Problem Set 2

FILL IN YOUR NAME

Conceptual Problems

1. What is the conditional independence assumption?
2. What is a bad control?

Applied Problems

1. Consider the following dataset, which is the result of a set of estimated treatment effects for nine studies. The designs were reported without covariates (ATE_NoCovars) and with covariate adjustment (ATE_YesCovars).

```
data_cp3 <- tibble(
  study = 1:9,
  ATE_NoCovars = c(5,3,2,6,1,0, -3, -5, 0),
  ATE_YesCovars = c(4,3,2,5,1,0,-1,-4, -1)
)
```

a) Verify that the ATE is 1 in both the No Covariates and Covariates groups either with R or by hand.

b) Suppose researchers use the following decision rule when deciding which estimate to report, “Estimate the ATE using both estimators and report whichever estimate is larger.” Under this rule, are the reported estimates unbiased? Why or why not?

2. Effective Weights and Randomized Assignment

Consider the following dataset. Calculate the effective weights of regression of Y on D, X_1, X_2,and X_3. Plot the effective weights on a graph. Include a vertical line at .05.

```
set.seed(8675309)
data_a2 <- tibble(
  N = 1000,
  D = sample(0:1, 1000, replace = T),
  X1 = rnorm(1000, 12, 4),
  X2 = runif(1000, 10, 100),
  X3 = rnorm(1000, 10, 5),
  Y = 1.5*D + 2*X1 + X2 + 5*X3 + rnorm(1000)
)
```

Characterize the resulting graph. Do any points fall on or below 0.05? What does this tell us about the benefits of randomization?

3. Regression Mechanics with Covariate adjustment.

Write a function called `fwl` that performs the following steps:

a) Regresses a treatment variable on all covariates and computes the residual for each observation. For this problem, you should use `lm_robust()` to calculate the residuals. A residual is the difference between an observed value, and the **fitted values** of the regression. Store the values in a variable called `ex`

b) Regress the outcome variable Y on all covariates, not including treatment assignment, and compute the residual for each observation. Store the values in a variable called `ey`

c) Regress the residuals found in part (b) on the residuals found in part (a). Store the ex **coefficient** from the regression estimate in a variable called `resid_coef`

d) Run a regression of Y on D and all covariates. Store the treatment **coefficient** in a variable called `treat_coef`

e) Return the result of the Boolean comparison of treat_coef and resid_coef rounded to four decimal places.

Run your function with the following dataset where Y is the outcome variable, D is the treatment variable, and X1, X2, and X3 are additional covariates. The full regression is of the form

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \epsilon$$

What result do you get?

```r
set.seed(42)
data_a3 <- tibble(
  N = 1000,
  D = sample(0:1, 1000, replace = T),
  X1 = rnorm(1000),
  X2 = runif(1000),
  X3 = rnorm(1000, 10, 5),
  Y = 2*D + 4*X1 + X2 + .5*X3 + rnorm(1000)
)
```

4. Regression and Difference in Means

Using the following dataset

```r
set.seed(720)
data_a4 <- fabricate(
  N = 1000,
  Y = rnorm(N, 100, 10),
  D = complete_ra(N)
)
```

a) Either apply a function from a package we use in this class or write your own difference in means function. If you do the latter, you are not required to calculate standard errors. For either option, compute the difference in means estimator of Y on D.

b) Using a robust estimator, run the regression of Y on D. Are your results the same? Is this reasonable based on what we've discussed in class and why?

5. Regression and Block Randomization

Using the following dataset, based on an actual Get out the Vote (GOTV) experiment.

```r
set.seed(22)
data_a5 <- fabricate(
  N = 10000,
  in_poverty = sample(c(rep(1,5000), rep(0,5000)), N, replace = F),
  college_grad = sample(c(rep(0,5000), rep(1,5000)), N, replace = F),
  # Code Democrats as 1 and Republicans as 0
  partyID = sample(c(rep(1,5000),rep(0,5000)), N, replace = F),
  Y0 = rnorm(N, mean = (2*college_grad + -5*in_poverty + 4*partyID),sd = 5),
  Y1 = Y0 + 15,
  D = sample(c(rep(0,5000), rep(1,5000)), N, replace = F),
  Y = if_else(D == 1, Y1, Y0)
)
```

a) Run a robust regression without taking into account potentially predictive covariates.

b) Run a robust regression taking into account all predictive covariates you think matter.

c) In the latter regression, which of the coefficients has a causal interpretation? Why?

d) Based on this problem, what do you observe to be the advantage of covariate adjustment?

6. The following is based on data drawn from Montegomery, Nyhan, and Torres (2018).

The survey experiment was conducted as follows. Subjects signed a consent form to participate. They were then asked to provide demographic information, including their age, gender, and party affiliation. Following the demographic questions, subjects were randomized into a treatment condition that provided information about a judge and whether a politician endorsed them. The control condition did not include an endorsement. Subjects were then asked whether they approved of the judge or not, which was measured on a scale of 1-4. After questions about approval, subjects were asked to evaluate on a scale of 1-7 how ideological the judge was likely to be.

a) Using regression, estimate the treatment effect with appropriate covariates?

b) Show what happens if you include an inappropriate covariate into the regression model. Why does this occur based on the information given in the problem description?