Causal Inference
Problem Set 2

Due: April 18th

Problem 1

This problem provides a quick review of the basic potential outcomes notation we discussed in class.

a) For a binary treatment $D \in \{0, 1\}$, outcome variable Y, and individual units indexed by i, what is the meaning of Y_{1i} and Y_{0i}? Describe both in words, and choose an example to illustrate.

The quantity Y_{1i} designates the outcome unit i would have on variable Y *were it to receive the treatment* (regardless of whether it actually receives the treatment). Similarly, Y_{0i} is the non-treatment outcome for unit i, regardless of whether or not it got the treatment. For example, Y_{1i} could be the number of papers student i will publish during her career if she does take 350B, whereas Y_{0i} is the number of publications student i would have without taking the class (clearly $Y_{0i} \ll Y_{1i}$, $\forall i$).

b) What is the difference between Y_{1i} and "Y_i for a unit that actually received the treatment"? Explain the difference using the example you began in part (a).

The quantity "Y_i for a unit that actually recieved the treatment" represents an observed outcome for unit i under treatment. However Y_{i1} is the "what-if" outcome for unit i whether it received the treatment for not in reality. For our example, Y_i for a student who took the class may be 42 papers, but Y_{j1} for other students j would exist as a hypothetical quantity even for those students j who did not take the class.

c) Define the average treatment effect (ATE) and average treatment effect among the treated (ATT) using potential outcomes notation. Describe in words what each quantity means.

The ATE is $E[Y_1 - Y_0]$, which compares the hypothetical outcome under treatment and under non-treatment for *each unit* i, and computes an expectation across these units. The ATT is similar, but only averages together the individual treatment effects among those who actually happened to take the treatment, $E[Y_1 - Y_0|D = 1]$.

d) When will the ATT and the ATE be equal to each other? Prove it.

Let's see when the ATE and ATT are the same:

$$\begin{aligned} ATE &= ATT \\ E[Y_1 - Y_0] &= E[Y_1 - Y_0|D = 1] \\ E[Y_1] - E[Y_0] &= E[Y_1|D = 1] - E[Y_0|D = 1] \end{aligned}$$

This is guaranteed to be true when $E[Y_1] = E[Y_1|D = 1]$ and $E[Y_0] = E[Y_0|D = 1]$. This happens when $Y_1 \perp D$ and $Y_0 \perp D$. Such is the case under randomized experiments, where we ensure that both Y_1 and Y_0 are independent of D.

e) Show formally that $\tau_{ATE(x)} = E[Y_1 - Y_0|X = x]$ (i.e. a subgroup average treatment effect) is identifiable given random assignment.

$$\begin{aligned} ATE_X &= E[Y_1 - Y_0|X = x] \\ &= E[Y_1|X = x] - E[Y_0|X = x] \\ &= E[Y_1|X = x, D = 1] - E[Y_0|X = x, D = 0] \text{ by random assignment of D} \end{aligned}$$

Since the last line can be estimated from the data, identification has been achieved.

$$\begin{aligned} &E[Y_1|X = x, D = 1] - E[Y_0|X = x, D = 0] \\ &= E[Y|X = x, D = 1] - E[Y|X = x, D = 0] \end{aligned}$$

There are two reasons to have you go through this. The first is that it shows you can estimate the ATE within various sub-groups, and it will still be identified. The second is that while you have here a different quantity for every choice of X, you could also take the expectation of ATE_X over these different quantities of X, and get back to a single number. When we get to blocking, this is effectively what you are doing. Being in a particular block is like choosing $X = x$. You get a treatment effect at that block, and then compute the average such effect over all the blocks.

Problem 2

For this problem we will use data from Benjamin A. Olken. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy.* 115: 300-249. The paper and the data set are available on Piazza.

The objective of this experiment was to evaluate two interventions aimed at reducing corruption in road building projects in Indonesian villages. One treatment was audits by engineers; the other was

encouraging community participation in monitoring. This problem focuses on the latter intervention, which consisted of inviting villagers to public meetings where project officials accounted for budget expenditures. The main dependent variable is *pct_missing*, a measure of the difference between what the villages claimed they spent on road construction and an independent estimate of what the villages actually spent. Treatment status is indicated by the dummy variable *treat_invite*, which takes a value of 1 if the village received the intervention and 0 if it did not.

The variables in the data set are:

- *pct_missing*: Percent expenditures missing
- *treat_invite*: Treatment assignment
- *head_edu*: Village head education
- *mosques*: Mosques per 1,000
- *pct_poor*: Percent of households below the poverty line
- *total_budget*: Total budget (Rp. million) (determined prior to intervention)

a) Estimate the average treatment effect in this new dataset, using the difference in means estimator.

```
> library(dplyr)
> dataolk <- read.csv("olken_data.csv")
> dataolk <- filter(dataolk, !is.na(pct_missing))
>
> SATE <- mean(dataolk$pct_missing[dataolk$treat_invite==1]) -
+               mean(dataolk$pct_missing[dataolk$treat_invite==0])
> SATE
[1] -0.02314737
```

b) Derive an expression for a conservative estimator of the standard error of the above difference-in-means. Why is this estimator conservative? The conservative estimator is simply the traditional estimator of the variance of the difference in means estimator under the assumption that all observations are iid. Recall that this is something we used and derived early on last quarter.

$$\begin{aligned}
\widehat{var}(\bar{Y}_{treated} - \bar{Y}_{control}) &= \widehat{var}(\bar{Y}_{treated}) + \widehat{var}(\bar{Y}_{control}) \\
&= \frac{\hat{\sigma}_T^2}{N_T} + \frac{\hat{\sigma}_C^2}{N_C} \\
\widehat{se}(\bar{Y}_{treated} - \bar{Y}_{control}) &= \sqrt{\widehat{var}(\bar{Y}_{treated} - \bar{Y}_{control})}
\end{aligned}$$

where $\hat{\sigma}_T^2$ and $\hat{\sigma}_C^2$ are the population variance estimators with $n-1$ in the denominator.

See the lecture slides for proof of the conservatism of this estimator relative to the true variance.

c) Use the data to estimate the standard error you derived in (b).

```
> SE <- sqrt( var(dataolk$pct_missing[dataolk$treat_invite==1])/
+                 sum(dataolk$treat_invite==1) +
+               var(dataolk$pct_missing[dataolk$treat_invite==0])/
+                  sum(dataolk$treat_invite==0) )
> SE
[1] 0.03285978
```

d) Check the covariate balance in this dataset on all covariates (all variables that are not the treatment assignment or the outcome). Decide on sensible balance test statistics and report them in a table. How do the treatment and control group differ?

```
> d<-dataolk
> head(d)
  treat_invite pct_missing head_edu   mosques  pct_poor total_budget
1            0  0.38527447        6 0.9083831 0.4001222     40.56500
2            1 -0.09574836       14 1.0666667 0.1856149     69.32150
3            1  0.14771932       12 0.7117438 0.4000000     41.10650
4            1 -0.18259122        9 0.9489917 0.4379366     17.06200
5            0 -0.29304767        9 1.6233766 0.3126954     72.08600
6            1 -0.09736358        9 0.7381890 0.4076087     69.83308
>
> covars<-c("head_edu","mosques","pct_poor", "total_budget")
>
> balance<-as.data.frame(matrix(nrow=length(covars), ncol=3))
> colnames(balance)<-c("control","treatment","p.val.diff")
> rownames(balance)<-covars
>
> for(i in 1:length(covars)){
+
+ t<-t.test(d[d$treat_invite==0,covars[i]],d[d$treat_invite==1,covars[i]], na.rm=T )
+ balance[covars[i],"control"]<-t$estimate[1]
+ balance[covars[i],"treatment"]<-t$estimate[2]
+ balance[covars[i],"p.val.diff"]<-t$p.value
+
+ }
```

```
>
> balance
                control  treatment p.val.diff
head_edu     11.5026178 11.4339623  0.7758529
mosques       1.4738225  1.4121246  0.4051287
pct_poor      0.4051651  0.4139121  0.6398512
total_budget 81.9831580 80.2227790  0.6738513
```

The covariates appear well balanced across the treatment and control groups.

e) Compare the number of treated to untreated units. Comment on the result, and whether it is good or bad in your view.

There are more treated than control units. This will not bias our estimates because the treatment was randomly assigned. However, all else equal, the maximally efficient allocation would be a 50/50 split, so this may harm the precision of our estimates.

```
> table(d$treat_invite)
  0   1
191 376
```

f) Now use regression to estimate the $SATE$ (sample average treatment effect). Is this estimate different from the difference-in-means estimate?

```
> library(lmtest)
> library(sandwich)
> regmod1 <- lm(pct_missing ~ treat_invite, data=dataolk)
> coeftest(regmod1, vcov = vcovHC(regmod1,"HC2"))

t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.252106   0.026393  9.5521   <2e-16 ***
treat_invite -0.023147   0.032860 -0.7044   0.4815
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The estimate of the SATE is exactly the same. However, depending upon the covariance matrix specified for the regression, the estimated standard error is different.

g) Using your answer from part (c) (not from part (f)), conduct a t-test of the null hypothesis that $SATE = 0$. You may use a normal approximation to determine the critical value. Choose an α

level and whether you want to conduct a one-sided or two-sided test, based on what you think is sensible.

Using the conventional α of 0.05, a two-sided test, and the normal approximation, the critical value is 1.96.

$$t = \frac{-0.02314737}{0.03285978} = -0.7044286$$

$|t|$ ¡ 1.96. Thus, fail to reject the null that $SATE = 0$.

h) Is the conventional estimated standard error of the OLS estimate of the $SATE$ (i.e. $\sqrt{\hat{\sigma}_u^2 (X'X)^{-1}_{(2,2)}}$) different than the estimated standard error of the difference-in-means estimate? Why or why not?

The conventional standard error of the OLS estimate is different than the standard error calculated above because the former assumes equal variances across groups, while the latter allows variance to differ by treatment status.

i) Re-estimate the $SATE$ using three additional regression models: (1) one in which you include all pre-treatment covariates as additional linear predictors, (2) another in which you include arbitrary functions of the covariates (polynomials, logs, interactions, etc.) as additional linear predictors, and (3) a third in which you include demeaned versions of the covariates $(X_i - \bar{X})$ as well as the interactions between each of them and the treatment. Report the treatment effect estimates and their robust standard errors. How do these results vary across the regressions?

```
> reg2<-lm(pct_missing~treat_invite+head_edu+mosques+pct_poor+total_budget,data=a)
> summary(reg2)
\end{Sinput}
\begin{Soutput}
> regmod1 <- lm(pct_missing ~ treat_invite, data=dataolk)
>
> regmod2 <- lm(pct_missing ~ treat_invite + head_edu + mosques + pct_poor +
total_budget, data=dataolk)
>
> regmod3 <- lm(pct_missing ~ treat_invite + head_edu + mosques*pct_poor +
total_budget + I(total_budget^2), data=dataolk)
>
> dataolk$head_edu.dem <- dataolk$head_edu - mean(dataolk$head_edu,na.rm=T)
> dataolk$mosques.dem <- dataolk$mosques - mean(dataolk$mosques,na.rm=T)
> dataolk$pct_poor.dem <- dataolk$pct_poor - mean(dataolk$pct_poor,na.rm=T)
> dataolk$total_budget.dem <- dataolk$total_budget - mean(dataolk$total_budget,na.rm=T)
>
```

```
> regmod4 <- lm(pct_missing ~ treat_invite + treat_invite*head_edu.dem +
+                 treat_invite*mosques.dem + treat_invite*pct_poor.dem +
+                 treat_invite*total_budget.dem, data=dataolk)
>
> coeftest(regmod1, vcov = vcovHC(regmod1,"HC2"))["treat_invite",]
   Estimate  Std. Error     t value    Pr(>|t|)
-0.02314737  0.03285978 -0.70442856  0.48151114
> coeftest(regmod2, vcov = vcovHC(regmod2,"HC2"))["treat_invite",]
   Estimate  Std. Error     t value    Pr(>|t|)
-0.02641825  0.03263124 -0.80959992  0.41858335
> coeftest(regmod3, vcov = vcovHC(regmod3,"HC2"))["treat_invite",]
   Estimate  Std. Error     t value    Pr(>|t|)
-0.02202635  0.03250039 -0.67772557  0.49828351
> coeftest(regmod4, vcov = vcovHC(regmod4,"HC2"))["treat_invite",]
   Estimate  Std. Error     t value    Pr(>|t|)
-0.02696935  0.03276203 -0.82318934  0.41082503
```

Point estimates for the treatment effects are nearly identical across functional forms. In this case, standard errors are also virtually unchanged, though that's not always the case.

j) Show formally why the variance differs between the controlled estimate of the treatment effect (i.e. the estimate from the regression including pre-treatment covariates) and the uncontrolled estimate. To do this, use a simplified setup in which you compare the estimator of the variance of $\hat{\beta}_1$ in the model $y_i = \alpha + \beta_1 D_i + \epsilon_{1i}$ to the estimator of the variance of $\hat{\beta}_1$ in the model $y_i = \alpha + \beta_1 D_i + \beta_2 X_{1i} + \epsilon_{2i}$.

Recall that $var(\hat{\beta}_{OLS}) = \sigma^2_\epsilon (X'X)^{-1}$. Then, for the model without covariates,

$$\widehat{var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2_{\epsilon_1}}{\sum_i (D_i - \bar{D}_i)^2}$$

While, in the model with covariates:

$$\widehat{var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2_{\epsilon_2}}{(1 - R^2_D) \sum_i (D_i - \bar{D}_i)^2}$$

where R^2_D is the R^2 from regressing the treatment indicator on all the other covariates.

Thus covariates exert two opposing effects on the variance of β_1. If the covariates effectively explain change in the outcome variable, then the residuals will be much smaller, and thus $\hat{\sigma}^2_2 < \hat{\sigma}^2_1$, reducing the overall variance of the coefficient. On the other hand, if the included covariates

happen to be highly predictive of the treatment variable—by chance, of course, since treatment is randomized—then including covariates can increase the variance of $\hat{\beta}_1$ (because of the $(1 - R_D^2)$ term in the denominator). Finally, including covariates will also affect the degrees of freedom in the estimate of the error variance, which in an extreme case where the covariate explains none of the variation in the outcome could cause the estimate to increase.

Problem 3

- Consider the fictional data set: POdata_2.csv. In these fictional data, we observe an outcome for each unit both under treatment and under control (which, again, is usually impossible in the real world). In addition, we now have a covariate X. This problem uses the data to explore the assignment mechanism and blocking.

  Define a treatment vector as the $N \times 1$ vector T that contains each unit's treatment status. Consider the following experimental designs:

  1. Each unit i is treated if $Y_{1i} > 100$ and not treated otherwise.
  2. Each unit i has probability of receiving treatment, $Pr(T_i = 1) = 0.5$.
  3. Treatment is randomly assigned, with $N/2$ units fixed to appear in treatment and control groups.
  4. First, each unit is paired with the other unit with an identical X value. Then, within each pair, treatment assignment is randomized: one unit is treated and the other is not treated, with each possible outcome having equal probability.

  To answer the following questions, set the seed at 2 in **R**.

  (a) For each of these designs:

  (i) Compute the number of potential treatment vectors T.

  Design 1: one potential treatment vector
  Design 2: this requires ordered counting with replacement, i.e. each i's treatment assignment can take two values (one or zero), thus $2 \times 2 \times ... = 2^{1000}$
  Design 3: now the number of treated units is fixed (i.e., N/2), thus $\binom{N}{\frac{N}{2}} = \binom{1000}{500} = \frac{1000!}{500!(1000-500)!}$

Design 4: Each pair of i's with the same X have a treatment assignment can take two values (TC or CT), thus 2^{500}

(ii) Compute the probability of obtaining a particular treatment vector, and define the probability that any given i receives the treatment.

	$Pr(T = t)$	$Pr(T_i = 1)$
Design 1	1	1 if $Y_{1i} > 100$, and 0 otherwise
Design 2	$\frac{1}{2^{1000}}$	0.5
Design 3	$\frac{500!(1000-500)!}{1000!}$	0.5
Design 4	$\frac{1}{2^{500}}$	0.5

(iii) Implement the treatment assignment mechanism, and estimate $E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$. For each design: Is your estimator unbiased for the true ATE? If not, calculate the bias. How does the realization of the estimate compare to the true ATE using the individual-level potential outcomes?

ATE using individual-level potential outcomes

```
> ATE <- mean(data$Treat) - mean(data$Control)
[1] 1.941756
```

Design 1: Potential outcomes are related to the assignment by definition, so $E[Y_i|T_i = 1] - E[Y_i|T_i = 0] \neq ATE$. This is because the treatment assignment is correlated with the potential outcomes.

```
> set.seed(2)
> data$Tr1 <- 0
> for (i in 1:nrow(data)){
+   data$Tr1[i] <- if(data$Treat[i]>100){1}else{0}
+ }
> ATE1 <- mean(data$Treat[data$Tr1==1]) - mean(data$Control[data$Tr1==0])
[1] 51.95164
```

The bias term can be recovered exactly by calculating:
$Bias = (E[Y_{0i}|T_i = 1] - E[Y_{0i}|T_i = 0]) + (1 - \pi)\{E[\tau|T_i = 1] - E[\tau|T_i = 0]\}$:

```
> pi <- sum(data$Tr1==1)/nrow(data)
> bias <- (mean(data$Control[data$Tr1==1]) - mean(data$Control[data$Tr1==0]))
+   (1-pi)*(mean(data$Treat[data$Tr1==1] - data$Control[data$Tr1==1]) -
+             mean(data$Treat[data$Tr1==0] - data$Control[data$Tr1==0]))
> bias
[1] 50.00988
```

```
> ATE + bias
[1] 51.95164
> ATE1
[1] 51.95164
```

Design 2: Should be an unbiased estimate of the ATE because of randomization. The specific realization of the treatment vector draw places the estimated ATE somewhat higher than the "true" sample ATE. This happens because of the specific realization of the treatment vector.

```
> data$Tr2 <- sample(0:1, size=1000, replace=TRUE)
> ATE2 <- mean(data$Treat[data$Tr2==1]) - mean(data$Control[data$Tr2==0])
[1] 2.977697
```

Design 3: Should be an unbiased estimate of the ATE.

```
> K <- sample(1:1000, size=500, replace=FALSE)
> data$Tr3 <- 0
> data$Tr3[K] <- 1
> ATE3 <- mean(data$Treat[data$Tr3==1]) - mean(data$Control[data$Tr3==0])
[1] 2.180979
```

Design 4: Should be an unbiased estimate of the ATE.

```
> data <- ddply(data, .(X), transform, Tr4=sample(0:1, replace=FALSE))
> ATE4 <- mean(data$Treat[data$Tr4==1]) - mean(data$Control[data$Tr4==0])
[1] 1.9785
```

(b) Estimate the ATE for design 4 using a bivariate OLS regression and report the conservative estimate for the standard errors. Can you reject the null that the ATE is zero? How can efficiency be improved?

```
> data$outcome4 <- data$Treat*data$Tr4 + data$Control*(1-data$Tr4)
>
> library(lmtest)
> library(sandwich)
> modd4 <- lm(outcome4~Tr4, data=data)
> coeftest(modd4, vcov = vcovHC(modd4,"HC2"))["Tr4",]
  Estimate Std. Error    t value   Pr(>|t|)
 1.9784999  1.8336087  1.0790197  0.2808397
```

The ATE is exactly the same as in (a), where a difference of means was computed. Using the conservative standard errors we cannot reject the null that the ATE equals zero.

Efficiency can be improved by exploiting the block design, and introducing block fixed effects. Now the null can be rejected.

```
> modd4block <- lm(outcome4~Tr4+factor(X),data=data)
> coeftest(modd4block, vcov = vcovHC(modd4block,"HC2"))["Tr4",]
    Estimate   Std. Error      t value     Pr(>|t|)
1.978500e+00 1.015424e-01 1.948447e+01 2.681943e-63
```

Given the perfect treatment assignment balance within each block, the ATE estimate is the same. However, its standard error is greatly reduced.

(c) Suppose that we estimate the ATE with a difference-in-means between treatment and control groups. Design a Monte Carlo study to compare the Mean Square Error (MSE) of this estimator across the four experimental designs (in **R**, set the seed at 2, and draw 1000 samples to conduct the analysis). Rank order the experimental designs by MSE. Offer an explanation for the ranking. In particular, what component of the MSE drives the differences for each design? Why does this component vary across designs? (Hint: Plotting the individual-level potential outcomes, Y_{0i} vs. Y_{1i}, may be informative.)

```
> r <- 1000
> MSE1 <- matrix(NA, ncol=1, nrow=r)
> MSE2 <- matrix(NA, ncol=1, nrow=r)
> MSE3 <- matrix(NA, ncol=1, nrow=r)
> MSE4 <- matrix(NA, ncol=1, nrow=r)


 > # design 1
> MSE1 <- (ATE1-ATE)^2
> set.seed(2)
> for (i in 1:r){
+  # design 2
+  data$Tr2 <- sample(0:1, size=1000, replace=TRUE)
+  ATE2 <- mean(data$Treat[data$Tr2==1]) - mean(data$Control[data$Tr2==0])
+  MSE2[i] <- (ATE2-ATE)^2
+  print(i)
+ }


> set.seed(2)
> for (i in 1:r){
+  # design 3
+  K <- sample(1:1000, size=500, replace=FALSE)
```

```
+  data$Tr3 <- 0
+  data$Tr3[K] <- 1
+  ATE3 <- mean(data$Treat[data$Tr3==1]) - mean(data$Control[data$Tr3==0])
+  MSE3[i] <- (ATE3-ATE)^2
+  print(i)
+ }

> set.seed(1)
> for (i in 1:r){
+  # design 4
+  data <- ddply(data, .(X), transform, Tr4=sample(0:1, replace=FALSE))
+  ATE4 <- mean(data$Treat[data$Tr4==1]) - mean(data$Control[data$Tr4==0])
+  MSE4[i] <- (ATE4-ATE)^2
+  print(i)
+ }

> cbind(mean(MSE1,na.rn=TRUE),mean(MSE2,na.rn=TRUE),
      mean(MSE3,na.rn=TRUE),mean(MSE4,na.rn=TRUE))
         [,1]     [,2]   [,3]         [,4]
[1,] 2500.988 3.569471 3.1672 0.004947579
```

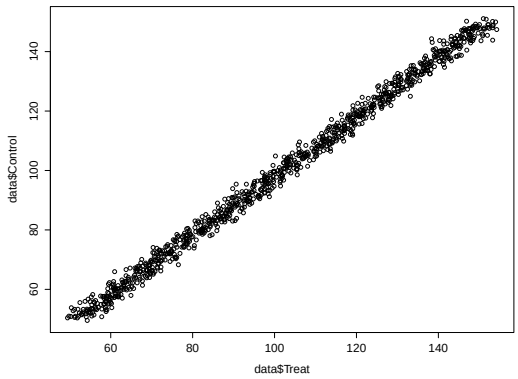
While there is not variance in the estimate of Design 1, it produces a very large bias, so its MSE is driven by this bias. The other three designs are unbiased because all induce variance in the ATE estimate. Therefore, the differences in the MSE for designs 2-4 are driven by differences in the variance of the estimate of the ATE.

The sampling variance produced by the potential treatment vector is highest in design 2, because it does not fix the number of treated units at 0.5, so it makes sense that its estimated variance (and thus MSE) is the largest after design 1.

Design 4 is the best because it is using the information contained in the covariate X to account for variation unrelated to the randomization. This leads to a very large reduction of the variance of the estimated ATE. Notice the high covariance between the individual potential outcomes:

```
> cov(data$Treat,data$Control)
[1] 834.6813
```

as well as the fact that X is correlated with the potential outcomes (i.e. as X gets larger, the potential outcomes get larger).


data$Control
60
80
100
120
140
data$Treat
60
80
100
120
140