

Problem Set 3

Erin Choi - eyc321 - Section 2

Due Nov 27, 2022

Question 1 (35 points)

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 (5 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.)

```
library(tidyverse)
library(haven)
library(estimatr) # for lm with robust se
library(kableExtra) # balance table
library(ggpubr) # histograms
library(MatchIt) # matching
```

```

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

# run robust regression
regress_rob <- lm_robust(RUSTAND ~ TRCKNOW, data = TRC_data)
tidy(regress_rob)

##           term      estimate  std.error statistic      p.value  conf.low  conf.high
## 1 (Intercept)  2.5311438  0.02805761  90.212369  0.000000e+00  2.476131  2.5861565
## 2      TRCKNOW -0.2177317  0.04433111  -4.911488  9.491614e-07 -0.304652 -0.1308115
##      df outcome
## 1 3203 RUSTAND
## 2 3203 RUSTAND

# report
cat(paste("ATE:", regress_rob$coefficients[2],
          "\nCI: [", regress_rob$conf.low[2], ",", regress_rob$conf.high[2], "]" ))

## ATE: -0.217731736129198
## CI: [ -0.304651966013627 , -0.130811506244768 ]

```

The estimate for the ATE of TRC exposure on racial attitudes is about -0.218, indicating that TRC exposure decreases agreement with the prompt that the respondent finds it difficult to understand the customs of the opposite racial group and thus causes greater understanding of racial differences. The 95% confidence interval does not contain 0, which shows that the estimated causal effect is statistically significant.

Units all have the ability to know about TRC, so positivity is assumed. There is only one version of the “treatment,” which is any knowledge of the TRC, and as stated in the question, it is assumed that TRC exposure is ignorable and units do not select into knowing or not knowing about the TRC. As long as survey data was collected in a way that prevented spillover from occurring between units and TRC knowledge was not shared between respondents during the study, the ATE would be identified. However, if spillover was not prevented and participants could share TRC knowledge during the period of the study (2000-2001), the calculated ATE would not estimate a causal effect.

Part B (7 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?

```

balance = TRC_data %>% group_by(TRCKNOW) %>%
  summarise(RUSTAND = mean(RUSTAND),
            age = mean(age),
            female = mean(female),
            wealth = mean(wealth),
            religiosity = mean(religiosity),
            ethsalience = mean(ethsalience),
            rcblack = mean(rcblack),
            rcwhite = mean(rcwhite),
            rccol = mean(rccol),
            EDUC = mean(EDUC))
kbl(balance) %>% kable_paper("hover", full_width = F, font_size = 7)

```

TRCKNOW	RUSTAND	age	female	wealth	religiosity	ethsalience	rcblack	rcwhite	rccol	EDUC
0	2.531144	40.45470	0.4326161	5792.775	3.915628	2.713477	0.5130238	0.2531144	0.1574179	3.846546
1	2.313412	38.94024	0.5378735	6945.170	3.840167	2.734538	0.5517721	0.2696317	0.1104934	4.291869

Noticeable differences between the groups of exposed and nonexposed respondents occur in the age, gender, wealth, race (black and colored), and education covariates.

- The average age of the nonexposed group is greater than that of the exposed group.
- There is a greater proportion of female respondents in the exposed group compared to the nonexposed group.
- The exposed respondents are wealthier than nonexposed respondents on average.
- There is a greater proportion of black respondents in the exposed group than in the nonexposed group.
- There is a greater proportion of colored respondents in the nonexposed group than in the exposed group.
- Exposed respondents have a higher level of education than nonexposed respondents on average.

Part C (9 points)

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

1. Use an additive logistic regression model to estimate the propensity score for each observation.
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.
4. Plot the histograms of the propensity scores in treatment and control.

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

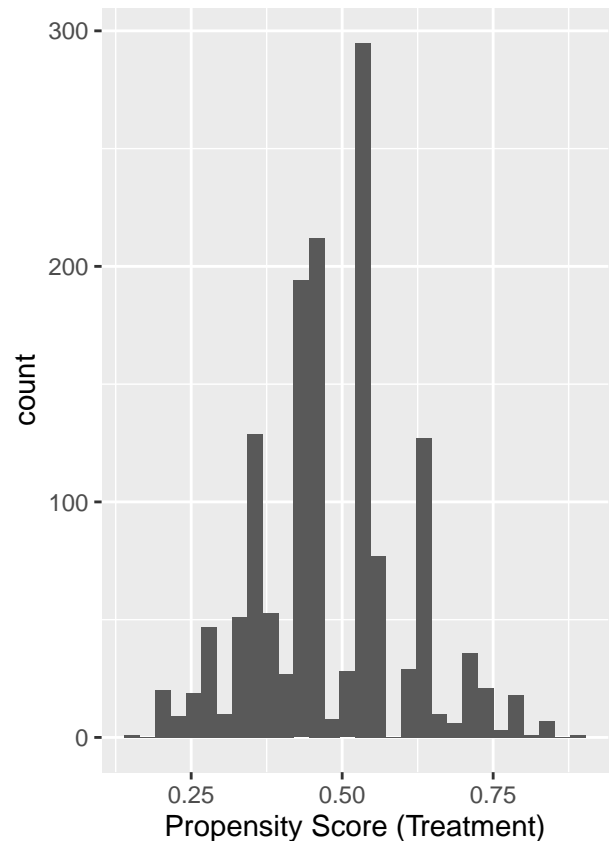
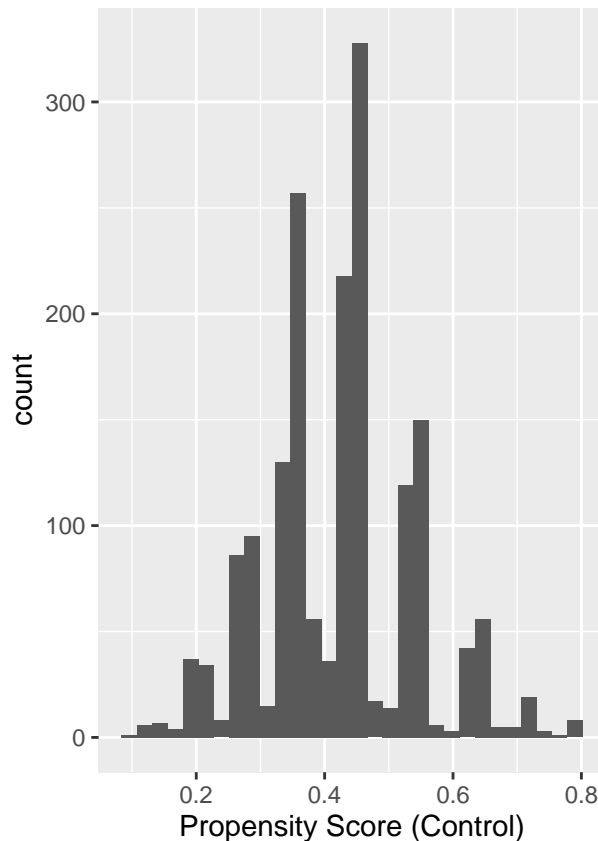
```
# 1. estimate propensity scores - logit model, predict probabilities
logit_model = glm(TRCKNOW ~ age + female + wealth + religiosity +
                  ethsalience + rcblack + rcwhite + rccol + EDUC,
                  data = TRC_data,
                  family = binomial(link = "logit"))
TRC_data$pscore=predict(logit_model, type = "response")

# 2. construct IPTW using unstabilized weights
# unstabilized weights
TRC_data = TRC_data %>% mutate(iptw=ifelse(TRC_data$TRCKNOW==1,1/TRC_data$pscore,
                                           1/(1-TRC_data$pscore)))

# 3. construct IPW estimator - point estimate for ATE
weighted_ate = mean(TRC_data$iptw*TRC_data$RUSTAND*TRC_data$TRCKNOW) -
                mean(TRC_data$iptw*TRC_data$RUSTAND*(1-TRC_data$TRCKNOW))
paste("Weighted ATE:", weighted_ate)
```

```
## [1] "Weighted ATE: -0.162328972831736"
```

```
# 4. plot histograms of propensity scores in treatment and control
ggarrange(TRC_data%>%
  filter(TRCKNOW == 0) %>%
  ggplot() + xlab("Propensity Score (Control)") +
  geom_histogram(aes(x = pscore), bins=30),
  TRC_data %>%
  filter(TRCKNOW == 1) %>%
  ggplot() + xlab("Propensity Score (Treatment)") +
  geom_histogram(aes(x = pscore), bins=30))
```



Part D (14 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, Std, 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(10003)

# IPTW Bootstrap
n_iter <- 1000 # Suggested number of iterations
boot_ate = c()
```

```

for (i in 1:n_iter) {
  # resample
  boot_data = TRC_data[sample(1:nrow(TRC_data),nrow(TRC_data),replace = T),]

  # get propensity scores
  boot_ps = glm(TRCKNOW ~ age + female + wealth + religiosity +
               ethsalience + rcblack + rcwhite + rccol + EDUC,
               data = boot_data,
               family = binomial(link = "logit"))
  boot_data$pscore = predict(boot_ps, type="response")

  # generate weights
  boot_data = boot_data %>% mutate(iptw=ifelse(boot_data$TRCKNOW==1,
                                             1/boot_data$pscore,
                                             1/(1-boot_data$pscore)))

  # compute ate estimate
  weighted_ate = mean(TRC_data$iptw*TRC_data$RUSTAND*TRC_data$TRCKNOW) -
                 mean(TRC_data$iptw*TRC_data$RUSTAND*(1-TRC_data$TRCKNOW))
  boot_ate[i] = mean(boot_data$iptw*boot_data$RUSTAND*boot_data$TRCKNOW)-
               mean(boot_data$iptw*boot_data$RUSTAND*(1-boot_data$TRCKNOW))
}

boot_se = sd(boot_ate)
boot_ci = c(weighted_ate-1.96*boot_se, weighted_ate+1.96*boot_se)
cat(paste("Bootstrapped SE:", boot_se,
          "\nBootstrapped CI: [", boot_ci[1], ",", boot_ci[2], "]""))

## Bootstrapped SE: 0.0454350553998852
## Bootstrapped CI: [ -0.251381681415511 , -0.0732762642479606 ]

```

The estimated weighted ATE, found in part C, is -0.162. This estimate is smaller in magnitude than the unweighted ATE estimate of -0.218. There is still a negative effect, which indicates that TRC exposure causes greater understanding of the customs of opposite racial groups. However, when propensity score weighting is applied, it is found that the actual effect of TRC exposure on understanding of racial differences is not as large as previously estimated.

From bootstrapping, the estimated standard error is 0.045, and the 95% confidence interval is [-0.251, -0.073]. The bootstrapped standard error is very similar to the unweighted standard error of 0.044, so the range of the bootstrapped confidence interval remains similar to the unweighted confidence interval. Since the weighted ATE estimate is closer to zero than the unweighted point estimate, the bootstrapped confidence interval as a whole is also closer to zero than before (compared to [-0.305, -0.131]). However, as with the unweighted confidence interval, the bootstrapped 95% confidence interval still does not contain zero, indicating that the weighted ATE represents a statistically significant effect.

Question 2 (35 points)

Part A (5 points)

Estimate the ATE 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 ATE. Report the 95% confidence interval of your estimate.

```

# check balance without using any matching
m.out0 <- matchit(TRCKNOW ~ age + female + wealth + religiosity +
                  ethsalience + rcblack + rcwhite + rccol + EDUC,
                  data = TRC_data,
                  method = NULL, distance = "glm")
summary(m.out0)

##
## Call:
## matchit(formula = TRCKNOW ~ age + female + wealth + religiosity +
##          ethsalience + rcblack + rcwhite + rccol + EDUC, data = TRC_data,
##          method = NULL, distance = "glm")
##
## Summary of Balance for All Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance           0.4831           0.4212           0.5013           1.1237           0.1431
## age                38.9402           40.4547          -0.1023           0.8637           0.0231
## female              0.5379           0.4326           0.2111              .           0.1053
## wealth             6945.1703          5792.7746           0.1513           1.0756           0.0603
## religiosity         3.8402           3.9156          -0.0408           1.0556           0.0113
## ethsalience        2.7345           2.7135           0.0374           0.9112           0.0053
## rcblack             0.5518           0.5130           0.0779              .           0.0387
## rcwhite             0.2696           0.2531           0.0372              .           0.0165
## rccol               0.1105           0.1574          -0.1497              .           0.0469
## EDUC               4.2919           3.8465           0.3730           1.1970           0.0557
##           eCDF Max
## distance           0.2237
## age                0.0619
## female              0.1053
## wealth             0.0877
## religiosity         0.0265
## ethsalience        0.0148
## rcblack             0.0387
## rcwhite             0.0165
## rccol               0.0469
## EDUC               0.1268
##
##
## Sample Sizes:
##           Control Treated
## All           1766     1439
## Matched       1766     1439
## Unmatched         0         0
## Discarded         0         0

# nearest neighbor matching
m.out1 <- matchit(TRCKNOW ~ age + female + wealth + religiosity +
                  ethsalience + rcblack + rcwhite + rccol + EDUC,
                  data = TRC_data,
                  method = "nearest", distance = "glm")
summary(m.out1, un = FALSE)

##

```

```
## Call:
## matchit(formula = TRCKNOW ~ age + female + wealth + religiosity +
##          ethsalience + rcblack + rcwhite + rccol + EDUC, data = TRC_data,
##          method = "nearest", distance = "glm")
##
## 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
```

```
# estimate the ATE and report 95% CI of the estimate
m.data1 <- match.data(m.out1)
# head(m.data1)

regress_nn = lm_robust(RUSTAND ~ TRCKNOW + age + female + wealth + religiosity +
                      ethsalience + rcblack + rcwhite + rccol + EDUC,
                      data = m.data1)
# tidy(regress_nn)

cat(paste("Nearest Neighbors Matching ATE:", regress_nn$coefficients[2],
          "\nCI: [", regress_nn$conf.low[2], ",", regress_nn$conf.high[2], "]" ))
```

```
## Nearest Neighbors Matching ATE: -0.181455327235525
## CI: [ -0.270559745689513 , -0.0923509087815363 ]
```

Part B (5 points)

Estimate the ATE 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 ATE. Report the 95% confidence interval of your estimate.

```
# exact matching
m.out2 <- matchit(TRCKNOW ~ age + female + wealth + religiosity +
                  ethsalience + rcblack + rcwhite + rccol + EDUC,
                  data = TRC_data,
                  method = "exact", distance = "glm")
summary(m.out2, 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

# estimate the ATE and report 95% CI of the estimate
m.data2 <- match.data(m.out2)
# head(m.data2)
```



```
regress_exact = lm_robust(RUSTAND ~ TRCKNOW + age + female + wealth + religiosity +
  ethsalience + rcblack + rcwhite + rccol + EDUC,
  data = m.data2)
# tidy(regress_exact)

cat(paste("Exact Matching ATE:", regress_exact$coefficients[2],
  "\nCI: [", regress_exact$conf.low[2], ",", regress_exact$conf.high[2], "]""))
```

```
## Exact Matching ATE: 0.102718731725748
## CI: [ -0.250101398116716 , 0.455538861568213 ]
```

Part C (5 points)

Estimate the ATE 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 ATE. Report the 95% confidence interval of your estimate.

```
# coarsened exact matching
m.out3 <- matchit(TRCKNOW ~ age + female + wealth + religiosity +
  ethsalience + rcblack + rcwhite + rccol + EDUC,
  data = TRC_data,
  method = "cem", estimand = 'ATE')
summary(m.out3, un = FALSE)
```

```
##
## Call:
## matchit(formula = TRCKNOW ~ age + female + wealth + religiosity +
##   ethsalience + rcblack + rcwhite + rccol + EDUC, data = TRC_data,
##   method = "cem", estimand = "ATE")
##
## Summary of Balance for Matched Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## age           35.6566      35.6970      -0.0026      0.9907      0.0025
## female         0.4548      0.4548      -0.0000      .      0.0000
## wealth        4556.3904    4524.9720      0.0042      0.9953      0.0129
## religiosity     4.1656      4.1656      0.0000      1.0002      0.0000
## ethsalience    2.8770      2.8770      0.0000      1.0002      0.0000
## rcblack         0.7205      0.7205      0.0000      .      0.0000
## rcwhite         0.1865      0.1865      0.0000      .      0.0000
## rccol           0.0753      0.0753      0.0000      .      0.0000
## EDUC            3.9195      3.9195      0.0000      1.0002      0.0000
##           eCDF Max Std. Pair Dist.
## age           0.0132      0.1065
## female         0.0000      0.0000
## wealth         0.0593      0.0405
## 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) 714.27 633.42
## Matched      802.   726.
## Unmatched    964.   713.
## Discarded      0.    0.

# estimate the ATE and report 95% CI of the estimate
m.data3 <- match.data(m.out3)
# head(m.data3)

regress_cem = lm_robust(RUSTAND ~ TRCKNOW + age + female + wealth + religiosity +
                        ethsalience + rcblack + rcwhite + rccol + EDUC,
                        data = m.data3)
# tidy(regress_cem)

cat(paste("Coarsened Exact Matching ATE:", regress_cem$coefficients[2],
        "\nCI: [", regress_cem$conf.low[2], ",", regress_cem$conf.high[2], "]""))

## Coarsened Exact Matching ATE: -0.126172198960051
## CI: [ -0.247317486352698 , -0.00502691156740454 ]
```

Part D (20 points)

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

In general, matching algorithms help induce balance in data and indirectly create new weights for units. Each method matches some number of control units to each treatment unit based on a metric of “closeness” of covariates. The matching method that is used determines what closeness/distance metric is used and how many control units are matched to one treatment unit. In this case, all algorithms determine distance/closeness between units based on logistic regression propensity scores (glm).

Nearest Neighbor Matching is a greedy algorithm, so it matches the one control unit nearest to the current treatment unit and removes this matched control from the sample so it cannot be matched to another treatment unit. Any control units that are not matched to treatment units are excluded. This means the order of units in the sample matters, so this method is not the best for minimizing covariate differences between the treatment and control groups, and it cannot effectively remove confounding from the data. This is evident in the results for Part A; some covariates, such as ethsalience, race, and education, show improved balance, but others like wealth (6945.1703 versus 6317.7762) are still not very balanced. The balance for religiosity actually worsens upon applying NN Matching, with the difference in means increasing from about 0.0754 to 0.1042. Ultimately, NN Matching does a relatively poor job in balancing covariates, which is one of the main goals of matching.

In contrast, Exact Matching creates subclasses of covariate value combinations to match each treatment unit with all control units (multiple rather than just one) that have the same covariate values as it, for all covariates. Units in subclasses that don’t contain both treated and control units are discarded. Unlike NN Matching, this method gets rid of confounding completely since units are matched exactly - averages for the treatment and control groups are exactly the same for each covariate. However, it is difficult to find exact matches for units, so many units are usually discarded. Thus the elimination of confounding comes at the cost of precision and can make any conclusions impossible to generalize, since the sample may no longer be representative of the population it was originally drawn from. The loss of precision is very apparent in this case because the point estimate is positive compared to both other methods’ negative estimates, and the

confidence interval $[-0.2501, 0.4555]$ is much wider than that produced by the other algorithms - it is about 2 to 3 times the range of the other confidence intervals reported.

Coarsened Exact Matching (CEM) bins covariates before applying Exact Matching. A smaller number of bins means the bins are larger, so units with quite different covariate values can be matched together. In contrast, having more bins means each bin is smaller, making it more difficult to match units, and more units may be left unmatched/dropped as a result. In this case, while applying CEM leaves about half of the original sample unmatched, there are still many more matched units than when using regular Exact Matching, which matches only around 5% of the sample. Covariate balance after performing CEM is also much better than using NN Matching, with all covariates being well if not perfectly balanced. Taking into account the tradeoff between precision/matching many units and good covariate balance/eliminating confounding, among the three algorithms used, CEM seems to best preserve a fair proportion of the sample while balancing covariates well.

Question 3 (30 points)

Part A (10 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.)

```
# Fit a model among TRCKNOW == 1 to get  $E[Y_i(1) | X]$ 
# run regression only on treatment units
treatment_model <- lm_robust(RUSTAND ~ age + female + wealth + religiosity +
  ethsaliency + rcblack + rcwhite + rccol + EDUC,
  data=subset(TRC_data, TRCKNOW==1))

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

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

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

# take average of the differences between observed/treated outcomes:
# treated minus control for each unit
regression_ate <- mean(TRC_data$RUSTAND_treated - TRC_data$RUSTAND_control)
paste("ATE using regression method:", regression_ate)

## [1] "ATE using regression method: -0.174386603228822"

# bootstrap for SEs, get CI
set.seed(10003)
# n_iter <- 1000

regression_boot_ate = c()
```

```

for (i in 1:n_iter){
  # resample
  TRC_data_boot = TRC_data[sample(1:nrow(TRC_data),nrow(TRC_data),replace = T),]

  # fit regression on only treatment units
  treatment_model_boot <- lm_robust(RUSTAND ~ age + female + wealth + religiosity +
    ethsalience + rcblack + rcwhite + rccol + EDUC,
    data=subset(TRC_data_boot, TRCKNOW==1))

  # fit regression on only control units
  control_model_boot <- lm_robust(RUSTAND ~ age + female + wealth + religiosity +
    ethsalience + rcblack + rcwhite + rccol + EDUC,
    data=subset(TRC_data_boot, TRCKNOW==0))

  # predict potential outcome under treatment for all units
  TRC_data_boot$RUSTAND_treated_boot <- predict(treatment_model_boot,
    newdata = TRC_data_boot)

  # predict potential outcome under control for all units
  TRC_data_boot$RUSTAND_control_boot <- predict(control_model_boot,
    newdata = TRC_data_boot)

  # store bootstrapped estimate (average of differences)
  regression_boot_ate[i] <- mean(TRC_data_boot$RUSTAND_treated_boot -
    TRC_data_boot$RUSTAND_control_boot)
}

# estimate standard error
regression_boot_se = sd(regression_boot_ate)

# construct 95% confidence interval
regression_ci = c(regression_ate - 1.96*regression_boot_se,
  regression_ate + 1.96*regression_boot_se)

cat(paste("Regression Method \nBootstrapped SE:", regression_boot_se,
  "\nCI: [", regression_ci[1], ",", regression_ci[2], "]""))

## Regression Method
## Bootstrapped SE: 0.0445594228081805
## CI: [ -0.261723071932856 , -0.0870501345247887 ]

```

Using the regression method, the estimate for the ATE of TRC exposure on racial attitudes is about -0.1744. The negative value indicates that TRC exposure decreases agreement with the prompt that the respondent finds it difficult to understand the customs of the opposite racial group, so on average, TRC exposure causes greater understanding of racial differences by a value of 0.1744 (where the outcome, RUSTAND, is an integer value in [0,4] for each unit). The 95% confidence interval produced by bootstrapping and using the regression method does not contain 0, which shows that the estimated causal effect of TRC exposure on racial understanding is statistically significant.

Part B (10 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.

```

# calculate the ATT

# treatment model and control model for original data were fitted in Part A.
# potential outcomes were also predicted in Part A.

# average of differences between observed and imputed potential outcomes
# using only treatment units (average treatment effect among the *treated*)
regression_att <- mean(TRC_data$RUSTAND_treated[TRC_data$TRCKNOW==1] -
                      TRC_data$RUSTAND_control[TRC_data$TRCKNOW==1])
paste("ATT using regression method:", regression_att)

## [1] "ATT using regression method: -0.203373730161722"

# bootstrap for SEs, get CI
set.seed(10003)
# n_iter <- 1000

regression_boot_att = c()

for (i in 1:n_iter){
  # resample
  TRC_data_boot = TRC_data[sample(1:nrow(TRC_data),nrow(TRC_data),replace = T),]

  # fit regression on only treatment units
  treatment_model_boot <- lm_robust(RUSTAND ~ age + female + wealth + religiosity +
                                   ethsalience + rcblack + rcwhite + rccol + EDUC,
                                   data=subset(TRC_data_boot, TRCKNOW==1))

  # fit regression on only control units
  control_model_boot <- lm_robust(RUSTAND ~ age + female + wealth + religiosity +
                                   ethsalience + rcblack + rcwhite + rccol + EDUC,
                                   data=subset(TRC_data_boot, TRCKNOW==0))

  # predict potential outcome under treatment for all units
  TRC_data_boot$RUSTAND_treated_boot <- predict(treatment_model_boot,
                                                newdata = TRC_data_boot)

  # predict potential outcome under control for all units
  TRC_data_boot$RUSTAND_control_boot <- predict(control_model_boot,
                                                newdata = TRC_data_boot)

  # store bootstrapped estimate of ATT
  regression_boot_att[i] <- mean(TRC_data_boot$RUSTAND_treated_boot[TRC_data_boot$TRCKNOW==1]
                                - TRC_data_boot$RUSTAND_control_boot[TRC_data_boot$TRCKNOW==1])
}

# estimate standard error
regression_boot_se_2 = sd(regression_boot_att)

# construct 95% confidence interval
regression_ci_2 = c(regression_att - 1.96*regression_boot_se_2,
                    regression_att + 1.96*regression_boot_se_2)

```

```
cat(paste("Regression Method for ATT \nBootstrapped SE:", regression_boot_se_2,
        "\nCI: [", regression_ci_2[1], ",", regression_ci_2[2], "]" ))
```

```
## Regression Method for ATT
## Bootstrapped SE: 0.0457615327442961
## CI: [ -0.293066334340543 , -0.113681125982902 ]
```

Using the regression method, the estimate for the ATT of TRC exposure on racial attitudes is about -0.2034. This estimate is also negative, indicating that TRC exposure causes greater understanding of racial differences by a value of 0.2034. The 95% confidence interval constructed by bootstrapping with the regression method and ATT calculation does not contain 0, so this estimate of the causal effect of TRC exposure on racial understanding is statistically significant.

Part C (10 points)

Compare and contrast the ATE and ATT from the regression approach. Are they similar or different in terms of coefficients and standard errors. Speculate about why the similarities or dissimilarities exist?

The ATE calculated using the regression method (-0.1744) is smaller in magnitude (closer to 0) than the calculated ATT (-0.2034). The bootstrapped standard errors of the two estimates are very similar (0.0446 versus 0.0458), so the widths of their respective confidence intervals are also similar, though the confidence interval of the ATT is as a whole more negative than that of the ATE since the point ATT itself has a greater magnitude than the point ATE.

Since it was conditional ignorability of the treatment variable was assumed, the difference between the point estimates may be attributed to the difference in effect of the treatment on the treatment and control groups. The different observable and recorded covariates in the data were all controlled for, but there may still be unobservable covariates that were not included in the regressions and that still have an effect on the estimates.

On the other hand, the similarity in the bootstrapped standard errors - and thus the widths of the confidence intervals constructed from them - may come from the similarity in group sizes. ATT is often used when there is a much smaller number of treatment units than control units, but in this sample, the two groups are both large and fairly even in size, though the control group is somewhat larger (1766 control units and 1439 treatment units). We know that the size of a sample has a great impact on standard error, so standard errors could differ between the groups if the sizes were not as balanced.