Problem Set 2

Your Name - Net ID - Section Number

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")

linreg <- lm_robust(RUSTAND ~ TRCKNOW, data = TRC_data)

tidy(linreg)
```

```
##          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
```

The estimate for the average effect of knowing about TRC is -0.2177 on respondents' racial attitudes. The 95% confidence interval is equal to [-0.3047, -0.1308]. 0 is noticeably not within this interval, therefore we reject the null hypothesis of no effect since the effect is statistically significant at 95% confidence level.

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?

```
TRC_data <- 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))

means <- TRC_data %>%
  group_by(TRCKNOW) %>%
  summarize(age = mean(age_std),
            female = mean(female_std),
            wealth = mean(wealth_std),
            religiosity = mean(religiosity_std),
            ethsalience = mean(ethsalience_std),
            rcblack = mean(rcblack_std),
            rcwhite = mean(rcwhite_std),
            rccol = mean(rccol_std),
            EDUC = mean(EDUC_std))

# means
meansoutput <- as.data.frame(means[-c(1)])
rownames(meansoutput) <- c('Control Means', 'Treatment Means')

meansoutput
```

```
##                      age    female    wealth religiosity ethsalience  rcblack
## Control Means   2.618827 0.8657988 0.7736468    2.147880    4.691181 1.027792
## Treatment Means 2.520788 1.0764515 0.9275535    2.106486    4.727592 1.105420
##                    rcwhite     rccol     EDUC
## Control Means   0.5765801 0.4586597 3.317230
## Treatment Means 0.6142056 0.3219384 3.701274
```

On average, exposed subjects tend to be slighty younger, wealthier, less religious, slightly stronger ethnic identification, and better educated. Exposed subjects were also more likely to be female, black or white but not "colored".

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

```
logreg <- glm(TRCKNOW ~ age + female + religiosity + ethsalience + rcblack + rcwhite +
↪  rccol + EDUC, family = binomial(link = 'logit'), data = TRC_data)

TRC_data$scores <- predict(logreg, type = 'response')

TRC_data$wt <- TRC_data$TRCKNOW * (1 / TRC_data$scores) + (1 - TRC_data$TRCKNOW) * (1 /
↪  (1 - TRC_data$scores))


point_est <- mean(TRC_data$wt * TRC_data$RUSTAND * TRC_data$TRCKNOW - TRC_data$wt *
↪  TRC_data$RUSTAND * (1 - TRC_data$TRCKNOW))

point_est
```

```
## [1] -0.1694384
```

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)

n_iter <- 1000
IPTW_boot <- rep(NA, n_iter)

for (i in 1:n_iter) {
  TRC_boot <- TRC_data[sample(nrow(TRC_data), replace = T),]
  TRC_boot$boot_score <- glm(logreg, family = binomial(), data = TRC_boot)$fitted.values
  TRC_boot$wt <- TRC_boot$TRCKNOW * (1 / TRC_boot$boot_score) + (1 - TRC_boot$TRCKNOW) *
↪  (1 / (1 - TRC_boot$boot_score))
  IPTW_boot[i] <- summary(lm_robust(RUSTAND ~ TRCKNOW, weights = TRC_boot$wt, data =
↪  TRC_boot))$coefficients[2, 1]
}
```

```
SE <-sd(IPTW_boot)

tibble('Estimate' = point_est,
       'Standard Error' = SE,
       '95% CI Lower' = point_est - qnorm(0.975) * SE,
       '95% CI Upper' = point_est + qnorm(0.975) * SE)
```

```
## # A tibble: 1 x 4
##   Estimate `Standard Error` `95% CI Lower` `95% CI Upper`
##      <dbl>            <dbl>          <dbl>          <dbl>
## 1   -0.169           0.0440         -0.256        -0.0832
```

The 95% confidence interval is [-0.2569, -0.0819] with the point estimate of -0.1694 and standard error of 0.0446. Since 0 is not included inside the confidence interval we still reject the null hypothesis of no treatment effect. This estimate is somewhat closer to zero than the first estimate which would imply that the effect was overestimated in part A due to confounding effects of some or all covariates.

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()
m.out1 <- matchit(TRCKNOW ~ age + female + wealth + religiosity + ethsalience + rcblack +
↪  rcwhite + rccol + EDUC, data = TRC_data, method = 'nearest', link = 'logit')

m.data1 <- match.data(m.out1)

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

```
##           term      estimate    std.error  statistic      p.value      conf.low
## 1  (Intercept)  2.863798e+00 1.924724e-01 14.8790020 2.730161e-48  2.486399e+00
## 2      TRCKNOW -1.802434e-01 4.545871e-02 -3.9649914 7.519278e-05 -2.693784e-01
## 3          age  2.124821e-04 1.565603e-03  0.1357190 8.920530e-01 -2.857340e-03
## 4       female -1.224397e-01 4.632683e-02 -2.6429544 8.263332e-03 -2.132769e-01
## 5       wealth -1.228282e-05 4.332807e-06 -2.8348408 4.616964e-03 -2.077855e-05
## 6  religiosity -1.812679e-02 1.267233e-02 -1.4304225 1.527048e-01 -4.297459e-02
## 7  ethsalience  1.799750e-02 3.994206e-02  0.4505901 6.523191e-01 -6.032057e-02
## 8      rcblack  4.221736e-01 9.376444e-02  4.5024918 6.984141e-06  2.383211e-01
## 9      rcwhite  1.093332e-01 9.596039e-02  1.1393574 2.546493e-01 -7.882516e-02
## 10       rccol -1.079063e-01 1.038701e-01 -1.0388580 2.989584e-01 -3.115740e-01
## 11        EDUC -1.065639e-01 2.402210e-02 -4.4360766 9.506390e-06 -1.536662e-01
##        conf.high   df outcome
## 1   3.241196e+00 2867 RUSTAND
## 2  -9.110832e-02 2867 RUSTAND
## 3   3.282305e-03 2867 RUSTAND
## 4  -3.160243e-02 2867 RUSTAND
## 5  -3.787086e-06 2867 RUSTAND
## 6   6.721014e-03 2867 RUSTAND
## 7   9.631556e-02 2867 RUSTAND
## 8   6.060262e-01 2867 RUSTAND
## 9   2.974915e-01 2867 RUSTAND
## 10  9.576137e-02 2867 RUSTAND
## 11 -5.946155e-02 2867 RUSTAND
```

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
m.out2 <- matchit(TRCKNOW ~ age + female + wealth + religiosity + ethsalience + rcblack +
↪  rcwhite + rccol + EDUC, data = TRC_data, method = 'exact', link = 'logit')

m.data2 <- match.data(m.out2)

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

```
##           term      estimate    std.error  statistic      p.value      conf.low
## 1  (Intercept)  5.339649e+00 9.933369e-01  5.3754662 2.505836e-07  3.3787019714
## 2      TRCKNOW  1.027187e-01 1.787245e-01  0.5747322 5.662367e-01 -0.2501013981
## 3          age  1.951547e-03 9.940822e-03  0.1963165 8.445983e-01 -0.0176726334
## 4       female  1.644171e-01 2.057038e-01  0.7992906 4.252439e-01 -0.2416628433
## 5       wealth  1.227178e-05 2.953197e-05  0.4155422 6.782722e-01 -0.0000460273
## 6  religiosity -9.708998e-02 6.069108e-02 -1.5997406 1.115239e-01 -0.2169002704
## 7  ethsalience -4.504668e-01 3.575883e-01 -1.2597358 2.095016e-01 -1.1563821048
## 8      rcblack -1.213618e-01 4.595932e-01 -0.2640635 7.920527e-01 -1.0286448919
## 9      rcwhite -6.978282e-01 3.429611e-01 -2.0347156 4.344219e-02 -1.3748677894
## 10       rccol -1.380816e-01 7.976226e-01 -0.1731164 8.627671e-01 -1.7126687326
## 11        EDUC -2.433734e-01 1.423939e-01 -1.7091555 8.925777e-02 -0.5244733483
##         conf.high  df outcome
## 1   7.300595e+00 169 RUSTAND
## 2   4.555389e-01 169 RUSTAND
## 3   2.157573e-02 169 RUSTAND
## 4   5.704971e-01 169 RUSTAND
## 5   7.057087e-05 169 RUSTAND
## 6   2.272030e-02 169 RUSTAND
## 7   2.554484e-01 169 RUSTAND
## 8   7.859213e-01 169 RUSTAND
## 9  -2.078863e-02 169 RUSTAND
## 10  1.436506e+00 169 RUSTAND
## 11  3.772655e-02 169 RUSTAND
```

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.

```
m.out3 <- matchit(TRCKNOW ~ age + female + wealth + religiosity + ethsalience + rcblack +
↪  rcwhite + rccol + EDUC, data = TRC_data, method = 'cem', link = 'logit')

m.data3 <- match.data(m.out3)

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

```
##             term      estimate    std.error  statistic      p.value      conf.low
## 1  (Intercept)  3.170812e+00 4.004629e-01  7.9178660 4.630727e-15  2.3852920172
## 2      TRCKNOW -1.261722e-01 6.176064e-02 -2.0429225 4.123263e-02 -0.2473174864
## 3          age  1.604415e-03 2.560838e-03  0.6265196 5.310684e-01 -0.0034187421
## 4       female -1.077344e-01 6.444329e-02 -1.6717705 9.477582e-02 -0.2341417805
## 5       wealth -1.660956e-05 9.169960e-06 -1.8113011 7.029196e-02 -0.0000345967
## 6  religiosity -1.693928e-02 1.968193e-02 -0.8606517 3.895659e-01 -0.0555459540
## 7  ethsalience  2.672914e-02 8.404640e-02  0.3180283 7.505072e-01 -0.1381303092
## 8      rcblack -3.070173e-02 2.349849e-01 -0.1306541 8.960663e-01 -0.4916313644
## 9      rcwhite -3.336013e-01 2.241358e-01 -1.4883893 1.368561e-01 -0.7732502536
## 10       rccol -7.255778e-01 2.411758e-01 -3.0085013 2.668642e-03 -1.1986512005
## 11        EDUC -8.750641e-02 4.634598e-02 -1.8881122 5.920157e-02 -0.1784153933
##         conf.high   df outcome
## 1   3.956331e+00 1517 RUSTAND
## 2  -5.026912e-03 1517 RUSTAND
## 3   6.627572e-03 1517 RUSTAND
## 4   1.867299e-02 1517 RUSTAND
## 5   1.377583e-06 1517 RUSTAND
## 6   2.166739e-02 1517 RUSTAND
## 7   1.915886e-01 1517 RUSTAND
## 8   4.302279e-01 1517 RUSTAND
## 9   1.060476e-01 1517 RUSTAND
## 10 -2.525044e-01 1517 RUSTAND
## 11  3.402573e-03 1517 RUSTAND
```

part d (10 points)

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

The estimate using nearest neighbors was -0.1802 with a standard error of 0.0454 and 95% confidence interval of [-0.2638, -0.0911]. The estimate using exact matching was 0.1027 with a standard error of 0.1787 and 95% confidence interval of [-0.2501, 0.4555]. The estimate using coarsened exact matching was -0.1262 with a standard error of 0.0618 and 95% confidence interval of [-0.2473, -0.0050]. The exact matching estimate was not statistically significant, while the nearest neighbors and CEM estimates were. The nearest neighbors also had the smallest standard error. As a result, I would choose the nearest neighbors algorithm as it had a statistically significant effect (which agrees with all the other estimates besides the exact matching) while also having the smallest standard error, or the most precise estimate. This is likely because the nearest neighbor algorithm is able to match a lot more treatments so less data is dropped and thus the sample size is larger.

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 ~ TRCKNOW + age + female + wealth + religiosity +
↪  ethsalience + rcblack + rcwhite + rccol + EDUC, data = TRC_data, subset = TRCKNOW ==
↪  1)

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


## Predict the potential outcome under treatment for all units
TRC_data$treat_outcome <- predict(treatment_model, newdata = TRC_data)
## Predict the potential outcome under control for all units
TRC_data$cont_outcome <- predict(control_model, newdata = TRC_data)
## Average of the differences
avg_diff <- mean(TRC_data$treat_outcome - TRC_data$cont_outcome)
paste('Estimate:', avg_diff)
```

```
## [1] "Estimate: -0.174386603228822"
```

```r
### Bootstrap for SEs
set.seed(123)

n_iter <- 1000
SE_boot <- rep(NA, n_iter)

for (i in 1:n_iter) {
  boot_data <- TRC_data[sample(1:nrow(TRC_data), nrow(TRC_data), replace = T),]
  treat_boot_model <- lm_robust(RUSTAND ~ age + female + wealth + religiosity +
↪  ethsalience + rcblack + rcwhite + rccol + EDUC, data = boot_data, subset = TRCKNOW ==
↪  1)
  cont_boot_model <- lm_robust(RUSTAND ~ age + female + wealth + religiosity +
↪  ethsalience + rcblack + rcwhite + rccol + EDUC, data = boot_data, subset = TRCKNOW ==
↪  0)
  boot_data$treat_boot <- predict(treat_boot_model, newdata = boot_data)
  boot_data$cont_boot <- predict(cont_boot_model, newdata = boot_data)
  SE_boot[i] <- mean(boot_data$treat_boot - boot_data$cont_boot)
}

SE <- sd(SE_boot)
paste('Standard Error:', SE)
```

```
## [1] "Standard Error: 0.0437870715274353"
```

```
paste('confidence interval: [', avg_diff - qnorm(0.975) * SE, ',', avg_diff +
↪  qnorm(0.975) * SE, ']')
```

```
## [1] "confidence interval: [ -0.260207686411074 , -0.0885655200465694 ]"
```

The point estimate for the average treatment effect is -0.1744 with a standard error of 0.0438 and a confidence interval of [-0.2602, -0.0886]. Since 0 is outside of this confidence interval we still reject the null hypothesis of no effect.

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.

```
## 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 = TRC_data, subset = 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 = TRC_data, subset = TRCKNOW == 0)
## Predict the potential outcome under treatment for all units
TRC_data$treated <- predict(treatment_model, newdata = TRC_data)
## Predict the potential outcome under control for all units
TRC_data$control <- predict(control_model, newdata = TRC_data)
## Average of the differences
avg_diff <- mean(TRC_data[TRC_data$TRCKNOW == 1,]$treated - TRC_data[TRC_data$TRCKNOW ==
↪  1,]$control)
paste('Estimate:', avg_diff)
```

```
## [1] "Estimate: -0.203373730161722"
```

```
### Bootstrap for SEs
set.seed(123)
n_iter <- 1000
SE_boot <- rep(NA, n_iter)

for (i in 1:n_iter) {
  boot_data <- TRC_data[sample(1:nrow(TRC_data), nrow(TRC_data), replace = T),]
  treat_boot_model <- lm_robust(RUSTAND ~ age + female + wealth + religiosity +
↪  ethsalience + rcblack + rcwhite + rccol + EDUC, data = boot_data, subset = TRCKNOW ==
↪  1)
  cont_boot_model <- lm_robust(RUSTAND ~ age + female + wealth + religiosity +
↪  ethsalience + rcblack + rcwhite + rccol + EDUC, data = boot_data, subset = TRCKNOW ==
↪  0)
  boot_data$treat_boot <- predict(treat_boot_model, newdata = boot_data)
  boot_data$cont_boot <- predict(cont_boot_model, newdata = boot_data)
  SE_boot[i] <- mean(boot_data[boot_data$TRCKNOW == 1,]$treat_boot -
↪  boot_data[boot_data$TRCKNOW == 1,]$cont_boot)
}
```

```r
SE <- sd(SE_boot)
paste('Standard Error:', SE)
```

```
## [1] "Standard Error: 0.0452040966478705"
```

```r
paste('confidence interval: [', avg_diff - qnorm(0.975) * SE, ',', avg_diff +
↪  qnorm(0.975) * SE, ']')
```

```
## [1] "confidence interval: [ -0.291972131545216 , -0.114775328778228 ]"
```

The point estimate for ATT is -0.2034 with a standard error of 0.0452 and confidence interval of [-0.2920, -0.1148]. Since 0 is outside of this confidence interval we still reject the null hypothesis of no effect/

part c (+4 points)

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

The magnitude of the ATE estimate is lower with a slightly lower bootstrapped standard error. The ATT likely has a higher standard error because it involves a lower sample size (only treated subjects) and possibly a higher magnitude estimate because of some inherent difference (bias perhaps) between the treated and control groups.