Assignment 3 Notebook

Chiebuka Onwuzurike

Questions 1

1.a: Of the above variables, please identify which may be ‘good’ controls in a regression?

*** Note: By good control variables,I mean those which do not prevent a causal interpretation of the coefficient on treatment. ***

I think some good control variables would be anonymity score, log followers, and racism scores pre treatment. What makes these treatments good controls are they are before the treatment. There is also the possibility of a relationship between these control and the outcome or treatment. I hypothesize there could be a relationship with high anonymity score or higher amount of followers to be less affected by the treatment and continue racist comments.

1.b: Run a regression and a t.test of ‘racism.scores.post.2mon’ on ‘any_treatment’. What do we learn from this regression about the effect of the treatment? Please explain in words in addition to just returning the number.

After the experiment we can see there is no statistically significant effect on the outcome based on the treatment. Both the t.test and linear regression lead us to this interpretation.

```
summary(lm(racism.scores.post.2mon ~ any_treatment, data = tweets_data))
```

```
##
## Call:
## lm(formula = racism.scores.post.2mon ~ any_treatment, data = tweets_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25217 -0.16940 -0.12314  0.01502  2.74996
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.25217    0.05356   4.708 4.22e-06 ***
## any_treatment -0.08277    0.06041  -1.370    0.172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3862 on 241 degrees of freedom
## Multiple R-squared:  0.00773,    Adjusted R-squared:  0.003613
## F-statistic: 1.877 on 1 and 241 DF,  p-value: 0.1719
```

```
t.test(tweets_data[any_treatment==1,racism.scores.post.2mon], tweets_data[any_treatment==0,racis
m.scores.post.2mon])
```

```
##
##  Welch Two Sample t-test
##
## data:  tweets_data[any_treatment == 1, racism.scores.post.2mon] and tweets_data[any_treatment
== 0, racism.scores.post.2mon]
## t = -1.2003, df = 69.367, p-value = 0.2341
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.22033209  0.05478379
## sample estimates:
## mean of x mean of y
## 0.1693971 0.2521712
```

1.c Perform a randomization / balance check on this dataset for all of the variables identified in a). Hint: we can do this by comparing the pre-treatment outcomes between the treatment and control group. If there are significant differences, then there may be a problem with the experiment.

We can see the proportion for treated and control was 78.6% to 21.4%. After doing a t-test between the racism score prior to the experiment we can see we accurately performed randomization. The p-value of the two sample t-test for racism pre score, anonymity, and log followers is respectively 0.06535, 0.5779, and 0.3369 which are not statistically significant and further means we cannot reject the null. We did a check and found there is no difference between control and treatment group prior to the experiment.

```
#prop.test(tweets_data[any_treatment==1,.N],tweets_data[,.N])
#summary(lm(racism.scores.pre.2mon ~ any_treatment, data = tweets_data))
t.test(tweets_data[any_treatment==1,racism.scores.pre.2mon], tweets_data[any_treatment==0,racis
m.scores.pre.2mon])
```

```
##
##  Welch Two Sample t-test
##
## data:  tweets_data[any_treatment == 1, racism.scores.pre.2mon] and tweets_data[any_treatment
== 0, racism.scores.pre.2mon]
## t = -1.8821, df = 52.675, p-value = 0.06535
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.278647792  0.008881565
## sample estimates:
##  mean of x  mean of y
## 0.09278438 0.22766749
```

```
t.test(tweets_data[any_treatment==1,anonymity], tweets_data[any_treatment==0,anonymity])
```

```
##
##  Welch Two Sample t-test
##
## data:  tweets_data[any_treatment == 1, anonymity] and tweets_data[any_treatment == 0, anonymi
ty]
## t = -0.55886, df = 76.729, p-value = 0.5779
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2834826  0.1592377
## sample estimates:
## mean of x mean of y
##  1.534031  1.596154
```

```
t.test(tweets_data[any_treatment==1,log.followers], tweets_data[any_treatment==0,log.followers])
```

```
##
##  Welch Two Sample t-test
##
## data:  tweets_data[any_treatment == 1, log.followers] and tweets_data[any_treatment == 0, lo
g.followers]
## t = -0.96535, df = 90.545, p-value = 0.3369
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8172420  0.2827161
## sample estimates:
## mean of x mean of y
##  5.408653  5.675915
```

1.d Add the variables from a) as controls into the regression from b). What happens to our estimate of the effect of the treatment and it's standard error? Why does this happen in words?

After adding the variables anonymity, log.followers and racism score pre experiment to the previous regression of just any_treatment, we can see there is an increase of the coefficient(-0.08277 -> 0.022048) for any_treatment variable and there is a decrease of the standard error (0.06041 -> 0.053289). The p-value of of any_treatment also increased from (0.172 -> 0.679) which is further evidence the treatment had no statistically significant effect. The decrease in standard error would be a good sign if any_treatment would statically significant because then it mean our noise for the treatment decreased. What was statistically significant at 99.9% level was racism scores pre experiment. This means that a good indicator of current racism scores are previous racism scores, another indicaor that the treat had no effect.

```
summary(lm(racism.scores.post.2mon ~ any_treatment, data = tweets_data))
```

```
##
## Call:
## lm(formula = racism.scores.post.2mon ~ any_treatment, data = tweets_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25217 -0.16940 -0.12314  0.01502  2.74996
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.25217    0.05356   4.708 4.22e-06 ***
## any_treatment -0.08277    0.06041  -1.370    0.172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3862 on 241 degrees of freedom
## Multiple R-squared:  0.00773,    Adjusted R-squared:  0.003613
## F-statistic: 1.877 on 1 and 241 DF,  p-value: 0.1719
```

```
summary(lm(racism.scores.post.2mon ~ any_treatment + anonymity + log.followers + racism.scores.p
re.2mon,data = tweets_data))
```

```
##
## Call:
## lm(formula = racism.scores.post.2mon ~ any_treatment + anonymity +
##     log.followers + racism.scores.pre.2mon, data = tweets_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02355 -0.11934 -0.09273  0.03360  2.51055
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.056823   0.093223   0.610    0.543
## any_treatment            0.022048   0.053289   0.414    0.679
## anonymity                0.018068   0.031601   0.572    0.568
## log.followers           -0.001387   0.011235  -0.123    0.902
## racism.scores.pre.2mon   0.765946   0.082556   9.278   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3323 on 237 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.277,  Adjusted R-squared:  0.2648
## F-statistic:  22.7 on 4 and 237 DF,  p-value: 6.843e-16
```

1.e BONUS: we would like to know whether treatment arm 2 or treatment arm 3 is statistically significantly better at reducing racist behavior.

Perform a t.test or regression and test for the null hypothesis that treatment arm 2 has the same effect as treatment arm 3.

After doing a t-test between the treatment arm 2 and 3, We can see there isn't a statistical difference between the two. The p-value of the two sample t-test is 0.1091 which is not statistically significant and further means we cannot reject the null. However after running a regression we can see that treatment arm 2 isn't statistically significant but treatment are 3 is. It should be noted both have negative coefficients. In conclusion I will still say there is no difference between the two weighing my judgment more on the t-sample t-test.

```
summary(lm(racism.scores.post.2mon ~ treatment_arm, data = tweets_data))
```

```
##
## Call:
## lm(formula = racism.scores.post.2mon ~ treatment_arm, data = tweets_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25217 -0.20211 -0.07290  0.02379  2.70112
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.252171   0.053201   4.740 3.68e-06 ***
## treatment_arm1 -0.033936   0.076380  -0.444   0.6572
## treatment_arm2 -0.113242   0.078583  -1.441   0.1509
## treatment_arm3 -0.179268   0.075986  -2.359   0.0191 *
## treatment_arm4 -0.004187   0.076789  -0.055   0.9566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3836 on 238 degrees of freedom
## Multiple R-squared:  0.03312,    Adjusted R-squared:  0.01687
## F-statistic: 2.038 on 4 and 238 DF,  p-value: 0.08981
```

```
t.test(tweets_data[treatment_arm==2,racism.scores.post.2mon], tweets_data[treatment_arm==3,racis
m.scores.post.2mon])
```

```
##
##  Welch Two Sample t-test
##
## data:  tweets_data[treatment_arm == 2, racism.scores.post.2mon] and tweets_data[treatment_arm
== 3, racism.scores.post.2mon]
## t = 1.6246, df = 64.539, p-value = 0.1091
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0151519  0.1472047
## sample estimates:
##  mean of x  mean of y
## 0.13892962 0.07290323
```

Questions 2

2.a Describe the treatment in the first experiment and the unit of randomization. What share was randomized to the treatment?

(This refers to the experiment conducted in August 2015, the first experiment described in the introduction of the paper.)

For the experiment, the control had Up-Front prices(checkout price with additional fees was listing price) while the treatment had Back-End prices (listing price did not include additional fees seen at checkout). The unit of randomization was at the user level and experimental treatment conditions were assigned at the cookie level, which identifies a browser on a computer. Based on Table I the percentage of site in the sample was 66% and that percentage was randomized with 50.06% going to the Back End users. After adjusting for the two error the experiment couldn't reject the null at the 99% confidence level but still could at the 95% confidence level because the t-statistic was 1.99 which is less than 2.576 and more than 1.96.

2.b B. Table II displays a randomization / balance check. A randomization check is a regression where the dependent variable occurs before the experiment. The treatment should have no effect if the experiment was done properly. Suggest a variable not used by the authors that would be appropriate to include in a balance check.

Hint: consider whether your proposed variable is measured before the treatment happens or after the treatment happens.

I don't exactly know how it would be done but possibly determining what browser they use for example Chrome, Firefox, Safari, Microsoft Edge or determining if they are accessing the website from a mobile or desktop device. Unsure if you'd need IP address to get this information, but these factors would happen before the treatment and

and might give an indication of the user or user behavior..14

2.c What is the effect of the treatment on the Propensity to Purchase at least one product? Calculate the 95% confidence interval for this estimate.

There is an increased the transaction rate over the full course of the experiment by 14.1%. The 95% confidence interval for this estimate is with a standard error (.09%) is 0.139236 (13.9%) -> 0.142764 (14.3%).

```
.141 -(.0009 * 1.96)
```

```
## [1] 0.139236
```

```
.141 +(.0009 * 1.96)
```

```
## [1] 0.142764
```

2.d Suppose the authors randomized by city of the event. Name one benefit that may occur as a result of this randomization strategy and one harm.

If the authors chose to randomize by the city level, they might be better able to capture the spillovers across user for the same city. However we would be loosing statistical power by randomizing at the city level because there are less cities than users.

2.e Suppose that you are the product manager for the monetization team at Stubhub. Based on the evidence presented above, would you launch the treatment to the entire site? The answer should be less than 1 paragraph. It should consist of an answer (Yes, no), and two pieces of evidence relating to that recommendation.

Yes, we should move forward and launch the treatment on the entire site. There is an increase in transaction rate of 14.1% compared from Back-End user to Upfront users. Additionally consumers identified with cookies in the Back-end Fee group spent 20.64% more than those assigned to the Upfront Fee group. Both supporting evidence we are confident at the 1% level.

```
print("transaction rate")
```

```
## [1] "transaction rate"
```

```
.141 -(.0009 * 2.58)
```

```
## [1] 0.138678
```

```
.141 +(.0009 * 2.58)
```

```
## [1] 0.143322
```

```
print("10 day cookie revenue")
```

```
## [1] "10 day cookie revenue"
```

```
.2064 -(.0138 * 2.58)
```

```
## [1] 0.170796
```

```
.2064 +(.0138 * 2.58)
```

```
## [1] 0.242004
```

How long did this assignment take you to do (hours)? How hard was it (easy, reasonable, hard, too hard)?

Took me around 4-6 hours. The coding part was hard and the reading was long, boring, and semi complicated which made the assignment harder and take longer. Final answer it was pretty hard.