POLS 207 - Problem Set #2

Cori Lopazanski

4/15/2021

Problem 1

a) Estimate the average treatment effect in the dataset, using the difference in means estimator. (Make sure to first remove any units with missing value(s) for one or more variables.)

```r
### Read Olken Data & Initial Cleaning
olken <- read_csv("olken_data.csv") %>%
  drop_na() %>%  # Remove all observations with NA
  mutate(treat_control = case_when(treat_invite == 0 ~ "control",
                                   treat_invite == 1 ~ "treatment"))

# Calculate average percent missing for each group (treatment and control):
means <- olken %>%
  group_by(treat_invite) %>%
  summarize(mean = mean(pct_missing))

# Find difference in means:
ate <- means$mean[means$treat_invite == 1] - means$mean[means$treat_invite == 0]
```

Encouraging community participation in monitoring had an average treatment effect of -0.0249495 on the percent expenditures missing for road building projects.

b) On what characteristics (i.e., covariates) do the treatment and control group differ, if any?

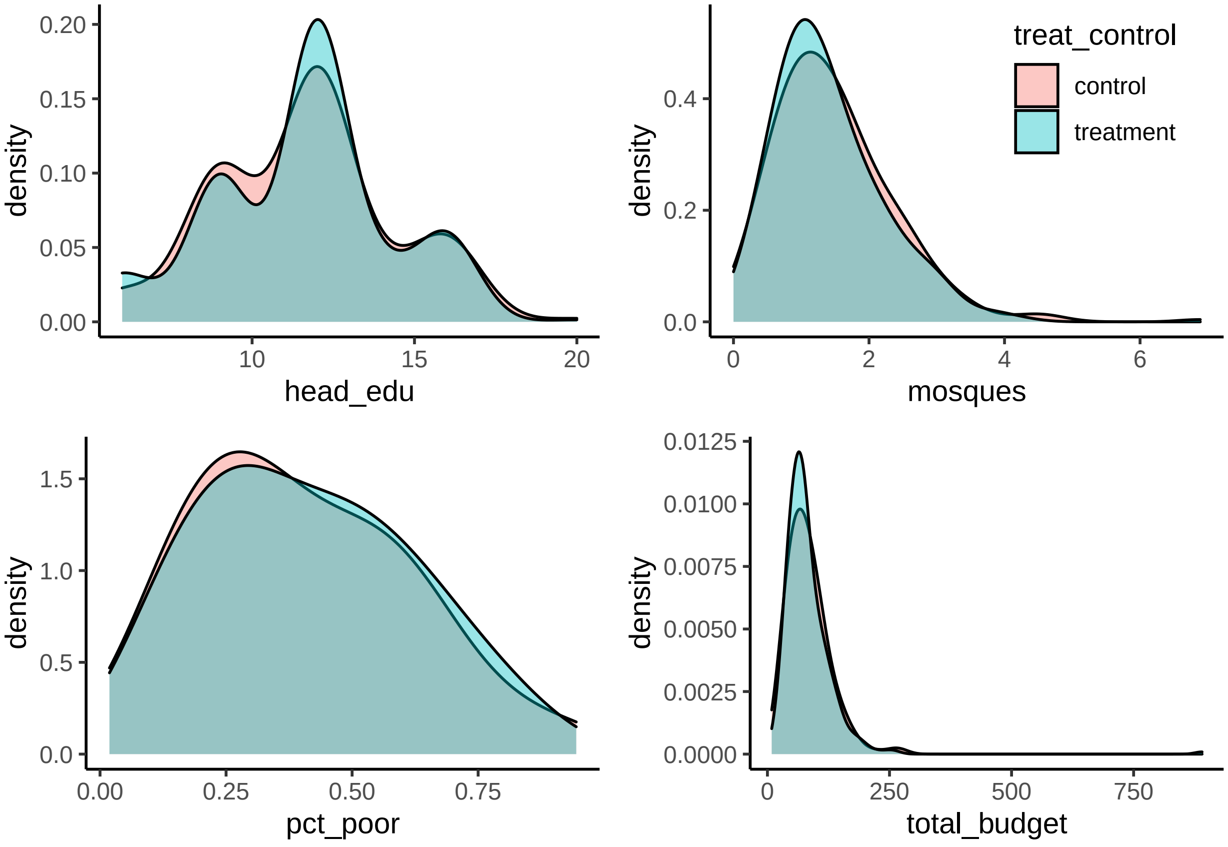
```r
### Visualizations of covariate distributions
###################################################
## Village Head Education
edu <- ggplot(data = olken) +
  geom_density(aes(x = head_edu, group = treat_control, fill = treat_control),
               adjust = 1.5, alpha = 0.4) +
  theme_classic()+
  theme(legend.position = "NONE")

## Mosques Per 1000
mosque <- ggplot(data = olken) +
  geom_density(aes(x = mosques, group = treat_control, fill = treat_control),
               adjust = 1.5, alpha = 0.4)  +
  theme_classic()+
  theme(legend.position = c(.75, .75))

## Percent Households Below Poverty Line
poverty <- ggplot(data = olken) +
  geom_density(aes(x = pct_poor, group = treat_control, fill = treat_control),
               adjust = 1.5, alpha = 0.4) +
    theme_classic()+
    theme(legend.position = "NONE")


## Total Budget
budget <- ggplot(data = olken) +
  geom_density(aes(x = total_budget, group = treat_control, fill = treat_control),
               adjust = 1.5, alpha = 0.4) +
  theme_classic()+
  theme(legend.position = "NONE")


## Arrange Covariate Plots
grid.arrange(arrangeGrob(edu, mosque, ncol = 2),
             arrangeGrob(poverty, budget, ncol = 2),
             nrow = 2)
```



```r
### Use t-test to compare difference in means
##########################################################
edu.t <- t.test(olken$head_edu[olken$treat_invite == 0],
                olken$head_edu[olken$treat_invite == 1])

mosque.t <- t.test(olken$mosques[olken$treat_invite == 0],
                   olken$mosques[olken$treat_invite == 1])

poverty.t <- t.test(olken$pct_poor[olken$treat_invite == 0],
                    olken$pct_poor[olken$treat_invite == 1])
```

Table 1: Difference in means for covariate chracteristics

Covariate	Control Mean	Treatment Mean	Difference	Statistic	P-Value
head_edu	11.583851	11.5466238	0.0372271	0.1409161	0.8880239
mosques	1.472887	1.4195176	0.0533699	0.6624145	0.5081709
pct_poor	0.400366	0.4106243	-0.0102584	-0.4968260	0.6196482
total_budget	83.354209	83.1643437	0.1898659	0.0394438	0.9685551

```r
budget.t <- t.test(olken$total_budget[olken$treat_invite == 0],
                   olken$total_budget[olken$treat_invite == 1])

balance <- map_df(list(edu.t, mosque.t, poverty.t, budget.t), tidy) %>%
  mutate(Covariate = c("head_edu", "mosques", "pct_poor", "total_budget")) %>%
  select(Covariate, estimate1, estimate2, estimate, statistic, p.value) %>%
  rename("Difference" = estimate,
         "Control Mean" = estimate1,
         "Treatment Mean" = estimate2,
         "Statistic" = statistic,
         "P-Value" = p.value)

kable(balance, caption = "Difference in means for covariate chracteristics")
```

The treatment and control groups are fairly balanced across the different characteristics in the dataset.

c) Now use regression to estimate the ATE (average treatment effect). Use heteroskedastic-consistent standard errors. Is this estimate different from the difference-in-means estimate?

```r
## Basic model of percent missing as a function of treatment status:
ate_lm <- lm(data = olken, pct_missing ~ treat_invite)

## Compute robust standard errors:
lm_robust <- coeftest(ate_lm, vcov = vcovHC(ate_lm, type = "HC1"))

stargazer(lm_robust)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Apr 20, 2021 - 16:26:39

Table 2:

	Dependent variable:
treat_invite	−0.025
	(0.033)
Constant	0.253***
	(0.027)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimated ATE generated through basic regression does not differ from the difference-in-

means estimate.

d) Re-estimate the ATE 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 heteroskedastic-consistent standard errors. How do these results vary across the regressions?

```r
### Additional Regressions
####################################
## 1. All covariates as additional linear predictors
d1 <- lm(data = olken,
         pct_missing ~ treat_invite + head_edu + mosques + pct_poor + total_budget)

## 2. Arbitrary functions as additional protdictors
d2 <- lm(data = olken,
         pct_missing ~ treat_invite + I(head_edu^2) + mosques + pct_poor + log(total_budget))

## 3. Demenaed versions of the covariates (X_i - X) and their interactions.
demeaned <- olken %>%
  mutate_at(vars(head_edu, mosques, pct_poor, total_budget),
            function(x){x- mean(x)})

d3 <- lm(data = demeaned,
         pct_missing ~ treat_invite*(head_edu+mosques+pct_poor+total_budget))


### Compiled Results
########################################
models <- list(d1, d2, d3)

stargazer(models,
          se = estimatr::starprep(models, se_type = "HC1"),
          header = F,
          single.row = T,
          title = "Comparison of Regression Estimates of Average Treatment Effect")
```

e) These results suggest that there is not a strong effect of community monitoring on corruption. The data duggest that on average, encouraging community participation in monitoring reduced the missing expenditures by ~0.025 percentage points.

Problem 2

a) A key identification assumption in a randomized experiment is that there is no interference between units (the SUTVA assumption). To make one's inference robust to the violation of that assumption within clusters of units, one must correct standard errors using methods such as cluster-robust standard errors.

False - I'm unsure about this one, but I thought that the only way to correct for interference is in the experimental design, and that standard error adjustments only help to correct for heteroskedasticity within clusters.

b) The potential outcomes framework assumes that treatment assignment of unit i does not affect the potential outcomes of unit j.

True - This is called the non-interference, SUTVA assumption.

Table 3: Comparison of Regression Estimates of Average Treatment Effect

	Dependent variable:		
	pct_missing		
	(1)	(2)	(3)
treat_invite	−0.026 (0.033)	−0.025 (0.033)	−0.027 (0.033)
head_edu	−0.006 (0.006)		−0.008 (0.011)
I(head_edu^2)		−0.0002 (0.0003)	
mosques	−0.048*** (0.018)	−0.045** (0.018)	−0.069** (0.027)
pct_poor	−0.118 (0.073)	−0.121* (0.073)	−0.096 (0.129)
total_budget	0.001* (0.0003)		0.0004 (0.001)
log(total_budget)		0.076** (0.030)	
treat_invite:head_edu			0.004 (0.013)
treat_invite:mosques			0.030 (0.036)
treat_invite:pct_poor			−0.030 (0.157)
treat_invite:total_budget			0.0001 (0.001)
Constant	0.390*** (0.092)	0.073 (0.144)	0.255*** (0.026)
Observations	472	472	472
R^2	0.029	0.034	0.031
Adjusted R^2	0.019	0.024	0.012
Residual Std. Error	0.341 (df = 466)	0.340 (df = 466)	0.342 (df = 462)
F Statistic	2.823** (df = 5; 466)	3.303*** (df = 5; 466)	1.635 (df = 9; 462)
Note:			*p<0.1; **p<0.05; ***p<0.01

c) The following identity generally holds: $ATE = 1/2ATT + 1/2ATC$ where ATE is the average treatment effect, ATT is the average treatment effect for the treated, and ATC is the average treatment effect for the controls.

False

Recall that:

$$\begin{aligned} ATC &= \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 0] \\ ATT &= \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1] \end{aligned}$$

Since:

$$\begin{aligned} ATE &= \mathbb{E}[Y_{1i} - Y_{0i}] \\ &= \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] \\ &= \mathbb{E}[Y_{1i}|D_i = 1] + \mathbb{E}[Y_{1i}|D_i = 0] - (\mathbb{E}[Y_{0i}|D_i = 1] + \mathbb{E}[Y_{0i}|D_i = 0]) \\ &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 1] + \mathbb{E}[Y_{1i}|D_i = 0] - \mathbb{E}[Y_{0i}|D_i = 0] \end{aligned}$$

This can be rewritten:

$$\begin{aligned} ATE &= \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] + \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 0] \\ ATE &= ATT + ATC \end{aligned}$$

Problem 3

You are on a team of researchers who want to estimate the effect of having completed college in the past ($C_i = 1$) on an individuals' current earnings (Y_i). Some individuals may have taken up a white collar job after finishing college ($W_i = 1$), but before you measure earnings. Assume that whether a subject completes college is randomly assigned. Your fellow researcher says that you need to control for the occupation of a person in your model. To do this they write the difference in means of potential outcomes for all individuals who have a white collar job as follows:

$$\mathbb{E}(Y_{1i}|W_{1i} = 1, C_i = 1) - \mathbb{E}(Y_{0i}|W_{0i} = 1, C_i = 0)$$

a) There are two subscripts for W because it is also a variable with potential outcomes respective to each individual - for this scenario, there are units that both went to college and took a white collar job, individuals which went to college and did not take a white collar job, individuals who did not go to college and took a white collar job, and individuals who did not go to college and did not take a white collar job. The specific difference in means above describes the difference in expectation for individuals who went to college and took a white collar job to individuals who did not go to college and took a white collar job.

b) This difference is *not* unbiased, because you have conditioned on a post-treatment variable.