Problem Set 1

Your name here

2025-01-27

0.1 General guidelines

Please answer the questions directly in this document, stating what assumptions you are making if necessary, and attempt the questions even if you are not totally sure of your answers.

You can insert in-line equations using LaTeXnotation like this `$y_i = \alpha + \beta_\tau \tau + \epsilon_i$` and they will show like this: $y_i = \alpha + \beta_\tau \tau + \epsilon_i$. If you want equations to be separated from the main body, just wrap them around double dollar signs instead of one, like this: `$$y_i = \alpha + \beta_\tau \tau + \epsilon_i$$` for the following result:

$$y_i = \alpha + \beta_\tau \tau + \epsilon_i$$

Some exercises require calculations. Unless it is a simple computation that can be made on a physical calculator, you should document them directly in here. To insert `R` code (that will be rendered directly in this document), use the following syntax:

```
# This is R code
# Code and results will be shown in the compiled document

# Set seed for reproduciblity
set.seed(1492)

# Let's made a simple calculation
# assuming the following model
# Y0 ~ Norm(10,2)
# Y1 = Y0 + \tau + epsilon_i
# epsilon ~ Norm(0,1)
# tau = 1.5

n <- 1000
epsilon <- rnorm(n)
```

```r
y0 <- rnorm(n, mean = 10, sd = 2)
y1 <- y0 + 1.5 + epsilon

# Now, let's calculate the ATE
# These two ways are equivalent (why?)

mean(y1-y0)
```

```
[1] 1.560768
```

```r
mean(y1) - mean(y0)
```

```
[1] 1.560768
```

```r
# Now, let's create one possible random assignment vector
# for a binary treatment with p=0.5
d <- rbinom(n, 1, 0.5)

# Let's create the observed outcome variable, using the switching equation
y <- y1*d+(1-d)*y0

# Now, let's approximate the ATE with a difference-in-means
# Think: why cannot we simply use mean(y1-y0 anymore?)
# Think: why does this ATEhat differs from the ATE calculated above?

mean(y[d==1]) - mean(y[d==0])
```

```
[1] 1.439973
```

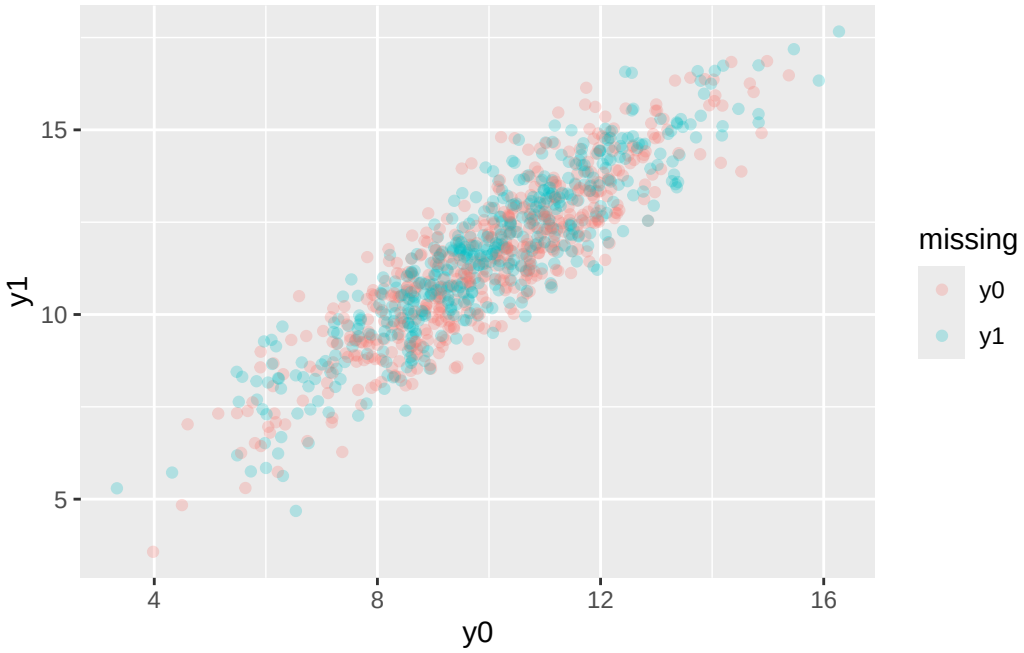
Similarly, plots will be rendered in the final document if you include their code in this document:

```r
# Simple plot example
# First, let's combine our data in a single data.frame

data <- data.frame(y1,y0,y,d)
data$missing <- ifelse(d==1, "y0","y1")

library(ggplot2) # Load ggplot2 package
```

```r
# Create a scatterplot, highlighting observed vs missing outcomes
ggplot(data, aes(x = y0, y = y1, color = missing)) +
  geom_point(alpha = 0.25)
```



For more details, please consult the `Quarto` documentation linked in the syllabus.

1 Potential Outcomes and Experiments

Data from the book needed to answer some questions, and `R` code that could be useful, can be found in https://isps.yale.edu/FEDAI

1.1

Answer Exercises from Chapter 1 in Gerber and Green's *Field Experiments*

Here your answers

1.2

Answer Exercises from Chapter 2 in Gerber and Green's *Field Experiments*

Here your answers.

1.3 Extra credit

Answer Exercises 1 to 8 from Chapter 3 in Gerber and Green's *Field Experiments*

Here your answers

2 Structural Causal Model

Credit: Problems adapted from Onyi Arah's "Logic, causation, and probability".

2.1

Imagine we want to estimate the effect $P(Y|do(1)) - P(Y|do(0)) = P(Y_1) - P(Y_0)$ adjusting for some observed covariates. Please explain both the *backdoor criterion* of covariate selection, and the resulting *adjustment formula* (also known as *backdoor fomula* or *g-fomula*).

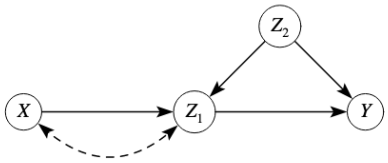
2.2

Please explain the difference between an observational (or pre-intervention) DAG, an interventional DAG, and a testing DAG. Include figures to illustrate the difference.

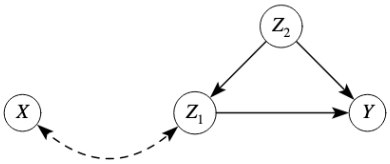
2.3

Now, consider the following DAGs. Assume all name nodes are measured. For each DAG, briefly discuss or justify whether $P(Y|do(1)) - P(Y|do(0))$ is identifiable or not (i.e., whether we can obtain a unique solution or answer to our causal query if we had infinite data on the measure variables on the DAGs).

Figure 1: DAG (a)



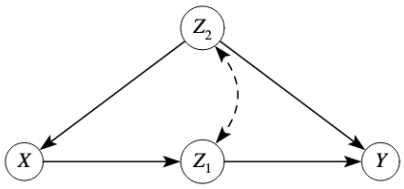
(A) A pre-intervention DAG G



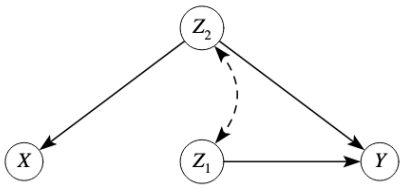
(B) A test DAG $G_{\underline{X}}$

2.4

Figure 2: DAG (b)



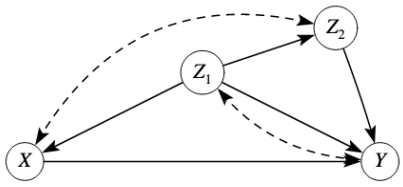
(A) A pre-intervention DAG G



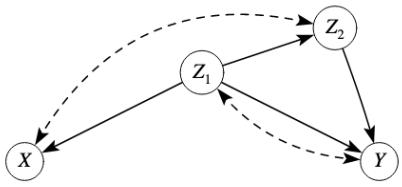
(B) A test DAG $G_{\underline{X}}$

2.5

Figure 3: DAG (c)



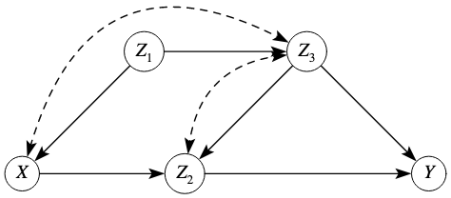
(A) A pre-intervention DAG G



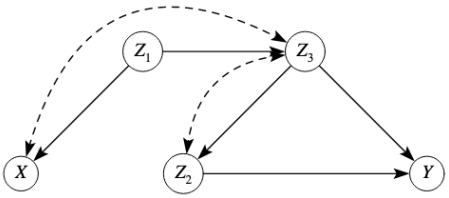
(B) A test DAG $G_{\underline{X}}$

2.6

Figure 4: DAG (d)



(A) A pre-intervention DAG G

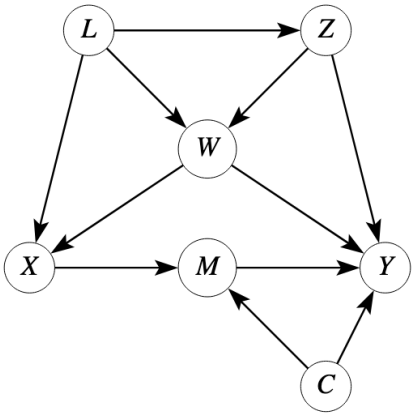


(B) A test DAG $G_{\underline{X}}$

2.7

Consider DAG 2 below

Figure 5: DAG 2



a. Please write out (factorize) the joint probability distribution for the nodes in DAG 2.

b. Please write out the (non-parametric) structural equations according to DAG 2. Define any variables you introduce.

c. Write an empirical analogue of the average treatment effect of X on Y based on DAG 2. Assume that all variables are binary in this question.

d. Now consider intervening on Z so that $Z = z^*$. Please show the intervention DAG, and rewrite the structural equations from part b to reflect this intervention scenario.