Problem Set 2

Ghina Al Shdaifat - gha2009 - 003

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)

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

# Estimate the average treatment effect of TRC exposure
# HC1 -> to ensure standard error is robust
model <- lm_robust(RUSTAND ~ TRCKNOW, data = TRC_data, se_type = "HC1")

# Output the results with a 95% confidence interval
tidy(model, conf.int = TRUE)
```

```
##          term   estimate  std.error statistic      p.value   conf.low
## 1 (Intercept)  2.5311438 0.02805842 90.209762 0.000000e+00  2.4761295
## 2     TRCKNOW -0.2177317 0.04433068 -4.911536 9.489329e-07 -0.3046511
##    conf.high   df outcome
## 1  2.5861581 3203 RUSTAND
## 2 -0.1308123 3203 RUSTAND
```

Interpretation:

- The coefficient for the treatment dummy ('TRCKNOW) is -0.2177, with a standard error of 0.04433. In this case, the negative sign indicates that as knowledge about the TRC increases, the level of agreement decreases, indicating an inverse relationship between both the treatment and outcome and hence suggesting an improvement in racial attitudes.
- The corresponding p-value is approximately 0.0000009489, which is much lower than the alpha threshold of 0.05, indicating that this is a statistically significant result.
- The 95% confidence interval for the `TRCKNOW` coefficient ranges from -0.3047 to -0.1308, which means we can say with 95% confidence that the true effect of TRC knowledge on racial attitudes lies within this interval. Since the interval does not include 0, this further validates that the effect is statistically significant.

Overall, the results indicate that there is a statistically significant association between knowledge of the TRC and racial attitudes, with greater knowledge correlating with more understanding towards customs and ways of the opposite racial group. In other words, TRC reduced respondent's racial attitudes by 0.2177.

Part b (15 points)

Examine whether exposed and non-exposed 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?

```r
balance_results <- data.frame(Covariate = character(),
                              Mean_Exposed = numeric(),
                              Mean_Non_Exposed = numeric(),
                              T_Value = numeric(),
                              P_Value = numeric(),
                              stringsAsFactors = FALSE)

covariates <- c("age", "female", "wealth", "religiosity", "ethsalience",
                "rcblack", "rcwhite", "rccol", "EDUC")

for (covariate in covariates){
  exposed_data <- TRC_data %>% filter(TRCKNOW == 1) %>% pull(!!sym(covariate))
  non_exposed_data <- TRC_data %>% filter(TRCKNOW == 0) %>%
    pull(!!sym(covariate))

  # Mean
  mean_exposed <- mean(exposed_data, na.rm = TRUE)
  mean_non_exposed <- mean(non_exposed_data, na.rm = TRUE)

  # t-test
  test_result <- t.test(exposed_data, non_exposed_data, var.equal = TRUE)

# Append data frame with values
  balance_results <- balance_results %>%
    add_row(Covariate = covariate,
            Mean_Exposed = mean_exposed,
            Mean_Non_Exposed = mean_non_exposed,
            T_Value = test_result$statistic,
            P_Value = test_result$p.value)
```

```r
}

# Print output
print(balance_results)
```

```
##     Covariate Mean_Exposed Mean_Non_Exposed   T_Value      P_Value
## 1         age   38.9402363       40.4546999 -2.763489 5.751392e-03
## 2      female    0.5378735        0.4326161  5.963605 2.736259e-09
## 3      wealth 6945.1702571     5792.7746319  4.345886 1.430278e-05
## 4 religiosity    3.8401668        3.9156285 -1.165659 2.438392e-01
## 5 ethsalience    2.7345379        2.7134768  1.025303 3.052978e-01
## 6     rcblack    0.5517721        0.5130238  2.187198 2.880017e-02
## 7     rcwhite    0.2696317        0.2531144  1.059503 2.894507e-01
## 8       rccol    0.1104934        0.1574179 -3.858219 1.164664e-04
## 9        EDUC    4.2918694        3.8465459 11.015352 1.007117e-27
```

Interpretation:

The analysis of the dataset reveals significant imbalances between the treated and control groups. Utilizing a p-value threshold of 0.05, notable differences are observed in several key attributes: age, female gender, wealth, rcblack, rccol, and EDUC. Considering that out of nine examined attributes, six exhibit statistically significant differences in their group averages, it is clear that the groups are imbalanced.

Specifically, the age group exposed to the treatment is, on average, approximately a year and a half younger than their non-exposed counterparts, indicated by an absolute t-value of 2.76. The table above also demonstrates that the exposed group contains approximately 10% more females, with a t-value of 5.96, while simultaneously being wealthier. The proportion of black respondents is 55.18% in the exposed group and 51.30% in the non-exposed group. The positive t-value (2.187198) and p-value (0.02880017) indicate a significant difference; on the other hand, white respondents have no significant difference with a p-value of 0.2894507 and a t-value of 1.059503. Generally, there is a lower proportion of colored respondents in the exposed group and this difference (11.05% of the exposed group and 15.74%) is significant with a p-value of 0.0001164664 and a negative t-value of -3.858219. Finally, the mean education level is higher for the exposed group (4.2919) compared to the non-exposed group (3.8465). The high positive t-value (11.015352) and very low p-value (1.007117e-27) suggest a significant difference, indicating higher education levels among the exposed group.

Overall, we can observe that there are indeed significant differences between the exposed and non-exposed respondents in regard to specific covariates. The groups that are not balanced on the other covariates could potentially influence the assumption of ignorability and hence lead to inaccuracies in identifying casual inferences that are drawn from the treatment effect analysis. Since there are differences in statistical significance between both groups, we can assume that the groups are not balanced.

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

```r
# 1. Logistic regression model
propensity_model <- glm(TRCKNOW ~ age + female + wealth + religiosity +
                        ethsalience + rcblack + rcwhite + rccol + EDUC, data =
                          TRC_data, family = binomial(link = "logit"))
# Extract predicted propensity scores
TRC_data$propensity_score <- predict(propensity_model, type = "response")

# 2. IPTW using unstabalized weights
TRC_data$iptw <- ifelse(TRC_data$TRCKNOW == 1,
                        1 / TRC_data$propensity_score,
                        1 / (1 - TRC_data$propensity_score))

# 3. Calculate the IPTW ATE using weighted means
TRC_data <- TRC_data %>%
  mutate(weighted_outcome_treat = ifelse(TRCKNOW == 1, RUSTAND * iptw, 0),
         weighted_outcome_control = ifelse(TRCKNOW == 0, RUSTAND * iptw, 0))

# Sum the weighted outcomes and divide by the sum of weights for each group
ate_treat <- sum(TRC_data$weighted_outcome_treat) /
  sum(TRC_data$iptw[TRC_data$TRCKNOW == 1])
ate_control <- sum(TRC_data$weighted_outcome_control) /
  sum(TRC_data$iptw[TRC_data$TRCKNOW == 0])

ate_point_estimate <- ate_treat - ate_control

# Print the ATE point estimate
print("Average Treatment Effect Point Estimate")
```

```
## [1] "Average Treatment Effect Point Estimate"
```

```r
ate_point_estimate
```

```
## [1] -0.1631028
```

Interpretation

- The negative sign of the estimate suggests that exposure to the TRC is associated with a decrease in the level of agreement with the statement. In other words, those who are aware of the TRC are estimated to find it less difficult or challenging to understand the customs and ways of the opposite racial group, compared to those who are not aware of the TRC.
- The estimate has a relatively small magnitude, suggesting that while there is an association, the effect of the TRC exposure on racial attitudes or behavior is modest.
- In this scenario, we are assuming that the TRC exposure is conditionally ignorable given the covariates (i.e, no unmeasured confounders that affect both the treatment assignment and the outcome), which means that the point estimate can be interpreted causally. Therefore, it suggests that the TRC exposure has a causal impact in reducing negative racial attitudes.

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.)

```r
# Set random seed
set.seed(123)
nBoot <- 1000 # Number of iterations
ate_boot <- rep(NA, nBoot) # Placeholder to store estimates

# For each iteration
for(boot in 1:nBoot){
  # Re-sample rows with replacement
  TRC_boot <- TRC_data[sample(1:nrow(TRC_data), nrow(TRC_data), replace=T),]

  # 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))}

# SD of the ate_boot to get our estimated SE
sd(ate_boot)
```

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

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

```
## [1] -0.25197297 -0.07423258
```

```r
# Estimate
mean(ate_boot)
```

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

Interpretation:

In part d, we implemented the bootstrap method to identify ATE using IPTW. The value of -0.1631028 represents the estimates difference in racial attitudes between those who were and were not exposed to the TRC. Similar to our interpretation in part a, the negative value of -0.1596519 indicates that TRC leads to a reduction in negative racial attitudes. To identify the significance of the effect, the confidence interval ranges from -0.25197297 to -0.0742325 and does not include 0, suggesting that the effect is indeed statistically significant at the 95% confidence level.

Comparison:

The estimated ATE in part c implements a more directed approach without bootstrapping. As shown in the output, the point estimate from part c is close to the one in part d (-0.1631028). On the other hand, part a measures the effect of TRC exposure on racial attitudes using a direct linear regression model. The coefficient for `TRCKNOW` was -0.2177, which is slightly larger in magnitude than the ATE estimates in parts c and d. The difference might be due to the varying estimate methods. In other words, the coefficient in part a represents the change in the outcome variable for each unit change in the predictor, while part c and d (IPTW and bootstrapping) adjust for potential confounders by weighting observations, which might lead to a more nuanced estimate of the treatment effect.

Overall, all parts consistently show that knowledge about the TRC is associated with a decrease in the difficulty of understanding customs and behaviors of the opposite racial group, suggesting a positive impact of TRC exposure on racial attitudes. The bootstrapped confidence intervals in part d provide a more solid understanding of the estimate's variability and further validates the statistical significance of the effect. Although the effect sizes are fairly small, they are consistent and statistically significant across the different methods, reinforcing the importance of TRC exposure in influencing racial attitudes and that those with TRC exposure find it less difficult to understand the customs and ways of the opposite racial group.

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)
library(estimatr)

# Perform nearest neighbor matching
# For each treated unit, one nearest control unit will be selected as its match
matchit_model <- matchit(TRCKNOW ~ age + female + wealth + religiosity +
                         ethsalience + rcblack + rcwhite + rccol + EDUC,
                         data = TRC_data, method = "nearest", estimand = "ATT",
                         distance = "glm")

# Estimate the ATT
matched_data <- match.data(matchit_model, weights = "new_weights")

att_model <- lm_robust(RUSTAND ~ TRCKNOW, data = matched_data, se_type = "HC1")

# Obtain the point estimate and robust standard error for ATT
att_estimate <- coef(summary(att_model))["TRCKNOW", "Estimate"]
att_se <- coef(summary(att_model))["TRCKNOW", "Std. Error"]

# Construct the 95% confidence interval
ci_lower <- att_estimate - qnorm(0.975) * att_se
ci_upper <- att_estimate + qnorm(0.975) * att_se
```

```r
# Report the ATT and confidence interval
list(ATT = att_estimate, SE = att_se, CI_lower = ci_lower, CI_upper = ci_upper)
```

```
## $ATT
## [1] -0.2140375
##
## $SE
## [1] 0.04663497
##
## $CI_lower
## [1] -0.3054404
##
## $CI_upper
## [1] -0.1226347
```

Interpretation:

The value of -0.2140375 for the ATT estimate represents the average effect of TRC exposure on the racial attitudes of those respondents who were actually exposed to the TRC. A negative value indicates that, on average, exposure to the TRC is associated with a decrease in the level of agreement with the statement regarding difficulty in understanding customs and ways of the opposite racial group. In other words, those exposed to the TRC, on average, report a more favorable or understanding attitude towards the customs and ways of the opposite racial group compared to what they would have reported had they not been exposed. The confidence interval ranges from -0.3054404 to -0.1226347 suggesting that we can be 95% confident that the true ATT lies within this range. Considering that the entire interval is below zero strengthens the evidence that TRC exposure has a statistically negative effect on the level of agreement with the statement regarding the difficulty of understanding customs of the opposite racial group.

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.

```r
library(MatchIt)
library(estimatr)

# Perform nearest neighbor matching
# For each treated unit, one nearest control unit will be selected as its match
matchit_model <- matchit(TRCKNOW ~ age + female + wealth + religiosity +
                         ethsalience + rcblack + rcwhite + rccol + EDUC,
                         data = TRC_data, method = "exact", estimand = "ATT",
                         distance = "glm")

# Estimate the ATT
matched_data <- match.data(matchit_model, weights = "new_weights")

att_model <- lm_robust(RUSTAND ~ TRCKNOW, data = matched_data, se_type = "HC1")

# Obtain the point estimate and robust standard error for ATT
att_estimate <- coef(summary(att_model))["TRCKNOW", "Estimate"]
att_se <- coef(summary(att_model))["TRCKNOW", "Std. Error"]
```

```r
# Construct the 95% confidence interval
ci_lower <- att_estimate - qnorm(0.975) * att_se
ci_upper <- att_estimate + qnorm(0.975) * att_se

# Report the ATT and confidence interval
list(ATT = att_estimate, SE = att_se, CI_lower = ci_lower, CI_upper = ci_upper)
```

```
## $ATT
## [1] 0.1067194
##
## $SE
## [1] 0.181528
##
## $CI_lower
## [1] -0.249069
##
## $CI_upper
## [1] 0.4625077
```

Interpretation:

Unlike the previous estimate with nearest neighbor matching, the ATT estimate value is positive (0.1067194). This suggests that, on average, exposure to the TRC is associated with a slight increase in the level of agreement with the statement regarding difficulty in understanding customs and ways of the opposite racial group. However, given the context of the study, this might imply that those exposed to the TRC are slightly more likely to acknowledge differences or complexities in understanding the customs and ways of the opposite racial group. While the standard error is relatively large in comparison to the ATT estimate (0.181528) indicating a higher level of uncertainty or variability in the estimate, the wide confidence interval (ranging from -0.249069 to 0.4625077) which crosses zero suggests that the effect of TRC exposure on racial attitudes is actually not statistically significant at the 95% confidence level. Such a wide range in both positive and negative values indicates that the data is also consistent with a wide range of effects of TRC exposure on racial attitudes.

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
library(MatchIt)
library(estimatr)

# Perform nearest neighbor matching
# For each treated unit, one nearest control unit will be selected as its match
matchit_model <- matchit(TRCKNOW ~ age + female + wealth + religiosity +
                          ethsalience + rcblack + rcwhite + rccol + EDUC,
                          data = TRC_data, method = "cem", estimand = "ATT",
                          distance = "glm")

# Estimate the ATT
matched_data <- match.data(matchit_model, weights = "new_weights")
```

```r
att_model <- lm_robust(RUSTAND ~ TRCKNOW, data = matched_data, se_type = "HC1")

# Obtain the point estimate and robust standard error for ATT
att_estimate <- coef(summary(att_model))["TRCKNOW", "Estimate"]
att_se <- coef(summary(att_model))["TRCKNOW", "Std. Error"]

# Construct the 95% confidence interval
ci_lower <- att_estimate - qnorm(0.975) * att_se
ci_upper <- att_estimate + qnorm(0.975) * att_se

# Report the ATT and confidence interval
list(ATT = att_estimate, SE = att_se, CI_lower = ci_lower, CI_upper = ci_upper)
```

```
## $ATT
## [1] -0.1375487
##
## $SE
## [1] 0.06311022
##
## $CI_lower
## [1] -0.2612424
##
## $CI_upper
## [1] -0.01385493
```

Interpretation:

An ATT estimate of -0.1375487 indicates that, on average, exposure to the TRC is associated with a decrease in the level of agreement with the statement about difficulty understanding customs and ways of the opposite racial group among those respondents who were actually exposed to the TRC. A negative ATT value implies that TRC exposure leads to more positive or understanding attitudes towards the opposite racial group. The standard error value of 0.06311022 and a 95% confidence interval ranging from -0.2612424 to -0.01385493 suggests that the negative effect of TRC exposure on the stated measure of racial attitudes is statistically significant at the 95% confidence level. The interval also indicates the range within which the true ATT is likely to fall.

part d (10 points)

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

Part a: Nearest Neighbor Matching - ATT Estimate: -0.2140375 - Standard Error: 0.04663497 - Confidence Interval: [-0.3054404, -0.1226347]

The nearest neighbor matching method often yields a good balance between both the treatment and control groups since it is flexible, and therefore does not require exact matches on covariates. This can be beneficial with continuous variables or large datasets, yet simultaneously can lead to sensitivity to the choice of distance metric. As a result, this could lead to inaccuracies in identifying the number of neighbors chosen, allowing for bad matches to take place if the nearest neighbor is not very close, where there are not enough close control units for each treated unit.

Part b: Exact Matching - ATT Estimate: 0.1067194 - Standard Error: 0.181528 - Confidence Interval: [-0.249069, 0.4625077]

The exact matching method ensures a perfect match on covariates, where treatment and control units that have identical values on all covariates are paired. As a result, this eliminates the possibility of model dependency in matching. However simultaneously, this also leads to a significant loss of data, resulting in a very wide range of confidence intervals (as seen in our output), indicating lower precision in our estimates.

Part c: Coarsened Exact Matching: - ATT Estimate: -0.1375487 - Standard Error: 0.06311022 - Confidence Interval: [-0.2612424, -0.01385493]

The coarsened exact matching method provides a balance between exact matching and nearest neighbor matching as it is a method that groups treatment and control units based on covariates and matches within these categories. It overall reduces the dimensionality of the covariate space, potentially leading to better matches and a reduced loss of data. However in reality, it might still lose some observations though generally less than exact matching.

Considering the pros and cons of each method, the Nearest Neighbor Matching provides a balance between match quality and retaining enough data. The negative ATT estimate aligns with the expected impact of the TRC and has a relatively small confidence interval, indicating high precision. While Exact Matching ensures perfect matches, it resulted in a less precise estimate with a wider confidence interval. The positive ATT estimate is counterintuitive given the context of TRC exposure, potentially reflecting the loss of valuable data. Coarsened Exact Matching seems to offer a middle ground. It provides a statistically significant negative ATT estimate, and has a moderately precise confidence interval. This method appears effective in balancing the data retention and match quality, making it a robust choice for estimating the ATT.

In conclusion, **Coarsened Exact Matching** might be the preferred approach for this analysis, considering its balance in data retention, match quality, and the intuitiveness and statistical significance of its results. However, by implementing a variety of matching algorithms, as done here, is necessary to identify the most appropriate choice and simultaneously ensure that the robustness of our approach is accurate in identifying causal inferences in our data.

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
library(tidyverse)
library(broom)
library(boot)
## Fit a model among TRCKNOW == 1 to get E[Y_i(1) | X]
# Subset the data into treated and control groups
treated_data <- TRC_data %>% filter(TRCKNOW == 1)
control_data <- TRC_data %>% filter(TRCKNOW == 0)

## Fit a model among TRCKNOW == 0 to get E[Y_i(0) | X]
model_treated <- lm(RUSTAND ~ age + female + wealth + religiosity + ethsalience
                    + rcblack + rcwhite + rccol + EDUC, data = treated_data)
model_control <- lm(RUSTAND ~ age + female + wealth + religiosity + ethsalience
                    + rcblack + rcwhite + rccol + EDUC, data = control_data)

## Predict the potential outcome under treatment for all units
```

```r
TRC_data$predicted_treated <- predict(model_treated, newdata = TRC_data)
## Predict the potential outcome under control for all units
TRC_data$predicted_control <- predict(model_control, newdata = TRC_data)

## Average of the differences
ATE <- mean(TRC_data$predicted_treated - TRC_data$predicted_control)

# Bootstrap
boot_func <- function(data, indices) {
  # Sample with replacement from the dataset
  bs_data <- data[indices, ]

  # Fit models and predict as above
  model_treated <- lm(RUSTAND ~ age + female + wealth + religiosity +
                      ethsalience + rcblack + rcwhite + rccol + EDUC, data =
                        bs_data[bs_data$TRCKNOW == 1, ])
  model_control <- lm(RUSTAND ~ age + female + wealth + religiosity +
                      ethsalience + rcblack + rcwhite + rccol + EDUC, data =
                        bs_data[bs_data$TRCKNOW == 0, ])

  bs_data$predicted_treated <- predict(model_treated, newdata = bs_data)
  bs_data$predicted_control <- predict(model_control, newdata = bs_data)

  # Return the ATE estimate
  mean(bs_data$predicted_treated - bs_data$predicted_control)
}

# Set the seed for reproducibility
set.seed(123)

# Perform bootstrap
boot_results <- boot(TRC_data, boot_func, R = 1000)

# Calculate bootstrapped standard error and confidence interval
boot_se <- sd(boot_results$t)
conf_int <- boot.ci(boot_results, type = "perc")$percent[4:5]

# Results
list(
  ATE = ATE,
  SE = boot_se,
  CI_lower = conf_int[1],
  CI_upper = conf_int[2]
)
```

```
## $ATE
## [1] -0.1743866
##
## $SE
## [1] 0.04403547
##
## $CI_lower
## [1] -0.2593992
```

```
##
## $CI_upper
## [1] -0.085592
```

Interpretation:

An ATE estimate of -0.1743866 suggests that exposure to the TRC is associated with a decrease in the level of agreement with the statement about difficulty understanding customs and ways of the opposite racial group. In other words, TRC exposure leads, on average, to more favorable or understanding attitudes towards the opposite racial group. The standard error of 0.04403547 reflects the variability of the estimated effect size across different samples indicating higher precision in our estimate. As for the 95% confidence interval, the values range from -0.2593992 to -0.085592, suggesting that we can be 95% confident that the true ATE lies within this range. The entire interval is below zero, further reinforcing the finding that the effect of TRC exposure on racial attitudes is negative as well as statistically significant.

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
# Fit model to the control group
model_control <- lm_robust(RUSTAND ~ age + female + wealth + religiosity +
                           ethsalience + rcblack + rcwhite + rccol + EDUC,
                           data = subset(TRC_data, TRC_data$TRCKNOW == 0))
model_treatment <- lm_robust(RUSTAND ~ age + female + wealth + religiosity +
                           ethsalience + rcblack + rcwhite + rccol + EDUC,
                           data = subset(TRC_data, TRC_data$TRCKNOW == 1))


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

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

# Average of the differences between observed and imputed potential outcomes
point_ATT <- mean(TRC_data$trc_treated[TRC_data$TRCKNOW==1] -
                  TRC_data$trc_control[TRC_data$TRCKNOW==1])

# Bootstrap for SEs
set.seed(123)

nBoot <- 2000 # Number of iterations

boot_results_ATT <- rep(NA, 2000)

for (iter in 1:nBoot){
  # Resample w/ replacement
  treatment_boot <- TRC_data[sample(1:nrow(TRC_data), nrow(TRC_data),
                                    replace=T),]

  # Fit a model to get E[Y_i(1) | X]
  treatment_model_boot <- lm_robust(RUSTAND ~ age + female + wealth +
                                    religiosity + ethsalience + rcblack +
```

```r
                                    rcwhite + rccol + EDUC, data =
                                    subset(treatment_boot, TRCKNOW == 1))
  # Fit a modelto get E[Y_i(0) | X]
  control_model_boot <- lm_robust(RUSTAND ~ age + female + wealth +
                                    religiosity + ethsalience + rcblack +
                                    rcwhite + rccol + EDUC, data =
                                    subset(treatment_boot, TRCKNOW == 0))
  # Predict the potential outcome under treatment for all units
  treatment_boot$trc_treated_boot <- predict(treatment_model_boot, newdata =
                                               treatment_boot)
  # Predict the potential outcome under control for all units
  treatment_boot$trc_control_boot <- predict(control_model_boot, newdata =
                                              treatment_boot)
  # Store bootstrapped estimate
  boot_results_ATT[iter] <- mean(
    treatment_boot$trc_treated_boot[treatment_boot$TRCKNOW==1] -
    treatment_boot$trc_control_boot[treatment_boot$TRCKNOW==1])
}

# Standard error
standard_error = sd(boot_results_ATT)

# 95% confidence interval
ci = c(point_ATT - 1.96*sd(boot_results_ATT), point_ATT + 1.96*
         sd(boot_results_ATT))
# Results
list(
  ATT = point_ATT,
  SE = standard_error,
  CI_lower = ci[1],
  CI_upper = ci[2]
)
```

```
## $ATT
## [1] -0.2033737
##
## $SE
## [1] 0.04641271
##
## $CI_lower
## [1] -0.2943426
##
## $CI_upper
## [1] -0.1124048
```

Interpretation:

The ATT estimate has a value of -0.2033737, which suggests that those exposed to the TRC, on average, exhibit more positive or understanding racial attitudes compared to what they would have exhibited had they not been exposed. The standard error value of 0.04641271 and our confidence interval, which ranges from -0.2943426 to -0.1124048, provides a range of values within which the true ATT is likely to lie with 95% confidence. Considering that the entire confidence interval is below zero and does not cross 0 further reinforces the evidence that the effect of TRC exposure on racial attitudes is negative and statistically significant.

part c (+4 points)

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

The ATE estimate, as shown previously, is -0.1743866, while the ATT estimate is -0.2033737. Considering that the ATT is more negative than the ATE suggests that the treatment (TRC exposure) had a slightly stronger effect on those who were treated compared to the average effect on the entire population. This could indicate that those who were exposed to the TRC were more receptive or influenced by the exposure, leading to a more significant change in their racial attitudes.

In the context of the study and reality, the ATE is useful for understanding the overall impact of the intervention on the entire population. On the other hand, the ATT is more relevant when considering the continuation or modification of the TRC, as it focuses on the impact on those who are actually receiving the treatment. In other words, the negative ATE value suggests that overall, TRC exposure is associated with more positive racial attitudes in the general population. The more negative ATT value additionally indicates that this effect is even more prominent or stronger among those who actually received the treatment. In other words, this might indicate that the TRC was particularly effective for those who were exposed to the TRC were perhaps more open to change.