CHDV Homework 2

Yuyang Jiang

0: Data Preparation

```
library(haven)
library(tidyr)
df_origin <- read_dta('CHDV 30102_ECLSK98_class size.dta')

# 1. deal with null values in treatment and outcome
df <- drop_na(df_origin, c(A4CLSIZE, C4R2MSCL))
df$X <- df$A4CLSIZE
df$X[which(df$X < 19)] <- 1
df$X[which(df$X > 18)] <- 0

# 2. deal with null values in pretreatment covariates
sum(is.na(df$GENDER))
```

```
[1] 0
```

```
df$gender3 <- as.factor(df$GENDER)
levels(df$gender3)[levels(df$gender3)=='-9'] <- '3'

sum(is.na(df$RACE))
```

```
[1] 0
```

```
df$race6 <- as.factor(df$RACE)
levels(df$race6)[levels(df$race6)=='3'] <- 'Hispanic'
levels(df$race6)[levels(df$race6)=='4'] <- 'Hispanic'
levels(df$race6)[levels(df$race6)=='6'] <- 'Indigenous or Native Americans'
levels(df$race6)[levels(df$race6)=='7'] <- 'Indigenous or Native Americans'
```

```r
levels(df$race6)[levels(df$race6)=='-9'] <- 'Other Races'
levels(df$race6)[levels(df$race6)=='8'] <- 'Other Races'

# sum(is.na(df$C1RRSCAL))
df$missing_C1RR <- as.numeric(is.na(df$C1RRSCAL))
df$C1RRSCAL[is.na(df$C1RRSCAL)] <- mean(df$C1RRSCAL, na.rm = TRUE)
# sum(is.na(df$C1RRSCAL))

# sum(is.na(df$C2RRSCAL))
df$missing_C2RR <- as.numeric(is.na(df$C2RRSCAL))
df$C2RRSCAL[is.na(df$C2RRSCAL)] <- mean(df$C2RRSCAL, na.rm = TRUE)
# sum(is.na(df$C2RRSCAL))

# sum(is.na(df$C1R2MSCL))
df$missing_C1R2 <- as.numeric(is.na(df$C1R2MSCL))
df$C1R2MSCL[is.na(df$C1R2MSCL)] <- mean(df$C1R2MSCL, na.rm = TRUE)
# sum(is.na(df$C1R2MSCL))

# sum(is.na(df$C2R2MSCL))
df$missing_C2R2 <- as.numeric(is.na(df$C2R2MSCL))
df$C2R2MSCL[is.na(df$C2R2MSCL)] <- mean(df$C2R2MSCL, na.rm = TRUE)
# sum(is.na(df$C2R2MSCL))

# sum(is.na(df$B4YRSTC))
df$missing_B4 <- as.numeric(is.na(df$B4YRSTC))
df$B4YRSTC[is.na(df$B4YRSTC)] <- mean(df$B4YRSTC, na.rm = TRUE)
# sum(is.na(df$B4YRSTC))
```

1: Descriptive Analysis

```r
model1 <- lm(C4R2MSCL~X, data = df)
summary(model1) # Estimate of X is mean difference
```

```
Call:
lm(formula = C4R2MSCL ~ X, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-64.507 -10.272  -1.057   9.383  51.920
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.50657    0.15945  348.12   <2e-16 ***
X           -0.00608    0.30011   -0.02    0.984
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.63 on 13382 degrees of freedom
Multiple R-squared:  3.067e-08, Adjusted R-squared:  -7.47e-05
F-statistic: 0.0004105 on 1 and 13382 DF,  p-value: 0.9838
```

```r
# effect size
cat("Effect Size:")
```

```
Effect Size:
```

```r
coef(model1)[2] / sd(df[df$X==0, ]$C4R2MSCL)
```

```
            X
-0.0003902936
```

1.

- mean difference: -0.00608
- standard error: 0.30011
- hypothesis testing: 0.984
- effect size: -0.0003902936

2. Potential Confounders

```r
library(tableone)

cov <- c("gender3", "race6", "C1RRSCAL", "C2RRSCAL",
         "C1R2MSCL", "C2R2MSCL", "B4YRSTC")
trt <- "X"

tab1 <- CreateTableOne(vars = cov, data = df, strata = trt,
                       factorVars = c("gender3", "race6"),
```

```
                      test = TRUE)
tab1
```

```
                                Stratified by X
                                 0             1             p      test
n                                 9606          3778
gender3 = 2 (%)                   4778 (49.7)   1809 (47.9)   0.055
race6 (%)                                                     <0.001
   Other Races                     263 ( 2.7)     76 ( 2.0)
   1                              5438 (56.6)   2364 (62.6)
   2                              1275 (13.3)    528 (14.0)
   Hispanic                       1648 (17.2)    561 (14.8)
   5                               682 ( 7.1)    152 ( 4.0)
   Indigenous or Native Americans  300 ( 3.1)     97 ( 2.6)
C1RRSCAL (mean (SD))              21.79 (10.34) 21.70 (9.58)   0.652
C2RRSCAL (mean (SD))              32.38 (12.86) 31.91 (12.50)  0.060
C1R2MSCL (mean (SD))              21.64 (9.13)  21.34 (8.91)   0.086
C2R2MSCL (mean (SD))              31.98 (11.85) 31.79 (11.73)  0.411
B4YRSTC (mean (SD))               14.49 (10.31) 13.71 (9.75)   <0.001
```

2a. Here for smaller p-values, it means that pre-existing differences between students in small classes and those in regular classes will be more significant.

- Grade 1 teacher's teaching experience significantly differs between two groups. More specifically, teachers of regular classes has longer teaching experience.
- The races of students between two groups are also quite different. Regular classes tend to have more races like Hispanic, Asian, Indigenous or Native Americans while small classes tend to have more races like Black and White.

2b.

Based on the table above, we can see that reading score and math score in kindergarten are almost the same between two groups while teachers of regular classes tend to have longer teaching experience. And experienced teachers will be better at giving classes and with better classes, students will have better grades. Therefore, the control seems to be relatively advantaged.

3. Propensity score and common support

3a.

Not necessary to include.

- Theoretically, misspecification of propensity score model is less consequential than misspecification of outcome model.
- Empirically, there don't exist many significant terms in quadratic / interaction / nonlinear terms.

Logistic Regression Model:

$$\log \frac{\theta}{1-\theta} = \beta_0 + \sum \beta_i * X_i + \epsilon_i.$$

```
df_new <- df[, c("X", "gender3", "race6",
                 "C1RRSCAL", "C2RRSCAL", "C1R2MSCL",
                 "C2R2MSCL", "B4YRSTC", "missing_C1RR",
                 "missing_C2RR", "missing_C1R2",
                 "missing_C2R2", "missing_B4")]

# summary(glm(X ~ .*., data = df_new, family = binomial(link="logit")))
model2 <- glm(X ~ ., data = df_new, family = binomial(link="logit"))
summary(model2)
```

```
Call:
glm(formula = X ~ ., family = binomial(link = "logit"), data = df_new)

Coefficients: (2 not defined because of singularities)
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.893534   0.147058  -6.076 1.23e-09 ***
gender32                        -0.067002   0.038926  -1.721  0.08521 .
race61                           0.418707   0.133048   3.147  0.00165 **
race62                           0.317508   0.140608   2.258  0.02394 *
race6Hispanic                    0.103533   0.140895   0.735  0.46245
race65                          -0.293268   0.158808  -1.847  0.06479 .
race6Indigenous or Native Americans  0.066401   0.175567   0.378  0.70527
C1RRSCAL                         0.006377   0.003753   1.699  0.08924 .
C2RRSCAL                        -0.006647   0.002768  -2.402  0.01633 *
C1R2MSCL                        -0.012423   0.004147  -2.996  0.00274 **
C2R2MSCL                         0.004481   0.002966   1.511  0.13077
B4YRSTC                         -0.008450   0.001928  -4.384 1.17e-05 ***
missing_C1RR                     0.156628   0.068162   2.298  0.02157 *
missing_C2RR                    -0.093580   0.254075  -0.368  0.71264
missing_C1R2                           NA         NA      NA       NA
missing_C2R2                           NA         NA      NA       NA
```

```
missing_B4                          0.135866   0.149151   0.911  0.36233
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15929  on 13383  degrees of freedom
Residual deviance: 15800  on 13369  degrees of freedom
AIC: 15830

Number of Fisher Scoring iterations: 4
```

```r
df$theta <- predict(model2, type = "response")
```

3b.

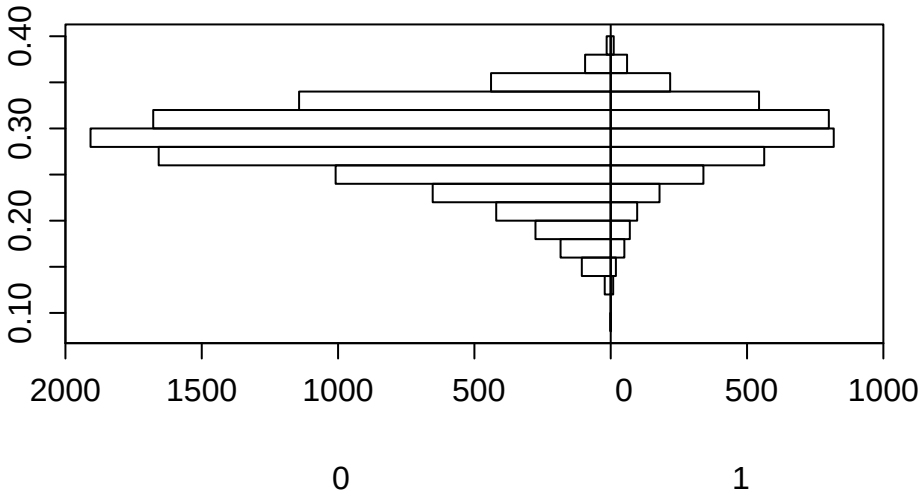
```r
library(Hmisc)
```

```
Attaching package: 'Hmisc'

The following objects are masked from 'package:base':

    format.pval, units
```

```r
histbackback(split(df$theta, df$X))
```




```
# examining the between-group differences in the mean and variance
# of the logit propensity score
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v purrr     1.0.2
v forcats   1.0.0     v readr     2.1.4
v ggplot2   3.4.3     v stringr   1.5.0
v lubridate 1.9.3     v tibble    3.2.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter()    masks stats::filter()
x dplyr::lag()       masks stats::lag()
x dplyr::src()       masks Hmisc::src()
x dplyr::summarize() masks Hmisc::summarize()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
tab2 <- df %>%
  group_by(X) %>%
  summarize(
    mean_theta = mean(theta, na.rm = TRUE),
    var_theta = var(theta, na.rm = TRUE)
```

```
  )
tab2
```

```
# A tibble: 2 x 3
      X mean_theta var_theta
  <dbl>      <dbl>     <dbl>
1     0      0.280   0.00196
2     1      0.289   0.00168
```

```
# common support with a 20% caplier
tab3 <- df %>%
  group_by(X) %>%
  summarise(
    min_theta = min(theta, na.rm = TRUE),
    max_theta = max(theta, na.rm = TRUE)
  )
lower <- max(tab3$min_theta) - 0.2 * sd(df$theta)
upper <- min(tab3$max_theta) + 0.2 * sd(df$theta)

# Extreme case
df$support <- 1
df$support[which(df$theta < lower | df$theta > upper)] <- 0
df$CHILDID[which(df$support == 0)]
```

```
[1] "0685001C"
```

4. Propensity Score Matching

4a. We choose ATT. Because according to the table below, there are many unmatched cases in the control group. In this way, we would gain a more accurate estimation with ATT.

4b. 4c.

```
library(MatchIt)

# Performing one-to-one matching without replacement
matchit_obj <- matchit(X~gender3+race6+C1RRSCAL+C2RRSCAL+C1R2MSCL+C2R2MSCL
                       +B4YRSTC+missing_C1RR+missing_C2RR+missing_C1R2
                       +missing_C2R2+missing_B4, data = df,
                       method = "nearest", ratio = 1,
```

```
                         distance = "glm", caliper = 0.2)
summary(matchit_obj, standardize = TRUE)
```

```
Call:
matchit(formula = X ~ gender3 + race6 + C1RRSCAL + C2RRSCAL +
    C1R2MSCL + C2R2MSCL + B4YRSTC + missing_C1RR + missing_C2RR +
    missing_C1R2 + missing_C2R2 + missing_B4, data = df, method = "nearest",
    distance = "glm", caliper = 0.2, ratio = 1)

Summary of Balance for All Data:
                                  Means Treated Means Control Std. Mean Diff.
distance                             0.2891        0.2796          0.2313
gender31                             0.5212        0.5026          0.0372
gender32                             0.4788        0.4974         -0.0372
race6Other Races                     0.0201        0.0274         -0.0517
race61                               0.6257        0.5661          0.1232
race62                               0.1398        0.1327          0.0203
race6Hispanic                        0.1485        0.1716         -0.0649
race65                               0.0402        0.0710         -0.1566
race6Indigenous or Native Americans  0.0257        0.0312         -0.0351
C1RRSCAL                            21.6986       21.7862         -0.0091
C2RRSCAL                            31.9148       32.3762         -0.0369
C1R2MSCL                            21.3369       21.6361         -0.0336
C2R2MSCL                            31.7902       31.9765         -0.0159
B4YRSTC                             13.7061       14.4900         -0.0804
missing_C1RR                         0.0937        0.0794          0.0490
missing_C2RR                         0.0058        0.0060         -0.0028
missing_C1R2                         0.0937        0.0794          0.0490
missing_C2R2                         0.0058        0.0060         -0.0028
missing_B4                           0.0175        0.0167          0.0062
                                  Var. Ratio eCDF Mean eCDF Max
distance                              0.8551    0.0635   0.1030
gender31                                   .    0.0186   0.0186
gender32                                   .    0.0186   0.0186
race6Other Races                           .    0.0073   0.0073
race61                                     .    0.0596   0.0596
race62                                     .    0.0070   0.0070
race6Hispanic                              .    0.0231   0.0231
race65                                     .    0.0308   0.0308
race6Indigenous or Native Americans        .    0.0056   0.0056
C1RRSCAL                              0.8591    0.0117   0.0245
```

C2RRSCAL	0.9448	0.0169	0.0354
C1R2MSCL	0.9524	0.0104	0.0316
C2R2MSCL	0.9800	0.0053	0.0145
B4YRSTC	0.8950	0.0228	0.0457
missing_C1RR	.	0.0143	0.0143
missing_C2RR	.	0.0002	0.0002
missing_C1R2	.	0.0143	0.0143
missing_C2R2	.	0.0002	0.0002
missing_B4	.	0.0008	0.0008

Summary of Balance for Matched Data:

	Means Treated	Means Control	Std. Mean Diff.
distance	0.2891	0.2891	0.0002
gender31	0.5212	0.5156	0.0111
gender32	0.4788	0.4844	-0.0111
race6Other Races	0.0201	0.0193	0.0057
race61	0.6257	0.6276	-0.0038
race62	0.1398	0.1392	0.0015
race6Hispanic	0.1485	0.1466	0.0052
race65	0.0402	0.0426	-0.0121
race6Indigenous or Native Americans	0.0257	0.0246	0.0067
C1RRSCAL	21.6986	21.5310	0.0175
C2RRSCAL	31.9148	31.7901	0.0100
C1R2MSCL	21.3369	21.3602	-0.0026
C2R2MSCL	31.7902	31.6685	0.0104
B4YRSTC	13.7061	13.4696	0.0242
missing_C1RR	0.0937	0.0905	0.0109
missing_C2RR	0.0058	0.0066	-0.0104
missing_C1R2	0.0937	0.0905	0.0109
missing_C2R2	0.0058	0.0066	-0.0104
missing_B4	0.0175	0.0199	-0.0182

	Var. Ratio	eCDF Mean	eCDF Max
distance	1.0002	0.0001	0.0021
gender31	.	0.0056	0.0056
gender32	.	0.0056	0.0056
race6Other Races	.	0.0008	0.0008
race61	.	0.0019	0.0019
race62	.	0.0005	0.0005
race6Hispanic	.	0.0019	0.0019
race65	.	0.0024	0.0024
race6Indigenous or Native Americans	.	0.0011	0.0011
C1RRSCAL	1.0154	0.0037	0.0130
C2RRSCAL	1.0406	0.0059	0.0238

```
C1R2MSCL                                  1.0530    0.0040   0.0164
C2R2MSCL                                  1.0338    0.0066   0.0161
B4YRSTC                                   0.9460    0.0097   0.0209
missing_C1RR                                   .    0.0032   0.0032
missing_C2RR                                   .    0.0008   0.0008
missing_C1R2                                   .    0.0032   0.0032
missing_C2R2                                   .    0.0008   0.0008
missing_B4                                     .    0.0024   0.0024
                                       Std. Pair Dist.
distance                                       0.0006
gender31                                       0.9352
gender32                                       0.9352
race6Other Races                               0.2507
race61                                         0.6689
race62                                         0.6641
race6Hispanic                                  0.5114
race65                                         0.1010
race6Indigenous or Native Americans            0.2778
C1RRSCAL                                       0.9718
C2RRSCAL                                       0.9907
C1R2MSCL                                       0.9724
C2R2MSCL                                       1.0524
B4YRSTC                                        1.0081
missing_C1RR                                   0.4705
missing_C2RR                                   0.1496
missing_C1R2                                   0.4705
missing_C2R2                                   0.1496
missing_B4                                     0.2849

Sample Sizes:
          Control Treated
All          9606    3778
Matched      3778    3778
Unmatched    5828       0
Discarded       0       0
```

```
# distance <-> propensity score
```

4d.

```
library(lmtest)
```

```
Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```r
library(Hmisc)
library(sandwich)

matched_data <- match.data(matchit_obj)
model2 <- lm(C4R2MSCL~X+gender3+race6+C1RRSCAL+C2RRSCAL+C1R2MSCL+C2R2MSCL
                        +B4YRSTC+missing_C1RR+missing_C2RR+missing_C1R2
                        +missing_C2R2+missing_B4, data = matched_data)
( test <- coeftest( model2, vcov = vcovHC( model2 ) ) )
```

```
t test of coefficients:

                                  Estimate Std. Error t value  Pr(>|t|)
(Intercept)                     22.836354   1.015339 22.4914 < 2.2e-16 ***
X                                0.287432   0.227432  1.2638 0.2063363
gender32                        -1.047166   0.230283 -4.5473 5.519e-06 ***
race61                           1.552723   0.904534  1.7166 0.0860933 .
race62                          -2.401667   0.922854 -2.6024 0.0092745 **
race6Hispanic                    0.392050   0.947824  0.4136 0.6791559
race65                           3.040349   1.159734  2.6216 0.0087696 **
race6Indigenous or Native Americans -1.982477   1.074449 -1.8451 0.0650606 .
C1RRSCAL                        -0.103879   0.024302 -4.2744 1.940e-05 ***
C2RRSCAL                         0.086538   0.017516  4.9405 7.959e-07 ***
C1R2MSCL                         0.398164   0.028433 14.0038 < 2.2e-16 ***
C2R2MSCL                         0.729909   0.019462 37.5043 < 2.2e-16 ***
B4YRSTC                         -0.018703   0.011635 -1.6075 0.1079853
missing_C1RR                     1.534575   0.413761  3.7088 0.0002097 ***
missing_C2RR                    -5.450454   2.201183 -2.4761 0.0133026 *
missing_B4                       2.831957   0.801901  3.5316 0.0004156 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coef(model2)[2] / sd(matched_data[matched_data$X==0, ]$C4R2MSCL)
```

```
        X
0.01880601
```

Conclusion: Class size type change from regular class to small class can exert an increase on Grade 1 studens' average math score with a size effect of 0.0188.

5. Propensity Score Stratification for estimating the ATE

```r
# Continue with data in common support
df_str <- df[df$support == 1, ]
strata_quintile <- quantile(df_str$theta,
                            probs = c( 0.2, 0.4, 0.6, 0.8 ) ) # full sample for ATE

strata_data <- df_str %>%
  mutate( strata1 = as.numeric( theta <= strata_quintile[ 1 ] ),
          strata2 = as.numeric( theta > strata_quintile[ 1 ]
                                & theta <= strata_quintile[ 2 ] ),
          strata3 = as.numeric( theta > strata_quintile[ 2 ]
                                & theta <= strata_quintile[ 3 ] ),
          strata4 = as.numeric( theta > strata_quintile[ 3 ]
                                & theta <= strata_quintile[ 4 ] ),
          strata5 = as.numeric( theta >= strata_quintile[ 4 ] ) ,
          strata=case_when(( theta <= strata_quintile[ 1 ] )~1,
                           ( theta > strata_quintile[ 1 ] &
                               theta <= strata_quintile[ 2 ] ) ~ 2,
                           ( theta > strata_quintile[ 2 ] &
                               theta <= strata_quintile[ 3 ] ) ~ 3,
                           ( theta > strata_quintile[ 3 ] &
                               theta <= strata_quintile[ 4 ] ) ~ 4,
                           TRUE ~ 5))

# Check observations of treatment and control in each strata
tab3 <- strata_data %>%
  group_by(strata, X) %>%
  summarise(
    mean = mean(theta, na.rm = T),
    var = var(theta, na.rm = T),
    .groups = 'drop' # Drop grouping for pivot_wider compatibility
```

```
  ) %>%
  pivot_wider(
    names_from = X,
    values_from = c(mean, var)
  )
tab3 <- tab3 %>%
  mutate(
    std_diff = (mean_1 - mean_0) / sqrt((var_1 + var_0) / 2),
    var_ratio = var_1 / var_0)
tab3 <- tab3 %>%
  mutate(
    mean_std_diff = mean(std_diff),
    mean_vr = mean(var_ratio)
  )
tab3
```

```
# A tibble: 5 x 9
  strata mean_0 mean_1    var_0   var_1 std_diff var_ratio mean_std_diff mean_vr
   <dbl>  <dbl>  <dbl>    <dbl>   <dbl>    <dbl>     <dbl>         <dbl>   <dbl>
1      1  0.214  0.217  8.38e-4 8.38e-4   0.102       1.00        0.0530    1.04
2      2  0.265  0.266  5.90e-5 5.97e-5   0.0225      1.01        0.0530    1.04
3      3  0.288  0.288  3.23e-5 3.33e-5   0.0432      1.03        0.0530    1.04
4      4  0.308  0.308  3.57e-5 3.82e-5   0.0377      1.07        0.0530    1.04
5      5  0.335  0.336  1.88e-4 2.01e-4   0.0597      1.07        0.0530    1.04
```

5b. There is no need for further subdividing. Because all variance ratios fall in the range of $[4/5, 5/4]$.

5c.

```
# Balance Checking for each of the covariates
library(RItools)
library(ggplot2)

xBalance(X ~ gender3+race6+C1RRSCAL+C2RRSCAL+C1R2MSCL+C2R2MSCL
                      +B4YRSTC+missing_C1RR+missing_C2RR+missing_C1R2
                      +missing_C2R2+missing_B4,
         strata=factor(strata_data$strata),
         data=strata_data,
         report=c("adj.means", "adj.mean.diffs",
                  "std.diffs",
                  "z.scores", "p.values"))
```

```
                                strata():     strat
                                stat      Treatment   Control  adj.diff  std.diff      z
vars
gender31                                      0.517     0.516     0.0012     0.00      0.13
gender32                                      0.483     0.484    -0.0012     0.00     -0.13
race6Other Races                              0.0221    0.0214    0.000731   0.00      0.27
race61                                        0.610     0.611    -0.000497   0.00     -0.06
race62                                        0.138     0.139    -0.000976   0.00     -0.15
race6Hispanic                                 0.158     0.149     0.00853    0.02      1.35
race65                                        0.0445    0.0544   -0.00991   -0.04     -2.71
race6Indigenous or Native Americans           0.0276    0.0255    0.00212    0.01      0.71
C1RRSCAL                                      21.7      21.8     -0.0825    -0.01     -0.43
C2RRSCAL                                      32.0      32.1     -0.105     -0.01     -0.44
C1R2MSCL                                      21.4      21.5     -0.0352     0.00     -0.21
C2R2MSCL                                      31.8      31.9     -0.0558     0.00     -0.25
B4YRSTC                                       13.9      13.9     -0.0123     0.00     -0.07
missing_C1RR                                  0.0900    0.0872    0.0028     0.01      0.53
missing_C2RR                                  0.00592   0.00588   3.86e-05   0.00      0.03
missing_C1R2                                  0.0900    0.0872    0.0028     0.01      0.53
missing_C2R2                                  0.00592   0.00588   3.86e-05   0.00      0.03
missing_B4                                    0.0175    0.0174    0.000125   0.00      0.05
```

```r
# ps.makestrata is not available in the new version of R,
# so need to calculate variance ratio manually
```

```r
vr_c1rr <- strata_data %>%
  group_by(strata, X) %>%
  summarise(
    var = var(C1RRSCAL, na.rm = T),
    .groups = 'drop'
  ) %>%
  spread(X, var) %>%
  mutate(variance_ratio = `1` / `0`)
cat("C1RRSCAL:", mean(vr_c1rr$variance_ratio), "\n")
```

```
C1RRSCAL: 0.9696596
```

```r
vr_c2rr <- strata_data %>%
  group_by(strata, X) %>%
  summarise(
```

```
    var = var(C2RRSCAL, na.rm = T),
    .groups = 'drop'
  ) %>%
  spread(X, var) %>%
  mutate(variance_ratio = `1` / `0`)
cat("C2RRSCAL:", mean(vr_c2rr$variance_ratio), "\n")
```

```
C2RRSCAL: 1.029666
```

```
vr_c1r2 <- strata_data %>%
  group_by(strata, X) %>%
  summarise(
    var = var(C1R2MSCL, na.rm = T),
    .groups = 'drop'
  ) %>%
  spread(X, var) %>%
  mutate(variance_ratio = `1` / `0`)
cat("C1R2MSCL:", mean(vr_c1r2$variance_ratio), "\n")
```

```
C1R2MSCL: 1.075943
```

```
vr_c2r2 <- strata_data %>%
  group_by(strata, X) %>%
  summarise(
    var = var(C2R2MSCL, na.rm = T),
    .groups = 'drop'
  ) %>%
  spread(X, var) %>%
  mutate(variance_ratio = `1` / `0`)
cat("C2R2MSCL:", mean(vr_c2r2$variance_ratio), "\n")
```

```
C2R2MSCL: 1.046452
```

```
vr_b4 <- strata_data %>%
  group_by(strata, X) %>%
  summarise(
    var = var(B4YRSTC, na.rm = T),
    .groups = 'drop'
```

```
  ) %>%
  spread(X, var) %>%
  mutate(variance_ratio = `1` / `0`)
cat("B4YRSTC:", mean(vr_b4$variance_ratio), "\n")
```

```
B4YRSTC: 0.9666786
```

```
# we cannot calculate variance ratios for categorical variables
```

5d.

```
# Examine the within-stratum mean difference in the outcome
tab4 <- strata_data %>%
  group_by(strata, X) %>%
  dplyr::summarize(average.outcome=mean(C4R2MSCL, na.rm=TRUE)) %>%
  spread(X, average.outcome) %>% ungroup() %>%
  mutate(mean.difference = `1`-`0`)
```

```
`summarise()` has grouped output by 'strata'. You can override using the
`.groups` argument.
```

```
tab4
```

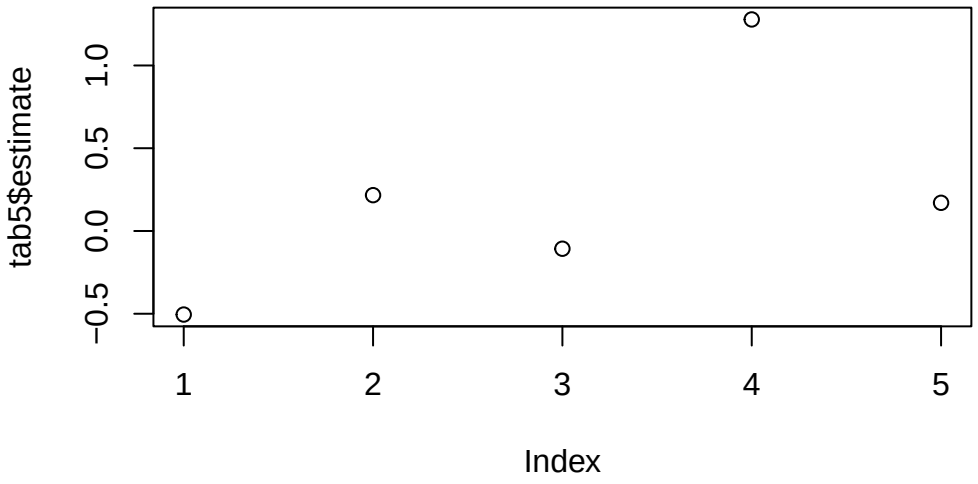
```
# A tibble: 5 x 4
  strata   `0`   `1` mean.difference
   <dbl> <dbl> <dbl>           <dbl>
1      1  56.3  56.5           0.216
2      2  55.9  57.2           1.28
3      3  56.6  56.1          -0.505
4      4  55.2  55.0          -0.107
5      5  53.3  53.5           0.170
```

```
# Estimate stratum-specific treatment effects
library(broom)

tab5 <- strata_data %>%
  group_by(strata) %>%
  nest() %>%
```

```
  mutate(model_summary = map(data, ~ tidy(lm(C4R2MSCL ~ X, data = .))
                             %>% filter(term == "X"))) %>%
  select(-data) %>%
  unnest(model_summary)

plot(tab5$estimate)
```



Generally, with higher treatment probability, estimated effect of class size on Grade 1 math tends to be higher.

5e.

```
# Pooled effects including stratum and covariance adjustment
model3 <- lm(C4R2MSCL ~ X + gender3+race6+C1RRSCAL+C2RRSCAL+C1R2MSCL+
           C2R2MSCL+B4YRSTC+missing_C1RR+missing_C2RR+
           missing_C1R2+missing_C2R2+missing_B4+factor(strata),
           data = strata_data)
( test <- coeftest( model3, vcov = vcovHC( model3 ) ) )
```

```
t test of coefficients:
```

```
                                  Estimate Std. Error t value  Pr(>|t|)
(Intercept)                      26.461196   0.893040 29.6305 < 2.2e-16 ***
X                                 0.253670   0.192023  1.3210 0.1865110
gender32                         -1.750388   0.203358 -8.6074 < 2.2e-16 ***
race61                            4.148835   0.897200  4.6242 3.796e-06 ***
race62                           -0.840755   0.800822 -1.0499 0.2937991
race6Hispanic                     0.052772   0.631294  0.0836 0.9333804
race65                            2.159495   0.715298  3.0190 0.0025408 **
race6Indigenous or Native Americans -2.465421   0.722493 -3.4124 0.0006459 ***
C1RRSCAL                         -0.030286   0.019964 -1.5170 0.1292834
C2RRSCAL                          0.042804   0.016771  2.5523 0.0107125 *
C1R2MSCL                          0.270453   0.028269  9.5672 < 2.2e-16 ***
C2R2MSCL                          0.754273   0.016256 46.4011 < 2.2e-16 ***
B4YRSTC                          -0.086953   0.015925 -5.4601 4.843e-08 ***
missing_C1RR                      2.417331   0.393676  6.1404 8.464e-10 ***
missing_C2RR                     -3.601353   1.841310 -1.9559 0.0505019 .
missing_B4                        2.700174   0.705582  3.8269 0.0001304 ***
factor(strata)2                  -0.900086   0.446828 -2.0144 0.0439887 *
factor(strata)3                  -1.869145   0.607669 -3.0759 0.0021027 **
factor(strata)4                  -3.305289   0.747971 -4.4190 9.994e-06 ***
factor(strata)5                  -4.704710   0.932596 -5.0447 4.601e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coef(model3)[2] / sd(strata_data[strata_data$X==0, ]$C4R2MSCL)
```

```
         X
0.01628881
```

Conclusion: Class size type change from regular class to small class can exert an increase on Grade 1 studens' average math score with a size effect of 0.01629.

6. Inverse-probability-of-treatment weighting

6a. 6b.

```r
library(cobalt)
```

```
cobalt (Version 4.5.3, Build Date: 2024-01-09)
```

```
Attaching package: 'cobalt'

The following object is masked from 'package:MatchIt':

    lalonde
```

```r
df_str$W_ATE <- ifelse(df_str$X == 1, mean( df_str$X ) / df_str$theta,
                           (1 - mean(df_str$X)) / (1 - df_str$theta))

# Balance Checking
tab6 <- bal.tab( X ~ gender3+race6+C1RRSCAL+C2RRSCAL+C1R2MSCL+
           C2R2MSCL+B4YRSTC+missing_C1RR+missing_C2RR+
           missing_C1R2+missing_C2R2+missing_B4+theta,
         data = df_str, estimand = "ATE", m.threshold = 0.05,
         disp.v.ratio = TRUE )

tab7 <- bal.tab( X ~ gender3+race6+C1RRSCAL+C2RRSCAL+C1R2MSCL+
           C2R2MSCL+B4YRSTC+missing_C1RR+missing_C2RR+
           missing_C1R2+missing_C2R2+missing_B4+theta,
         data = df_str, estimand = "ATE", m.threshold = 0.05,
         disp.v.ratio = TRUE, weights = df_str$W_ATE, method = "weighting" )

df_sum <- data.frame(
  Diff.before = tab6$Balance$Diff.Un,
  Diff.after = tab7$Balance$Diff.Adj
)
rownames(df_sum) <- rownames(tab6$Balance)
df_sum
```

```
                                  Diff.before    Diff.after
gender3_2                       -0.0185203577 -5.470141e-04
race6_Other Races               -0.0072651084  3.548853e-04
race6_1                          0.0595644418 -1.758427e-03
race6_2                          0.0070131221 -5.057356e-04
race6_Hispanic                  -0.0230860383  8.865779e-04
race6_5                         -0.0306676451  1.452749e-03
race6_Indigenous or Native Americans -0.0055587721 -4.300500e-04
C1RRSCAL                        -0.0082938997  4.230468e-03
C2RRSCAL                        -0.0360007723  5.207148e-03
C1R2MSCL                        -0.0325706374  6.522870e-03
```

```
C2R2MSCL                                -0.0153928321  5.014261e-03
B4YRSTC                                 -0.0778955522  1.816730e-04
missing_C1RR                             0.0142625778 -9.832981e-04
missing_C2RR                            -0.0002153347 -8.916875e-05
missing_C1R2                             0.0142625778 -9.832981e-04
missing_C2R2                            -0.0002153347 -8.916875e-05
missing_B4                               0.0008115700  5.682216e-04
theta                                    0.2217451320 -4.740689e-03
```

6c.

```
# Doubly robust estimate - weights and covariance adjustment in the output model
model5 <- lm( C4R2MSCL ~ X + gender3+race6+C1RRSCAL+C2RRSCAL+C1R2MSCL+
            C2R2MSCL+B4YRSTC+missing_C1RR+missing_C2RR+
            missing_C1R2+missing_C2R2+missing_B4,
              weights = W_ATE, data = df_str )
( test <- coeftest( model5, vcov = vcovHC( model5 ) ) )
```

```
t test of coefficients:

                                   Estimate Std. Error t value  Pr(>|t|)
(Intercept)                      23.5306694  0.6829388 34.4550 < 2.2e-16 ***
X                                 0.2314181  0.1958502  1.1816 0.2373822
gender32                         -1.1718866  0.1747841 -6.7048 2.098e-11 ***
race61                            1.5234788  0.6274640  2.4280 0.0151957 *
race62                           -2.6152991  0.6344787 -4.1220 3.779e-05 ***
race6Hispanic                     0.0103194  0.6474851  0.0159 0.9872844
race65                            2.8198239  0.7590452  3.7150 0.0002041 ***
race6Indigenous or Native Americans -2.1391861  0.7463060 -2.8664 0.0041586 **
C1RRSCAL                         -0.0809742  0.0173870 -4.6572 3.237e-06 ***
C2RRSCAL                          0.0884107  0.0129726  6.8152 9.819e-12 ***
C1R2MSCL                          0.3643713  0.0230444 15.8117 < 2.2e-16 ***
C2R2MSCL                          0.7202170  0.0153645 46.8755 < 2.2e-16 ***
B4YRSTC                          -0.0134617  0.0085401 -1.5763 0.1149835
missing_C1RR                      1.0133247  0.3279770  3.0896 0.0020082 **
missing_C2RR                     -3.0964343  1.7740182 -1.7454 0.0809321 .
missing_B4                        1.6307645  0.6935264  2.3514 0.0187169 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coef(model5)[2] / sd(df_str[df_str$X==0, ]$C4R2MSCL)
```

```
        X
0.01485997
```

Conclusion: Class size type change from regular class to small class can exert an increase on Grade 1 studens' average math score with a size effect of 0.01486.

7. Marginal mean weighting through stratification (MMWS)

7a.

```r
# Mac cannot download mmws.exe file
library(WeightIt)

n.strata = 5
# MMWS calculation
tab6 <- df_str %>%
  mutate(LGP_STR = cut(theta, quantile(theta, prob = seq(0, 1, 1/n.strata)),
                       include.lowest = TRUE, labels = FALSE)) %>%
  group_by(LGP_STR) %>%
  mutate(NMZ1 = sum(X),
         NMZ0 = n() - NMZ1,
         PZ1 = mean(X),
         PZ0 = 1 - PZ1,
         MMWS = ifelse(X == 1, PZ1 / (NMZ1 / n()), PZ0 / (NMZ0 / n())))

# Balance Checking
tab8 <- bal.tab( X ~ theta+gender3+race6+C1RRSCAL+C2RRSCAL+C1R2MSCL+
           C2R2MSCL+B4YRSTC+missing_C1RR+missing_C2RR+
           missing_C1R2+missing_C2R2+missing_B4,
         data = df_str, estimand = "ATE", m.threshold = 0.05,
         disp.v.ratio = TRUE )

mmws.weighit <- weightit(X ~ gender3+race6+C1RRSCAL+C2RRSCAL+C1R2MSCL+
           C2R2MSCL+B4YRSTC+missing_C1RR+missing_C2RR+
           missing_C1R2+missing_C2R2+missing_B4,
           data = df_str, method = "ps",
           estimand = "ATE", subclass = n.strata)
tab9 <- bal.tab(mmws.weighit, m.threshold = 0.05,
```

```
        disp.v.ratio = TRUE) #

df_sum2 <- data.frame(
  Diff.before = tab8$Balance$Diff.Un,
  Diff.after = tab9$Balance$Diff.Adj,
  Var.R.before = tab8$Balance$V.Ratio.Un,
  Var.R.after = tab9$Balance$V.Ratio.Adj
)
rownames(df_sum2) <- rownames(tab9$Balance)
df_sum2
```

```
                                  Diff.before    Diff.after Var.R.before
prop.score                       0.2217451320  0.0211543404    0.8565186
gender3_2                       -0.0185203577 -0.0030098927           NA
race6_Other Races               -0.0072651084  0.0008689980           NA
race6_1                          0.0595644418 -0.0012420668           NA
race6_2                          0.0070131221 -0.0013414791           NA
race6_Hispanic                  -0.0230860383  0.0115016880           NA
race6_5                         -0.0306676451 -0.0118047134           NA
race6_Indigenous or Native Americans -0.0055587721  0.0020175732           NA
C1RRSCAL                        -0.0082938997 -0.0080041280    0.8609743
C2RRSCAL                        -0.0360007723 -0.0087367763    0.9460248
C1R2MSCL                        -0.0325706374 -0.0019160965    0.9559169
C2R2MSCL                        -0.0153928321 -0.0032107265    0.9815929
B4YRSTC                         -0.0778955522 -0.0030974865    0.8952697
missing_C1RR                     0.0142625778  0.0018276496           NA
missing_C2RR                    -0.0002153347  0.0001075018           NA
missing_C1R2                     0.0142625778  0.0018276496           NA
missing_C2R2                    -0.0002153347  0.0001075018           NA
missing_B4                       0.0008115700  0.0007475883           NA
                                Var.R.after
prop.score                        0.9714351
gender3_2                                NA
race6_Other Races                        NA
race6_1                                  NA
race6_2                                  NA
race6_Hispanic                           NA
race6_5                                  NA
race6_Indigenous or Native Americans     NA
C1RRSCAL                          0.9723100
C2RRSCAL                          1.0382215
C1R2MSCL                          1.0728385
```

```
C2R2MSCL                                1.0344469
B4YRSTC                                 0.9352480
missing_C1RR                                   NA
missing_C2RR                                   NA
missing_C1R2                                   NA
missing_C2R2                                   NA
missing_B4                                     NA
```

7b.

```r
# outcome model
model6 <- lm( C4R2MSCL ~ X + gender3+race6+C1RRSCAL+C2RRSCAL+C1R2MSCL+
           C2R2MSCL+B4YRSTC+missing_C1RR+missing_C2RR+
           missing_C1R2+missing_C2R2+missing_B4,
             weights = MMWS, data = tab6 )
( test <- coeftest( model6, vcov = vcovHC( model6 ) ) )
```

```
t test of coefficients:

                                  Estimate Std. Error t value  Pr(>|t|)
(Intercept)                     23.6353183  0.6704693 35.2519 < 2.2e-16 ***
X                                0.2250087  0.1924325  1.1693 0.2423091
gender32                        -1.1586423  0.1740540 -6.6568 2.908e-11 ***
race61                           1.2941667  0.5739598  2.2548 0.0241618 *
race62                          -2.8287736  0.5938207 -4.7637 1.921e-06 ***
race6Hispanic                   -0.1711852  0.6033612 -0.2837 0.7766299
race65                           2.4975234  0.7289871  3.4260 0.0006143 ***
race6Indigenous or Native Americans -2.4053408  0.7127904 -3.3745 0.0007415 ***
C1RRSCAL                        -0.0849187  0.0170429 -4.9826 6.351e-07 ***
C2RRSCAL                         0.0914812  0.0128956  7.0940 1.369e-12 ***
C1R2MSCL                         0.3731509  0.0210858 17.6968 < 2.2e-16 ***
C2R2MSCL                         0.7172693  0.0153695 46.6684 < 2.2e-16 ***
B4YRSTC                         -0.0137460  0.0084948 -1.6182 0.1056517
missing_C1RR                     1.0937921  0.3263239  3.3519 0.0008049 ***
missing_C2RR                    -2.9194318  1.8245394 -1.6001 0.1096017
missing_B4                       1.6651534  0.6877547  2.4211 0.0154850 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coef(model6)[2] / sd(df_str[df_str$X==0, ]$C4R2MSCL)
```

```
        X
0.01444841
```

Conclusion: Class size type change from regular class to small class can exert an increase on Grade 1 studens' average math score with a size effect of 0.01448.

8. Identification Assumption

8a.

Conditioning on the propensity score, the potential outcome is independent of treatment assignment.

$$Y(0), Y(1) \perp\!\!\!\perp Z \,|\, \theta$$

$$where\ \theta = pr(Z = 1|X = x)$$

- $Y(0)$ stands for math achievement score of grade 1 student who is assigned to a regular class;
- $Y(1)$ stands for math achievement score of grade 1 student who is assigned to a small class;
- $Z = 0$ stands for grade 1 student who is assigned to a small class;
- $Z = 1$ stands for grade 1 student who is assigned to a regular class;
- X stands for pre-treatment covariates;
- θ stands for propensity score of each individual.

8b.

We need to consider how wealthy or educated the students' families are, also called socioeconomic status (SES).

- **Richer Schools Might Have Smaller Classes:** Schools in wealthier areas might have more money to make classes smaller. If we don't consider SES, we might wrongly think that smaller classes alone are making the big difference in math scores, when actually, the wealthier area and all the benefits that come with it play a big role too.
- **Wealthier Kids Might Already Do Better in School:** Kids from wealthier families often have more support and resources, like books and help with homework. This means they might get better math scores not just because of smaller classes, but because of these extra advantages.

8c.

In causal inference, particularly when using quasi-experimental data, the purpose of a sensitivity analysis is to assess how robust the estimated causal effects are to potential unobserved confounders or assumptions made during the analysis.

Conditions for Sensitivity to Potential Bias:

- **Magnitude of Omission:** If the omitted variable (e.g., SES) has a strong influence on both the treatment (class size reduction) and the outcome (math achievement), the study's results might be highly sensitive to this omission. The sensitivity analysis would show that even a small correlation between the omitted variable and both the treatment and outcome could substantially bias the estimated effect of class size reduction.
- **Direction of Bias:** The analysis could reveal whether omitting a confounder would likely lead to an overestimation or underestimation of the true effect. For example, if higher SES is associated with both smaller classes and better math outcomes independently, omitting SES might overestimate the effect of class size reduction.