Problem Set 2 Solutions

Conceptual Problems

1. What is the conditional independence assumption?

The CIA asserts that conditional on observed characteristics, selection bias disappears. Formally $\{Y_{0i}, Y_{1i}\} \perp D_i | X_i$ *where* X_i *is an observed characteristic. The CIA allows us to give our regression a causal interpretation. It is trivially satisfied when* D_i *is known to be randomly assigned, either in general or conditional on covariates.*

2. What is a bad control?

A bad control is a variable that is itself an outcome variable in the notional or actual experiment at hand. It is a variable that is measured after treatment has occurred.

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)
)
```

```

data_cp3 %>%
  summarise(NC_avg = mean(ATE_NoCovars),
            C_avg = mean(ATE_YesCovars))
```

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

```
## # A tibble: 1 x 2
##   NC_avg C_avg
##    <dbl> <dbl>
## 1      1     1
```

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? *A way to think about this problem is to generate a new variable that implements this rule and then compare the average of that variable to the true ATE of 1.*

```
data_cp3 %>%
  mutate(reported = if_else(ATE_NoCovars > ATE_YesCovars, ATE_NoCovars, ATE_YesCovars))%>%
  summarise(report_avg = mean(reported))
```

```
## # A tibble: 1 x 1
##   report_avg
##        <dbl>
## 1       1.33
```

As can be seen, we get biased estimates. Even though it is the case that each estimator is unbiased, the greater of two unbiased estimates is not unbiased. The average ATE of the reported group is 1.33, which is greater than the true ATE of 1.

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)
)
```

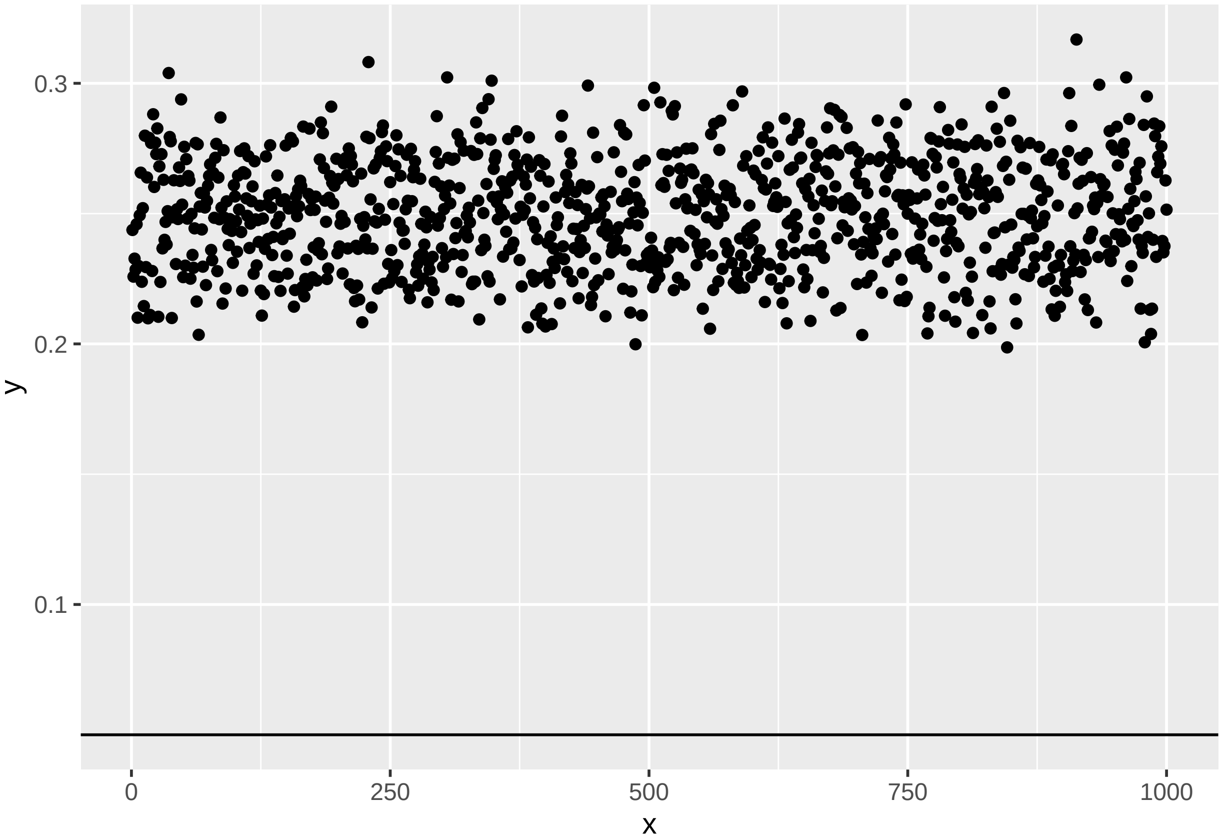
Characterize the resulting graph. Do any points fall on or below 0.05? What does this tell us about the benefits of randomization?

We can apply our Effective Weights function from section.

```
effectiveWeights <- function(Y, controls, data){
  # Make the OLS formula call
  treat_formula <- reformulate(termlabels = c(controls),
                               response = Y)
  # Run a regression of the treatment on the controls
  treat.model <- lm(as.formula(treat_formula), data = data)
  # Extract the residuals from that regression
  d.tilde <- as.numeric(residuals(treat.model))
  # Square the residuals.
  weights <- d.tilde^2
  return(weights)
}

w <- effectiveWeights("D", controls = c("X1","X2", "X3"), data = data_a2)

tibble(x = 1:1000, y = w)%>%
  ggplot(aes(x, y))+
  geom_point()+
  geom_hline(yintercept = 0.05)
```



As explained in lecture, with randomization, the data points for the regression contribute equally to the regression up to some disturbance terms. The graph made shows a uniform distribution of weights across all points. No points are below the line, which is at a weight where we might start to worry about overlap problems.

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

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

What result do you get?

```
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)
)
```

```
## A way to solve this
fwl <- function(Y, D, controls, data){
  a_form <- reformulate(termlabels = controls,
                        response = D)
  b_form <- reformulate(termlabels = c(controls),
                        response = Y)
  d_form <- reformulate(termlabels = c(D, controls),
                        response = Y)
  m1 <- lm_robust(as.formula(a_form), data = data)
  ex <- data[[D]] - m1$fitted.values

  m2 <- lm_robust(as.formula(b_form), data = data)
  ey <- data[[Y]] - m2$fitted.values

  m3 <- lm_robust(ey ~ ex)
  resid_coef <- unname(m3$coefficients[2])

  m4 <- lm_robust(as.formula(d_form), data = data)
  treat_coef <- unname(m4$coefficients[2])

  return(round(treat_coef,4)==round(resid_coef,4))
}

fwl("Y", "D", controls = c("X1", "X2", "X3"), data = data_a3)
```

```
## [1] TRUE
```

This function is demonstrating the Frisch-Waugh-Lovell Theorem

4. Regression and Difference in Means

Using the following dataset

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

```

## If you've looked up the help documents on estimatr
difference_in_means(Y ~ D, data = data_a4)%>%
  tidy()
```

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.

```
##   term   estimate std.error statistic   p.value  conf.low conf.high       df
## 1    D -0.7431616 0.6319082 -1.176059 0.2398518 -1.983184 0.4968606 997.3927
##   outcome
## 1       Y
```

```
## Alternatively calculating by hand
diff_means <- function(Y, D, data){
  Y1 <- mean(data[[Y]][data[[D]]==1], na.rm = T)
  Y0 <- mean(data[[Y]][data[[D]]==0], na.rm = T)
  return(Y1 - Y0)
}

diff_means("Y", "D", data = data_a4)
```

```
## [1] -0.7431616
```

```

lm_robust(Y~D, data = data_a4)%>%
  tidy()
```

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?

```
##          term    estimate std.error  statistic   p.value  conf.low   conf.high
## 1 (Intercept) 100.4960787 0.4412795 227.737929 0.0000000 99.630137 101.3620208
## 2           D  -0.7431616 0.6319082  -1.176059 0.2398516 -1.983183   0.4968597
##    df outcome
## 1 998       Y
## 2 998       Y
```

You should get the same answer. As we discussed in lecture, a regression is equivalent to the difference of means estimator in the bivariate case.

5. Regression and Block Randomization

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

```
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)
)
```

```

lm_a <- lm_robust(Y ~ D , data = data_a5, se_type = "stata")

lm_a %>% tidy()
```

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

```
##          term   estimate  std.error statistic      p.value   conf.low
## 1 (Intercept)  0.3950622 0.08511062   4.64175 3.498689e-06  0.2282282
## 2           D 15.1336634 0.12007439 126.03572 0.000000e+00 14.8982934
##    conf.high   df outcome
## 1  0.5618961 9998       Y
## 2 15.3690334 9998       Y
```

```
lm_out <- lm_robust( Y~ D+in_poverty + college_grad + partyID, data = data_a5, se_type = "stata")

lm_out %>% tidy()
```

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

```
##           term   estimate  std.error  statistic      p.value   conf.low
## 1  (Intercept) -0.1147887 0.11043186  -1.039452 2.986195e-01 -0.3312574
## 2            D 15.0208613 0.09892359 151.843071 0.000000e+00 14.8269512
## 3   in_poverty -5.0325584 0.09889859 -50.886049 0.000000e+00 -5.2264195
## 4 college_grad  2.1006275 0.09889261  21.241501 5.666788e-98  1.9067780
## 5      partyID  4.0644347 0.09892272  41.086969 0.000000e+00  3.8705262
##   conf.high   df outcome
## 1  0.101680 9995       Y
## 2 15.214771 9995       Y
## 3 -4.838697 9995       Y
## 4  2.294477 9995       Y
## 5  4.258343 9995       Y
```

c) In the latter regression, which of the coefficients has a causal interpretation? Why? *The only coefficient that ever has a causal interpretation is the treatment coefficient. All other parameters are nuisance parameters that exist to increase precision. In this specific case, we are blocking on a pre-treatment covariate of interest because we think it is relevant, but the blocks do not themselves have any interpretation. As a general rule, every regression has only one coefficient that could matter from a causal perspective. If you want another one, that's another research project.*

d) Based on this problem, what do you observe to be the advantage of covariate adjustment? *The coefficient estimate gets closer to the true ATE of 15. More importantly, the standard errors around the treatment coefficient get smaller with covariate adjustment with pre-treated variables.*

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.

```
lm_robust(approval ~ endorsement + age, data = judges)%>%tidy()
```

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

```
##          term    estimate  std.error statistic      p.value     conf.low
## 1 (Intercept)  2.21250925 0.13328859 16.599390 1.420022e-41  1.949927533
## 2 endorsement -0.19276688 0.10140184 -1.901019 5.851261e-02 -0.392530946
## 3         age  0.09113549 0.04780957  1.906219 5.783154e-02 -0.003050504
##     conf.high  df  outcome
## 1 2.475090977 237 approval
## 2 0.006997191 237 approval
## 3 0.185321488 237 approval
```

The estimated treatment effect of the endorsement is approximate a -.19 shift in approval. Including age is helpful for the precision of the estimate because it is pre-treatment

```
lm_robust(approval ~ endorsement + age + ideology, data = judges)%>% tidy()
```

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?

```
##          term    estimate  std.error  statistic      p.value    conf.low
## 1 (Intercept)  2.47943555 0.13965587 17.7538942 2.312536e-45  2.20430415
## 2 endorsement -0.05857954 0.09439898 -0.6205527 5.354928e-01 -0.24455183
## 3         age  0.11722398 0.04473673  2.6203072 9.355145e-03  0.02908963
## 4    ideology -0.20891886 0.03148898 -6.6346668 2.203111e-10 -0.27095425
##    conf.high  df  outcome
## 1  2.7545670 236 approval
## 2  0.1273928 236 approval
## 3  0.2053583 236 approval
## 4 -0.1468835 236 approval
```

The inappropriate variable is ideology. The problem description shows that it is measured after the treatment is implemented. That makes it a post-treatment variable. Here is biases our estimates upward by .15 units.