PLSC 504 – Fall 2022
Endogenous Selection and Potential Outcomes

September 19, 2022

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

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

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

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

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

Define:

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

Then

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

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:

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

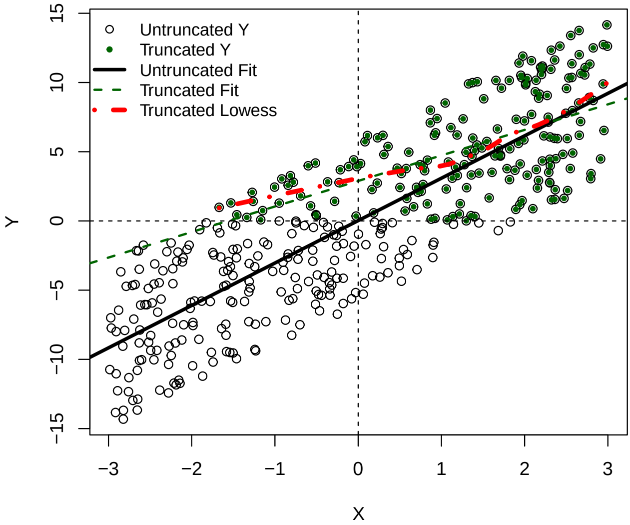
What we get:

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

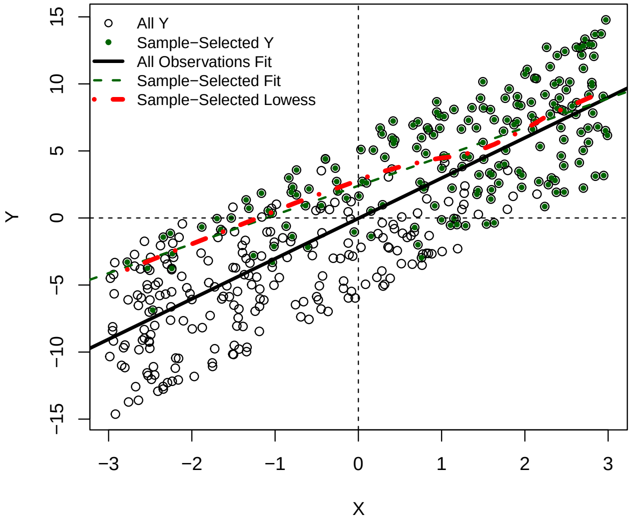
Without conditioning on $\mathbf{Z}$:

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

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
- 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 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

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

- MLE (above)
- Or, reconsider:

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

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

Heckman's Two-Step Estimator

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

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

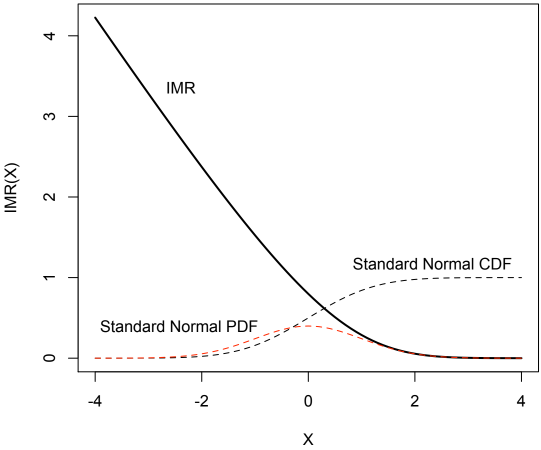
and calculate the estimated inverse Mills' ratio:

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

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?



- 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)

- 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
    - $\implies$ $\mathbf{X}, \mathbf{Z}$ highly correlated
    - ...just makes things worse (Stolzenberg and Relles 1997)

Some Practical Things

- In practice, few people use two-step anymore,
- Sensitive to joint normality of $\{u_i, u_2\}$,
- Very sensitive to model specification...
- Key issue: endogeneity of selection...

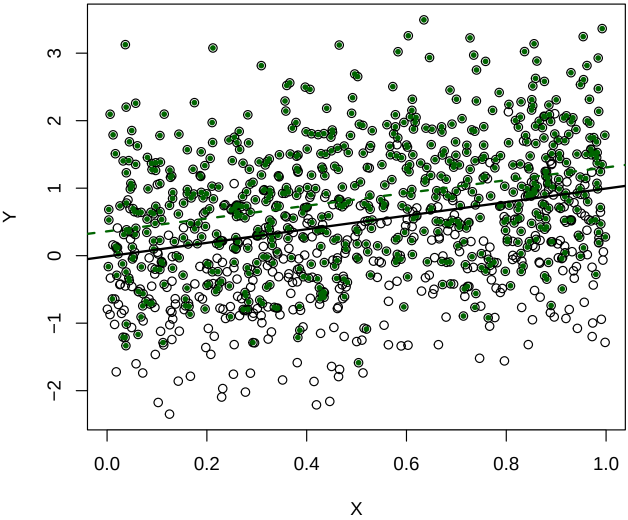
Simulated Example I: $\text{Cov}(X, Z) = 0$

```
> 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")
```

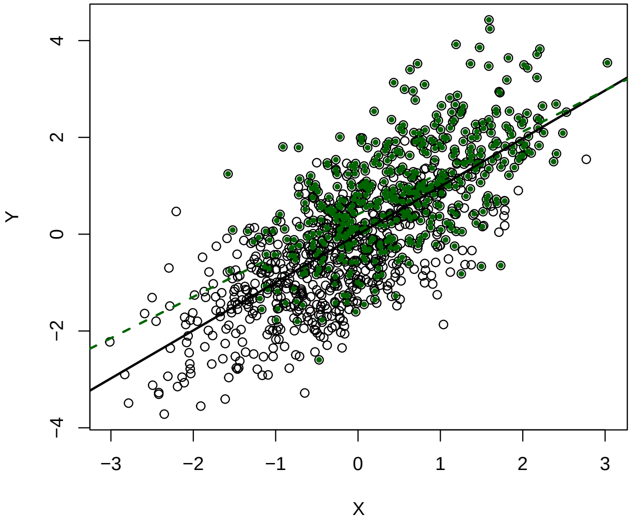
Simulation I (continued)

	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: $\text{Cov}(X, Z) > 0$

```
> 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:

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

- Different outcome stages:
  - Poisson (Greene 1995)
  - Durations (Boehmke et al. 2006)
  - Count/binary/ordinal (Mirand and Rabe-Hesketh 2005)
- Selection stage is <u>ordered</u> (Chiburis & Lokshin 2007)
- Multiple-stage models (not much... work in finance + Signorino and others)

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)
  - `dursel` (binary-duration model)
  - `xtheckman` (selection models for panel data)
  - Also Bayesian versions, using the `bayes:` prefix

Further Readings: 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.

Further Readings: Applications

- Berinsky, Adam J. 1999. "The Two Faces of Public Opinion." *American Journal of Political Science* 43:1209-1230.
- Blanton, Shannon Lindsey. 2000. "Promoting Human Rights and Democracy in the Developing World: U.S. Rhetoric versus U.S. Arms Exports." *American Journal of Political Science* 44:123-131.
- Hart, David M. 2001. "Why Do Some Firms Give? Why Do Some Give a Lot?: High-Tech PACs, 1977-1996." *The Journal of Politics* 63:1230-1249.
- Jensen, Nathan M. 2003. "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment." *International Organization* 57:587-616.
- Jo, Hyeran. 2008. "Taming the Selection Bias: An Application to Compliance with International Agreements." 2008 *Visions in Methodology* conference, Columbus, OH.
- Nooruddin, Irfan. 2002. "Modeling Selection Bias in Studies of Sanctions Efficacy." *International Interactions* 28: 57-74.
- Timpone, Richard J. 1998. "Structure, Behavior and Voter Turnout in the United States." *American Political Science Review* 92: 145-158.
- Vance, Colin, and Nolan Ritter. 2014. "Is Peace a Missing Value or a Zero? On Selection Models in Political Science." *Journal of Peace Research* 51:528-540.
- Von Stein, Jana. 2005. "Do Treaties Constrain or Screen? Selection Bias and Treaty Compliance." *American Political Science Review* 99:611-622.

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*:

  <u>*It is impossible to observe the causal effect of a treatment / predictor on a single unit.*</u>
- Specific challenges:
  - *Confounding*
  - *Selection bias*
  - *Heterogenous treatment effects*

Causal statements imply counterfactual 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:

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

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)

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

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.

[press "pause"]

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.

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

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...

- "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

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

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 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

$$\begin{tikzcd} & \text{Pam} \arrow[dl] \arrow[dr] & \\ \text{Michael} \arrow[rr] & & \text{Dwight} \end{tikzcd}$$

A DAG

$$\begin{tikzcd} & \text{Pam} \arrow[dl] & \\ \text{Michael} \arrow[rr] & & \text{Dwight} \arrow[ul] \end{tikzcd}$$

Not a DAG

DAGs and Confounding

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

No Confounding

$$\begin{tikzcd} & X \arrow[dl] \arrow[dr] & \\ W \arrow[rr, dashed] & & Y \end{tikzcd}$$

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	$(\bar{Y}\mid W=1) - (\bar{Y}\mid W=0)$
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: 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$_{\text{obs}}$	-	8.67	13.33	-	11	4.67
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	-
Mean_{obs}	-	11.33	10.67	-	11	-0.67
Mean_{all}	-	(10)	(12)	(2)	-	-

$t = 0.71,\ p = 0.74$

Next time: How to make causal(-ish) inferences from observational data...