Problem Set 2

Yiyun (Leo) Yao - yy3959 - (Recitation) 002

Due Nov 10, 2023

This homework must be turned in on Brightspace by Nov. 10, 2023. It must be your own work, and your own work only – you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone.

Your homework submission must be written and submitted using Rmarkdown. No handwritten solutions will be accepted. **No zip files will be accepted. Make sure we can read each line of code in the pdf document.** You should submit the following:

1. A compiled PDF file named yourNetID_solutions.pdf containing your solutions to the problems.
2. A .Rmd file containing the code and text used to produce your compiled pdf named yourNetID_solutions.Rmd.

Note that math can be typeset in Rmarkdown in the same way as Latex. Please make sure your answers are clearly structured in the Rmarkdown file:

1. Label each question part
2. Do not include written answers as code comments.
3. The code used to obtain the answer for each question part should accompany the written answer. Comment your code!

Question 1 (Total: 50)

In new democracies and post-conflict settings, Truth and Reconciliation Commissions (TRCs) are often tasked with investigating and reporting about wrongdoing in previous governments. Depending on the context, institutions such as TRCs are expected to reduce hostilities (e.g. racial hostilities) and promote peace.

In 1995, South Africa's new government formed a national TRC in the aftermath of apartheid. [Gibson 2004] uses survey data collected from 2000-2001 to examine whether this TRC promoted inter-racial reconciliation. The outcome of interest is respondent racial attitudes (as measured by the level of agreement with the prompt: "I find it difficult to understand the customs and ways of [the opposite racial group]".) The treatment is "exposure to the TRC" as measured by the individual's level of self-reported knowledge about the TRC.

You will need to use the trc_data.dta file for this question. The relevant variables are:

- RUSTAND - Outcome: respondent's racial attitudes (higher values indicate greater agreement)
- TRCKNOW - Treatment dummy (1 = if knows about the TRC, 0 = otherwise)
- age - Respondent age (in 2001)
- female - Respondent gender
- wealth - Measure of wealth constructed based on asset ownership (assets are fridge, floor polisher, vacuum cleaner, microwave oven, hi-fi, washing machine, telephone, TV, car)
- religiosity - Self-reported religiosity (7 point scale)
- ethsalience - Self-reported ethnic identification (4 point scale)
- rcblack - Respondent is black
- rcwhite - Respondent is white
- rccol - Respondent is coloured (distinct multiracial ethnic group)
- EDUC - Level of education (9 point scale)

Part a (15 points)

Estimate the average treatment effect of TRC exposure on respondents' racial attitudes under the assumption that TRC exposure is ignorable. Report a 95% confidence interval for your estimate and interpret your results. (Use robust standard errors throughout.)

```r
library(tidyverse)
library(haven)
library(estimatr) # for lm with robust se : ?lm_robust()

# Load in the TRC data (it's a STATA .dta so we use the haven package)
TRC_data <- haven::read_dta("trc_data.dta")

# subgroup: treated & control
treated = TRC_data$RUSTAND[TRC_data$TRCKNOW == 1]
control = TRC_data$RUSTAND[TRC_data$TRCKNOW == 0]
# Point Estimate
ateTRC <- mean(treated) - mean(control)
# Standard Error
seTRC <- sqrt(var(treated)/length(treated) + var(control)/length(control))
# 95% CI
ci95TRC <- c(ateTRC - qnorm(.975)*seTRC,
              ateTRC + qnorm(.975)*seTRC)
cat("Point estimate of ATE", ateTRC, "\n")
```

```
## Point estimate of ATE -0.2177317
```

```
cat("Standard error:", seTRC, "\n")
```

```
## Standard error: 0.04433111
```

```
cat("95% confidence interval:", ci95TRC, "\n")
```

```
## 95% confidence interval: -0.3046191 -0.1308444
```

The average treatment effect of TRC exposure on respondents' racial attitudes is approximately -0.22, which indicates that exposure to TRC reduces level of agreement with the prompt. This can be seen from the fact that the control group has higher level of racial attitudes, which is equivalent to higher level of agreement. Therefore, from the average treatment effect, we can conclude that TRC exposure helps reduce racism.

The 95% confidence interval is from -0.305 to -0.131, which does not include 0 (the null hypothesis value for the ATE). We could reject the null hypothesis at a significance level $\alpha = 0.05$. This implies that there is convincing evidence that the treatment has a statistically distinguishable effect. The rejection of the null hypothesis ensures that we are confident in our conclusion.

Part b (15 points)

Examine whether exposed and nonexposed respondents differ on the full set of observed covariates using a series of balance tests. Briefly discuss, in which ways do exposed and nonexposed respondents differ?

```
# Standardize the covariates
TRC_data_Standardized <- TRC_data %>%
mutate(age_std = age/sd(age),
       female_std = female/sd(female),
       wealth_std = wealth/sd(wealth),
       religiosity_std = religiosity/sd(religiosity),
       ethsalience_std = ethsalience/sd(ethsalience),
       rcblack_std = rcblack/sd(rcblack),
       rcwhite_std = rcwhite/sd(rcwhite),
       rccol_std = rccol/sd(rccol),
       EDUC_std = EDUC/sd(EDUC))

# Balance between treated and control
balance_table <- TRC_data_Standardized %>% group_by(TRCKNOW) %>%
  summarize(age_std = mean(age_std),
            female_std = mean(female_std),
            wealth_std = mean(wealth_std),
            religiosity_std = mean(religiosity_std),
            ethsalience_std = mean(ethsalience_std),
            rcblack_std = mean(rcblack_std),
            rcwhite_std = mean(rcwhite_std),
            rccol_std = mean(rccol_std),
            EDUC_std = mean(EDUC_std))
balance_table
```

```
## # A tibble: 2 x 10
##   TRCKNOW age_std female_std wealth_std religiosity_std ethsalience_std
##     <dbl>   <dbl>      <dbl>      <dbl>           <dbl>           <dbl>
## 1       0    2.62      0.866      0.774            2.15            4.69
## 2       1    2.52      1.08       0.928            2.11            4.73
## # i 4 more variables: rcblack_std <dbl>, rcwhite_std <dbl>, rccol_std <dbl>,
## #   EDUC_std <dbl>
```

```
# Take the absolute differences
abs_balance_diff <- abs(balance_table[1, 2:ncol(balance_table)] -
                          balance_table[2, 2:ncol(balance_table)])
abs_balance_diff
```

```
##     age_std female_std wealth_std religiosity_std ethsalience_std rcblack_std
## 1 0.0980385  0.2106527  0.1539068      0.04139381      0.03641136  0.07762829
##   rcwhite_std rccol_std  EDUC_std
## 1  0.03762548 0.1367213 0.3840434
```

We can see that none of the absolute differences are 0. So all covariates differ from exposed to nonexposed respondents. Specifically, respondents exposed to TRC are younger, more female, welathier, less religious, and more educated.

Part c (10 points)

Now assume that TRC exposure is conditionally ignorable given the set of observed covariates:

1. Use a logistic regression model to estimate the propensity score for each observation. (For purposes of this question, do not include any interactions.)
2. With this model, construct inverse propensity of treatment weights (IPTW) for each observation using the unstabilized weights.
3. Use the propensity score to construct an IPW estimator and report the point estimate for the ATE.

Use the following covariates: age, female, wealth, religiosity, ethsalience, rcblack, rcwhite, rccol, EDUC

```
library(broom)
# the logistic regression model to estimate the propensity score for each observation
pscore_model <- glm(TRCKNOW ~ age + female + wealth + religiosity +
                      ethsalience + rcblack + rcwhite + rccol + EDUC,
                    data=TRC_data_Standardized,
                    family=binomial(link="logit"))
tidy(pscore_model)
```

```
## # A tibble: 10 x 5
##    term          estimate  std.error statistic  p.value
##    <chr>            <dbl>      <dbl>     <dbl>    <dbl>
##  1 (Intercept) -2.52      0.313         -8.03  9.81e-16
##  2 age          0.000371  0.00254        0.146 8.84e- 1
##  3 female       0.388     0.0751         5.17  2.32e- 7
##  4 wealth       0.0000244 0.00000685     3.56  3.75e- 4
##  5 religiosity  0.0113    0.0209         0.540 5.89e- 1
##  6 ethsalience  0.0601    0.0650         0.925 3.55e- 1
##  7 rcblack      0.472     0.152          3.10  1.93e- 3
##  8 rcwhite     -0.280     0.163         -1.71  8.66e- 2
##  9 rccol       -0.215     0.171         -1.26  2.08e- 1
## 10 EDUC         0.392     0.0396         9.91  3.64e-23
```

```
# Get the propensity scores for each observation
TRC_data_Standardized$e <- predict(pscore_model, type = "response")

# Generate the weights (unstabilized)
TRC_data_Standardized$wt <- NA
TRC_data_Standardized$wt[TRC_data_Standardized$TRCKNOW == 1] <-
  1/TRC_data_Standardized$e[TRC_data_Standardized$TRCKNOW==1]
TRC_data_Standardized$wt[TRC_data_Standardized$TRCKNOW == 0] <-
  1/(1 -TRC_data_Standardized$e[TRC_data_Standardized$TRCKNOW==0])
point_wtd <-
  mean(TRC_data_Standardized$wt * TRC_data_Standardized$RUSTAND *
         TRC_data_Standardized$TRCKNOW - TRC_data_Standardized$wt *
         TRC_data_Standardized$RUSTAND * (1-TRC_data_Standardized$TRCKNOW))
point_wtd
```

```
## [1] -0.162329
```

Part d (10 points)

Using the bootstrap method (resampling individual rows of the data with replacement), obtain an estimate for the standard error of your IPTW estimator for the ATE. Compute a 95% confidence interval and interpret your findings. (You should report estimate, standard error, 95% CI lower, 95% CI upper, for interpretation, compare your results in Part C/D to your estimate from Part A and briefly discuss your findings.)

```
# Set random seed
set.seed(123)

#IPTW Bootstrap
n_iter <- 1000 # Suggested number of iterations
ate_boot <- rep(NA, n_iter) # Placeholder to store estimates

# For each iteration
for(boot in 1:n_iter){
  # Resample rows with replacement
  TRC_boot <- TRC_data_Standardized[sample(1:nrow(TRC_data_Standardized),
                                          nrow(TRC_data_Standardized),
                                          replace=T),] #replace = T is key!
  # Fit the propensity score model on the bootstrapped data
  pscore_model_boot <- glm(TRCKNOW ~ age + female + wealth + religiosity +
                             ethsalience + rcblack + rcwhite + rccol + EDUC,
                           data=TRC_boot, family=binomial(link="logit"))
  # Save the propensities
  TRC_boot$e <- predict(pscore_model_boot, type = "response")
  # Calculate the weights
  TRC_boot$wt <- NA
  TRC_boot$wt[TRC_boot$TRCKNOW == 1] <- 1/TRC_boot$e[TRC_boot$TRCKNOW==1]
  TRC_boot$wt[TRC_boot$TRCKNOW == 0] <- 1/(1 - TRC_boot$e[TRC_boot$TRCKNOW==0])
  # Compute and store the ATE
  ate_boot[boot] <-
    mean(TRC_boot$wt * TRC_boot$RUSTAND * TRC_boot$TRCKNOW -
         TRC_boot$wt * TRC_boot$RUSTAND * (1-TRC_boot$TRCKNOW))
}
mean(ate_boot)
```

```
## [1] -0.1596519
```

```r
# Take the SD of the ate_boot to get our estimated SE - can do asymptotic inference
sd(ate_boot)
```

```
## [1] 0.04534277
```

```r
# Asymptotic 95\% CI
c(point_wtd - qnorm(.975)*sd(ate_boot),
point_wtd + qnorm(.975)*sd(ate_boot))
```

```
## [1] -0.25119917 -0.07345878
```

The average of the ATE estimates obtained from the bootstrap resampling procedure is -0.16. In the bootstrap, each iteration of involves resampling the data, fitting a propensity score model, calculating weights, and then computing the ATE. The mean of these ATE estimates provides a point estimate at -0.16, which, like our estimate from Part A (-0.22), tells us exposure to TRC reduces agreement. The difference between the two values is caused by whether or not taking into account of variability and uncertainty.

The standard error from the bootstrap is 0.045, very close to our results from Part A (0.044). And the 95% confidence interval is from -0.25 to -0.07. Like the previous one from Part A, it does not include 0, so we can reject the null and claim that the treatment has a statistically distinguishable effect. There's one difference that this new confidence interval result is closer to 0.

Question 2 (Total: 50 points)

Use the same data set as in Question 1.

Part a (15 points)

Estimate the ATT of TRC exposure on respondents' racial attitudes using the MatchIt approach. You can use the matchit function from MatchIt package in R. Implement the nearest neighbor matching algorithm and estimate the ATT. Report the 95% confidence interval of your estimate.

```r
library(MatchIt)
```

```
## Warning: package 'MatchIt' was built under R version 4.3.2
```

```r
# Read the help file first! Check out the default settings
# ?matchit()
library(estimatr)
trc_m_nn <- matchit(TRCKNOW ~ age + female + wealth + religiosity +
                           ethsalience + rcblack + rcwhite + rccol + EDUC,
                         data = TRC_data, method = "nearest",
                         link ="logit", distance = "glm")
# Checking balance after NN matching
summary(trc_m_nn, un = FALSE)
```

```
##
## Call:
## matchit(formula = TRCKNOW ~ age + female + wealth + religiosity +
##     ethsalience + rcblack + rcwhite + rccol + EDUC, data = TRC_data,
##     method = "nearest", distance = "glm", link = "logit")
##
## Summary of Balance for Matched Data:
##             Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance           0.4831        0.4511          0.2593     1.3885    0.0699
## age               38.9402       39.2231         -0.0191     0.9094    0.0140
## female             0.5379        0.4899          0.0962          .    0.0479
## wealth          6945.1703     6317.7762          0.0824     1.0291    0.0373
## religiosity        3.8402        3.9444         -0.0563     1.0746    0.0149
## ethsalience        2.7345        2.7359         -0.0025     1.0065    0.0007
## rcblack            0.5518        0.5476          0.0084          .    0.0042
## rcwhite            0.2696        0.2550          0.0329          .    0.0146
## rccol              0.1105        0.1279         -0.0554          .    0.0174
## EDUC               4.2919        4.0361          0.2142     1.3978    0.0327
##             eCDF Max Std. Pair Dist.
## distance      0.1612          0.2595
## age           0.0368          1.1572
## female        0.0479          0.8544
## wealth        0.0674          0.8969
## religiosity   0.0306          1.1307
## ethsalience   0.0021          0.7030
## rcblack       0.0042          0.8328
## rcwhite       0.0146          0.8315
## rccol         0.0174          0.5475
## EDUC          0.0827          0.4738
##
## Sample Sizes:
##           Control Treated
## All          1766    1439
## Matched      1439    1439
## Unmatched     327       0
## Discarded       0       0
```

```
m_data_nn <- match.data(trc_m_nn)
treated_nn <- m_data_nn$RUSTAND[m_data_nn$TRCKNOW == 1]
control_nn <- m_data_nn$RUSTAND[m_data_nn$TRCKNOW == 0]
# Point Estimate
ate_nn <- mean(treated_nn) - mean(control_nn)
cat("Point estimate of ATE", ate_nn, "\n")
```

```
## Point estimate of ATE -0.2140375
```

```
# Standard Error
se_nn <- sqrt(var(treated_nn)/length(treated_nn) +
                var(control_nn)/length(control_nn))
cat("Standard error:", se_nn, "\n")
```

```
## Standard error: 0.04663497
```

```
# 95% CI
ci95_nn <- c(ate_nn - qnorm(.975)*se_nn, ate_nn + qnorm(.975)*se_nn)
cat("95% confidence interval:", ci95_nn, "\n")
```

```
## 95% confidence interval: -0.3054404 -0.1226347
```

The 95% confidence interval [-0.3054404, -0.1226347] does not include 0, so we could reject the null hypothesis and claim that the treatment has statistically distinguishable effect.

Part b (15 points)

Estimate the ATT of TRC exposure on respondents' racial attitudes using the MatchIt approach. You can use the matchit function from MatchIt package in R. Implement the exact matching algorithm and estimate the ATT. Report the 95% confidence interval of your estimate.

```
trc_m_exa <- matchit(TRCKNOW ~ age + female + wealth + religiosity
                     + ethsalience + rcblack + rcwhite + rccol + EDUC,
                     data = TRC_data, method = "exact", distance = "glm")
summary(trc_m_exa, un = FALSE)
```

```
##
## Call:
## matchit(formula = TRCKNOW ~ age + female + wealth + religiosity +
##     ethsalience + rcblack + rcwhite + rccol + EDUC, data = TRC_data,
##     method = "exact", distance = "glm")
##
## Summary of Balance for Matched Data:
##              Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## age                31.8523       31.8523               0      0.998         0
## female              0.5455        0.5455               0          .         0
## wealth           3852.2727     3852.2727               0      0.998         0
## religiosity         4.1818        4.1818               0      0.998         0
## ethsalience         2.9659        2.9659               0      0.998         0
## rcblack             0.7955        0.7955               0          .         0
## rcwhite             0.1705        0.1705               0          .         0
## rccol               0.0227        0.0227               0          .         0
## EDUC                4.0114        4.0114               0      0.998         0
##              eCDF Max Std. Pair Dist.
## age                 0               0
## female              0               0
## wealth              0               0
## religiosity         0               0
## ethsalience         0               0
## rcblack             0               0
## rcwhite             0               0
## rccol               0               0
## EDUC                0               0
##
## Sample Sizes:
##                Control Treated
## All            1766.      1439
## Matched (ESS)   75.06       88
## Matched          92.        88
## Unmatched      1674.      1351
## Discarded         0.         0
```

```r
m_data_exa <- match.data(trc_m_exa)
treated_exa <- m_data_exa$RUSTAND[m_data_exa$TRCKNOW == 1]
control_exa <- m_data_exa$RUSTAND[m_data_exa$TRCKNOW == 0]
# Point Estimate
ate_exa <- mean(treated_exa) - mean(control_exa)
cat("Point estimate of ATE", ate_exa, "\n")
```

```
## Point estimate of ATE 0.1067194
```

```r
# Standard Error
se_exa <- sqrt(var(treated_exa)/length(treated_exa) +
               var(control_exa)/length(control_exa))
cat("Standard error:", se_exa, "\n")
```

```
## Standard error: 0.1815278
```

```r
# 95% CI
ci95_exa <- c(ate_exa - qnorm(.975)*se_exa, ate_exa + qnorm(.975)*se_exa)
cat("95% confidence interval:", ci95_exa, "\n")
```

```
## 95% confidence interval: -0.2490687 0.4625074
```

The 95% confidence interval [-0.2490687, 0.4625074] includes 0. Therefore, we could not reject the null hypothesis and we do not have convincing evidence that the treatment has statistically distinguishable effect.

Part c (10 points)

Estimate the ATT of TRC exposure on respondents' racial attitudes using the MatchIt approach. You can use the matchit function from MatchIt package in R. Implement the **coarsened exact matching** algorithm and estimate the ATT. Report the 95% confidence interval of your estimate.

```r
trc_m_cem <- matchit(TRCKNOW ~ age + female + wealth + religiosity +
                       ethsalience + rcblack + rcwhite + rccol + EDUC,
                     data = TRC_data, method = "cem", distance = "glm")
summary(trc_m_cem, un = FALSE)
```

```
##
## Call:
## matchit(formula = TRCKNOW ~ age + female + wealth + religiosity +
##     ethsalience + rcblack + rcwhite + rccol + EDUC, data = TRC_data,
##     method = "cem", distance = "glm")
##
## Summary of Balance for Matched Data:
##             Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## age               35.3733       35.4106         -0.0025     0.9923    0.0027
## female            0.5083       0.5083          0.0000          .    0.0000
## wealth         4555.9780    4519.6104          0.0048     0.9922    0.0156
## religiosity       4.1019       4.1019          0.0000     0.9995    0.0000
## ethsalience       2.8760       2.8760          0.0000     0.9995    0.0000
## rcblack           0.7287       0.7287          0.0000          .    0.0000
## rcwhite           0.1791       0.1791          0.0000          .    0.0000
## rccol             0.0744       0.0744          0.0000          .    0.0000
## EDUC              3.9587       3.9587          0.0000     0.9995    0.0000
##              eCDF Max Std. Pair Dist.
## age            0.0140          0.1107
## female         0.0000          0.0000
## wealth         0.0698          0.0398
## religiosity    0.0000          0.0000
## ethsalience    0.0000          0.0000
## rcblack        0.0000          0.0000
## rcwhite        0.0000          0.0000
## rccol          0.0000          0.0000
## EDUC           0.0000          0.0000
##
## Sample Sizes:
##                Control Treated
## All             1766.     1439
## Matched (ESS)    519.4     726
## Matched          802.      726
## Unmatched        964.      713
## Discarded          0.        0
```

```
m_data_cem <- match.data(trc_m_cem)
treated_cem <- m_data_cem$RUSTAND[m_data_cem$TRCKNOW == 1]
control_cem <- m_data_cem$RUSTAND[m_data_cem$TRCKNOW == 0]
# Point Estimate
ate_cem <- mean(treated_cem) - mean(control_cem)
cat("Point estimate of ATE", ate_cem, "\n")
```

```
## Point estimate of ATE -0.1375487
```

```
# Standard Error
se_cem <- sqrt(var(treated_cem)/length(treated_cem) +
                var(control_cem)/length(control_cem))
cat("Standard error:", se_cem, "\n")
```

```
## Standard error: 0.06311059
```

```
# 95% CI
ci95_cem <- c(ate_cem - qnorm(.975)*se_cem, ate_cem + qnorm(.975)*se_cem)
cat("95% confidence interval:", ci95_cem, "\n")
```

```
## 95% confidence interval: -0.2612432 -0.01385421
```

The 95% confidence interval [-0.2612432, -0.01385421] does not include 0, so we could reject the null hypothesis and claim that the treatment has statistically distinguishable effect.

part d (10 points)

Compare and contrast the three different matching algorithms. Provide evidence and an argument about which one we should use.

Exact matching algorithm is a strong matching method for its distribution is exactly the same for treated and matched controls. However, the weakness of exact matching regarding is its sample size. For exact matching algorithm, there are only 92 respondents matched in the control group and 88 in the treated. On the other hand, for coarsened exact matching, we have 802 respondents matched in the control group and 726 in the treated, and for nearest neighbor matching algorithm, there are only 327 unmatched respondents left in the control group. The sample size is too small for the exact matching algorithm, comparing to the other two algorithms, to get a good estimate. Nearest neighbor matching algorithm is good at reducing bias because it's constantly making the closest matches. However,it might also cause uncertainty. Personally, I would use coarsened exact matching as it is a balance of the exact matching algorithm and nearest neighbor matching algorithm.

BONUS ONLY: Question 3 (Total: Up to +12)

Question 3 is for bonus points. (See forthcoming lecture on Nov. 7th)

part a (+4 points)

Using the regression method to predict potential outcomes for all individuals in the dataset and calculate the ATE with bootstrapped standard errors. Report and interpret your results. (Hint: Start by fitting the treatment and control model with subsets of the data.)

```r
## Fit a model among TRCKNOW == 1 to get E[Y_i(1) | X]
treatment_model <- lm_robust(RUSTAND ~ age + female + wealth + religiosity
                             + ethsalience + rcblack + rcwhite + rccol + EDUC,
                             data=subset(TRC_data, TRCKNOW == 1))

## Fit a model among TRCKNOW == 0 to get E[Y_i(0) | X]
control_model <- lm_robust(RUSTAND ~ age + female + wealth + religiosity
                             + ethsalience + rcblack + rcwhite + rccol + EDUC,
                             data=subset(TRC_data, TRCKNOW == 0))

## Predict the potential outcome under treatment for all units
TRC_data$RUSTAND_treated <- predict(treatment_model, newdata = TRC_data)

## Predict the potential outcome under control for all units
TRC_data$RUSTAND_control <- predict(control_model, newdata = TRC_data)

## Average of the differences
ate_reg = mean(TRC_data$RUSTAND_treated - TRC_data$RUSTAND_control)
cat("Average of the differences:", ate_reg, "\n")
```

```
## Average of the differences: -0.1743866
```

```r
### Bootstrap for SEs
set.seed(123)
nBoot <- 2000 # Number of iterations
boot_results <- rep(NA, nBoot)
```

```r
for (i in 1:nBoot){
  # Resample with replacement
  TRC_data_boot <- TRC_data[sample(1:nrow(TRC_data), nrow(TRC_data), replace=T),]
  # Fit a model among TRCKNOW == 1 to get E[Y_i(1) | X]
  treatment_model_boot <- lm_robust(RUSTAND ~ age + female + wealth
                                    + religiosity + ethsalience + rcblack
                                    + rcwhite + rccol + EDUC,
                                    data=subset(TRC_data_boot, TRCKNOW==1))
  # Fit a model among TRCKNOW == 0 to get E[Y_i(0) | X]
  control_model_boot <- lm_robust(RUSTAND ~ age + female + wealth
                                  + religiosity + ethsalience + rcblack
                                  + rcwhite + rccol + EDUC,
                                  data=subset(TRC_data_boot, TRCKNOW==0))
  # Predict the potential outcome under treatment for all units
  TRC_data_boot$RUSTAND_treated_boot <- predict(treatment_model_boot,
                                                newdata = TRC_data_boot)
  # Predict the potential outcome under control for all units
  TRC_data_boot$RUSTAND_control_boot <- predict(control_model_boot,
                                                newdata = TRC_data_boot)
  # Store bootstrapped estimate
  boot_results[i] <- mean(TRC_data_boot$RUSTAND_treated_boot -
                            TRC_data_boot$RUSTAND_control_boot)
}

# ATE
mean(boot_results)
```

```
## [1] -0.1740045
```

```r
# Take the SD of the ate_boot to get our estimated SE - can do asymptotic inference
sd(boot_results)
```

```
## [1] 0.04465872
```

The ATE of TRC exposure is approximately -0.17, which means exposure reduces agreement. The control group has higher level of agreement. The TRC exposure reduces racism. The standard error is approximately 0.045.

part b (+4 points)

Using the regression method to predict potential outcomes for all individuals and calculate the ATT with bootstrapped standard errors. Report and interpret your results.

```r
ATT_reg = mean(TRC_data$RUSTAND_treated[TRC_data$TRCKNOW == 1]-
TRC_data$RUSTAND_control[TRC_data$TRCKNOW == 1])
ATT_reg
```

```
## [1] -0.2033737
```

```r
### Bootstrap for SEs
set.seed(123)
nBoot <- 2000 # Number of iterations
boot_results_ATT <- rep(NA, nBoot)
for (iter in 1:nBoot){
  # Resample w/ replacement
  TRC_ATT_boot <- TRC_data[sample(1:nrow(TRC_data), nrow(TRC_data), replace=T),]
  ## Fit a model among TRCKNOW == 1 to get E[Y_i(1) | X]
  treatment_ATT_boot <- lm_robust(RUSTAND ~ age + female + wealth + religiosity
                                  + ethsalience + rcblack + rcwhite + rccol + EDUC,
                                  data=subset(TRC_ATT_boot, TRCKNOW==1))
  ## Fit a model among TRCKNOW == 0 to get E[Y_i(0) | X]
  control_model_boot <- lm_robust(RUSTAND ~ age + female + wealth + religiosity
                                  + ethsalience + rcblack + rcwhite + rccol + EDUC,
                                  data=subset(TRC_ATT_boot, TRCKNOW==0))
  ## Predict the potential outcome under treatment for all units
  TRC_ATT_boot$RUSTAND_treated_boot <- predict(treatment_ATT_boot,
                                               newdata = TRC_ATT_boot)
  ## Predict the potential outcome under control for all units
  TRC_ATT_boot$RUSTAND_control_boot <- predict(control_model_boot,
                                               newdata = TRC_ATT_boot)
  ## Store bootstrapped estimate
  boot_results_ATT[iter] <- mean(TRC_ATT_boot$RUSTAND_treated_boot[TRC_ATT_boot$TRCKNOW==1]
                                 - TRC_ATT_boot$RUSTAND_control_boot[TRC_ATT_boot$TRCKNOW==1])
}

### ATT
mean(boot_results_ATT)
```

```
## [1] -0.2030773
```

```r
### Standard error
sd(boot_results_ATT)
```

```
## [1] 0.04641271
```

The ATE of TRC exposure is approximately -0.20, which means exposure reduces agreement. The control group has higher level of agreement. The TRC exposure reduces racism. The standard error is approximately 0.046.

part c (+4 points)

Compare and contrast the ATE and ATT from the regression approach.

ATE has an average of differences of -0.17 with a standard error of 0.045. ATT has an average of differences of -0.20 with a standard error of 0.046. There is a 0.03 difference in the average of differences and a 0.001 difference in the standard errors. we could see that the coefficient of ATE and ATT is relatively different, and the standard errors of ATE and ATT is relatively similar.

The difference between ATT and ATE exist because ATE is the average of the individual treatment effects of the population, where we examine the difference between the effect of the treated and control group, whereas ATT is the average of the individual treatment effects of the treated. We know that standard deviation is an indicator of variability between observations. Thus, the similarity is probably caused by the fact that we are using the same dataset when estimating ATE and ATT.