PE-PSet3

Drazzel Feliu - 12174100

For this assignment, provide a write-up where you answer the questions below, selectively cutting and pasting output where needed. Be concise in your write-up; excess wordiness will be penalized. Also, submit a log file that includes commands and results for your entire analysis. The assignment makes use of AganStarrQJEData.dta, which you can find on Canvas.

In this problem set we will reproduce some of Amanda Agan and Sonja Starr's basic results, so start by reading their paper (Ban the box, criminal records, and racial discrimination: A field experiment), which you can find on Canvas.

```
# load data set
data <- read_dta("AganStarrQJEData.dta")
# create data table identifying class and labels for each variable
datainfo <- data.frame(variable=colnames(data),
           class=sapply(data, class)
           )
label=unlist(lapply(data, function (x) attr(x, "label")))
label <- as.data.frame(label)
label$variable <- rownames(label)
label <- label[c(2,1)]
datainfo <- left_join(datainfo, label, by = "variable")
```

```
## Warning: Column `variable` joining factor and character vector, coercing
## into character vector
```

```
# summary statistics of variables
summary(data)
```

Question 1:

For this question, restrict your analysis to the set of job applications that asked about criminal records ("Box" applications) in the before period ("pre-BTB" period). (Note: there are some applications that did not have a box in the pre-BTB period, but then added them in the post- period. Agan and Star code these as "remover = -1" in their data and call them "reverse compliers." Exclude these observations from your analysis throughout this assignment.)

```
# Permanently filter data for all reverse compliers
data <- filter(data, remover!=-1)
```

A)

What is the average callback rate for people who committed crimes? For those who didn't? Is the difference statistically significant?

Answer:

```
# Summary of Means across both groups
data %>% group_by(crime) %>% summarise(mean = mean(response)*100)
```

```
## # A tibble: 2 x 2
##   crime      mean
##   <dbl+lbl> <dbl>
## 1 0          12.5
## 2 1          10.9
```

```
# Significance test across two groups for the callback rate
t.test(response~crime, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  response by crime
## t = 2.9346, df = 14547, p-value = 0.003345
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.005183336 0.026034902
## sample estimates:
## mean in group 0 mean in group 1
##       0.1250518       0.1094426
```

The average callback rate for people who committed crimes is 10.94% and is 12.51% for individuals who did not commit crimes. The difference is statistically significant at a 95% confidence level.

B)

Can we interpret this as a causal effect? Explain briefly.

Answer:

Interpretating this relationship as causal is premature at the moment. We haven't controlled for the type of application individuals are receiving nor across other demographic controls that may have an influence on the rate of callbacks (level of education, interviewer bias through perceptions of race, geographic location of individuals, hiring needs across time). Several variables may disentangle the impact of criminal history on positive responses to applications.

Question 2:

Now consider just the "Box" applications but include both the pre- and post-BTB periods.

A)

Regress callback rates on race, GED, and employment gap. Include "chain 1" and "center" fixed effects. Does race appear to have an effect on callback rates? Does this coefficient have a causal interpretation?

```
reg1 <- data %>% filter(., crimbox==1) %>%
  lm(response~white + ged + empgap + chain_id + center, data = .)
```

```
stargazer(reg1, type = "latex", title = "Call Back Rates With Fixed Effects",
          covariate.labels = c("Race = White", "GED Acquired",
                               "Employment Gap Exists", "Chain", "Center"),
          digits = 6, single.row = TRUE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Nov 02, 2018 - 21:45:30

Table 1: Call Back Rates With Fixed Effects

	Dependent variable:
	response
Race = White	−0.001483 (0.011427)
GED Acquired	0.016029 (0.011452)
Employment Gap Exists	0.010460 (0.011445)
Chain	0.000142 (0.000097)
Center	0.000370*** (0.000117)
Constant	0.050609** (0.019738)
Observations	2,918
R^2	0.005263
Adjusted R^2	0.003555
Residual Std. Error	0.308518 (df = 2912)
F Statistic	3.081134*** (df = 5; 2912)
Note:	*p<0.1; **p<0.05; ***p<0.01

Answer:

The coefficient on race (Table 1) is not statistically significant. While being white seems to have a negative impact on the response rate, it cannot be distinguished from 0 and so we have to identify that it is likely not an effect on response rates. Subsequently, we cannot say it has a causal impact on response rate for applications that feature a box.

B)

Estimate the model again, but without the chain and center fixed effects. Does the coefficient on "white" change? Why is it important to include chain and center fixed effects?

```
reg2 <- data %>% filter(., crimbox==1) %>%
  lm(response~white + ged + empgap, data = .)
```

```
stargazer(reg2, type = "latex", title = "Call Back Rates Without Fixed Effects",
          covariate.labels = c("Race = White", "GED Acquired",
                               "Employment Gap Exists"),
          digits = 6, single.row = TRUE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Nov 02, 2018 - 21:45:30

Answer:

Removing the fixed effects minimizes the impact of race as a factor on response rate (from .

Table 2: Call Back Rates Without Fixed Effects

	Dependent variable:
	response
Race = White	−0.001140 (0.011448)
GED Acquired	0.015345 (0.011469)
Employment Gap Exists	0.011098 (0.011465)
Constant	0.094551*** (0.011629)
Observations	2,918
R^2	0.000893
Adjusted R^2	−0.000135
Residual Std. Error	0.309089 (df = 2914)
F Statistic	0.868506 (df = 3; 2914)
Note:	*p<0.1; **p<0.05; ***p<0.01

C)

Now add the "conviction" variable. What happens to the coefficient on "white"? If the coefficient changes, does this mean that the previous regression was subject to omitted variable bias?

```
reg3 <- data %>% filter(., crimbox==1) %>%
  lm(response~white + ged + empgap + crime, data = .)
```

```
stargazer(reg3, type = "latex", title = "Call Back Rates With Crime",
          covariate.labels = c("Race = White", "GED Acquired",
                               "Employment Gap Exists", "Crime Committed"),
          digits = 6, single.row = TRUE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Nov 02, 2018 - 21:45:30

Table 3: Call Back Rates With Crime

	Dependent variable:
	response
Race = White	−0.001117 (0.011413)
GED Acquired	0.014035 (0.011438)
Employment Gap Exists	0.009976 (0.011432)
Crime Committed	−0.049618*** (0.011416)
Constant	0.120794*** (0.013072)
Observations	2,918
R^2	0.007331
Adjusted R^2	0.005968
Residual Std. Error	0.308144 (df = 2913)
F Statistic	5.378062*** (df = 4; 2913)
Note:	*p<0.1; **p<0.05; ***p<0.01

Answer:

Question 3:

The authors estimate the following model for different subsets of the data, where "Box" is an indicator for whether the application had a box asking about employment2, and X is a vector of covariates:

$$Callback_{ij} = \alpha + \beta_1 Box_j + \beta_2 White_i + \beta_3 Box_j * White_i + X_i\gamma + \epsilon_{ij}$$

A)

Suppose they run this regression on the full sample, which includes both Box and non- Box applications, but only in the pre-period (don't actually do this yet). What do α, β_1, β_2, and β_3 tell you?

Answer:

α is going to tell you the baseline response rate for all individuals who submitted applications in the sample.

B)

Do you think "Box" and "non-Box" stores might differ in systematic ways, besides their decision to include a box asking about criminal history? In other words, do we think this variable is "as-if" randomly assigned?

Answer:

This variable is not "as-if" randomly assigned. Any store that willingly chooses to avoid using the box in principal expects to have a markedly different applicant pool, given the lack of the presence of the box. They are uniquely aware of the selection effect of knowingly removing the box and as such are probably more amenable to selection independent of criminal history.

C)

Suppose they run the regression on just the "Box" applications in both periods (again, don't do this yet). What is the interpretation of the coefficients now?

Answer:

Question 4:

For the below estimations, include controls for employment gap and ged, as well as center fixed effects. Again, exclude the so-called "reverse compliers."

A)

Estimate the model from question 3 on both "Box" and non-"Box" applications in just the pre-period.

Answer:

```
reg4 <- data %>% filter(., pre==1) %>%
  lm(response~crimbox + white + box_white + crime + ged + empgap, data = .)
summary(reg4)
```

```
##
## Call:
## lm(formula = response ~ crimbox + white + box_white + crime +
##     ged + empgap, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14407 -0.12108 -0.10591 -0.08971  0.92610
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.112052   0.009107  12.304  < 2e-16 ***
## crimbox      0.015811   0.010803   1.464 0.143357
## white        0.032016   0.009206   3.478 0.000509 ***
## box_white   -0.029622   0.015210  -1.948 0.051510 .
## crime       -0.028975   0.007333  -3.951 7.84e-05 ***
## ged         -0.008040   0.007333  -1.096 0.272982
## empgap      -0.001142   0.007335  -0.156 0.876326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3118 on 7238 degrees of freedom
## Multiple R-squared:  0.003874,   Adjusted R-squared:  0.003049
## F-statistic: 4.692 on 6 and 7238 DF,  p-value: 8.978e-05
```

B)

What kind of standard errors should you use, and why?

Answer:

C)

Is the coefficient on "crimbox" statistically significant? What about "white" and the interaction of "crimbox" and "white"? Interpret these findings.

Answer:

D)

Now estimate the model from question 3 on just "Box" applications in both periods. Interpret the coefficients.

Answer:

```
reg5 <- data %>% filter(., crimbox==1) %>%
  lm(response~crimbox + white + box_white + crime + ged + empgap, data = .)
summary(reg5)
```

```
##
## Call:
## lm(formula = response ~ crimbox + white + box_white + crime +
##     ged + empgap, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14480 -0.13077 -0.09407 -0.08004  0.92994
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.120794   0.013072   9.241  < 2e-16 ***
## crimbox            NA         NA      NA       NA
## white       -0.001117   0.011413  -0.098    0.922
## box_white          NA         NA      NA       NA
## crime       -0.049618   0.011416  -4.346 1.43e-05 ***
## ged          0.014035   0.011438   1.227    0.220
## empgap       0.009976   0.011432   0.873    0.383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3081 on 2913 degrees of freedom
## Multiple R-squared:  0.007331,   Adjusted R-squared:  0.005968
## F-statistic: 5.378 on 4 and 2913 DF,  p-value: 0.0002589
```

Question 5:

Based on the above analysis, what are your conclusions about the effects of BTB?

Answer: