CHDV HW3

Yuyang Jiang

Problem 1

- **kindergarten class status (Z):** Z=1 if assigned to small classes in kindergarten; 0 otherwise.
- **grade 1 class status (D):** D=1 if assigned to small classes in Grade 1; 0 otherwise.

1a.

1. **Compliers ($D_i(Z_i = 1) = 1,\ D_i(Z_i = 0) = 0$):** Students who follow their initial assignment throughout the study. Those assigned to small classes in kindergarten (Z=1) and stay in small classes in Grade 1, and those assigned to regular classes (Z=0) and stay in regular classes in Grade 1.
2. **Always-takers ($D_i(Z_i = 1) = D_i(Z_i = 0) = 1$):** Students who end up in small classes in Grade 1 regardless of their initial assignment in kindergarten. This means they would be in small classes in Grade 1 even if they were initially assigned to regular classes in kindergarten.
3. **Never-takers ($D_i(Z_i = 1) = D_i(Z_i = 0) = 0$):** Students who end up in regular classes in Grade 1 regardless of their initial assignment in kindergarten. This means they would be in regular classes in Grade 1 even if they were initially assigned to small classes in kindergarten.
4. **Defiers ($D_i(Z_i = 1) = 0,\ D_i(Z_i = 0) = 1$):** Students who do the opposite of their assignment. Those assigned to small classes (Z=1) in kindergarten but end up in regular classes in Grade 1, and those assigned to regular classes (Z=0) in kindergarten but end up in small classes in Grade 1.

1b.

```
library(haven)
df <- read_dta("CHDV 30102 webstar_Winter2024_rv.dta")
df <- df[which(df$tmathss1 != 999 & df$tmathssk != 999
```

```
                & df$cltypek < 9 & df$cltype1 < 9), ]
df$Z <- 1
df[(df$cltypek == 2 | df$cltypek == 3), ]$Z <- 0
df$D <- 1
df[(df$cltype1 == 2 | df$cltype1 == 3), ]$D <- 0
nrow(df)
```

```
[1] 3623
```

```
table(df$D, df$Z)
```

```
     0    1
0 2291   95
1  192 1045
```

Under the assumed absence of defiers:

- $Pr(compliers) = ITT = E[D|Z=1] - E[D|Z=0] = \frac{1045}{1045+99} - \frac{192}{192+2291} = 0.839$
- $Pr(always-takers) = \frac{192}{3623} = 0.053$
- $Pr(never-takers) = \frac{95}{3623} = 0.026$

Problem 2

2a.

Assumptions	The LATE
1. Stable unit treatment value assumption (SUTVA)	x
2. Exogeneity of Z	x
3. Exclusion restriction	x
4. Nonzero effect of Z on D	x
5. Monotonicity	x
6. Constant treatment effects; or zero covariance between the effect of Z on D and the effect of D on Y	

Figure 1: tab

1. **SUTVA:** $D_i(\mathbf{z}) = D_i(z_i)$: One's class assignment in Grade 1 was unaffected by the kindergarten class status of others. $Y_i(\mathbf{z, d}) = Y_i(z_i, d_i)$: One's grade 1 math achievement was unaffected by the grade 1 class status and kindergarten class status of others.
2. **Exogeneity of Z:** In the context of IV methods, it specifically refers to the instrument being independent of the error term in the outcome equation. This means the initial class status in kindergarten affects the math achievement in grade 1 only through its influence on the class status in grade 1.
3. **Exclusion restriction:** $(Y(z, d) = Y(d))$ Grade 1 math achievement was no longer affected by initial class status in kindergarten once grade 1 class status is determined; thus the causal effect of Z on Y is zero for always-takers and never-takers.
4. **Nonzero effect of Z on D:** $(E[D_i(1) - D_i(0)] \neq 0)$ Changing initial class assignment in kindergarten from regular classes to small classes affects the probability of being assigned to small classes in grade 1 on average. In other words, there is at least one complierin the population.
5. **Monotonicity:** $(D_i(1) \geq D_i(0)$ **for all** $i = 1, 2, ..., N)$ There is no defier who would have been assigned to the opposite class type in grade 1 against his/her initial class assignment in kindergarten.

2b.+2c.

1. SUTVA

```
library(dplyr)
```

```
Attaching package: 'dplyr'

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

    filter, lag

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

    intersect, setdiff, setequal, union
```

```
tab1 <- df %>%
  group_by(Z, D) %>%
  summarise(pr = n() / nrow(df), .groups = 'drop')
tab1
```

```
# A tibble: 4 x 3
      Z     D     pr
  <dbl> <dbl>  <dbl>
1     0     0 0.632
2     0     1 0.0530
3     1     0 0.0262
4     1     1 0.288
```

```
tab2 <- df %>%
  group_by(Z, D) %>%
  summarise(mean_Y = mean(tmathss1))
```

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

```
tab2
```

```
# A tibble: 4 x 3
# Groups:   Z [2]
      Z     D mean_Y
  <dbl> <dbl>  <dbl>
1     0     0   530.
2     0     1   537.
3     1     0   530.
4     1     1   541.
```

The tables above do not directly test the SUTVA assumptions but offer some insights:

- The no interference assumption is supported by the study's design if we assume that class assignments are independent across students.
- The consistency assumption is more about the treatment application and less about the observed outcomes; it is generally supported by the controlled nature of the study's treatment assignments.

To fully address SUTVA, we must rely on the study design and contextual knowledge about how treatments were administered and whether there could be any crossover effects between units (students). Below is theoretical justification in this context.

- Random Assignment: Students were randomly assigned to small or regular classes in kindergarten. This process helps ensure that the potential outcomes across different treatment groups are comparable, as randomization tends to balance both observed and unobserved factors across groups.
- No Evidence of Non-Compliance in Kindergarten: The documentation indicating no evidence of non-compliance during the kindergarten year supports the SUTVA assumption. It implies that the treatment (class size) was administered as intended, and students remained in their assigned conditions, thereby satisfying the consistency aspect of SUTVA for that year.
- Evidence of Non-Compliance in Grade 1: Finding evidence of non-compliance in Grade 1, where the treatment received (D) does not match the initial assignment (Z), complicates the SUTVA assumption in subsequent years. It indicates a breach in the consistency of treatment—students did not receive the treatment as assigned beyond kindergarten. However, this non-compliance does not inherently violate the no-interference assumption unless the switching of treatments between students affects the outcomes of others, which is not directly suggested by the information provided.

2. Exogeneity of Z

```
model2.1 <- lm(tmathss1~D, data = df)
cor(df$Z, model2.1$residuals)
```

```
[1] 0.008782155
```

Considering that the correlation between instrument variable and error term in outcome equation is vvery close to zero, we can conclude that Z is uncorrelated with error terms, ensuring the exogeneity of Z.

3. Exclusion Restriction

Since we have only one instrument for one endogenous regressor, our model is just-identified, and we cannot perform overidentification tests (such as the Hansen J test) to directly test the validity of the exclusion restriction assumption. In just-identified models, the number of instruments matches the number of endogenous variables, which means there are no extra instruments to test for the validity of the instruments. The exclusion restriction—that the instrument affects the dependent variable only through the endogenous variable—must be justified theoretically and cannot be tested directly with statistical tests in this scenario. Thus, below is theoretical justification in this context.

The exclusion restriction for using kindergarten class size assignment as an instrument for actual class size in analyzing educational outcomes is theoretically justified by the random nature of the assignment, the controlled study environment, the mechanisms through which

class size is presumed to affect educational outcomes, and the external validation of the study's implementation. These factors collectively argue that the instrument affects educational outcomes only through its impact on class size, without other direct pathways, thereby satisfying the exclusion restriction.

4. Nonzero effect of Z on D from 1b, we can see that $Pr(compiers) > 0$. Thus, we can conclude that nonzero average causal effect of Z on D holds.

5. Monotonicity

```
tab3 <- df %>%
  group_by(Z) %>%
  summarise(mean_D = mean(D))
tab3
```

```
# A tibble: 2 x 2
      Z mean_D
  <dbl>  <dbl>
1     0 0.0773
2     1 0.917
```

The data supports the monotonicity assumption. For individuals assigned to regular classes (Z=0), almost all remain in regular classes (mean_D close to 0), and for those assigned to small classes (Z=1), the vast majority attend small classes (mean_D close to 1). This pattern suggests there's a consistent direction in the effect of Z on D, with no evidence of individuals moving in the opposite direction of their assignment, which would violate monotonicity.

Problem 3

```
df$G <- 9
df[which(df$hdegk == 2), ]$G <- 0
df[which(df$hdegk == 3 | df$hdegk == 4 |
           df$hdegk == 5 | df$hdegk == 6), ]$G <- 1
# exclude the sample with hdegk = 1 or 9
df <- df[which(df$G != 9), ]
tab4 <- df %>%
  group_by(Z) %>%
  summarise(mean_G = mean(G))
tab4
```

```
# A tibble: 2 x 2
      Z mean_G
  <dbl>  <dbl>
1     0  0.384
2     1  0.282
```

3a. Almost equally likely. But still kindergarteners in small classes are a bit more likely to be taught by teachers with a master's degree or above.

3b. The suggestion to use the randomized treatment assignment to a small or regular class in kindergarten (Z) as an instrument for identifying the effect of kindergarten teacher degree (G) on math achievement raises concerns regarding its validity due to three main criteria:

1. Relevance: It's not evident that the assignment to class size would systematically influence the teacher's degree level, making Z potentially irrelevant to G.
2. Independence: While Z is likely independent of unobserved confounders due to randomization, satisfying this criterion.
3. Exclusion Restriction: Z likely violates this criterion because the class size can directly impact student math achievement through avenues other than the teacher's degree, such as the level of individual attention students receive.

In conclusion, while Z satisfies the independence criterion, it likely fails both the relevance and exclusion restriction criteria, making it an invalid instrument for G in this context.

```r
# baseline regression
model3.1 <- lm(tmathssk ~ G, data = df)
summary(model3.1)
```

```
Call:
lm(formula = tmathssk ~ G, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-137.190  -32.190   -3.977   28.810  134.810

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 491.1896     0.9527 515.581   <2e-16 ***
G             1.7875     1.6066   1.113    0.266
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 46.05 on 3601 degrees of freedom
Multiple R-squared:  0.0003436, Adjusted R-squared:  6.604e-05
F-statistic: 1.238 on 1 and 3601 DF,  p-value: 0.266
```

```r
library(AER)
```

```
Loading required package: car
```

```
Loading required package: carData
```

```

Attaching package: 'car'
```

```
The following object is masked from 'package:dplyr':

    recode
```

```
Loading required package: lmtest
```

```
Loading required package: zoo
```

```

Attaching package: 'zoo'
```

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

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

```
Loading required package: sandwich
```

```
Loading required package: survival
```

```r
model3.2 <- ivreg(tmathssk ~ G | Z, data = df)
summary(model3.2)
```

```
Call:
ivreg(formula = tmathssk ~ G | Z, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-162.489  -43.489   -3.489   37.669  179.669

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  516.489      7.214  71.598  < 2e-16 ***
G            -70.158     20.333  -3.451 0.000566 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.45 on 3601 degrees of freedom
Multiple R-Squared: -0.5564,    Adjusted R-squared: -0.5568
Wald test: 11.91 on 1 and 3601 DF,  p-value: 0.0005659
```

3c. The IV regression (using Z as an instrument for G) yields a significantly negative estimate for the effect of kindergarten teacher degree (G) on math scores, with a large and significant point estimate of -70.158 and a p-value of 0.000566. In contrast, the simple linear regression of math scores on teacher degree without using the instrument shows a small and statistically insignificant effect of G on math scores, with an estimate of 1.7875 and a p-value of 0.266. The standard errors in the IV regression are larger (20.333) compared to those in the simple regression (1.6066), reflecting the potentially increased variability when using instrumental variables. This comparison suggests that the IV estimate, which attempts to address endogeneity, reveals a substantial negative impact of teacher degree on student math achievement that is not apparent in the naive regression analysis.

Problem 4

4a. The suggestion to use a fuzzy regression discontinuity design (RDD) with a class size cutoff of 16 in the Project STAR data is innovative but faces practical challenges. The key concerns are:

1. **Overlap in Class Sizes**: The ranges for small (12-17 students) and regular (15-28 students) classes overlap, complicating the use of 16 as a clear cutoff and potentially blurring the distinction between treatment and control groups.
2. **Fuzzy Design Appropriateness**: While the fuzzy RDD approach suits situations where treatment compliance isn't perfect, the effectiveness of this method depends on a

sharp change in the probability of receiving the treatment at the cutoff. The overlap in class sizes questions the presence of such a sharp discontinuity at 16.

3. **Assumption Strength**: The method relies on strong assumptions, particularly that the running variable (class size) influences the outcome (math achievement) only through the treatment (being in a small class). This might not hold if the cutoff was not strictly used to define treatment groups.

4. **Cutoff Justification**: The validity of an RDD hinges on a clear rationale for the chosen cutoff to represent a discontinuity in treatment assignment. If the cutoff of 16 does not have a strong basis in how the study was conducted, the RDD analysis may not accurately estimate the causal effect.

5. **Limited Generalizability**: Any causal effect identified would be most relevant to students near the cutoff, limiting the findings' applicability to a broader range of class sizes.

In essence, while Statistician 1's approach is potentially valid for estimating the causal effect of class size on math achievement, the overlap in class sizes and questions about the cutoff's justification pose significant challenges to its implementation and interpretation.

4b. This suggestion aligns with the principles of instrumental variable analysis, where Z serves as a tool to address potential endogeneity between D and Y. The rationale and considerations include:

1. Relevance: The randomly assigned class type (Z) is expected to be strongly correlated with the actual class size (D), fulfilling the relevance criterion of an IV. Small classes are designed to have fewer students, and regular classes more, making Z a plausible instrument for D.

2. Independence: Since Z is randomly assigned, it should be independent of confounders that could affect kindergarten math achievement (Y), satisfying the independence criterion.

3. Exclusion Restriction: For Z to be a valid instrument, its effect on Y must operate entirely through D (actual class size), and not through other pathways. Given the design of Project STAR, this assumption is reasonable but needs to be carefully assessed.

Using Z as an IV for D in this context allows for a causal interpretation of the effect of actual class size on math achievement by leveraging the random assignment to address issues of non-random selection into class sizes and potential omitted variable bias. This method provides a way to isolate the effect of class size from other factors that might simultaneously influence class size and student outcomes.

However, the success of this approach hinges on the strength of the assumptions, especially the exclusion restriction. If the class type influences math achievement through channels other than class size (e.g., teacher quality or resources), the validity of the IV analysis could be

compromised. Overall, assuming these conditions are met or adequately addressed, Statistician 2's suggestion is appropriate.

4c.

```
library(tidyr)
df_2 <- read_dta("CHDV 30102 webstar_gkclasssize.dta")
df_2 <- drop_na(df_2, c(gkclasssize, gktmathss, gkclasstype))
df_2$Z <- 1 # small classes
df_2[which(df_2$gkclasstype == 2 |
             df_2$gkclasstype == 3), ]$Z <- 0 # regular classes

model4 <- ivreg(gktmathss ~ gkclasssize | Z, data = df_2)
summary(model4)
```

```
Call:
ivreg(formula = gktmathss ~ gkclasssize | Z, data = df_2)

Residuals:
    Min      1Q  Median      3Q     Max
-203.06  -32.78   -4.50   28.94  147.78

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 507.0953     3.7587 134.911  < 2e-16 ***
gkclasssize  -1.0693     0.1825  -5.859 4.92e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.57 on 5869 degrees of freedom
Multiple R-Squared: 0.00555,    Adjusted R-squared: 0.005381
Wald test: 34.32 on 1 and 5869 DF,  p-value: 4.924e-09
```

```
cat('estimated average effect: \n')
```

```
estimated average effect:
```

```
coef(model4)[2]*10
```

```
gkclasssize
  -10.69343
```

```
  cat("Effect size: (scaled by std of class size) \n")
```

```
Effect size: (scaled by std of class size)
```

```
  coef(model4)[2] / sd(df_2$gkclasssize) ##
```

```
gkclasssize
 -0.2700677
```

Problem 5

```
  df_origin <- read_dta('CHDV 30102_ECLSK98_class size_rv.dta')

  # 1. deal with null values in treatment, outcome and iv
  # Y: C4R2MSCL
  # Z: s4anumch (IV)
  # D: A4CLSIZE
  df_3 <- drop_na(df_origin, c(A4CLSIZE, C4R2MSCL, s4anumch))
  df_3 <- df_3[df_3$s4pupri == 1, ]

  # 2. deal with null values in pretreatment covariates
  sum(is.na(df_3$GENDER))
```

```
[1] 0
```

```
  sum(is.na(df_3$RACE))
```

```
[1] 0
```

```
  sum(is.na(df_3$w1sesl))
```

```
[1] 741
```

```r
sum(is.na(df_3$B4YRSTC))
```

```
[1] 162
```

```r
df_3$gender3 <- as.factor(df_3$GENDER)
levels(df_3$gender3)[levels(df_3$gender3)=='-9'] <- '3'

df_3$race6 <- as.factor(df_3$RACE)
levels(df_3$race6)[levels(df_3$race6)=='3'] <- 'Hispanic'
levels(df_3$race6)[levels(df_3$race6)=='4'] <- 'Hispanic'
levels(df_3$race6)[levels(df_3$race6)=='6'] <- 'Indigenous or Native Americans'
levels(df_3$race6)[levels(df_3$race6)=='7'] <- 'Indigenous or Native Americans'
levels(df_3$race6)[levels(df_3$race6)=='-9'] <- 'Other Races'
levels(df_3$race6)[levels(df_3$race6)=='8'] <- 'Other Races'

df_3$missing_B4 <- as.numeric(is.na(df_3$B4YRSTC))
df_3$B4YRSTC[is.na(df_3$B4YRSTC)] <- mean(df_3$B4YRSTC, na.rm = TRUE)
sum(is.na(df_3$B4YRSTC))
```

```
[1] 0
```

```r
df_3$missing_ses <- as.numeric(is.na(df_3$w1sesl))
df_3$w1sesl[is.na(df_3$w1sesl)] <- mean(df_3$w1sesl, na.rm = TRUE)
sum(is.na(df_3$w1sesl))
```

```
[1] 0
```

5a.

1. Relevance: Z should strongly predict D, which is plausible since higher enrollment could lead to larger classes if the number of teachers doesn't increase at the same rate.
2. Exogeneity: Z must not be correlated with unobserved factors that also affect math achievement. This assumption might be challenging since factors like community wealth or education policies could influence both enrollment numbers and achievement outcomes.
3. Exclusion Restriction: The impact of Z on math achievement should only be through D. This condition could be violated if, for instance, higher enrollments bring additional resources that directly improve achievement, independent of class size.

In essence, while public school enrollment might influence class size (relevance), ensuring it only affects math achievement through class size (exclusion) and is not correlated with other achievement-influencing factors (exogeneity) are critical yet challenging conditions to satisfy. These requirements make the validity of using public school enrollment as an instrumental variable for class size in this context debatable.

5b. If schools that increase Grade-1 class size in response to increased enrollment are also those where larger classes are more harmful to students' math learning, the instrumental variable (IV) analysis could indeed be biased. This situation implies a violation of the exogeneity assumption for valid IV analysis, as the instrument (enrollment) is not independent of unobserved factors that influence the outcome (math achievement).

The bias in the IV estimate would likely be negative, meaning the analysis would overestimate the harmful effects of larger class sizes on math achievement. This occurs because the instrument disproportionately selects schools where the negative impact of increased class size is more pronounced, leading to a more negative estimate of class size effects than what might be true on average across all schools.

Problem 6

6a.

```r
# Part 1
model6.1 <- lm(C4R2MSCL ~ A4CLSIZE + gender3 +
                 race6 + B4YRSTC + missing_B4 +
                 w1sesl + missing_ses, data = df_3)
summary(model6.1)
```

```
Call:
lm(formula = C4R2MSCL ~ A4CLSIZE + gender3 + race6 + B4YRSTC +
    missing_B4 + w1sesl + missing_ses, data = df_3)

Residuals:
    Min      1Q  Median      3Q     Max
-69.698  -8.973  -0.486   8.543  50.644

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     53.87104    1.23108  43.759  < 2e-16 ***
A4CLSIZE                         0.02602    0.03937   0.661 0.508672
gender32                        -1.23055    0.28954  -4.250 2.16e-05 ***
```

```
race61                                4.00685    0.91691   4.370 1.26e-05 ***
race62                               -3.50036    0.97318  -3.597 0.000324 ***
race6Hispanic                        -0.64973    0.96664  -0.672 0.501504
race65                                0.57335    1.06843   0.537 0.591537
race6Indigenous or Native Americans -2.54669    1.20563  -2.112 0.034684 *
B4YRSTC                              -0.01218    0.01422  -0.857 0.391735
missing_B4                            5.09373    1.12512   4.527 6.05e-06 ***
w1sesl                                6.36594    0.20891  30.473  < 2e-16 ***
missing_ses                          -4.32749    0.54882  -7.885 3.49e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.12 on 9531 degrees of freedom
Multiple R-squared:  0.1792,    Adjusted R-squared:  0.1783
F-statistic: 189.2 on 11 and 9531 DF,  p-value: < 2.2e-16
```

- Estimates for Part 1: 0.02602

```r
# Part 2
model6.2 <- lm(A4CLSIZE ~ s4anumch + gender3 +
                 race6 + B4YRSTC + missing_B4 +
                 w1sesl + missing_ses, data = df_3)
D_hat <- predict(model6.2)
model6.2.2 <- lm(C4R2MSCL ~ D_hat + gender3 +
                 race6 + B4YRSTC + missing_B4 +
                 w1sesl + missing_ses, data = df_3)
summary(model6.2.2)
```

```
Call:
lm(formula = C4R2MSCL ~ D_hat + gender3 + race6 + B4YRSTC + missing_B4 +
    w1sesl + missing_ses, data = df_3)

Residuals:
    Min      1Q  Median      3Q     Max
-69.555  -8.991  -0.489   8.548  50.151

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         43.84331    4.18322  10.481  < 2e-16 ***
D_hat                                0.51529    0.19901   2.589 0.009630 **
gender32                            -1.31338    0.29132  -4.508 6.61e-06 ***
```

```
race61                                4.27756    0.92294   4.635 3.62e-06 ***
race62                               -3.52140    0.97290  -3.619 0.000297 ***
race6Hispanic                        -0.54245    0.96727  -0.561 0.574946
race65                                0.74851    1.07036   0.699 0.484379
race6Indigenous or Native Americans -1.98071    1.22618  -1.615 0.106269
B4YRSTC                              -0.02080    0.01462  -1.423 0.154860
missing_B4                            5.24374    1.12634   4.656 3.27e-06 ***
w1sesl                                6.16698    0.22340  27.606  < 2e-16 ***
missing_ses                          -4.37110    0.54891  -7.963 1.87e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.12 on 9531 degrees of freedom
Multiple R-squared:  0.1798,    Adjusted R-squared:  0.1788
F-statistic: 189.9 on 11 and 9531 DF,  p-value: < 2.2e-16
```

- Estimates for Part 2: 0.51529

6b. The conditions under which the results from the second analysis (IV estimation) would contain less bias than those from the first (simple regression) are:

1. Relevance: The instrument (Z) must be correlated with the treatment variable (D). That is, variations in public school enrollment should predict variations in class size.
2. Exogeneity: The instrument must be exogenous, meaning that it is not correlated with the error term in the outcome equation. This implies that the instrument affects the outcome (Y) only through the treatment variable (D), and not through any other unobserved pathways.
3. Exclusion Restriction: The instrument should not have a direct effect on the outcome variable (Y) that is not through the treatment variable (D).

If these conditions are met, then the IV estimate is consistent for the causal effect of the treatment on the outcome, whereas the simple regression estimate may be biased due to omitted variable bias or simultaneity.

Problem 7

7a. The pair of regression models can be written as:

$$Y_i = \beta_0 + \beta_1^{IV} * D_i + X_i\beta + \epsilon_i$$

$$D_i = \alpha_0 + \alpha_1 * Z_i + X_i\alpha + v_i$$

```r
model7 <- ivreg(C4R2MSCL ~ A4CLSIZE | (s4anumch + gender3 +
                  race6 + B4YRSTC + missing_B4 +
                  w1sesl + missing_ses) + gender3 +
                  race6 + B4YRSTC + missing_B4 +
                  w1sesl + missing_ses, data = df_3)
summary(model7)
```

```
Call:
ivreg(formula = C4R2MSCL ~ A4CLSIZE | (s4anumch + gender3 + race6 +
    B4YRSTC + missing_B4 + w1sesl + missing_ses) + gender3 +
    race6 + B4YRSTC + missing_B4 + w1sesl + missing_ses, data = df_3)

Residuals:
     Min       1Q   Median       3Q      Max
-64.2558 -10.7381  -0.8583  10.0212  55.6891

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.4068     4.1484   5.160 2.52e-07 ***
A4CLSIZE      1.6096     0.2033   7.919 2.67e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.56 on 9541 degrees of freedom
Multiple R-Squared: -0.1298,    Adjusted R-squared: -0.1299
Wald test: 62.71 on 1 and 9541 DF,  p-value: 2.666e-15
```

- sample estimate: 1.6096
- standard error: 0.2033
- t statistic: 7.919
- p value: 2.67e-15

7b.

```r
effect <- coef(model6.2.2)['D_hat']
cat('estimated average effect: \n')
```

```
estimated average effect:
```

```
effect*10
```

```
   D_hat
5.152941
```

```
cat('effect size: (scaled by std of D hat) \n')
```

```
effect size: (scaled by std of D hat)
```

```
effect / sd(D_hat) ##
```

```
    D_hat
0.6177531
```

Problem 8

8a.

- DID model can be written as:

$$Y_i = \alpha_0 + \alpha_1 T_i + \beta_0 G_i + \beta_1 G_i * T_i + \epsilon_i.$$

where Y_i stands for student i's math achievement; $T_i = 1$ stands for the end of Grade 1 while $T_i = 0$ stands for the end of kindergarten; $G_i = 1$ means that the student i is assigned to small classes while $G_i = 0$ means that the student i is assigned to regular classes.

- Identification Assumption:

$$E[Y_1(0)|G = 1] - E[Y_0(0)|G = 1] = E[Y_1(0)|G = 0] - E[Y_0(0)|G = 0]$$

- Explanation: Here's why DID can identify the Average Treatment Effect on the Treated (ATT) when certain assumptions are met:

1. Parallel Trends Assumption: The core assumption behind DID is that the treatment and control groups would have followed the same trend over time in the absence of the treatment. This implies that any difference in trends between the two groups after the treatment can be attributed to the treatment effect.
2. No Anticipation Effect: It is assumed that the subjects (students) did not change their behavior in anticipation of the treatment (being placed in a smaller class).

3. No Spillover Effects: The treatment applied to the treatment group should not affect the control group.

Given these assumptions, the DID analysis compares the change in outcomes over time for the treatment group to the change in outcomes for the control group. The coefficient β_1 on the interaction term is the ATT, which measures the average causal effect of the treatment on those who receive it (students in small classes).

```r
df_3$G <- df_3$A4CLSIZE
df_3$G[which(df_3$G < 19)] <- 1
df_3$G[which(df_3$G > 18)] <- 0

df_3_1 <- df_3[, c('G', 'C2R2MSCL')]
colnames(df_3_1)[2] <- 'Y'
df_3_1$T <- 0
df_3_2 <- df_3[, c('G', 'C4R2MSCL')]
colnames(df_3_2)[2] <- 'Y'
df_3_2$T <- 1
df_a <- rbind(df_3_1, df_3_2)

model8.1 <- lm(Y~G*T, data = df_a)
summary(model8.1)
```

```
Call:
lm(formula = Y ~ G * T, data = df_a)

Residuals:
    Min      1Q  Median      3Q     Max
-63.378  -8.709  -1.100   7.612  66.301

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.9198     0.1660 186.220   <2e-16 ***
G            -0.6912     0.3136  -2.204   0.0275 *
T            23.4585     0.2344 100.063   <2e-16 ***
G:T           0.1607     0.4428   0.363   0.7166
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.72 on 19021 degrees of freedom
  (61 observations deleted due to missingness)
```

```
Multiple R-squared:  0.4235,    Adjusted R-squared:  0.4234
F-statistic:  4658 on 3 and 19021 DF,  p-value: < 2.2e-16
```

8b.

- DID modified model can be written as:

$$Y_i = \alpha_0 + \alpha_1 T_i + \beta_0 G_i + \beta_1 G_i * T_i + \gamma X + \epsilon_i.$$

- Modified Identification Assumption:

$$E[Y_1(0)|G=1, X=x] - E[Y_0(0)|G=1, X=x] = E[Y_1(0)|G=0, X=x] - E[Y_0(0)|G=0, X=x]$$

```
sum(is.na(df_3$C1R2MSCL))
```

```
[1] 665
```

```
df_3$missing_C1 <- as.numeric(is.na(df_3$C1R2MSCL))
df_3$C1R2MSCL[is.na(df_3$C1R2MSCL)] <- mean(df_3$C1R2MSCL, na.rm = TRUE)
sum(is.na(df_3$C1R2MSCL))
```

```
[1] 0
```

```
df_3_1 <- df_3[, c('G', 'C2R2MSCL', 'gender3', 'race6', 'w1sesl',
                   'missing_ses', 'C1R2MSCL', 'missing_C1')]
colnames(df_3_1)[2] <- 'Y'
df_3_1$T <- 0
df_3_2 <- df_3[, c('G', 'C4R2MSCL', 'gender3', 'race6', 'w1sesl',
                   'missing_ses', 'C1R2MSCL', 'missing_C1')]
colnames(df_3_2)[2] <- 'Y'
df_3_2$T <- 1
df_b <- rbind(df_3_1, df_3_2)

model8.2 <- lm(Y~G*T + gender3 + race6 + w1sesl + missing_ses +
                 C1R2MSCL + missing_C1, data = df_b)
summary(model8.2)
```

```
Call:
lm(formula = Y ~ G * T + gender3 + race6 + w1sesl + missing_ses +
    C1R2MSCL + missing_C1, data = df_b)

Residuals:
    Min      1Q  Median      3Q     Max
-74.818  -5.674  -0.580   5.152  57.532

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       9.441285   0.481179  19.621  < 2e-16 ***
G                                 0.027500   0.218915   0.126 0.900036
T                                23.461419   0.163152 143.801  < 2e-16 ***
gender32                         -0.693354   0.138554  -5.004 5.66e-07 ***
race61                            1.504263   0.439765   3.421 0.000626 ***
race62                           -1.861012   0.466631  -3.988 6.68e-05 ***
race6Hispanic                     0.434513   0.463394   0.938 0.348424
race65                            2.846647   0.512680   5.552 2.85e-08 ***
race6Indigenous or Native Americans  0.041076   0.577405   0.071 0.943289
w1sesl                            1.616710   0.105991  15.253  < 2e-16 ***
missing_ses                      -1.243176   0.263494  -4.718 2.40e-06 ***
C1R2MSCL                          1.021844   0.008816 115.907  < 2e-16 ***
missing_C1                        0.695991   0.274619   2.534 0.011272 *
G:T                               0.190952   0.308151   0.620 0.535482
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.545 on 19011 degrees of freedom
  (61 observations deleted due to missingness)
Multiple R-squared:  0.721, Adjusted R-squared:  0.7208
F-statistic:  3778 on 13 and 19011 DF,  p-value: < 2.2e-16
```

8c.

```r
df_3_1 <- df_3[, c('G', 'C2R2MSCL', 'gender3', 'race6', 'w1sesl',
                   'missing_ses', 'C1R2MSCL', 'missing_C1', 's4anumch')]
colnames(df_3_1)[2] <- 'Y'
df_3_1$T <- 0
df_3_2 <- df_3[, c('G', 'C4R2MSCL', 'gender3', 'race6', 'w1sesl',
                   'missing_ses', 'C1R2MSCL', 'missing_C1', 's4anumch')]
colnames(df_3_2)[2] <- 'Y'
```

```r
df_3_2$T <- 1
df_c <- rbind(df_3_1, df_3_2)

model8.3 <- lm(Y~G*T + gender3 + race6 + w1sesl + missing_ses +
                 C1R2MSCL + missing_C1 + s4anumch, data = df_c)
summary(model8.3)
```

```
Call:
lm(formula = Y ~ G * T + gender3 + race6 + w1sesl + missing_ses +
    C1R2MSCL + missing_C1 + s4anumch, data = df_c)

Residuals:
    Min      1Q  Median      3Q     Max
-74.752  -5.679  -0.573   5.146  57.636

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        9.1665607  0.5043413  18.175  < 2e-16 ***
G                                  0.0771053  0.2205969   0.350 0.726695
T                                 23.4616742  0.1631419 143.811  < 2e-16 ***
gender32                          -0.6923903  0.1385466  -4.998 5.86e-07 ***
race61                             1.5111697  0.4397549   3.436 0.000591 ***
race62                            -1.8664105  0.4666120  -4.000 6.36e-05 ***
race6Hispanic                      0.3790732  0.4643693   0.816 0.414328
race65                             2.8144069  0.5129559   5.487 4.15e-08 ***
race6Indigenous or Native Americans  0.0329062  0.5773879   0.057 0.954553
w1sesl                             1.6069746  0.1061199  15.143  < 2e-16 ***
missing_ses                       -1.2446414  0.2634788  -4.724 2.33e-06 ***
C1R2MSCL                           1.0214824  0.0088177 115.844  < 2e-16 ***
missing_C1                         0.7267516  0.2751241   2.642 0.008260 **
s4anumch                           0.0005457  0.0003003   1.817 0.069199 .
G:T                                0.1905017  0.3081321   0.618 0.536420
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.545 on 19010 degrees of freedom
  (61 observations deleted due to missingness)
Multiple R-squared:  0.721, Adjusted R-squared:  0.7208
F-statistic:  3509 on 14 and 19010 DF,  p-value: < 2.2e-16
```

Including school enrollment (s4anumch) as a baseline covariate in a modified difference-in-

differences (DID) analysis could be justified if it meets certain conditions.

First, let's consider whether school enrollment may be related to both the treatment (class size) and the outcome (math achievement):

- **Relation to treatment**: School enrollment could determine class size; schools with higher enrollment may have larger classes if they do not proportionally increase the number of teachers or classrooms. Conversely, schools with lower enrollment may be more likely to have smaller classes.
- **Relation to outcome**: The size of the school can be an indicator of resources, student diversity, teacher qualifications, extracurricular offerings, and other factors that can influence student outcomes, including math achievement.

Given these considerations, school enrollment has the potential to be a confounding variable that is associated with both the treatment and the outcome. Including it as a covariate could therefore control for these additional sources of variation and make the groups more comparable in terms of the observed baseline characteristics. This, in turn, can help to fulfill the parallel trends assumption more credibly by accounting for the ways in which schools of different sizes might affect student achievement over time, irrespective of class size.

Additionally, if school enrollment is relatively stable over time and not influenced by the treatment (i.e., a school's enrollment is unlikely to change due to the assignment of students to small or regular classes within the time frame of the study), it can be considered a valid pre-treatment covariate to include in the DID analysis.

However, it's crucial to ensure that school enrollment is indeed a pre-treatment characteristic and not influenced by other factors in the post-treatment period. If school enrollment is determined after the treatment assignment or influenced by the treatment itself, its inclusion as a covariate would not be appropriate and could introduce bias.

Based on the given context and if the above conditions hold, including school enrollment as a covariate in the DID analysis could be justified to account for unobserved factors at the school level that could bias the estimated effect of class size on math achievement.

Bonus Question

Please turn to the new page.

Bonus Question A

$$\mathbb{E}[D_i(1) - D_i(0)] = \mathbb{E}[D_i(1)] - \mathbb{E}[D_i(0)]$$

$$= P(D_i(1)=1) \times 1 - P(D_i(0)=1) \times 1$$

$$\overset{\text{Monotonicity}}{=\!=} P(i \text{ is a complier or always-taker}) - P(i \text{ is an always-taker})$$

$$= P(i \text{ is a complier})$$

$$\mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))] = \mathbb{E}[Y_i(1,1) - Y_i(0,0) \mid D_i(1)=1, D_i(0)=0] \times P(D_i(1)=1, D_i(0)=0)$$

$\mathbb{E}[Y_i(z,1) - Y_i(z,0) \mid i \text{ is a complier}]$ — $P(i \text{ is a complier})$

$$+ \mathbb{E}[Y_i(1,1) - Y_i(0,1) \mid D_i(1) = D_i(0) = 1] \times P(D_i(1)=1, D_i(0)=1)$$

$\mathbb{E}[Y_i(1,d) - Y_i(0,d) \mid d=1]$ — $P(i \text{ is an always-taker})$

$$+ \mathbb{E}[Y_i(1,0) - Y_i(0,0) \mid D_i(1) = D_i(0) = 0] \times P(D_i(1)=0, D_i(0)=0)$$

$\mathbb{E}[Y_i(1,d) - Y_i(0,d) \mid d=0]$ — $P(i \text{ is a never-taker})$

$$+ \mathbb{E}[Y_i(1,0) - Y_i(0,1) \mid D_i(1)=0, D_i(0)=1] \times P(D_i(1)=0, D_i(0)=1)$$

$P(i \text{ is a defier}) \overset{\text{Monotonicity}}{=\!=} 0$

$$= \mathbb{E}[G_i \mid i \text{ is a complier}] \times P(i \text{ is a complier})$$

$$+ \mathbb{E}[H_i \mid d=1] \times P(d=1) + \mathbb{E}[H_i \mid d=0] \times P(d=0)$$

$$= \mathbb{E}[G_i \mid i \text{ is a complier}] \times P(i \text{ is a complier}) + \mathbb{E}[H_i]$$

$$\Rightarrow \frac{\mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]}{\mathbb{E}[D_i(1) - D_i(0)]} = \mathbb{E}[G_i \mid i \text{ is a complier}] + \frac{\mathbb{E}[H_i]}{P(i \text{ is a complier})}$$

Bias for the violation of Exclusion Restriction.

Bonus Question B

$$\mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))] = \mathbb{E}[Y_i(d=1) - Y_i(d=0) \mid i \text{ is a complier}] \times P(i \text{ is a complier})$$

$$+ \mathbb{E}[Y_i(1,d) - Y_i(0,d) \mid i \text{ is an always-taker}] \times P(i \text{ is an always-taker})$$

$$+ \mathbb{E}[Y_i(1,d) - Y_i(0,d) \mid i \text{ is a never-taker}] \times P(i \text{ is a never-taker})$$

Under exclusion restriction $= 0$

$$+ \mathbb{E}[Y_i(d=0) - Y_i(d=1) \mid i \text{ is a defier}] \times P(i \text{ is a defier})$$

$$\mathbb{E}[D_i(1) - D_i(0)] = \mathbb{E}[D_i(1)] - \mathbb{E}[D_i(0)]$$

$$= P(D_i(1)=1) \times 1 - P(D_i(0)=1) \times 1$$

$$= P(i \text{ is a complier or always-taker}) - P(i \text{ is a defier or always-taker})$$

$$= P(i \text{ is a complier}) - P(i \text{ is a defier})$$

$$\frac{\mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]}{\mathbb{E}[D_i(1) - D_i(0)]} - \mathbb{E}[Y_i(1) - Y_i(0) \mid i \text{ is a complier}]$$

$$= \mathbb{E}[Y_i(1) - Y_i(0) \mid i \text{ is a complier}] \times \left(\frac{P(i \text{ is a complier})}{P(i \text{ is a complier}) - P(i \text{ is a defier})} - 1 \right)$$

$$+ \mathbb{E}[Y_i(0) - Y_i(1) \mid i \text{ is a defier}] \times \frac{P(i \text{ is a defier})}{P(i \text{ is a complier}) - P(i \text{ is a defier})}$$

$$= \underbrace{\frac{-P(i \text{ is a defier})}{P(i \text{ is a complier}) - P(i \text{ is a defier})}}_{\lambda} \times \underbrace{\{\mathbb{E}[Y_i(1) - Y_i(0) \mid i \text{ is a defier}] - \mathbb{E}[Y_i(1) - Y_i(0) \mid i \text{ is a complier}]}$$

Bias for violation of monotonicity.