define:

$$\rho = \text{corr}(u_1, u_2).$$

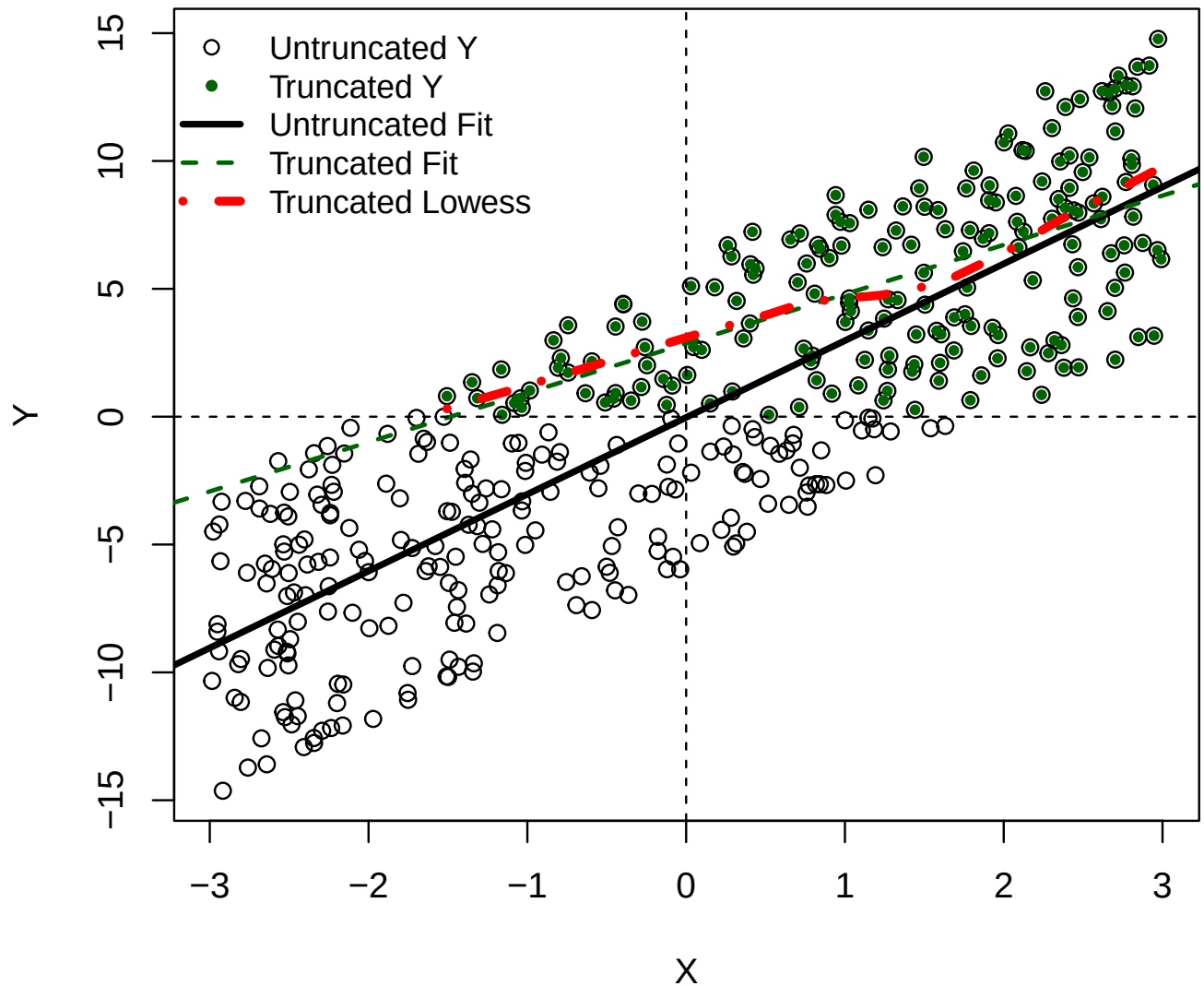
What we get:

$$\mathrm{E}(Y_{1i}|\mathbf{X}_i, \mathbf{Z}_i, D_i = 1) = \mathbf{X}_i\boldsymbol{\beta} + \rho\sigma_1 \left[\frac{\phi(\mathbf{Z}_i\gamma)}{\Phi(\mathbf{Z}_i\gamma)}\right]$$

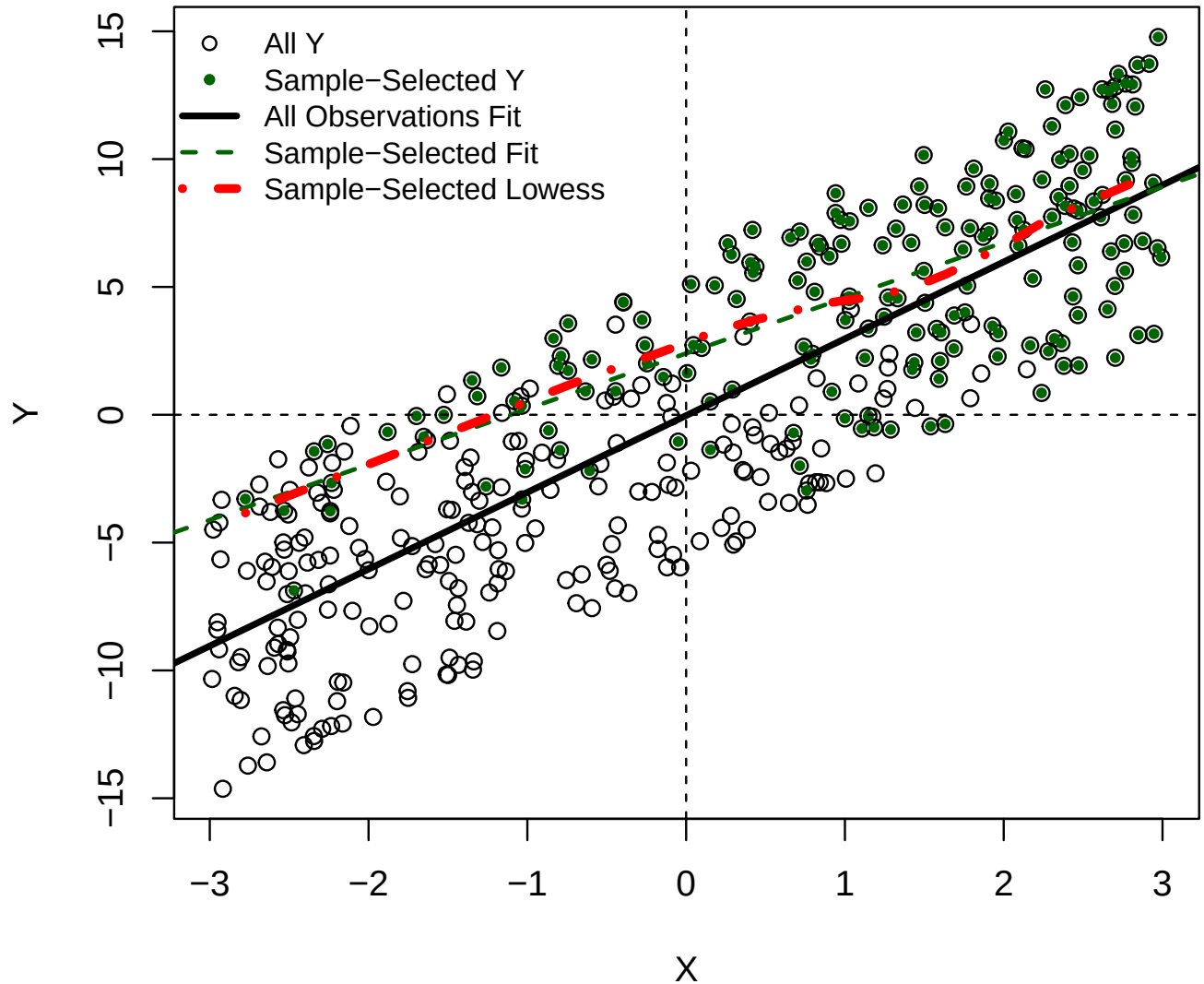
Without conditioning on $\mathbf{Z}$:

$$\mathrm{E}(Y_{1i}|\mathbf{X}_i, D_i = 1) = \mathbf{X}_i\boldsymbol{\beta} + \mathrm{E}\left\{\rho\sigma_1 \left[\frac{\phi(\mathbf{Z}_i\gamma)}{\Phi(\mathbf{Z}_i\gamma)}\right] \middle| \mathbf{X}_i\right\}$$

Truncation Bias



Sample Selection Bias



Selection Bias: Substantive Effects

- Specification Error (unless $\rho = 0$)
- Indeterminate bias in $\hat{\beta}$
- Including $\mathbf{Z}_i$ will not generally* remove the bias:

  "With quasi-experimental data derived from nonrandomized assignments, controlling for additional variables in a regression may worsen the estimate of the treatment effect, even when the additional variables improve the specification." – Achen (1986, p. 27)

- Bias remains even if inference is limited to the "selected" group.
  [This point is made nicely in Berk (1983)...]

* Unless sample selection is completely deterministic (i.e., determined *perfectly* by $\mathbf{X}$, $\mathbf{Z}$) (Heckman & Robb 1985).

E(Y) Under Selection

Conditional Density:

$$h(Y|\mathbf{X},\mathbf{Z},\boldsymbol{\beta},\gamma,\sigma_1,\rho) = \frac{\phi\left(\frac{Y_{1i}-\mathbf{X}_i\boldsymbol{\beta}}{\sigma_1}\right)}{\sigma_1\Phi(\mathbf{Z}_i\gamma)} \cdot \Phi\left[\frac{\frac{\rho(Y_{1i}-\mathbf{X}_i\boldsymbol{\beta})}{\sigma_1}+\mathbf{Z}_i\gamma}{\sqrt{1-\rho^2}}\right]$$

Note: $\rho = 0$ yields

$$\begin{aligned} h(Y|\mathbf{X},\mathbf{Z},\boldsymbol{\beta},\gamma,\sigma_1,\rho=0) &= \frac{\phi\left(\frac{Y_{1i}-\mathbf{X}_i\boldsymbol{\beta}}{\sigma_1}\right)}{\sigma_1\Phi(\mathbf{Z}_i\gamma)} \cdot \Phi\left[\frac{0+\mathbf{Z}_i\gamma}{1}\right] \\ &= \frac{\phi\left(\frac{Y_{1i}-\mathbf{X}_i\boldsymbol{\beta}}{\sigma_1}\right)}{\sigma_1}. \end{aligned}$$

Likelihood Under Selection

Under sample selection, the full likelihood is:

$$
\begin{aligned}
\ln L(\boldsymbol{\beta}, \gamma, \sigma_1, \rho | Y_1) \quad = \quad & \sum_{i=1}^{N} (1 - D_i) \ln[1 - \Phi(\mathbf{Z}_i \gamma)] \\
& + \sum_{i=1}^{N} D_i \ln[\Phi(\mathbf{Z}_i \gamma)] \\
& + \sum_{i=1}^{N} D_i \ln \left\{ \frac{\phi\left(\frac{Y_{1i} - \mathbf{X}_i \boldsymbol{\beta}}{\sigma_1}\right)}{\sigma_1 \Phi(\mathbf{Z}_i \gamma)} \cdot \Phi \left[\frac{\frac{\rho(Y_{1i} - \mathbf{X}_i \boldsymbol{\beta})}{\sigma_1} + \mathbf{Z}_i \gamma}{\sqrt{1 - \rho^2}} \right] \right\}
\end{aligned}
$$

Model Fitting

Estimation can be via:

- MLE (above)
- Or, reconsider:

$$\mathrm{E}(Y_{1i}|\mathbf{X}_i, \mathbf{Z}_i, D_i = 1) = \mathbf{X}_i\boldsymbol{\beta} + \rho\sigma_1 \left[\frac{\phi(\mathbf{Z}_i\gamma)}{\Phi(\mathbf{Z}_i\gamma)}\right]$$

- Note that $\Phi(\mathbf{Z}_i\gamma) = \mathrm{Pr}(D_i = 1)$
- Suggests a two-step approach...

Heckman's Two-Step Estimator

1. Estimate $\hat{\gamma}$ from

$$\Pr(D_i = 1) = \Phi(\mathbf{Z}_i\gamma)$$

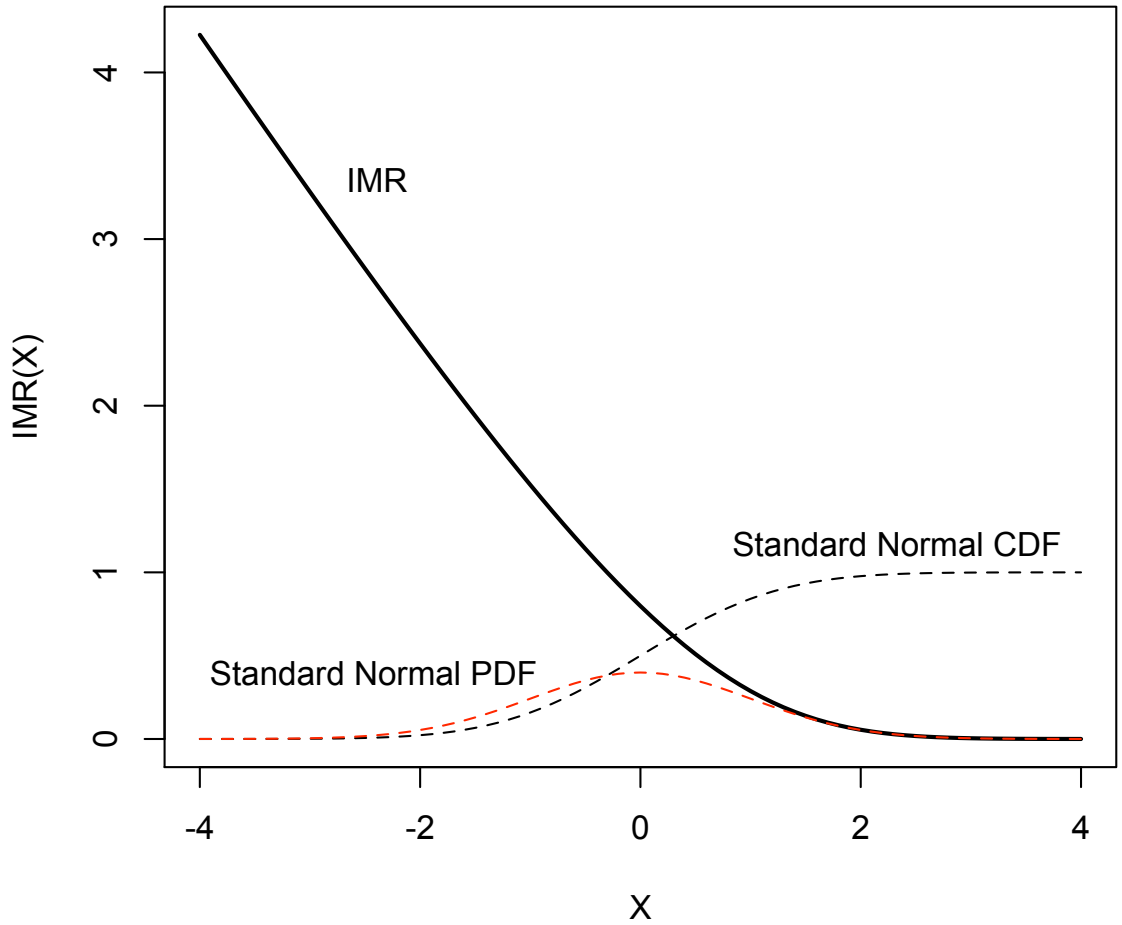
and calculate the estimated inverse Mills' ratio:

$$\hat{\lambda}_i = \frac{\phi(\mathbf{Z}_i\hat{\gamma})}{\Phi(\mathbf{Z}_i\hat{\gamma})}$$

2. Estimate $\boldsymbol{\beta}, \theta(\equiv \rho\sigma_1)$ as:

$$Y_{1i} = \mathbf{X}_i\boldsymbol{\beta} + \theta\hat{\lambda}_i + u_{1i}$$

What exactly *is* an "inverse Mills' ratio," anyway?



A Few Things...

In the two-step approach:

- Since $\sigma_1 > 0$, $\hat{\theta} = 0 \implies \rho = 0$
- Two-step approach:
  - Is "LIML"...
  - Consistent for $\hat{\boldsymbol{\beta}}$, but
  - Inconsistent estimating $\widehat{\mathbf{V}(\boldsymbol{\beta})}$; so
  - Standard errors require correction (e.g., bootstrap)
  - *Can* yield $\hat{\rho} \notin [-1, 1]$ (because $\hat{\rho} = \hat{\theta}/\hat{\sigma}_1$)
  - Sensitive to prediction of D_i (better prediction = better precision)

Identification, etc.

For any estimation approach:

- If $\mathbf{X} = \mathbf{Z}$, then $\boldsymbol{\beta}, \gamma, \rho$ (formally) identified by nonlinearity of $\Phi(\cdot)$
- (Much) better: $\geq$ one covariate in $\mathbf{Z}$ not in $\mathbf{X}$
- But...
  - Factors causing Y_1 also (often) cause D
  - $\Longrightarrow$ $\mathbf{X}, \mathbf{Z}$ highly correlated
  - ...just makes things worse (Stolzenberg and Relles 1997)

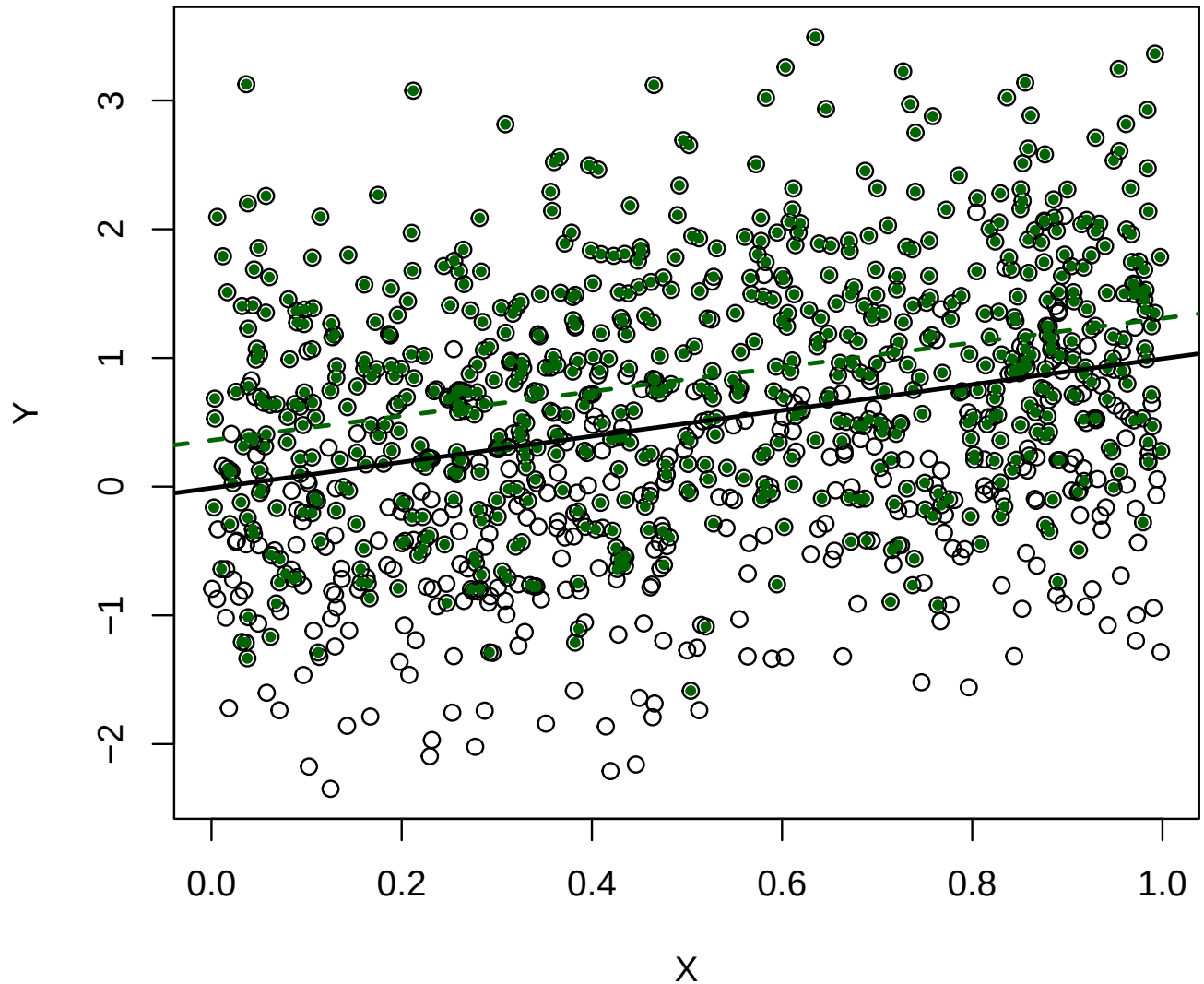
Some Practical Things

- In practice, few people use two-step anymore,
- Model is *always* sensitive to joint normality of $\{u_i, u_2\}$,
- It is also very sensitive to model specification...
- Key issue: endogeneity of selection...

Simulated Example I: $\text{Cov}(X, Z) = 0$

```
> set.seed(7222009)
> N  <- 1000           # N of observations

> # Bivariate normal us, correlated at r=0.7
> us <- rmvnorm(N,c(0,0),matrix(c(1,0.7,0.7,1),2,2))

> Z  <- runif(N)       # Sel. variable
> Sel<- Z + us[,1]>0   # Selection eq.
> X  <- runif(N)       # X
> Y  <- X + us[,2]     # B0=0, B1=1
> Yob<- ifelse(Sel==TRUE,Y,NA)     # Selected Y
>
> # OLSs:
>
> NoSel<-lm(Y~X)       # all data
> WithSel<-lm(Yob~X)   # sample-selected data
```

Simulation I (continued)



Simulation I (continued)

```
> # Two-Step:
>
> probit<-glm(Sel~Z,family=binomial(link="probit"))
> IMR<-((1/sqrt(2*pi))*exp(-((probit$linear.predictors)^2/2))) /
+   pnorm(probit$linear.predictors)
>
> OLS2step<-lm(Yob~X+IMR)
>
>
> # FIML:
>
> FIML<-selection(Sel~Z,Y~X,method="ml")
```

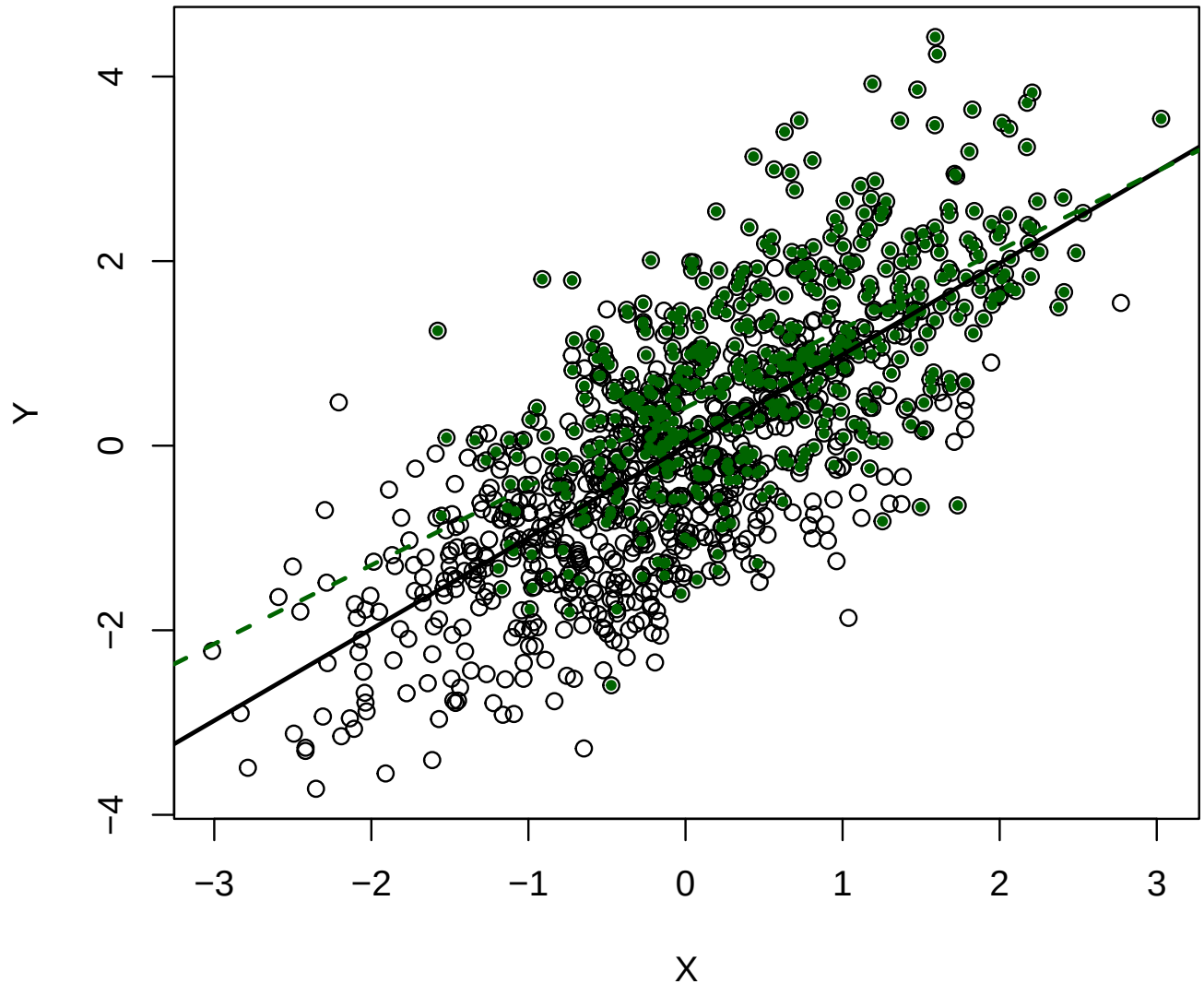
Simulation I (continued)

	OLS-All	OLS-Selected	Two-Stage	FIML
X (true OLS = 1)	1.000***	0.947***	0.948***	0.939***
	(0.106)	(0.114)	(0.114)	(0.112)
IMR			0.428*	
			(0.223)	
Constant (true = 0)	−0.011	0.360***	0.152	−0.007
	(0.062)	(0.068)	(0.128)	(0.092)
Observations	1,000	691	691	1,000
R^2	0.083	0.091	0.096	
Adjusted R^2	0.082	0.089	0.093	
Log Likelihood				−1,479.000
ρ				0.742*** (0.088)
Note:				*p<0.1; **p<0.05; ***p<0.01

Simulated Example II: $\mathrm{Cov}(X, Z) > 0$

```
> set.seed(9021970)
> N  <- 1000           # N of observations
>
> # Bivariate normal us & Xs, correlated at r=0.7 / 0.8
> us <- rmvnorm(N,c(0,0),matrix(c(1,0.7,0.7,1),2,2))
> Xs <- rmvnorm(N,c(0,0),matrix(c(1,0.8,0.8,1),2,2))
> Z <- Xs[,1]
> X <- Xs[,2]
> Sel<- Z + us[,1]>0       # Selection eq.
> Y  <- X + us[,2]         # B0=0, B1=1
> Yob<- ifelse(Sel==TRUE,Y,NA)  # Selected Y
>
> # OLSs:
>
> NoSel2<-lm(Y~X)      # all data
> WithSel2<-lm(Yob~X)  # sample-selected data
```

Simulation II (continued)



Simulation II (continued)

	OLS-All	OLS-Selected	Two-Stage	FIML
X (true OLS = 1)	0.991***	0.853***	1.020***	1.010***
	(0.029)	(0.046)	(0.061)	(0.056)
IMR			0.533***	
			(0.133)	
Constant (true = 0)	−0.005	0.412***	0.041	0.045
	(0.030)	(0.046)	(0.103)	(0.088)
Observations	1,000	511	511	1,000
R^2	0.533	0.403	0.421	
Adjusted R^2	0.532	0.401	0.419	
Log Likelihood				−1,146.000
ρ				0.560*** (0.097)
Note:				*p<0.1; **p<0.05; ***p<0.01

Extensions: "Probit-Probit"

- Selection + binary second stage ($Y_i \in \{0, 1\}$) (a/k/a "Heckit").
- Assume errors are bivariate standard Normal [so, $\{u_1, u_2 \sim \mathcal{BVN}(0, 0, 1, 1, \rho) \equiv \Phi_2(\cdot)]$
- Log-Likelihood:

$$\begin{aligned} \ln L(\boldsymbol{\beta}, \gamma, \sigma_1, \rho | Y_1) &= \sum_{Y_{1i}=1, D_i=1} \ln[\Phi_2(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\gamma, \rho)] \\ &+ \sum_{Y_{1i}=0, D_i=1} \ln[\Phi_2(-\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\gamma, -\rho)] \\ &+ \sum_{D_i=0} \ln \Phi(-\mathbf{Z}_i\gamma) \end{aligned}$$

More Extensions

- Different outcome stages:
  - Poisson (Greene 1995; Cameron & Trivedi 2013, Ch. 10)
  - Durations (Boehmke et al. 2006)
  - Count/binary/ordinal (Mirand and Rabe-Hesketh 2005)

- Selection stage is ordered (Chiburis & Lokshin 2007)

- Multiple-stage models (not much... work in finance + Signorino and others)

- Semi- and non-parametric variants (e.g., Liu and Yu (2019) on monotone control functions)

Sample Selection: Software

- R (`selection` and `heckit` in `sampleSelection`; robust estimation via `ssmrob`)
  - Binary selection
  - Continuous/binary outcomes
  - Also tobit, etc. models

- Stata
  - `heckman` (binary-continuous model)
  - `heckprob` (binary-binary model)
  - `heckoprobit` (ordinal Y)
  - `heckpoisson` (Poisson)
  - `dursel` (binary-duration model)
  - `xtheckman` (selection models for panel data)
  - Also Bayesian versions, using the `bayes:` prefix

References

Articles by Heckman (1974, 1976, 1979).

Breen, Richard. 1996. Regression Models for Censored, Sample Selected, or Truncated Data. Thousand Oaks, CA: Sage.

Stolzenberg, Ross M. and Daniel A. Relles. 1997. "Tools for Intuition about Sample Selection Bias and Its Correction." American Sociological Review 62:494-507.

Vella, Francis. 1998. "Estimating Models with Sample Selection Bias: A Survey." The Journal of Human Resources 33:127-169.

Winship, Christopher and Robert D. Mare. 1992. "Models for Sample Selection Bias." Annual Review of Sociology 18:327-350.

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
$\Pr(Y\|X = x)$	$\Pr[Y\|do(X = x)]$
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:

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

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.

So let's talk about missing data...

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"

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

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

Missing Data (continued)

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.

...Back To 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

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):

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

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

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

Corresponding estimates:

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

and

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

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).

Digression: 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

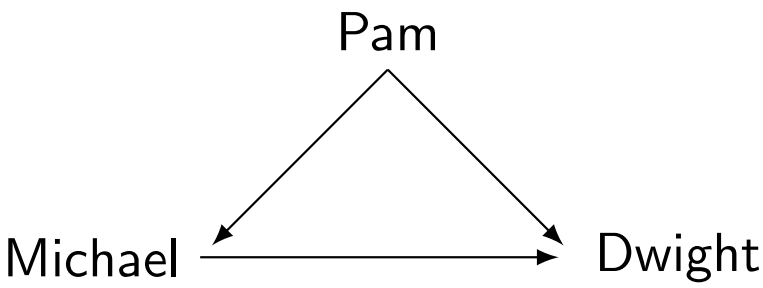


Figure: A DAG

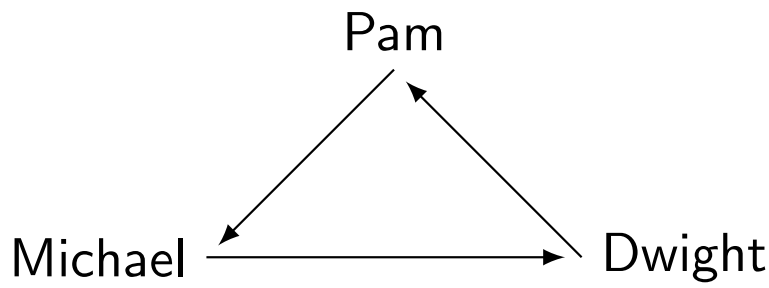


Figure: Not a DAG

DAGs and Confounding

Figure: No Confounding

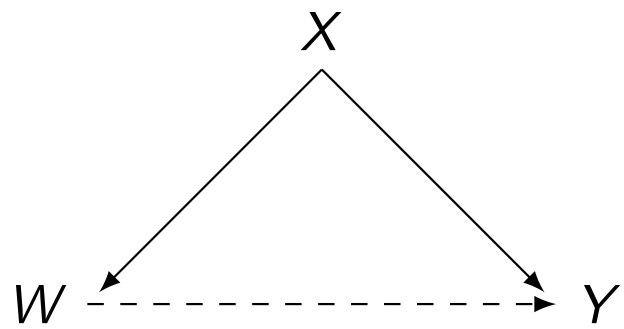


Figure: Confounding

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	-
Mean_{obs}	-	10	12	-	11	2
Mean_{all}	-	(10)	(12)	(2)	-	-

$t = -1.22,\ p = 0.14$

Confounding Bias: Some Toy Examples

Example Two: $\text{Cov}(W, Y) > 0$ (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	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	-
Mean_{obs}	-	8.67	13.33	-	11	4.67
Mean_{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	-
Mean_{obs}	-	11.33	10.67	-	11	-0.67
Mean_{all}	-	(10)	(12)	(2)	-	-

$t = 0.71$, $p = 0.74$

What We're On About

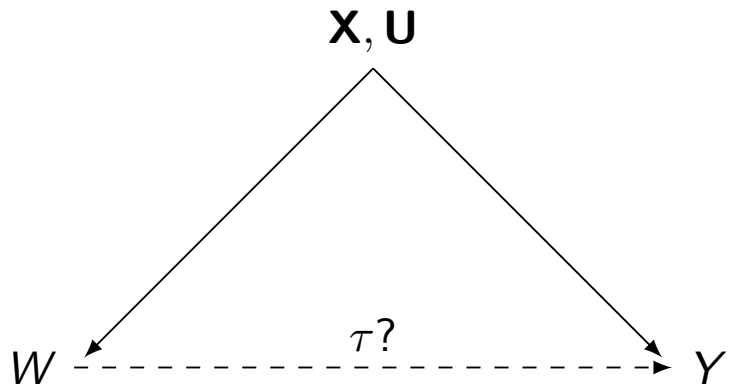


Figure: Potential Confounding

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,
- We're interested in estimating τ, the "treatment effect" of W on Y,
- $\mathbf{X}$ are observed confounders,
- $\mathbf{U}$ are unobserved confounders.

Things We Can Do

- **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

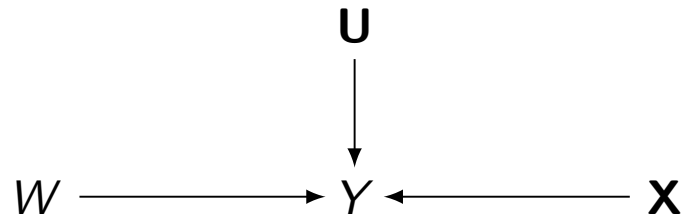


Figure: = no confounding!

Note:

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

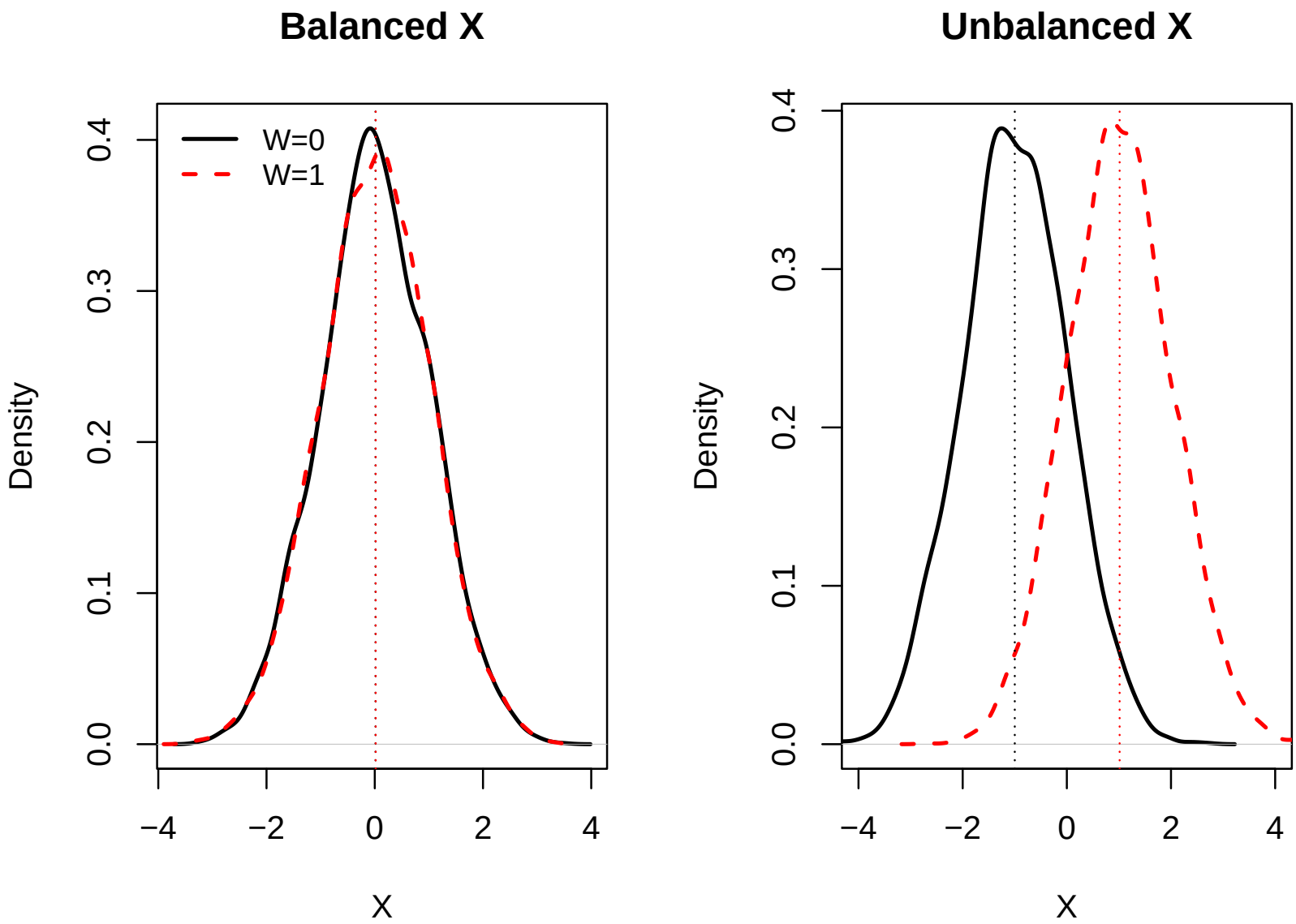
$$\mathrm{E}(\mathbf{X}_i, \mathbf{U}_i \mid W_i = 0) = \mathrm{E}(\mathbf{X}_i, \mathbf{U}_i \mid W_i = 1)$$

  or, more demandingly,

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

- Can 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 τ
  - 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 **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

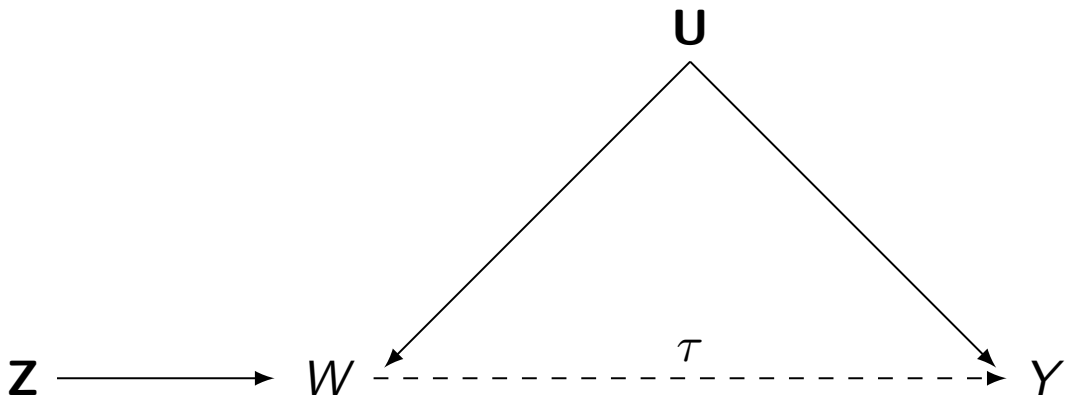


Figure: Instrumental Variables

As in the more general regression case where we have $\text{Cov}(\mathbf{X}, u) \neq 0$, 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., $\text{Cov}(\mathbf{Z}, \mathbf{U}) = 0$]
- 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
  - Shalizi (2020, chapters 20-21): Gathers all the issues together...
- 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)

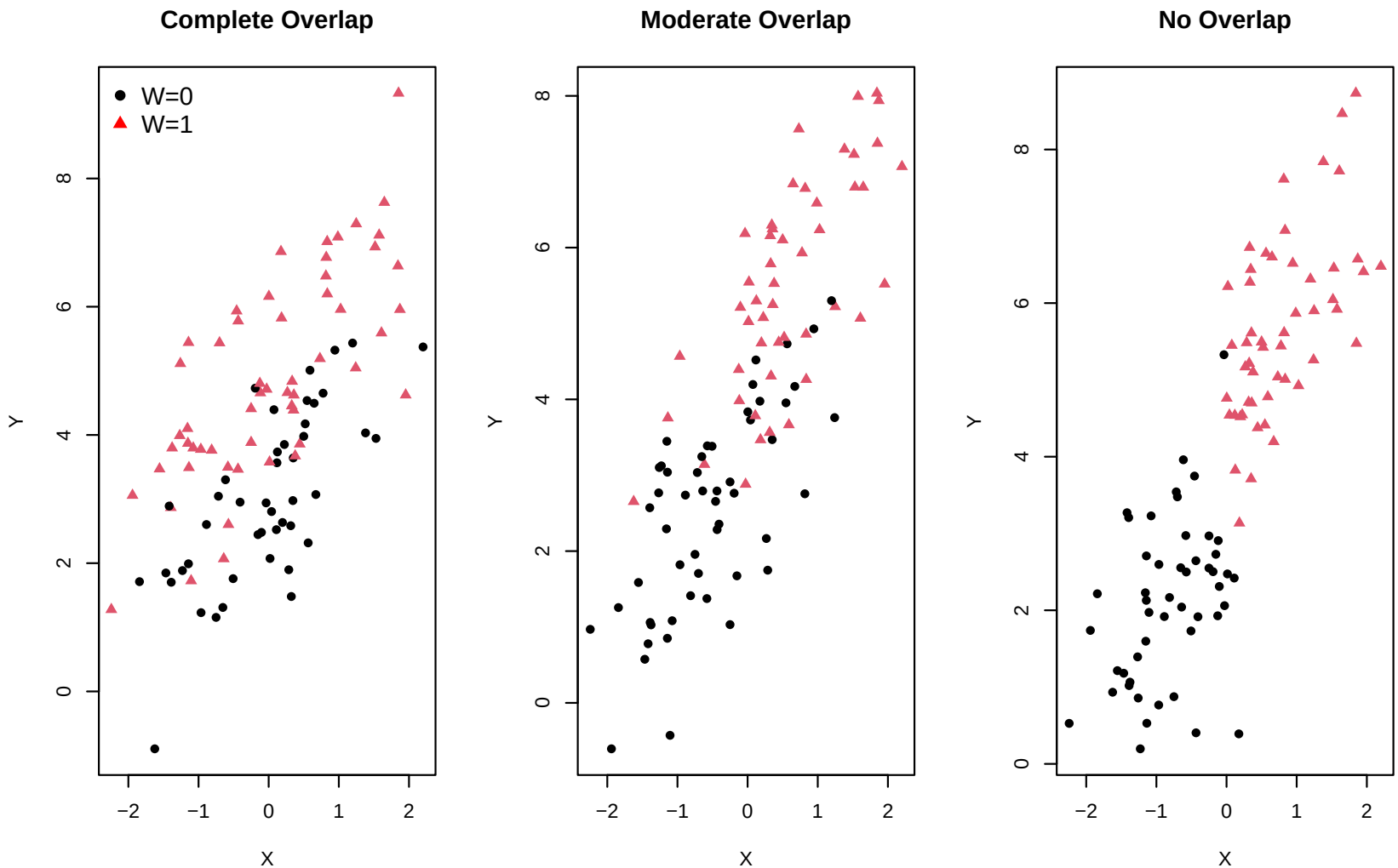
Nonrandom Assignment of W_i (continued)

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


Complete Overlap
Moderate Overlap
No Overlap
W=0
W=1
Y
X
−2
−1
0
1
2
0
2
4
6
8

Overlap and Balance

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 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...

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: Genetic matching (Diamond and Sekhon 2013), etc.[1]

[1] 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."

Matching Software

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":

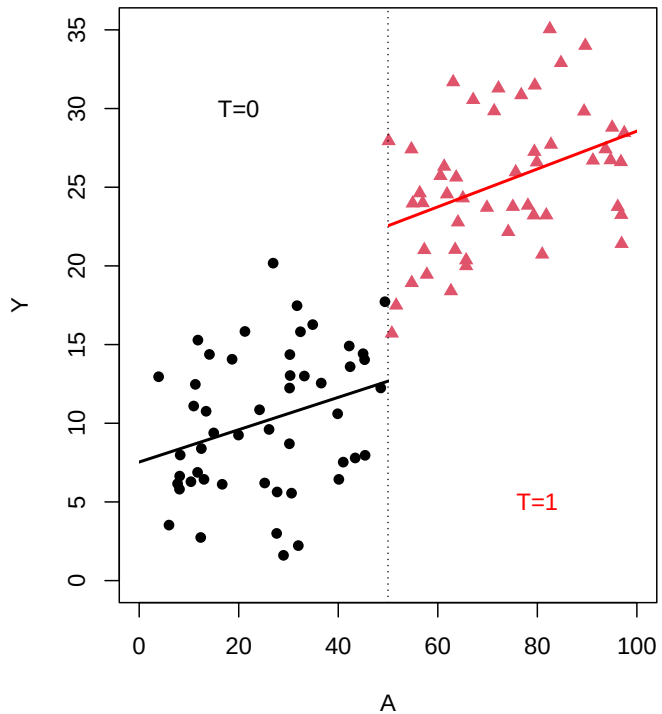
- Treatment changes abruptly [usually at some threshold(s)] according to the value(s) of some measured, continuous, pre-treatment variable(s)
  - This is known as the "assignment" or "forcing variable(s)," sometimes denoted **A**
  - Formally:

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

- Intuition: Observations near but on either side of the threshold(s) are highly comparable, and can be used to (locally) identify τ
- 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:

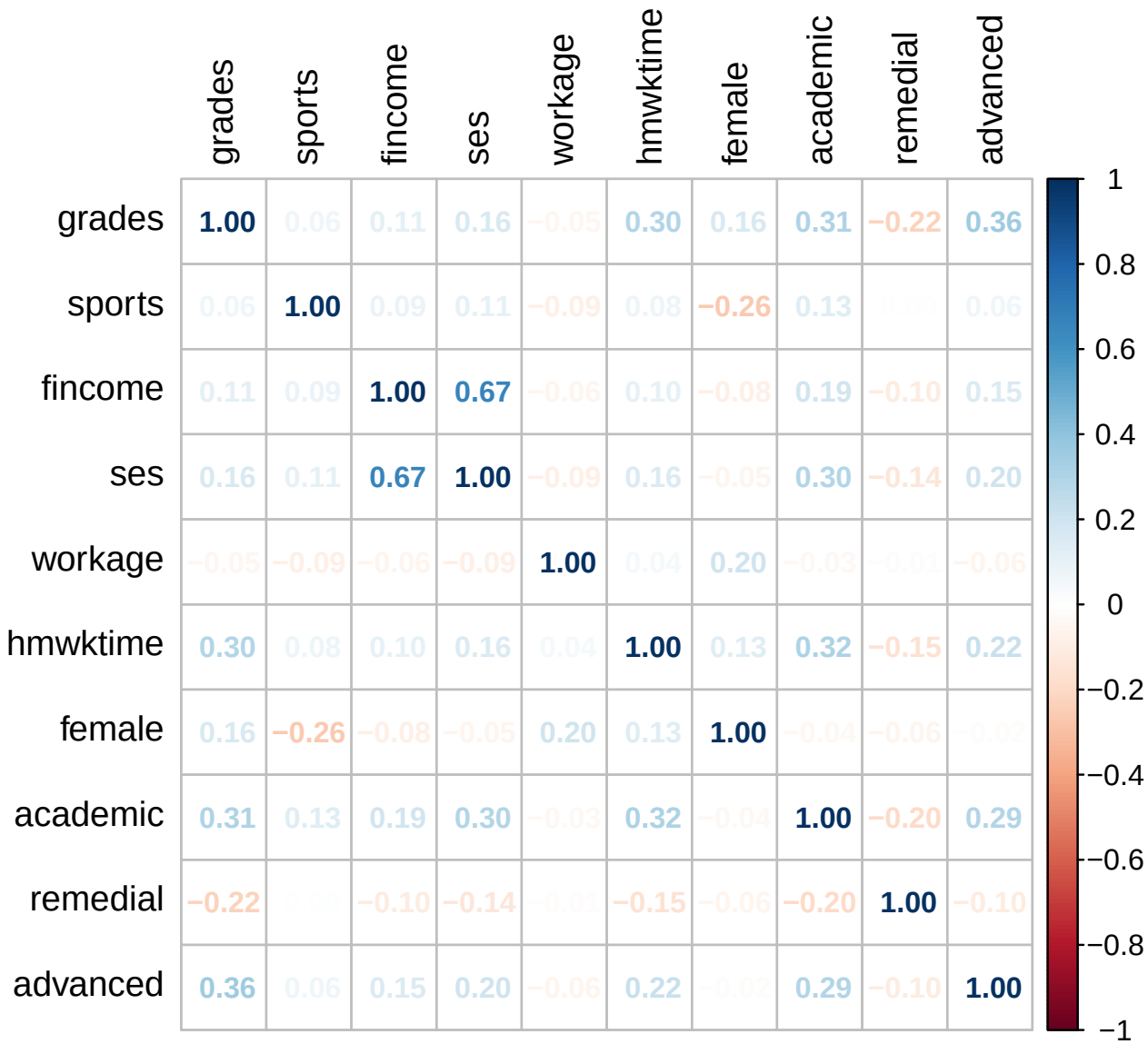
- `grades`: As=4, As & Bs=3.5, etc.
- `sports`: 1 if participated in varsity sports, 0 otherwise
- `fincome`: Family income (7-point scale)
- `ses`: Socioeconomic Status: 1=low, 2-middle, 3=high
- `workage`: Age at which started working
- `hmwktime`: Time spent on homework (7-point scale)*
- `female`: 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  
```

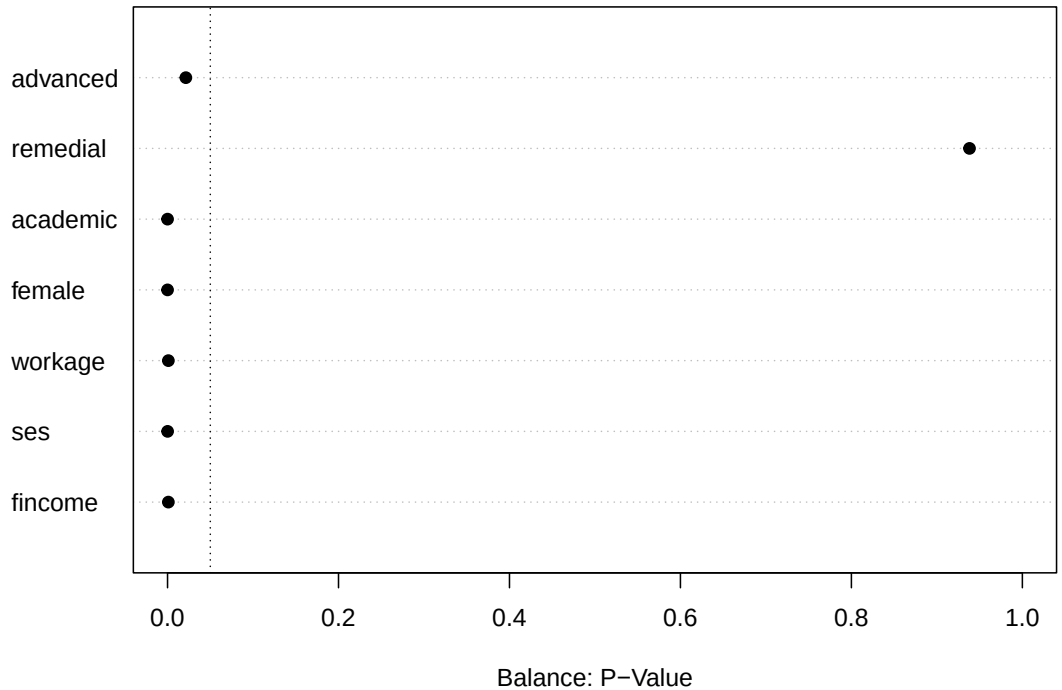
Correlation Plot



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)

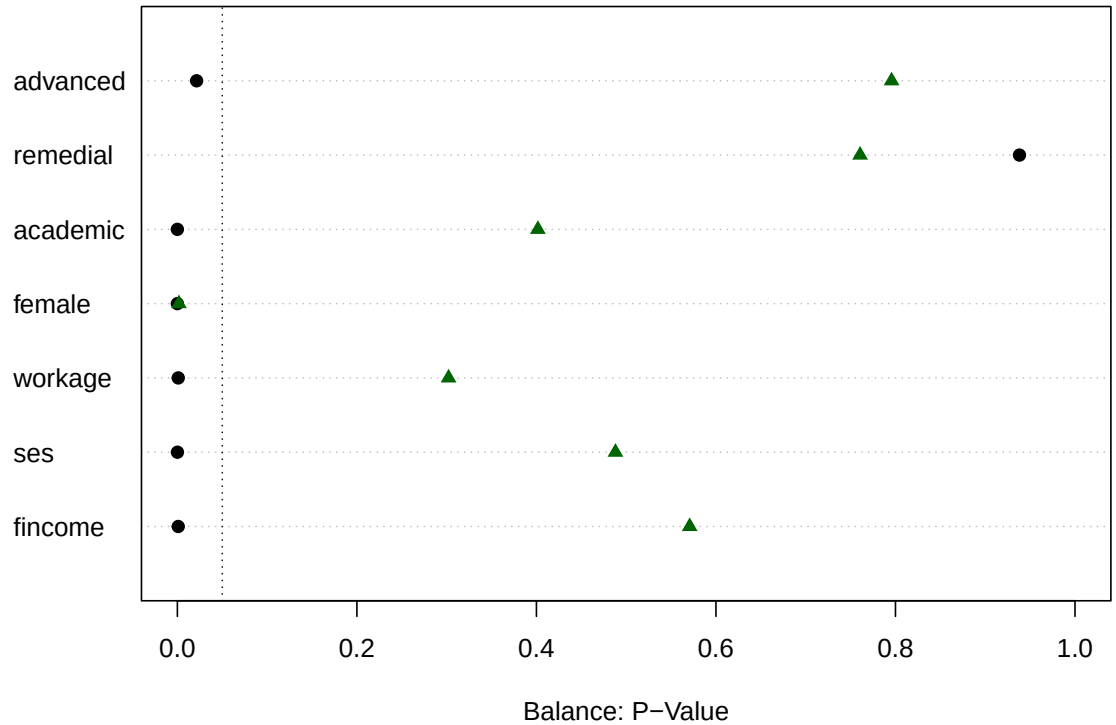


These are P-values associated with t-tests (for binary predictors) or Kolmogorov-Smirnov tests (for continuous predictors) for balance between $\texttt{sports} = 0$ and $\texttt{sports} = 1$.

Exact Matching

```
> 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
```

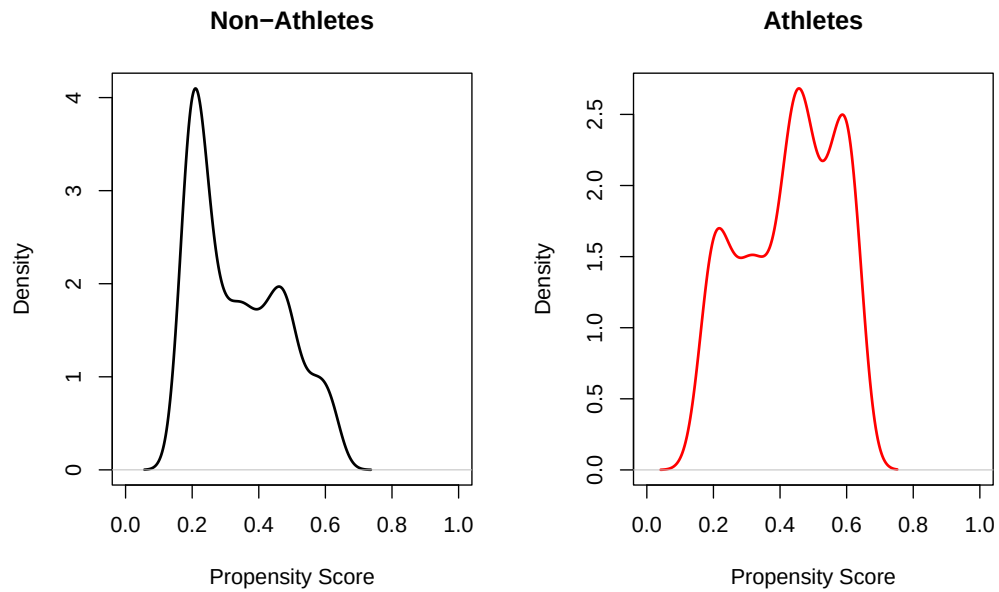
Exact Matching: Balance



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

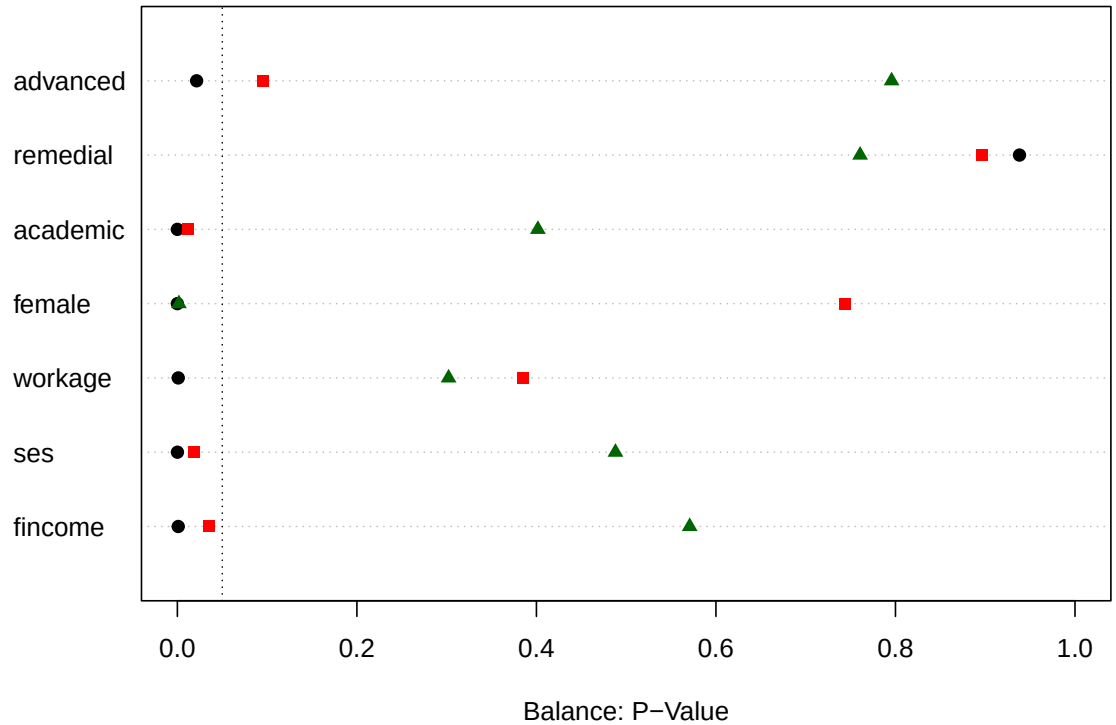
```
> PSfit <- glm(sports~fincome+ses+workage+female+academic+remedial+
+                advanced,data=sports,family=binomial(link="logit"))

> # 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.

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*	3.05*	2.84*	2.75*
	(0.13)	(0.23)	(0.16)	(0.17)
sports	0.10*	0.12*	0.09*	0.08
	(0.04)	(0.06)	(0.04)	(0.05)
fincome	0.00	0.05	−0.00	0.01
	(0.01)	(0.03)	(0.02)	(0.02)
ses	0.02	−0.14	0.05	0.03
	(0.03)	(0.07)	(0.04)	(0.05)
workage	−0.02*	−0.03*	−0.03*	−0.02*
	(0.01)	(0.01)	(0.01)	(0.01)
female	0.30*	0.34*	0.31*	0.29*
	(0.04)	(0.06)	(0.05)	(0.05)
academic	0.29*	0.24*	0.31*	0.31*
	(0.04)	(0.08)	(0.05)	(0.05)
remedial	−0.23*	−0.28*	−0.28*	−0.21*
	(0.04)	(0.06)	(0.05)	(0.05)
advanced	0.44*	0.51*	0.43*	0.40*
	(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	0.25	0.21
N	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)*
  - 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 (2021)
  - Simpson (2019)
  - Xu (2018)
  - Yamamoto (2018)
- Other useful things:
  - The Causal Inference Book
  - Some useful notes

* Especially good.