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 across both periods
data %>% filter(., crimbox==1) %>% filter(., pre==1) %>% group_by(crime) %>% summarise(mean = mean(resp
```

```
## # A tibble: 2 x 2
##   crime       mean
##   <dbl+lbl>  <dbl>
## 1 0         0.136
## 2 1         0.0846
```

```
# Significance test across two groups for the callback rate
data %>% filter(., crimbox==1) %>% filter(., pre==1) %>% t.test(response~crime, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  response by crime
## t = 4.2171, df = 2533.4, p-value = 2.561e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02735417 0.07490189
## sample estimates:
## mean in group 0 mean in group 1
##      0.13570887      0.08458084
```

The average callback rate for people who committed crimes is 8.46% and is 13.57% for individuals who did not commit crimes for all applications that ask about about prior criminal history in the pre-Ban The Box period. 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 + factor(chain_id) + factor(center), data = .)
```

```
stargazer(reg1, type = "latex", title = "Call Back Rates With Fixed Effects (Box Only, Both Periods)",
          covariate.labels = c("White", "GED",
                               "Employment Gap"), omit = c("chain_id", "center"),
```

```
          add.lines = list(c("Chain Fixed Effects", "Yes"),c("Center Fixed Effects", "Yes")),
          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: Thu, Nov 08, 2018 - 14:29:51

Table 1: Call Back Rates With Fixed Effects (Box Only, Both Periods)

	Dependent variable:
	response
White	−0.000879 (0.010764)
GED	0.011851 (0.011367)
Employment Gap	0.011322 (0.010976)
Constant	−0.025232 (0.128573)
Chain Fixed Effects	Yes
Center Fixed Effects	Yes
Observations	2,918
R^2	0.179273
Adjusted R^2	0.125298
Residual Std. Error	0.289057 (df = 2737)
F Statistic	3.321382*** (df = 180; 2737)
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 rates 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 (Box Only, Both Periods)
          covariate.labels = c("White", "GED",
                               "Employment Gap"),
          add.lines = list(c("Chain Fixed Effects", "No"),c("Center Fixed Effects", "No")),
          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: Thu, Nov 08, 2018 - 14:29:51

Table 2: Call Back Rates Without Fixed Effects (Box Only, Both Periods)

	Dependent variable:
	response
White	−0.001140 (0.011448)
GED	0.015345 (0.011469)
Employment Gap	0.011098 (0.011465)
Constant	0.094551*** (0.011629)
Chain Fixed Effects	No
Center Fixed Effects	No
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

Answer:

Removing the fixed effects increases the impact of race as a factor on response rate (from −0.000879 to −0.001140). The chain and center fixed effects normalize the impact of race on callbacks by the heterogeneous differences across jurisdictions, making them valuable tools for analyzing how race influences the response rate among applications that feature the box.

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 (Box Only, Both Periods)",
          covariate.labels = c("White", "GED",
                               "Employment Gap", "Crime"),
          add.lines = list(c("Chain Fixed Effects", "No"),c("Center Fixed Effects", "No")),
          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: Thu, Nov 08, 2018 - 14:29:51

Answer:

The coefficient on white decreases in magnitude without the fixed effects (Table 3). In the absence of the fixed effects, the white coefficient changes from −0.001140 to −0.001117. This is a clear example of the omitted variable bias influencing the white coefficient.

Table 3: Call Back Rates With Crime (Box Only, Both Periods)

	Dependent variable:
	response
White	−0.001117 (0.011413)
GED	0.014035 (0.011438)
Employment Gap	0.009976 (0.011432)
Crime	−0.049618*** (0.011416)
Constant	0.120794*** (0.013072)
Chain Fixed Effects	No
Center Fixed Effects	No
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

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 provides the baseline response rate for all individuals who submitted applications in the sample. β_1 is the percentage point impact that an application that has the box has on response rates across the board for all individuals. β_2 highlights the effect of being white on response rates across the board, independent of whether the box exists on the application. β_3 then is the impact that being white has on applications that do feature the box. All of these coefficients however are limited in that their impact would only be defined on applications submitted in the period before the box was banned, given the stipulation above.

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. However, given that the box effectively limits statistical discrimination, stores that do not feature the box may be more likely to be biased against black candidates and removing the box may be a tool to apply that bias broadly.

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:

Running this regression on only applications featuring the box requires removing two variables, Box_j & $Box_j * White_i$, as these variables no longer exhibit any meaningful variation. This does ultimately impact the coefficients α and β_2, where α is now the base response rate for all applications and β_2 is now the percentage point change in the response rate given the applicants race, under the presumption that all applications have the box.

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:

See Table 4.

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

```
stargazer(reg4, type = "latex", title = "Call Backs Rates (Pre-Period)",
          covariate.labels = c("Box","White", "Box x White","Crime",
                               "GED","Employment Gap"),
          add.lines = list(c("Chain Fixed Effects", "Yes"),c("Center Fixed Effects", "Yes")),
          omit = c("center","chain_id"),
          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: Thu, Nov 08, 2018 - 14:29:52

Table 4: Call Backs Rates (Pre-Period)

	Dependent variable:
	response
Box	−0.005354 (0.027502)
White	0.029763*** (0.008426)
Box x White	−0.027397** (0.013911)
Crime	−0.024550*** (0.006810)
GED	−0.004261 (0.006924)
Employment Gap	−0.001893 (0.006832)
Constant	−0.012837 (0.149354)
Chain Fixed Effects	Yes
Center Fixed Effects	Yes
Observations	7,245
R^2	0.216000
Adjusted R^2	0.172476
Residual Std. Error	0.284031 (df = 6863)
F Statistic	4.962800*** (df = 381; 6863)
Note:	*p<0.1; **p<0.05; ***p<0.01

B)

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

Answer:

```
ncvTest(reg4)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2356.25    Df = 1     p = 0
```

The p-value of the non-constant variance score test indicates that heteroskedasticity is present in our model and to address this, we should be applying robust standard errors to account for this going forward.

C)

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

Answer:

The coefficient on crimbox in this model is not statistically significant. The interaction of crimbox and white is statistically significant at the 95% level. The interaction coefficient signifies that being white and submitting for a position where the box is present slightly diminishes the likelihood (−0.0274) of a callback. This is simultaneously compounded by the crimbox coefficient (−0.0054), being cognizant of the fact that it's not statistically significant.

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~white + crime + ged + empgap + factor(chain_id) + factor(center), data = .)
robust_se <- as.vector(summary(reg5,robust = T)$coefficients[,"Std. Error"])
```

```
stargazer(reg5, type = "latex", title = "Call Backs Rates (Box Only, Both Periods)",
          covariate.labels = c("White", "Crime",
                               "GED","Employment Gap"),
          add.lines = list(c("Chain Fixed Effects", "Yes"),c("Center Fixed Effects", "Yes")),
          omit = c("chain_id","center"),
          digits = 6, single.row = TRUE,
          se = list(robust_se))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Nov 08, 2018 - 14:29:56

Table 5: Call Backs Rates (Box Only, Both Periods)

	Dependent variable:
	response
White	−0.000868 (0.010768)
Crime	−0.052030*** (0.011005)
GED	0.010484 (0.011018)
Employment Gap	0.010278 (0.011020)
Constant	−0.008812 (0.048595)
Chain Fixed Effects	Yes
Center Fixed Effects	Yes
Observations	2,918
R^2	0.186002
Adjusted R^2	0.132152
Residual Std. Error	0.287922 (df = 2736)
F Statistic	3.454086*** (df = 181; 2736)
Note:	*p<0.1; **p<0.05; ***p<0.01

In this model, subsetting the data for only applications that feature the box requires dropping the box and box * white interaction variables. Given that, being white does limit the response rate (−0.00087) to a certain degree, but this effect is not statistically significantly different from 0, and so does having previously committed a crime (−0.05203), which is significant. The response rate declines in greater magnitude given a crime as opposed to being white.

Question 5:

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

Answer:

The box is potentially equalizing the playing field for black applicants broadly while certainly reducing the opportunities for individuals with criminal histories in the same manner, after controlling for educational and employment histories. Black applicants, having the box present on applications, are met with less of the structural bias commonly present in non-box applications as evidenced by the decreased coefficient on white. However given the statistical significance of this coefficient, this relationship cannot be extrapolated broadly. For individuals who have been convicted of a crime before, the box reduces the likelihood that they will receive a callback. And this coefficient is indeed statistically significant.