CHDV 30102 Homework 1

Yuyang Jiang

1. Context

1a. Target population in each case:

1. For *Reasearch Question 1*: students who attend Project STAR Class in kindergarten.
2. For *Research Question 2*: students who attend Project STAR Class in first grade.

1b. Treatment Conditions:

1. For *Reasearch Question 1*: participants in kindergarten are assigned to a small class or a regular class.
2. For *Research Question 2*: participants in first grade are assigned to a small class or a regular class.

1c. ATE:

1. For *Reasearch Question 1*:

   $$\delta_K = E[Y_F(Z_K = 1)] - E[Y_F(Z_K = 0)]$$

   where $E[Y_F(Z_K = 1)]$ stands for the participant's expected math score in first grade when he/she was assigned to a regular class in kindergarten; $E[Y_F(Z_K = 0)$ stands for the participant's expected math score in first grade when he/she was assigned to a small class in kindergarten.

2. For *Reasearch Question 2*:

   $$\delta_F = E[Y_F(Z_F = 1)] - E[Y_F(Z_F = 0)]$$

   where $E[Y_F(Z_F = 1)]$ stands for the participant's expected math score in first grade when he/she was assigned to a regular class in first grade; $E[Y_F(Z_F = 0)$ stands for the participant's expected math score in first grade when he/she was assigned to a small class in first grade.

2. Setup for Research Question 1

```
library(tidyr)
library(lmtest)
```

```
Loading required package: zoo


Attaching package: 'zoo'

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

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

```
library(sandwich)
library(lme4)
```

```
Loading required package: Matrix


Attaching package: 'Matrix'

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

    expand, pack, unpack
```

Data Preparation: Restrict the analysis to the 6,258 students who had valid information about treatment group membership in grade 1.

```
df_original <- read.csv("webstar_Winter2024.csv")
df_original[df_original == "missing" |
                          df_original == "Missing" |
                          df_original == ""] <- NA

df_sample <- df_original[(df_original$star1 == "YES"), ]
df_sample <- drop_na(df_sample, tmathss1) # 6258*17
df_sample[df_sample == "small class"] <- 1
df_sample[df_sample == "regular class" | df_sample == "regular + aide class"] <- 0
```

```r
n <- sum(df_sample$stark == "NO") # 2313
n1 <- sum(is.na(df_sample$cltypek)) # 2313
m <- sum(is.na(df_sample$cltype1)) # 0
```

2a. Prima Facie Effect can be written as follows:

$$\delta_{PF_K} = E[Y_F(Z_K = 1)|Z_K = 1] - E[Y_F(Z_K = 0)|Z_K = 0].$$

It stands for the mean difference between kindergarten regular class and small class in students' first grade math score.

2b. Under Identification Assumption (written as below), $\delta_{PF_K} = \delta_K$.

$$E[Y_F(Z_K = 1)|Z_K = 1] = E[Y_F(Z_K = 1)|Z_K = 0] = E[Y_F(Z_K = 1)]$$

$$E[Y_F(Z_K = 0)|Z_K = 0] = E[Y_F(Z_K = 0)|Z_K = 1] = E[Y_F(Z_K = 0)]$$

2c. The reason why the assumption is plausible in the current study is that students involved in STAR project are a representative and random sample for the whole students and the study is also conducted as a randomization experiment.

3. Naive Analysis for Research Question 1

```r
df_1 <- df_sample[df_sample$stark == "YES", ]
model1 <- lm(tmathss1~cltypek, data = df_1)

# Calculate robust standard errors
vcov_cl <- vcovCL(model1, cluster = ~ schidkn, data = df_sample)
summary1 <- coeftest(model1, vcov = vcov_cl)
summary1
```

```
t test of coefficients:

            Estimate Std. Error  t value Pr(>|t|)
(Intercept) 540.7901     3.1058 174.1205  < 2e-16 ***
cltypek1      8.4302     3.4832   2.4202  0.01556 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3a.

1. Estimate: 8.4302
2. Robust Std. Error: 3.4832

3b. Standard Error quantifies the amount of variability or dispersion of $\hat{\delta}_{PF_K}$ from the true δ_{PF_K}. The lower standard error, the higher the statistical power for detecting a nonzero treatment effect.

3c.

1. t-value: 2.4202
2. p-value: 0.01556 *

Class size reduction in kindergarten will on average increase 8.430 points in students' math score in first grade at 95% significance level.

```r
# Extract the coefficient
coef_value <- summary1[2, "Estimate"]

# Extract the robust standard error
robust_se <- summary1[2, "Std. Error"]

# Effect Size
control_sd <- sd(df_1[df_1$cltypek == 0, ]$tmathssk)
eff_size <- coef_value / control_sd

# CI for Effect Size
# For a 95% confidence interval, z-value is approximately 1.96
z_value <- 1.96

lower_ci <- (coef_value - (z_value * robust_se)) / control_sd
upper_ci <- (coef_value + (z_value * robust_se)) / control_sd

ci <- c(lower_ci, upper_ci)
eff_size
```

```
[1] 0.06545182
```

```r
ci
```

```
[1] 0.01244648 0.11845716
```

3d.

1. Effect Size: 0.06545182
2. 95% CI: (0.01244648, 0.11845716)

3e. Class size reduction in kindergarten will exert positive effect on Grade 1 math achievement at 95% significance level and the effect size is 0.065.

4. Non-random Attrition and Non-compliance

4a.

1. Attrition rate:

```r
sum(df_1$cltypek == 0) / nrow(df_1)
```

```
[1] 0.6846641
```

```r
sum(df_1$cltypek == 1) / nrow(df_1)
```

```
[1] 0.3153359
```

Attrition rate differs between the treated and untreated group. But they do not violate identification assumptions.

2. Composition of Kindergarteners:

```r
table(df_1$cltypek, df_1$ssex)
```

```

   female male
 0   1354 1347
 1    613  631
```

```r
table(df_1$cltypek,  df_1$srace)
```

```

   asian black hispanic other white
 0     8   886        0     1  1806
 1     1   403        2     2   836
```

```r
table(df_1$cltypek, df_1$freelch1)
```

```
    free lunch missing free-lunch information non-free lunch
  0       1276                             73           1352
  1        606                             23            615
```

The composition is similar in terms of gender and race/ethnicity while different in terms of free-lunch status within each group.

As long as the expected outcome under certain treatment is independent of its treatment within each group, hence, identification assumption will hold and this is not affected by the size or composition of each group.

4b.

```r
# There's no missing in df_1
sum(is.na(df_1$cltype1))
```

```
[1] 0
```

```r
sum((df_1$cltypek == 1 & df_1$cltype1 == 0)) / sum(df_1$cltypek == 1)
```

```
[1] 0.085209
```

```r
sum((df_1$cltypek == 0 & df_1$cltype1 == 1)) / sum(df_1$cltypek == 0)
```

```
[1] 0.07811922
```

```r
# ZK = 1 -> ZF = 0
Y_1_0 <- mean(df_1[(df_1$cltypek == 1 & df_1$cltype1 == 0), ]$tmathssk)
Y_1_1 <- mean(df_1[(df_1$cltypek == 1 & df_1$cltype1 == 1), ]$tmathssk)
Y_1_0
```

```
[1] 520.5755
```

```r
Y_1_1
```

```
[1] 530.3436
```

```r
# ZK = 0 -> ZF = 1
Y_0_1 <- mean(df_1[(df_1$cltypek == 0 & df_1$cltype1 == 1), ]$tmathssk)
Y_0_0 <- mean(df_1[(df_1$cltypek == 0 & df_1$cltype1 == 0), ]$tmathssk)
Y_0_1
```

```
[1] 521.9621
```

```r
Y_0_0
```

```
[1] 519.4944
```

1. 8.52% of students initially assigned to small classes in kindergarten ($Z_K = 1$) switched to regular classes in Grade 1 ($Z_F = 0$).
2. 7.81% of students initially assigned to regular classes in kindergarten ($Z_K = 0$) switched to small classes in Grade 1 ($Z_F = 1$).
3. Yes. As we know from the question 3, small class is better than regular class in imrpoving math achievement. Therefore, for the initial student of small class, student with higher grades tend to stay in the same class while students with lower grades will find the class not suitable for them and will tend to convert to other classes. For the initial student of regular class, student with higher grades want to pursue a more challenging class so they want to convert while students with lower grades tend to stay the same class.
4. No, it wouldn't. Identification assumption holds for the independence between the expected outcome under certain treatment and its treatment within each group and each individual's future transition choice won't affect this equality.

5. Setup for Research Question 2

5a. Prima Facie Effect can be written as follows:

$$\delta_{PF_F} = E[Y_F(Z_F = 1)|Z_F = 1] - E[Y_F(Z_F = 0)|Z_F = 0].$$

It stands for the mean difference between Grade 1 regular class and small class in students' first grade math score.

5b. Under Identification Assumption (written as below), $\delta_{PF_F} = \delta_F$.

$$E[Y_F(Z_F = 1)|Z_F = 1] = E[Y_F(Z_F = 1)|Z_F = 0] = E[Y_F(Z_F = 1)]$$

$$E[Y_F(Z_F = 0)|Z_F = 0] = E[Y_F(Z_F = 0)|Z_F = 1] = E[Y_F(Z_F = 0)]$$

5c. The reason why the assumption is plausible in the current study is that students involved in STAR project are a representative and random sample for the whole students and the study is also conducted as a randomization experiment.

5d. No, they wouldn't. Identification assumption holds for the independence between the expected outcome under certain treatment and its treatment within each group. This won't be affected by the size or composition of each group and each individual's future transition choice within each group.

6. Naive Analysis for Research Question 2

```
model2 <- lm(tmathss1~cltype1, data = df_sample)
vcov_cl <- vcovCL(model2, cluster = ~ schid1n, data = df_sample)
summary2 <- coeftest(model2, vcov = vcov_cl)
summary2
```

```
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 543.6630     3.3092 164.291 < 2.2e-16 ***
cltype11      8.9048     3.1997   2.783  0.005402 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Extract the coefficient
coef_value <- summary2[2, "Estimate"]

# Extract the robust standard error
robust_se <- summary2[2, "Std. Error"]

# Effect Size
control_sd <- sd(df_sample[df_sample$cltype1 == 0, ]$tmathss1)
eff_size <- coef_value / control_sd
```

```
# CI for Effect Size
# For a 95% confidence interval, z-value is approximately 1.96
z_value <- 1.96

lower_ci <- (coef_value - (z_value * robust_se)) / control_sd
upper_ci <- (coef_value + (z_value * robust_se)) / control_sd

ci <- c(lower_ci, upper_ci)
eff_size
```

```
[1] 0.09083806
```

```
ci
```

```
[1] 0.02686259 0.15481353
```

6a.

1. Estimate: 8.9048
2. Robust Std. Error: 3.1997
3. t-value: 2.783
4. p-value: 0.005402 **
5. Effect Size: 0.09083806
6. 95% CI: (0.02686259, 0.15481353)

6b. Class size reduction in first grade will on average increase 8.905 points in students' math score in first grade at 99% significance level. And its effect size is 0.091.

7. Counfounding Analysis

Data Preparation: Check missingness of each potential confounder.

```
# check missingness
sum(is.na(df_sample$ssex))
```

```
[1] 13
```

```r
sum(is.na(df_sample$srace))
```

```
[1] 29
```

```r
sum(is.na(df_sample$freelch1))
```

```
[1] 0
```

```r
sum(is.na(df_sample$clad1))
```

```
[1] 42
```

```r
sum(is.na(df_sample$totexp1))
```

```
[1] 19
```

Data Preparation: For a categorical covariate, the missing cases may constitute an additional category. For a covariate measured on an interval scale, we create a missing indicator for the missing cases and then use the sample mean to replace the missing values in the covariate.

```r
# deal with missing values in totexp1
df_sample$totexp1[df_sample$totexp1=="first year teacher"] <- 0
df_sample$totexp1 <- as.numeric(df_sample$totexp1)
mean_totexp1 <- mean(df_sample$totexp1, na.rm = TRUE)
df_sample$totexp1[which(is.na(df_sample$totexp1))] <- mean_totexp1
df_sample$totexp1_miss_index <- as.numeric(df_sample$totexp1==mean_totexp1)

# deal with missing values in gender, race and clad level
df_sample$clad1[which(is.na(df_sample$clad1))] <- "missing"
df_sample$clad1 <- as.factor(df_sample$clad1)
df_sample$ssex[which(is.na(df_sample$ssex))] <- "missing"
df_sample$ssex <- as.factor(df_sample$ssex)
df_sample$srace[which(is.na(df_sample$srace))] <- "missing"
df_sample$srace <- as.factor(df_sample$srace)

df_sample$freelch1 <- as.factor(df_sample$freelch1)
```

```r
df_confound <- df_sample[,c("tmathss1", "cltype1",
                            "freelch1", "ssex",
                            "srace", "clad1",
                            "totexp1", "totexp1_miss_index")]
```

7a. We use χ^2 test to examine the relatioship between treatment and each covariate and for a confounder, it should not have a significant relationship with treatment. We also use ANOVA model to test whether there exists a significant association between outcome and each covariate. If it does, then this covariate might be a confounder.

```r
# Relationship b/w treatment and gender**
table(df_sample$cltype1, df_sample$ssex)
```

```
    female male missing
  0   2163 2350      12
  1    847  885       1
```

```r
chisq.test(as.factor(df_sample$cltype1), as.factor(df_sample$ssex))
```

```
Warning in chisq.test(as.factor(df_sample$cltype1), as.factor(df_sample$ssex)):
Chi-squared approximation may be incorrect
```

```

    Pearson's Chi-squared test

data:  as.factor(df_sample$cltype1) and as.factor(df_sample$ssex)
X-squared = 3.0792, df = 2, p-value = 0.2145
```

```r
# Relationship b/w outcome and gender
summary(aov(df_sample$tmathss1 ~ df_sample$ssex))
```

```
                 Df   Sum Sq Mean Sq F value Pr(>F)
df_sample$ssex    2  2697605 1348803   151.6 <2e-16 ***
Residuals      6255 55665825    8899
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(df_sample[df_sample$cltype1 == 1, ]$tmathss1 ~
              df_sample[df_sample$cltype1 == 1, ]$ssex))
```

```
                                            Df   Sum Sq Mean Sq F value   Pr(>F)
df_sample[df_sample$cltype1 == 1, ]$ssex     2   199617   99809   11.84 7.85e-06
Residuals                                 1730 14589710    8433

df_sample[df_sample$cltype1 == 1, ]$ssex ***
Residuals
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(df_sample[df_sample$cltype1 == 0, ]$tmathss1 ~
              df_sample[df_sample$cltype1 == 0, ]$ssex))
```

```
                                            Df   Sum Sq Mean Sq F value Pr(>F)
df_sample[df_sample$cltype1 == 0, ]$ssex     2  2528454 1264227   139.6 <2e-16
Residuals                                 4522 40946285    9055

df_sample[df_sample$cltype1 == 0, ]$ssex ***
Residuals
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Relationship b/w treatment and race**
table(df_sample$cltype1, df_sample$srace)
```

```
    am. indian asian black hispanic missing other white
  0          6    15  1584        6      25     7  2882
  1          3     6   581        3       4     3  1133
```

```r
chisq.test(as.factor(df_sample$cltype1), as.factor(df_sample$srace))
```

```
Warning in chisq.test(as.factor(df_sample$cltype1),
as.factor(df_sample$srace)): Chi-squared approximation may be incorrect
```

```
	Pearson's Chi-squared test

data:  as.factor(df_sample$cltype1) and as.factor(df_sample$srace)
X-squared = 4.468, df = 6, p-value = 0.6136
```

```
# Relationship b/w outcome and race
summary(aov(df_sample$tmathss1 ~ df_sample$srace))
```

```
                  Df   Sum Sq Mean Sq F value Pr(>F)    
df_sample$srace    6  6596443 1099407   132.8 <2e-16 ***
Residuals       6251 51766988    8281                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(df_sample[df_sample$cltype1 == 1, ]$tmathss1 ~
              df_sample[df_sample$cltype1 == 1, ]$srace))
```

```
                                            Df   Sum Sq Mean Sq F value Pr(>F)
df_sample[df_sample$cltype1 == 1, ]$srace    6  1057759  176293   22.16 <2e-16
Residuals                                 1726 13731568    7956               
                                             
df_sample[df_sample$cltype1 == 1, ]$srace ***
Residuals                                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(df_sample[df_sample$cltype1 == 0, ]$tmathss1 ~
              df_sample[df_sample$cltype1 == 0, ]$srace))
```

```
                                            Df   Sum Sq Mean Sq F value Pr(>F)
df_sample[df_sample$cltype1 == 0, ]$srace    6  5636588  939431   112.2 <2e-16
Residuals                                 4518 37838151    8375               
                                             
df_sample[df_sample$cltype1 == 0, ]$srace ***
Residuals                                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Relationship b/w treatment and clad1
table(df_sample$cltype1, df_sample$clad1)
```

```
    apprentice chose no to be on career ladder ladder level 1 level 2 level 3
  0        486                                 360            2903      81     146
  1        190                                 125            1155      15     101

    missing probation
  0      23       526
  1      19       128
```

```r
chisq.test(as.factor(df_sample$cltype1), as.factor(df_sample$clad1))
```

```

    Pearson's Chi-squared test

data:  as.factor(df_sample$cltype1) and as.factor(df_sample$clad1)
X-squared = 58.615, df = 6, p-value = 8.598e-11
```

```r
# Relationship b/w outcome and clad1
summary(aov(df_sample$tmathss1 ~ df_sample$clad1))
```

```
                  Df   Sum Sq Mean Sq F value   Pr(>F)    
df_sample$clad1    6   333113   55519    5.98 3.03e-06 ***
Residuals       6251 58030317    9283                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(df_sample[df_sample$cltype1 == 1, ]$tmathss1 ~
              df_sample[df_sample$cltype1 == 1, ]$clad1))
```

```
                                            Df   Sum Sq Mean Sq F value
df_sample[df_sample$cltype1 == 1, ]$clad1    6   654060  109010   13.31
Residuals                                 1726 14135268    8190        
                                            Pr(>F)    
df_sample[df_sample$cltype1 == 1, ]$clad1 8.56e-15 ***
```

```
Residuals
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(df_sample[df_sample$cltype1 == 0, ]$tmathss1 ~
              df_sample[df_sample$cltype1 == 0, ]$clad1))
```

```
                                           Df   Sum Sq Mean Sq F value Pr(>F)
df_sample[df_sample$cltype1 == 0, ]$clad1    6    95288   15881   1.654  0.128
Residuals                                 4518 43379451    9601
```

```
# Relationship b/w treatment and freelch
table(df_sample$cltype1, df_sample$freelch1)
```

```
    free lunch missing free-lunch information non-free lunch
  0       2387                            134           2004
  1        865                             36            832
```

```
chisq.test(as.factor(df_sample$cltype1), as.factor(df_sample$freelch1))
```

```
    Pearson's Chi-squared test

data:  as.factor(df_sample$cltype1) and as.factor(df_sample$freelch1)
X-squared = 9.3773, df = 2, p-value = 0.009199
```

```
# Relationship b/w outcome and freelch
summary(aov(df_sample$tmathss1 ~ df_sample$freelch1))
```

```
                     Df   Sum Sq Mean Sq F value   Pr(>F)
df_sample$freelch1    2   379310  189655   20.46 1.39e-09 ***
Residuals          6255 57984121    9270
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(df_sample[df_sample$cltype1 == 1, ]$tmathss1 ~
              df_sample[df_sample$cltype1 == 1, ]$freelch1))
```

```
                                              Df   Sum Sq Mean Sq F value
df_sample[df_sample$cltype1 == 1, ]$freelch1    2   276853  138426    16.5
Residuals                                   1730 14512474    8389
                                              Pr(>F)
df_sample[df_sample$cltype1 == 1, ]$freelch1 7.96e-08 ***
Residuals
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(df_sample[df_sample$cltype1 == 0, ]$tmathss1 ~
              df_sample[df_sample$cltype1 == 0, ]$freelch1))
```

```
                                              Df   Sum Sq Mean Sq F value
df_sample[df_sample$cltype1 == 0, ]$freelch1    2   197490   98745   10.32
Residuals                                   4522 43277249    9570
                                              Pr(>F)
df_sample[df_sample$cltype1 == 0, ]$freelch1 3.38e-05 ***
Residuals
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Relationship b/w treatment and totexp1
table(df_sample$cltype1, df_sample$totexp1_miss_index)
```

```
      0    1
  0 4525    0
  1 1714   19
```

```r
chisq.test(as.factor(df_sample$cltype1), df_sample$totexp1)
```

```
Warning in chisq.test(as.factor(df_sample$cltype1), df_sample$totexp1):
Chi-squared approximation may be incorrect
```

```
	Pearson's Chi-squared test

data:  as.factor(df_sample$cltype1) and df_sample$totexp1
X-squared = 646.63, df = 40, p-value < 2.2e-16
```

```r
# Relationship b/w outcome and totexp1
summary(aov(df_sample$tmathss1 ~
              df_sample$totexp1 + df_sample$totexp1_miss_index))
```

```
                               Df   Sum Sq Mean Sq F value   Pr(>F)
df_sample$totexp1               1     2242    2242   0.243    0.622
df_sample$totexp1_miss_index    1   540099  540099  58.427 2.43e-14 ***
Residuals                    6255 57821090    9244
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(df_sample[df_sample$cltype1 == 1, ]$tmathss1 ~
              df_sample[df_sample$cltype1 == 1, ]$totexp1 +
              df_sample[df_sample$cltype1 == 1, ]$totexp1_miss_index))
```

```
                                                    Df   Sum Sq Mean Sq
df_sample[df_sample$cltype1 == 1, ]$totexp1            1      978     978
df_sample[df_sample$cltype1 == 1, ]$totexp1_miss_index    1   503430  503430
Residuals                                         1730 14284919    8257
                                                  F value   Pr(>F)
df_sample[df_sample$cltype1 == 1, ]$totexp1          0.118    0.731
df_sample[df_sample$cltype1 == 1, ]$totexp1_miss_index  60.969 9.97e-15 ***
Residuals
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(df_sample[df_sample$cltype1 == 0, ]$tmathss1 ~
              df_sample[df_sample$cltype1 == 0, ]$totexp1 +
              df_sample[df_sample$cltype1 == 0, ]$totexp1_miss_index))
```

```
                                           Df   Sum Sq Mean Sq F value
df_sample[df_sample$cltype1 == 0, ]$totexp1    1     4256    4256   0.443
```

```
Residuals                                  4523 43470483    9611
                                           Pr(>F)
df_sample[df_sample$cltype1 == 0, ]$totexp1  0.506
Residuals
```

Among the above pre-treatment covariates, **gender** and **race** may have confounded our previous estimation.

7b.

```r
model3 <- lm(tmathss1 ~ cltype1 + ssex + srace, data = df_sample)
summary(model3)
```

```
Call:
lm(formula = tmathss1 ~ cltype1 + ssex + srace, data = df_sample)

Residuals:
    Min      1Q  Median      3Q     Max 
-423.41  -41.15  -13.30   15.70  476.16 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    768.815     30.345  25.336  < 2e-16 ***
cltype11         9.845      2.568   3.833 0.000128 ***
ssexmale         3.355      2.302   1.457 0.145132    
ssexmissing     63.185     33.966   1.860 0.062895 .  
sraceasian    -165.209     36.212  -4.562 5.16e-06 ***
sraceblack    -245.974     30.359  -8.102 6.44e-16 ***
sracehispanic -226.699     42.844  -5.291 1.26e-07 ***
sracemissing   166.242     37.869   4.390 1.15e-05 ***
sraceother    -230.082     41.757  -5.510 3.73e-08 ***
sracewhite    -220.867     30.329  -7.282 3.69e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 90.88 on 6248 degrees of freedom
Multiple R-squared:  0.1158,	Adjusted R-squared:  0.1146 
F-statistic: 90.95 on 9 and 6248 DF,  p-value: < 2.2e-16
```

I think we have underestimated the Grade 1 class size effect. On the one hand, from the above empirical evidence, the estimand of class size effect has increased compared with our previous findings in question 6. On the other hand, the existence of confounders (gender and race) might eliminate the class size effect when we fail to include confounders in our model.

8. Covariance Adjustment for Confounders

8a. Regression Model:

$$Y_{F,i} = \beta_0 + \beta_1 Z_{F,i} + \sum_{s \in \{male, missing\}} \beta_s \mathbf{1}_{s,i} + \sum_{r \in \{asian, black, hispanic, white, other, missing\}} \beta_r \mathbf{1}_{r,i} + \epsilon_i$$

1. Assumption 1: Linearity: The relationship between the covariates and the dependent variable should be linear.
2. Assumption 2: Homogeneity of variances: The variances in different groups should be similar.
3. Assumption 3: Normality: The residuals should be approximately normally distributed.

8b.

```
model4 <- aov(tmathss1 ~ cltype1 + ssex + srace, data = df_sample)
vcov_cl <- vcovCL(model4, cluster = ~ schid1n, data = df_sample)
summary4 <- coeftest(model4, vcov = vcov_cl)
summary4
```

```
t test of coefficients:

               Estimate Std. Error t value  Pr(>|t|)
(Intercept)    768.8152    88.8876  8.6493 < 2.2e-16 ***
cltype11         9.8450     3.1118  3.1637  0.001565 **
ssexmale         3.3546     2.2863  1.4673  0.142348
ssexmissing     63.1853    41.3251  1.5290  0.126320
sraceasian    -165.2087    98.3010 -1.6806  0.092883 .
sraceblack    -245.9742    88.3015 -2.7856  0.005359 **
sracehispanic -226.6990    82.1873 -2.7583  0.005827 **
sracemissing   166.2422    70.7711  2.3490  0.018854 *
sraceother    -230.0815    88.3673 -2.6037  0.009244 **
sracewhite    -220.8673    88.1315 -2.5061  0.012232 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Extract the coefficient
coef_value <- summary4[2, "Estimate"]

# Extract the robust standard error
robust_se <- summary4[2, "Std. Error"]

# Effect Size
control_sd <- sd(df_sample[df_sample$cltype1 == 0, ]$tmathss1)
eff_size <- coef_value / control_sd

# CI for Effect Size
# For a 95% confidence interval, z-value is approximately 1.96
z_value <- 1.96

lower_ci <- (coef_value - (z_value * robust_se)) / control_sd
upper_ci <- (coef_value + (z_value * robust_se)) / control_sd

ci <- c(lower_ci, upper_ci)
eff_size
```

```
[1] 0.1004293
```

```r
ci
```

```
[1] 0.03821133 0.16264720
```

1. Estimate: 9.8450
2. Robust Std. Error: 3.1118
3. t-value: 3.1637
4. p-value: 0.001565 **
5. Effect Size: 0.1004293
6. 95% CI: (0.03821133, 0.16264720)

8c. It is the same that class size reduction in first graden will exert positive effect on Grade 1 math achievement at 99% significance level but the effect size has increased to 0.100.

9. Restricted Student Sample

```
df_9 <- df_sample[df_sample$stark == "NO", ]
model5 <- lm(tmathss1~cltype1, data = df_9)
vcov_cl <- vcovCL(model5, cluster = ~ schid1n, data = df_9)
summary5 <- coeftest(model5, vcov = vcov_cl)
summary5
```

```
t test of coefficients:

            Estimate Std. Error  t value Pr(>|t|)
(Intercept) 548.1711     4.7657 115.0234  < 2e-16 ***
cltype11     15.2378     6.7243   2.2661  0.02354 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Extract the coefficient
coef_value <- summary5[2, "Estimate"]

# Extract the robust standard error
robust_se <- summary5[2, "Std. Error"]

# Effect Size
control_sd <- sd(df_9[df_9$cltype1 == 0, ]$tmathss1)
eff_size <- coef_value / control_sd

# CI for Effect Size
# Assuming a 95% confidence interval, z-value is approximately 1.96
z_value <- 1.96

lower_ci <- (coef_value - (z_value * robust_se)) / control_sd
upper_ci <- (coef_value + (z_value * robust_se)) / control_sd

ci <- c(lower_ci, upper_ci)
eff_size
```

```
[1] 0.1306243
```

```
ci
```

```
[1] 0.01764409 0.24360442
```

No, it doesn't. It exlcudes the confounders (eg. education background in kindergarten, gender, race), which may bias the estimation torwards population average causal effect.

10. Adding Z_K and Y_K as Confounders

10a. Adding Z_K as a confounder:

```
df_sample[is.na(df_sample$cltypek), ]$cltypek <- "missing"
df_sample$cltypek <- as.factor(df_sample$cltypek)

model6 <- aov(tmathss1 ~ cltype1 + ssex + srace + cltypek, data = df_sample)
vcov_cl <- vcovCL(model6, cluster = ~ schid1n, data = df_sample)
summary6 <- coeftest(model6, vcov = vcov_cl)
summary6
```

```
t test of coefficients:

                 Estimate Std. Error t value  Pr(>|t|)    
(Intercept)     764.47903   88.35193  8.6527 < 2.2e-16 ***
cltype11         11.27191    4.87352  2.3129  0.020761 *  
ssexmale          3.19193    2.31002  1.3818  0.167089    
ssexmissing      63.24137   41.40868  1.5272  0.126750    
sraceasian     -163.43138   98.18094 -1.6646  0.096044 .  
sraceblack     -243.35986   87.95679 -2.7668  0.005677 ** 
sracehispanic  -225.63357   82.16734 -2.7460  0.006049 ** 
sracemissing    166.44348   70.79229  2.3512  0.018746 *  
sraceother     -228.65650   88.34278 -2.5883  0.009668 ** 
sracewhite     -218.00046   87.81417 -2.4825  0.013072 *  
cltypek1         -0.98798    4.92467 -0.2006  0.841004    
cltypekmissing    3.96904    3.41264  1.1630  0.244856    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coefficients of Z_F stands for population average causal effect of class size reduction only in Grade 1 on Grade 1 math achievement.

10b. Yes, I agree. Because as discussed in 4b, kindergarten math achievement (Y_K) can influence the treatment in Grade 1 (Z_F). In this way, we should include it as a confounder.

```
model7 <- aov(tmathss1 ~ cltype1 + ssex + srace + tmathssk, data = df_sample)
vcov_cl <- vcovCL(model7, cluster = ~ schid1n, data = df_sample)
summary7 <- coeftest(model7, vcov = vcov_cl)
summary7
```

```
t test of coefficients:

                 Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)    7.5181e+02  8.7751e+01  8.5676 < 2.2e-16 ***
cltype11       1.1341e+01  3.2298e+00  3.5113  0.000449 ***
ssexmale       3.0884e+00  2.3030e+00  1.3411  0.179948    
ssexmissing    6.3184e+01  4.1354e+01  1.5279  0.126592    
sraceasian    -1.6161e+02  9.8159e+01 -1.6464  0.099734 .  
sraceblack    -2.4113e+02  8.7825e+01 -2.7456  0.006057 ** 
sracehispanic -2.2485e+02  8.1992e+01 -2.7424  0.006117 ** 
sracemissing   1.6645e+02  7.0801e+01  2.3509  0.018757 *  
sraceother    -2.2863e+02  8.8316e+01 -2.5888  0.009654 ** 
sracewhite    -2.1562e+02  8.7719e+01 -2.4581  0.013995 *  
tmathssk       1.6702e-02  6.2358e-03  2.6784  0.007416 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

11. Bonus Questions

B1. Pf:

$$\delta_{PF} - \delta = (E[Y|Z = 1] - E[Y|Z = 0]) - (E[Y(1) - Y(0)])$$

For the first term, we have:

$$E[Y|Z = 1] - E[Y|Z = 0] = E[Y(1)|Z = 1] - E[Y(0)|Z = 0]$$

For the second term, we have:

$$E[Y(1)-Y(0)] = (E[Y(1)|Z = 1]-E[Y(0)|Z = 1])\times Pr(Z = 1)+(E[Y(1)|Z = 0]-E[Y(0)|Z = 0])\times Pr(Z = 0)$$

Insert two items into the original equation, we get:

$$\delta_{PF}-\delta = (E[Y(1)|Z = 1]-E[Y(1)|Z = 0])\times Pr(Z = 0)+(E[Y(0)|Z = 1]-E[Y(0)|Z = 0])\times Pr(Z = 1)$$

Add first and then plus $(E[Y(0)|Z=1]-E[Y(0)|Z=0])\times Pr(Z=0)$, finally we can get:

$$\delta_{PF}-\delta=E[Y(0)|Z=1]-E[Y(0)|Z=0]+(E[\Delta|Z=1]-E[\Delta|Z=0])\times Pr(Z=0)$$

B2. Pf:

Under Independence, we have $E[Y(1)|Z=1]=E[Y(1)|Z=0]=E[Y(1)]$ and $E[Y(0)|Z=0]=E[Y(0)|Z=1]=E[Y(0)]$.

For bias 1 $E[Y(0)|Z=1]-E[Y(0)|Z=0]$, apparently, it turns into zero.

For bias 2, we can get $E[\Delta|Z=1]=E[\Delta|Z=0]=E[Y(1)]-E[Y(0)]$ from independence assumption. Hence, bias 2 also turns into zero.

B3. Pf:

$$\delta_{PF}-ATT=(E[Y|Z=1]-E[Y|Z=0])-(E[Y(1)-Y(0)|Z=1])$$

For the first term, we have:

$$E[Y|Z=1]-E[Y|Z=0]=E[Y(1)|Z=1]-E[Y(0)|Z=0]$$

Therefore, insert the first term and we get:

$$\delta_{PF}-ATT=E[Y(0)|Z=1]-E[Y(0)|Z=0]$$

In other words, only the first bias exists.