STAT 31900: Homework 2
Propensity Score-Based Methods

Robert Winter

Table of Contents

Throughout this document, ***bolded*** *language signifies the problem set's instructions. My responses are provided in non-bold text.*

1 Introduction

1.1 Research Question

In this assignment, you will apply the propensity score-based methods to quasi-experimental data in an evaluation of the effect of class size reduction in Grade 1 on math achievement by the end of Grade 1.You will use the nationally representative Early Childhood Longitudinal Study–Kindergarten cohort (ECLS-K) 1998 public-use data in the analysis. Please use the data file named `CHDV 30102_ECLSK98_class size.dta`.

1.2 Data

The treatment variable is Grade 1 class size (`A4CLSIZE`), which needs to be dichotomized. For this assignment, we consider a class of no more than 18 students (i.e., ≤ 18) to be relatively "small" and a class of 19 or more students to be "regular." The outcome is Grade 1 math achievement score (`C4R2MSCL`). Please exclude from your analysis any student who has missing information in the treatment indicator or the outcome.

The observed pretreatment covariates include:

- **student gender (you may create a three-category variable `gender3` by replacing `-9` with `3` for the missing category),**
- **student race (you may create a six-category variable `race6` by combining:**
  - **`3 hispanic, race specified` and `4 hispanic, race not specified` into one category named `Hispanic`,**
  - **`6 native Hawaiian, other pacific islander` and `7 American Indian or Alaska native` into one category named `Indigenous or Native Americans`, and**
  - **combining `-9 not ascertained` and `8 more than one race, non-hispanic` into one category named `Other Races`,**
- **reading score in fall kindergarten (`C1RRSCAL`),**
- **reading score in spring kindergarten (`C2RRSCAL`),**
- **math score in fall kindergarten (`C1R2MSCL`),**
- **math score in spring kindergarten (`C2R2MSCL`), and**
- **Grade 1 teacher's teaching experience in years (`B4YRSTC`).**

Note that `-1` and `-9` are missing values. As you did in Assignment 1, please create a missing indicator for the missing cases and then use the sample mean to replace the missing values in a continuous covariate (i.e., mean imputation along with a missing indicator).

This assignment has eight sets of questions in total.

2 Analysis

2.1 Question 1: Descriptive analysis

What is the mean difference in Grade 1 math achievement (C4R2MSCL) between students attending small classes (i.e., the treated) and those attending regular classes (i.e., the control)? Also report the standard error and the hypothesis testing result. How large is the effect size (as before, using the standard deviation of the control group students' math outcome as the scaling unit)?

We perform a simple linear regression of Grade 1 math scores (`C4R2MSCL`) on Grade 1 class size, treated as a categorical variable, with categories for small and regular classes. For this analysis, we also cluster standard errors by school, as students who attended the same school—and thus lived in the same community—likely had correlated scores.

This analysis gives an estimated population average treatment effect of approximately $\hat{\delta}_{PF} = -0.096$. Intuitively, this means that, on average, enrollment in a small Grade 1 classroom was associated with a 0.096-unit *decrease* in Grade 1 math scores compared to students enrolled in regular-sized Grade 1 classrooms. However, this estimate is *not* statistically significant. The standard error of our estimate of δ_{PF} is approximately 0.527, which conveys considerable uncertainty. Indeed, assuming that Grade 1 math score conditional on Grade 1 class size is normally distributed (a common assumption in regression analysis), our standard error of 0.527 means that we can be ~95% confident that the true change in Grade 1 math score associated with being in a small Grade 1 classroom is somewhere in the interval $(-0.096 - 2 \times 0.527, -0.096 + 2 \times 0.527) \approx (-1.132, 0.939)$. The fact that this interval contains 0 means that we cannot be confident that there even *is* a nonzero association between Grade 1 classroom size and Grade 1 math score. This is also reflected in the model's hypothesis testing result, wherein our null hypothesis is that Grade 1 classroom size is not associated with Grade 1 math outcomes. The p-value on our estimate of δ_{PF} is approximately 0.855, which means that there is a .855 probability of observing a disparity in Grade 1 math scores between students in regular-sized and small Grade 1 classrooms that is greater than the disparity we actually observed. In other words, the data we observed is not compelling enough to make us reject the null hypothesis.

We estimate the effect size using Glass's delta statistic, calculated as the difference in mean outcomes between the treatment and control groups divided by the standard deviation of the control group's outcomes. Here, Glass's delta equals $-0.006 \approx 0$, which tells us that the mean Grade 1 math score among students in small Grade 1 classrooms is barely more than zero standard deviations lower than the mean score among students in regular-sized Grade 1 classrooms. A 95% confidence interval for this effect size (calculated using the pivot method) is $(-0.04, 0.03)$. That this interval contains 0 means that at the $\alpha = 0.05$ significance level, Glass's delta is not significantly different from 0.

Factor	Coefficient Estimate	Standard Error	95% CI for Coef. Est.	p-value	Glass's Delta	95% CI for Glass's Delta
Treatment	-0.096	0.527	$(-1.132, 0.939)$	0.855	-0.006	$(-0.04, 0.03)$

In summary, we conclude that there is not statistical evidence that Grade 1 math outcomes are associated with Grade 1 class size.

2.2 Question 2: Potential confounders

2a. What are some major pre-existing differences between students in small classes and those in regular classes? Show your evidence (you may use `CreateTableOne()` from the `tableone` package in R).

As summarized in the table below, there are a number of statistically significant preexisting differences between students in small classes and those in regular classes in this study. In particular:

- There are significant ($p < 0.001$) differences in the racial/ethnic compositions of the two groups (`race6`). Small classrooms, on average, had higher proportions of Black and white students and smaller proportions of Asian, Hispanic, Indigenous or Native American students, as well as students of other or unidentified heritage, than regular-sized classrooms.
- There is a marginally significant ($p = 0.058$) difference in the gender composition of the two groups (`gender3`). Small classrooms, on average, were 47.9% female, while regular-sized classrooms, on average, were 49.7% female.
- Teachers in small classrooms had statistically significantly ($p < 0.001$) fewer years of experience (`B4YRSTC_impt`) than their counterparts in regular-sized classrooms, at least when restricting to teachers with non-missing years of experience data. Although the difference between the groups is statistically significant, the difference may not be substantively meaningful, as the difference in average years of experience is only $14.63 - 13.79 = 0.84$ years. Notably, there was not a significant difference ($p = 0.501$) in the proportion of teachers with missing years of experience data between the two groups.
- Students in small classrooms had statistically significantly ($p < 0.001$) lower reading scores as of both kindergarten fall (`C1RRSCAL_impt`) and kindergarten spring (`C2RRSCAL_impt`) than their peers in regular-sized classrooms, at least when restricting to students with non-missing scores. Once again, while these differences are statistically significant, they may not be substantively meaningful: as of kindergarten fall, the difference in average reading scores was only $23.70 - 23.14 = 0.56$ points, while as of kindergarten spring, the difference was only $34.04 - 33.28 = 0.76$ points. Notably, there were not significant differences in the proportion of students with

missing reading scores between the two groups, whether during kindergarten fall ($p = 0.509$) or during kindergarten spring ($p = 0.084$).

- Students in small classrooms had statistically significantly ($p = 0.005$) lower math scores as of kindergarten fall (`C1R2MSCL_impt`) than their peers in regular-sized classrooms, at least when restricting to students with non-missing scores. Once again, while these differences are statistically significant, they may not be substantively meaningful, as the difference in average scores is only $22.19 - 21.74 = 0.45$ points. Notably, there was not a significant difference in the proportion of students with missing math scores between the two groups as of kindergarten fall ($p = 0.185$). Moreover, there was not a significant difference in math scores between the two groups as of kindergarten spring ($p = 0.127$), nor was there a significant difference in the proportion of students with missing math scores as of kindergarten spring ($p = 0.187$).

In summary, on average, small Grade 1 classrooms in this study had less experienced teachers and consisted of students with lower past scores in reading and math than regular-sized Grade 1 classrooms. The two classroom sizes also had, on average, distinct racial/ethnic compositions, and marginally distinct gender compositions, which may be correlated with other socioeconomic factors that affect student outcomes.

```
                         Stratified by trt
                          0             1             p      test
n                          9583          3775
race6 (%)                                             <0.001
   White                   5438 (56.7)   2364 (62.6)
   Asian                    662 ( 6.9)    150 ( 4.0)
   Black                   1275 (13.3)    528 (14.0)
   Hispanic                1646 (17.2)    560 (14.8)
   Indigenous               300 ( 3.1)     97 ( 2.6)
   Other                    262 ( 2.7)     76 ( 2.0)
gender3 = 2 (%)            4766 (49.7)   1808 (47.9)   0.058
B4YRSTC_impt (mean (SD))  14.63 (10.15) 13.79 (9.67)  <0.001
B4YRSTC_miss = 1 (%)        218 ( 2.3)     78 ( 2.1)   0.501
C1RRSCAL_impt (mean (SD)) 23.70 (8.13)  23.14 (7.94)  <0.001
C1RRSCAL_miss = 1 (%)      1431 (14.9)    546 (14.5)   0.509
C2RRSCAL_impt (mean (SD)) 34.04 (10.50) 33.28 (10.61) <0.001
C2RRSCAL_miss = 1 (%)       492 ( 5.1)    166 ( 4.4)   0.084
C1R2MSCL_impt (mean (SD)) 22.19 (8.45)  21.74 (8.41)   0.005
C1R2MSCL_miss = 1 (%)       955 (10.0)    406 (10.8)   0.185
C2R2MSCL_impt (mean (SD)) 32.40 (11.25) 32.06 (11.33)  0.127
C2R2MSCL_miss = 1 (%)       156 ( 1.6)     49 ( 1.3)   0.187
```

2b. Which group (i.e., the treated or the control) seems to be relatively advantaged, that is, would likely have a higher math score on average even if

all students would have attended a small class or would have attended a regular class in Grade 1? Please provide your reasoning on the basis of empirical information.

To determine whether the pre-existing differences between the treated and control groups described above might advantage one group over the other, we perform a series of linear regressions, each measuring the relationship between Grade 1 math outcomes and the variable (or pair of variables) of interest. In each regression, we pool together all students, regardless of classroom size, and we use standard errors clustered at the school level. The results of these analyses are summarized as follows:

1. We examine the relationship between Grade 1 math outcomes and racial/ethnic heritage: $C4R2MSCL_i = \beta_0 + \beta_1 \cdot Asian_i + \beta_2 \cdot Black_i + \beta_3 \cdot Hispanic_i + \beta_4 \cdot Indigenous_i + \beta_5 \cdot Other_i + \varepsilon_i$, where "white" is treated as the reference category. Every non-white demographic has average Grade 1 math scores that are statistically significantly ($p < 0.001$) lower than that of their white peers—likely a reflection of socioeconomic differences between white and non-white families that result in differences in learning outcomes. Since the treated group consists of disproportionately many white students (62.6% vs 56.7%), racial/ethnic differences appear to advantage the treated group.
2. We examine the relationship between Grade 1 math outcomes and gender: $C4R2MSCL_i = \beta_0 + \beta_1 \cdot Female_i + \varepsilon_i$. We recover a statistically significant ($p < 0.001$) estimate of $\widehat{\beta}_1 = -1.667$, meaning that, on average, female students' Grade 1 math scores were 1.667 points lower than their male counterparts.
3. We examine the relationship between Grade 1 math outcomes and teacher years of experience, including a dummy variable for teachers with missing data: $C4R2MSCL_i = \beta_0 + \beta_1 \cdot B4YRSTC\_imputed_i + \beta_2 \cdot B4YRSTC\_missing_i + \varepsilon_i$. We recover a statistically significant ($p = 0.020$) estimate of $\widehat{\beta}_1 = 0.049$, meaning that, on average, a one-year increase in a teacher's experience is associated with a 0.049-point increase in their students' Grade 1 math scores. Since the control group's teachers averaged 0.84 more years of experience than the treated group's teachers, this advantages students in the control group. There was not a significant association between missing teacher data and student scores.
4. We examine the relationship between Grade 1 math outcomes and kindergarten fall reading scores, including a dummy variable for students with missing scores: $C4R2MSCL_i = \beta_0 + \beta_1 \cdot C1RRSCAL\_imputed_i + \beta_2 \cdot C1RRSCAL\_missing_i + \varepsilon_i$. We recover a statistically significant ($p < 0.001$) estimate of $\widehat{\beta}_1 = 0.968$, meaning that, on average, a one-point increase in a student's kindergarten fall reading score is associated with a 0.968-point increase in her Grade 1 math score. Since the control group's students averaged 0.56-point higher kindergarten fall reading scores than the treated group's students, this advantages the control group. Moreover, we recover a statistically significant ($p < 0.001$) estimate of $\widehat{\beta}_2 = -4.487$, meaning that, on average, having a missing kindergarten fall reading score is associated with a 4.487-point decrease in Grade 1 math score. However, since there was not a significant difference

in the prevalence of students with missing kindergarten fall reading scores across the two groups, this does not systematically advantage one group over the other.

5. We examine the relationship between Grade 1 math outcomes and kindergarten spring reading scores, including a dummy variable for students with missing scores: $C4R2MSCL_i = \beta_0 + \beta_1 \cdot C2RRSCAL\_imputed_i + \beta_2 \cdot C2RRSCAL\_missing_i + \varepsilon_i$. We recover a statistically significant ($p < 0.001$) estimate of $\widehat{\beta}_1 = 0.822$, meaning that, on average, a one-point increase in a student's kindergarten spring reading score is associated with a 0.822-point increase in her Grade 1 math score. Since the control group's students averaged 0.76-point higher kindergarten spring reading scores than the treated group's students, this advantages the control group. Moreover, we recover a statistically significant ($p < 0.001$) estimate of $\widehat{\beta}_2 = -10.652$, meaning that, on average, having a missing kindergarten spring reading score is associated with a 10.652-point decrease in Grade 1 math score. However, since there was not a significant difference in the prevalence of students with missing kindergarten spring reading scores across the two groups, this does not systematically advantage one group over the other.

6. We examine the relationship between Grade 1 math outcomes and kindergarten fall math scores, including a dummy variable for students with missing scores: $C4R2MSCL_i = \beta_0 + \beta_1 \cdot C1R2MSCL\_imputed_i + \beta_2 \cdot C1R2MSCL\_missing_i + \varepsilon_i$. We recover a statistically significant ($p < 0.001$) estimate of $\widehat{\beta}_1 = 1.237$, meaning that, on average, a one-point increase in a student's kindergarten fall math score is associated with a 1.237-point increase in her Grade 1 math score. Since the control group's students averaged 0.45-point higher kindergarten fall math scores than the treated group's students, this advantages the control group. There was not a significant association between missing kindergarten math score data and Grade 1 math scores.

7. We examine the relationship between Grade 1 math outcomes and kindergarten spring math scores, including a dummy variable for students with missing scores: $C4R2MSCL_i = \beta_0 + \beta_1 \cdot C2R2MSCL\_imputed_i + \beta_2 \cdot C2R2MSCL\_missing_i + \varepsilon_i$. We recover a statistically significant ($p < 0.001$) estimate of $\widehat{\beta}_1 = 1.041$, meaning that, on average, a one-point increase in a student's kindergarten spring math score is associated with a 1.041-point increase in her Grade 1 math score. However, since there was not a significant difference in kindergarten spring math scores across the two groups (at least among students with non-missing data), this does not systematically advantage one group over the other. We also recover a statistically significant ($p < 0.001$) estimate of $\widehat{\beta}_2 = -7.975$, meaning that, on average, having a missing kindergarten spring math score is associated with a 7.975-point decrease in Grade 1 math score. Once again, however, since there was not a significant difference in the prevalence of students with missing kindergarten spring math scores across the two groups, this does not systematically advantage one group over the other.

In summary, the control group consists of students with higher average kindergarten scores than the treated group, as well as more experienced teachers than the treated group. We have shown that both of these factors are associated with higher Grade 1 math scores. As such, it appears that, on balance, the control group is advantaged compared to the treated group: students in the control group would likely have obtained higher Grade 1 math scores regardless of their classroom size. This may be somewhat offset by racial/ethnic and gender differences between the treated and control groups, as the treated group contains disproportionate numbers of higher-scoring demographics, including male students and white students.

2.3 Question 3: Propensity score and common support

Use logistic regression to estimate every student's propensity of attending a small class.

3a. Explain how you decide whether it is necessary to include a quadratic term or other nonlinear forms of a continuous covariate or an interaction between two covariates. Write down your logistic regression model. Report the estimated coefficients and their standard errors in a table. Save the logit propensity score in the same dataset.

First, we considered a logistic regression model that regressed treatment on race, gender, teacher years of experience, reading score in fall kindergarten, reading score in spring kindergarten, math score in fall kindergarten, and math score in spring kindergarten, as well as missing data indicators for each of these factors ("Model 1"). Using a likelihood ratio test, the results of this model were not statistically different from those of a model with all the same regressors except for gender ("Model 2"), so we move forward without gender as a regressor. Inspired by the idea that more experienced teachers may be able to better support students of various racial/ethnic backgrounds, we then ran a regression that expanded Model 2 by including interactions of teacher years of experience (and an indicator of missing years of experience data) with each racial/ethnic category ("Model 3"). The results of Model 3 were statistically different from Model 2, so we moved forward with the more complex model. Finally, inspired by the idea that the relationship between teachers' years of experience and students' outcomes may be nonlinear (e.g., more experienced teachers are able to help their students learn in a superlinear way), we expanded Model 3 by including the square of teacher years of experience as a regressor ("Model 4"). The results of Model 4 were statistically different from Model 3, so we move forward with Model 4 as our final model.

That is, our final model is:

$$
\begin{aligned}
\eta_i = {} & \beta_0 + \beta_1 \cdot Asian_i + \beta_2 \cdot Black_i + \beta_3 \cdot Hispanic_i + \beta_4 \cdot Indigenous_i + \beta_5 \cdot OtherRace_i \\
& + \beta_6 \cdot B4YRSTC\_impt_i + \beta_7 \cdot B4YRSTC\_impt_i \times Asian_i \\
& + \beta_8 \cdot B4YRSTC\_impt_i \times Black_i + \beta_9 \cdot B4YRSTC\_impt_i \times Hispanic_i \\
& + \beta_{10} \cdot B4YRSTC\_impt_i \times Indigenous_i + \beta_{11} \cdot B4YRSTC\_impt_i \times OtherRace_i \\
& + \beta_{12} \cdot B4YRSTC\_impt_i^2 + \beta_{13} \cdot B4YRSTC\_miss_i \\
& + \beta_{14} \cdot B4YRSTC\_miss_i \times Asian_i + \beta_{15} \cdot B4YRSTC\_miss_i \times Black_i \\
& + \beta_{16} \cdot B4YRSTC\_miss_i \times Hispanic_i + \beta_{17} \cdot B4YRSTC\_miss_i \times Indigenous_i \\
& + \beta_{18} \cdot B4YRSTC\_miss_i \times OtherRace_i + \beta_{19} \cdot C1RRSCAL\_impt_i \\
& + \beta_{20} \cdot C1RRSCAL\_miss_i + \beta_{21} \cdot C2RRSCAL\_impt_i + \beta_{22} \cdot C2RRSCAL\_miss_i \\
& + \beta_{23} \cdot C1R2MSCL\_impt_i + \beta_{24} \cdot C1R2MSCL\_miss_i + \beta_{25} \cdot C2R2MSCL\_impt_i \\
& + \beta_{26} \cdot C2R2MSCL\_miss_i
\end{aligned}
$$

where η_i is individual i's log odds of being treated; where each "*_impt*" indicates that we have imputed missing data for the corresponding variable using the mean of the individuals with observed data, and each "*_miss*" is an indicator for whether the corresponding variable's data is missing; and where each non-racial/ethnic variable is named according to its name in the dataset.

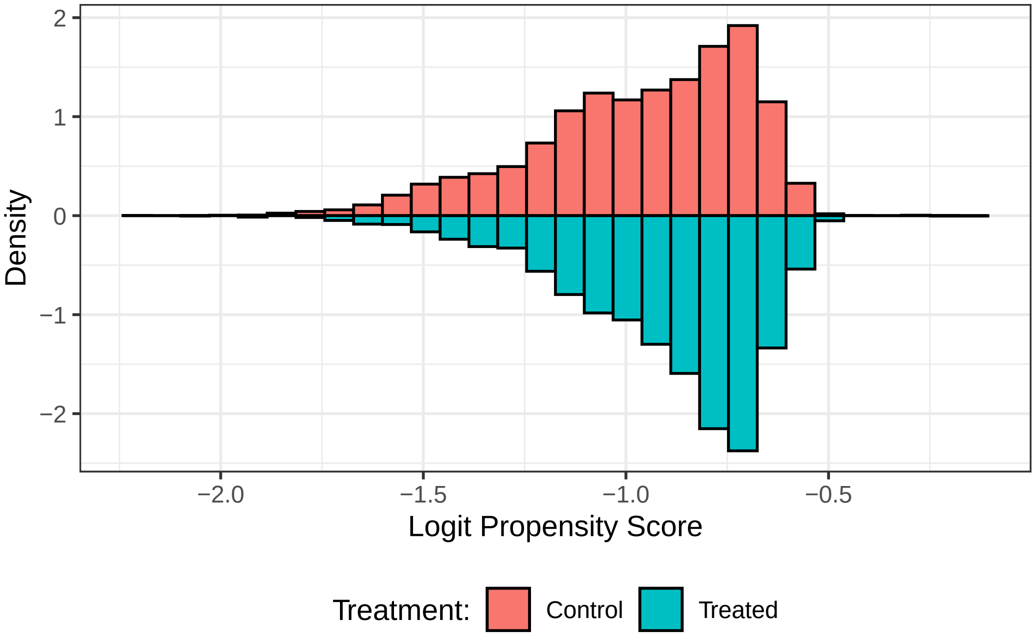
Propensity Score Model Results

Variable	Estimate	Std. Error
(Intercept)	−0.480	0.094
race6Asian	−0.902	0.173
race6Black	−0.009	0.096
race6Hispanic	−0.473	0.102
race6Indigenous	−0.283	0.199
race6Other	−0.638	0.244
B4YRSTC_impt	0.012	0.008
B4YRSTC_miss	−0.444	0.218
I(B4YRSTC_impt^2)	−0.001	0.000
C1RRSCAL_impt	0.000	0.004
C1RRSCAL_miss	−0.407	0.156
C2RRSCAL_impt	−0.007	0.003
C2RRSCAL_miss	0.365	0.169
C1R2MSCL_impt	−0.010	0.004
C1R2MSCL_miss	0.532	0.161
C2R2MSCL_impt	0.005	0.003
C2R2MSCL_miss	−0.408	0.234
race6Asian:B4YRSTC_impt	0.015	0.009
race6Black:B4YRSTC_impt	−0.007	0.006
race6Hispanic:B4YRSTC_impt	0.015	0.006

race6Indigenous:B4YRSTC_impt	−0.008	0.013
race6Other:B4YRSTC_impt	0.014	0.013
race6Asian:B4YRSTC_miss	0.677	0.417
race6Black:B4YRSTC_miss	0.424	0.366
race6Hispanic:B4YRSTC_miss	0.747	0.366
race6Indigenous:B4YRSTC_miss	0.840	1.251
race6Other:B4YRSTC_miss	0.732	0.724

3b. Compare the distribution of the logit propensity score between students in small classes and those in regular classes (a) by displaying the histograms for these two groups (you may use the `teffects overlap` in Stata or `histbackback()` in the `Hmisc` package in R) and (b) by examining the between-group differences in the mean and variance of the logit propensity score (you may use `tabstat` in Stata or the `group_by()` and `summarize()` functions in the `dplyr` package in R). To identify the region of common support (remember to allow for a caliper equal to 20% of a standard deviation of the logit propensity score), please report if there are any extreme cases in one treatment group that have no counterparts in the other group and therefore should be excluded from the analytic sample.

First, we display histograms of the logit propensity scores for the treated and control groups below. As shown, there appears to be a clear common support of logit propensity scores between roughly −1.75 and −0.5, though there are individuals with propensity scores above and below these levels who may not have counterparts in the opposite treatment group.

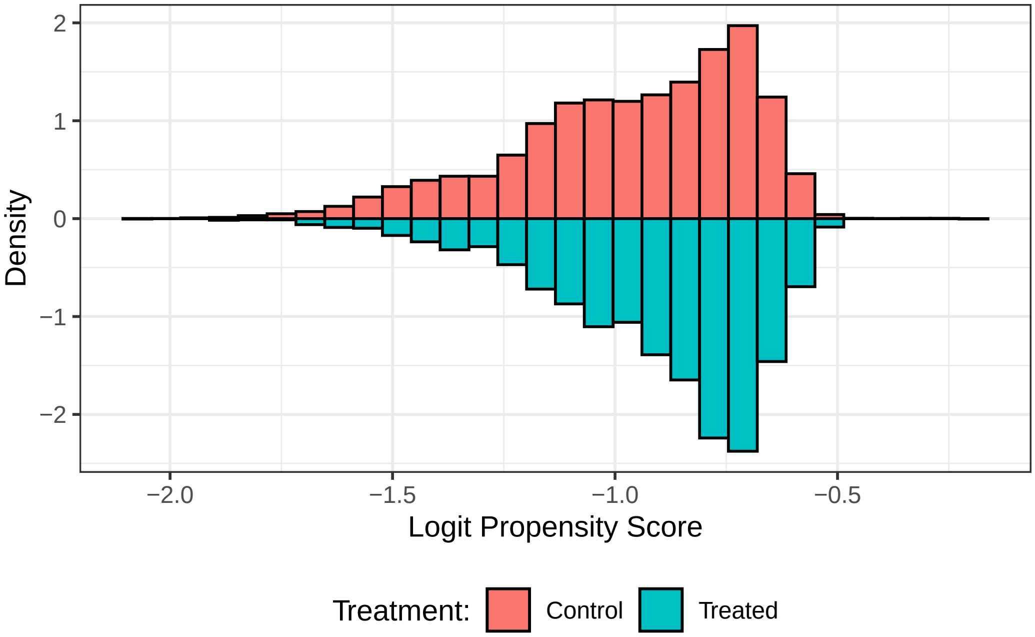


Next, we compare between-group differences in the mean and variance of logit propensity score by treatment group. As shown below, the two groups have fairly similar means: the average logit propensity score among students in regular-sized classrooms is approximately −0.964, while the average propensity score among students in small classrooms is approximately −0.902. The spread of scores in each group is also very similar: the variance of propensity scores among students in regular-sized classrooms is approximately 0.069, while the variance of propensity scores among students is approximately 0.056.

Logit Propensity Scores by Treatment Group

Treatment	Mean	Variance
0	−0.964	0.069
1	−0.902	0.056

To form the final dataset on which we perform analysis, we exclude all students whose propensity score was more than 20% of the propensity score's standard deviation above or below the common range of propensity scores across the two treatment groups. Fortunately, we lose very few students from our analysis pool: Student 0685001C (logit propensity score −2.219) and Student 0913013C (logit propensity score −0.156). As shown below, the distributions of logit propensity scores among the students in our final analytic sample are all but unchanged.



2.4 Question 4: Propensity score matching

4a. Decide whether you plan to estimate the population average treatment effect on the treated (ATT) or the population average treatment effect (ATE) and explain why.

We use propensity score matching to estimate the average treatment effect on the treated (ATT), since we are interested in determining the causal effect of being in a small classroom for students that actually were in (or actually could have been in) such classrooms. Estimating the ATE also imposes the additional requirement of needing to find matches for students in the control group, which may further narrow the number of matches that can be made.

4b. Apply propensity score matching in Stata or R (use `teffects psmatch` in Stata to conduct one-to-one matching with replacement which is the default option and takes into account that the propensity score is estimated in computing the standard error for the treatment effect estimator. R users may try the combination of `cobalt` and `MatchIt` packages or use `ps.match()`).

Below, we use the `matchit` function from the `MatchIt` package to conduct one-to-one matching with replacement:

```
A matchit object
 - method: 1:1 nearest neighbor matching with replacement
 - distance: Propensity score
             - estimated with logistic regression
 - number of obs.: 13356 (original), 6650 (matched)
 - target estimand: ATT
 - covariates: race6, B4YRSTC_impt, B4YRSTC_miss, I(B4YRSTC_impt^2), C1RRSCAL_impt, C1RRSCAL
```

4c. Check balance within the matched pairs in the logit propensity score. Also check balance in the pretreatment covariates (you may use the `tebalance summarize` command in Stata; in R, you may use `bal.tab()` from the `cobalt` package or `ps.match()` followed by `ps.balance()`). In each case, compare the standardized difference (you may again use the standard deviation of the control group students' math outcome as the scaling unit) and the variance ratio before and after matching. Summarize the results in a table.

First, we check balance in logit propensity scores and pretreatment covariates *before* weighting, which are summarized in the output below. As shown, the balance is not great. The standardized difference in logit propensity scores is 0.247, while the variance ratio is 0.820, each of which are just barely within their respective desired thresholds of 0.25 and $[\frac{4}{5}, \frac{5}{4}]$. The balance in the covariates is not great, either. On the one hand, the maximum standardized difference in a pretreatment covariate is 0.129—for the Asian indicator variable—which is not awful. However, there are several pretreatment covariates with variance ratios below $\frac{4}{5}$, including the interactions of each of the Asian, Indigenous, and Other Race indicators with (imputed, if missing) teacher years of experience. This imbalance illustrates the need for a propensity-score based procedure such as IPTW weighting.

Now, we check balance in logit propensity scores and pretreatment covariates *after* matching, again summarized in the output below. As shown, we have achieved excellent balance in the logit propensity scores ("`distance`") of our matched pairs. The standardized difference in logit propensity scores is $4.51 \times 10^{-5} \approx 0$, and the variance ratio is $1.0002 \approx 1$. Our matching procedure has very closely emulated random assignment to the treated and control conditions.

We have also achieved very good balance in each of our pretreatment covariates. The largest standardized difference for any pretreatment covariate was 0.04—for gender—which is well below the desired 0.25 threshold. And while the `cobalt` package's `bal.tab()` function does not compute variance ratios for binary variables, for continuous variables, variance ratios are all close to 1. Indeed, the smallest variance ratio for any pretreatment covariate was 0.993—for the square of teachers' years of experience—while the largest variance ratio was 1.09—for (imputed, in the case of missing data) kindergarten fall math scores. Thus, all variance ratios are well within the desired$[\frac{4}{5}, \frac{5}{4}]$ range.

Balance Checks: Post-Matching

Variable	Type	Std. Diff.	Var. Ratio
distance	Distance	0.000	1.000
race6_White	Binary	0.003	NA
race6_Asian	Binary	−0.002	NA
race6_Black	Binary	−0.027	NA
race6_Hispanic	Binary	0.017	NA
race6_Indigenous	Binary	−0.008	NA
race6_Other	Binary	0.023	NA
B4YRSTC_impt	Contin.	−0.020	1.017
B4YRSTC_miss	Binary	−0.002	NA
I(B4YRSTC_impt^2)	Contin.	−0.012	0.993
C1RRSCAL_impt	Contin.	−0.020	1.015
C1RRSCAL_miss	Binary	0.028	NA
C2RRSCAL_impt	Contin.	−0.004	1.086
C2RRSCAL_miss	Binary	0.020	NA
C1R2MSCL_impt	Contin.	−0.010	1.090
C1R2MSCL_miss	Binary	0.016	NA
C2R2MSCL_impt	Contin.	0.006	1.020
C2R2MSCL_miss	Binary	0.004	NA
gender3_2	Binary	−0.040	NA

4d. Use the matched sample to estimate the effect of class size type in Grade 1 math achievement. Report the standard error, the hypothesis testing result, and the effect size. What is your conclusion with regard to the class size effect?

Using our matched data, we now regress Grade 1 math scores (`C4R2MSCL`) on treatment (small or regular class size). As in Question 2, we cluster standard errors by school.

We estimate an ATT of approximately $\hat{\delta}_{ATT} = -0.153$. Intuitively, this means that, on average, enrollment in a small Grade 1 classroom caused a 0.153-unit *decrease* in Grade 1 math scores compared to students enrolled in regular-sized Grade 1 classrooms. However, this estimate is *not* statistically significant. The standard error of our estimate of δ_{ATT}

is approximately 0.570, which conveys considerable uncertainty. Indeed, assuming that Grade 1 math score conditional on Grade 1 class size is normally distributed (a common assumption in regression analysis), our standard error of 0.570 means that we can be ~95% confident that the true change in Grade 1 math score caused by being in a small Grade 1 classroom rather than a regular-sized Grade 1 classroom is somewhere in the interval $(-0.153 - 2 \times 0.570, -0.153 + 2 \times 0.570) \approx (-1.272, 0.966)$. The fact that this interval contains 0 means that we cannot be confident that there even *is* a nonzero relationship between Grade 1 classroom size and Grade 1 math score. This is also reflected in the model's hypothesis testing result, wherein our null hypothesis is that Grade 1 classroom size is not associated with Grade 1 math outcomes. The p-value on our estimate of δ_{ATT} is approximately 0.788, which means that there is a 0.788 probability of observing a disparity in Grade 1 math scores between students in regular-sized and small Grade 1 classrooms that is greater than the disparity we actually observed. In other words, the data we observed is not compelling enough to make us reject the null hypothesis.

We again estimate the effect size using Glass's delta statistic. Here, Glass's delta equals $-0.010 \approx 0$, which tells us that the mean Grade 1 math score among students in small Grade 1 classrooms is barely more than zero standard deviations lower than the mean score among students in regular-sized Grade 1 classrooms. A 95% confidence interval for this effect size (calculated using the pivot method) is $(-0.06, 0.04)$. That this interval contains 0 means that at the $\alpha = 0.05$ significance level, Glass's delta is not significantly different from 0.

Factor	Coefficient Estimate	Standard Error	95% CI for Coef. Est.	p-value	Glass's Delta	95% CI for Glass's Delta
Treatment	−0.153	0.570	$(-1.272, 0.966)$	0.788	−0.010	$(-0.06, 0.04)$

In summary, we conclude that there is not statistical evidence that Grade 1 class size has any effect on Grade 1 math outcomes.

2.5 Question 5: Propensity score stratification for estimating the ATE

5a. Within the common support as identified in Question 3, divide the analytic sample evenly into five strata (i.e., quintiles) on the logit propensity score estimated in Question 3. Within each stratum, compare the distribution of the logit propensity scores between the treated and the control by tabulating the corresponding means and variances. Also include in the table the average within-stratum standardized difference and average variance ratio in the logit propensity score that have been averaged over all the strata.

		Control			**Treated**		Standardized	Variance
Stratum	N	Mean	Variance	N	Mean	Variance	Difference	Ratio
1	2111	−1.354	0.026	561	−1.336	0.025	0.112	0.987

		Control			**Treated**		Standardized	Variance
Stratum	N	Mean	Variance	N	Mean	Variance	Difference	Ratio
2	2025	−1.062	0.003	646	−1.058	0.003	0.093	0.996
3	1875	−0.896	0.002	796	−0.893	0.002	0.060	0.997
4	1791	−0.767	0.001	880	−0.769	0.001	0.069	1.003
5	1780	−0.657	0.002	891	−0.654	0.002	0.062	0.995
Average								0.996

5b. Explain with reasons whether you decide to further subdivide the sample. If so, generate another table for the logit propensity score showing the results after you have modified your stratification. (Note: for this exercise, there is no need to modify your stratification more than once.)

As shown in the table above, all five strata are very nicely balanced in logit propensity scores. The largest within-stratum standardized difference in logit propensity score is 0.112 (for Stratum 1), which is below the desired 0.25 threshold. Meanwhile, the within-stratum variance ratios are all very close to 1: Stratum 5 has the lowest ratio at 0.995, and Stratum 4 has the highest at 1.003. Since we achieve such good balance in propensity scores with five strata, there is no there is no pressing need to subdivide the data into more strata, and we move forward with the stratification created in Part (a).

5c. Show the result of balance checking for each of the covariates by computing the average within-stratum standardized difference and the average variance ratio. (You may install the `pbalchk` package in Stata and display the results; in R, you may use `ps.makestrata()` followed by `ps.balance()` to generate the results of balance checking).

Below, we present the average within-stratum standardized differences between the treated and control groups for each covariate. As shown, we have achieved excellent balance along each covariate, with the magnitude of standardized differences never exceeding roughly 0.03.

Balance Checks

Variable	Treated	Control	Raw Diff.	Std Diff.
race6White	0.612	0.610	0.002	0.003
race6Asian	0.044	0.052	−0.008	−0.033
race6Black	0.138	0.139	−0.001	−0.003
race6Hispanic	0.157	0.151	0.006	0.017
race6Indigenous	0.027	0.027	0.000	0.002
race6Other	0.022	0.021	0.001	0.004
B4YRSTC_impt	14.029	13.978	0.052	0.005
race6White:B4YRSTC_impt	8.964	8.891	0.074	0.007
race6Asian:B4YRSTC_impt	0.661	0.760	−0.099	−0.023
race6Black:B4YRSTC_impt	1.755	1.756	−0.001	0.000

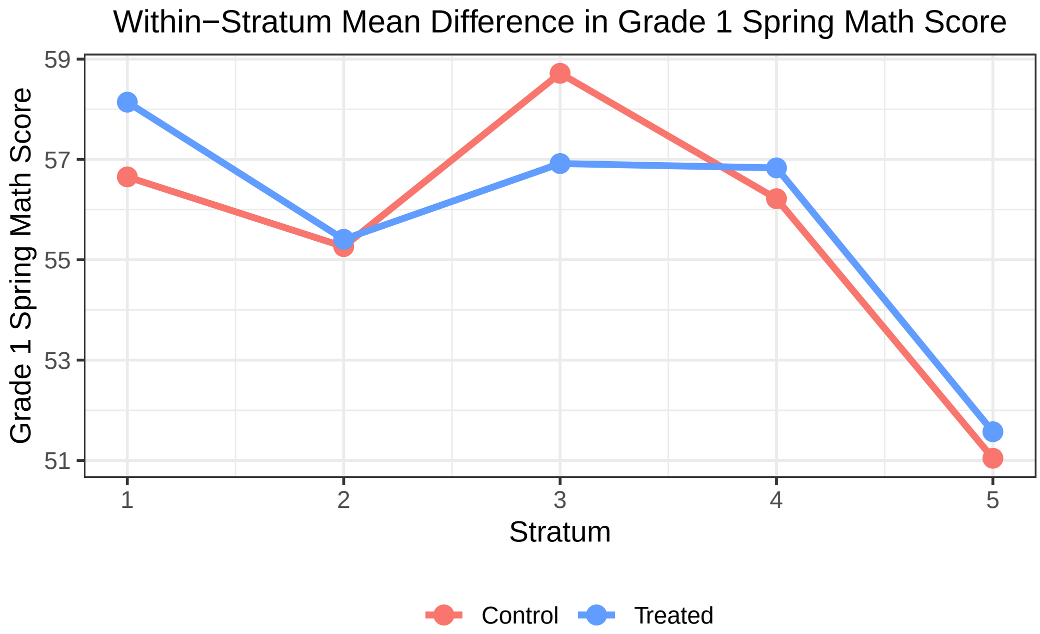
race6Hispanic:B4YRSTC_impt	1.981	1.907	0.073	0.012
race6Indigenous:B4YRSTC_impt	0.329	0.338	−0.008	−0.003
race6Other:B4YRSTC_impt	0.339	0.327	0.012	0.004
B4YRSTC_miss	0.021	0.021	0.001	0.003
race6White:B4YRSTC_miss	0.008	0.008	0.000	0.004
race6Asian:B4YRSTC_miss	0.003	0.003	0.000	0.001
race6Black:B4YRSTC_miss	0.004	0.005	0.000	−0.002
race6Hispanic:B4YRSTC_miss	0.005	0.004	0.000	0.003
race6Indigenous:B4YRSTC_miss	0.000	0.000	0.000	0.000
race6Other:B4YRSTC_miss	0.001	0.001	0.000	0.000
I(B4YRSTC_impt^2)	292.877	290.861	2.016	0.006
C1RRSCAL_impt	23.279	23.368	−0.089	−0.011
C1RRSCAL_miss	0.146	0.142	0.004	0.010
C2RRSCAL_impt	33.458	33.582	−0.124	−0.012
C2RRSCAL_miss	0.047	0.045	0.002	0.010
C1R2MSCL_impt	21.852	21.948	−0.096	−0.011
C1R2MSCL_miss	0.105	0.102	0.002	0.008
C2R2MSCL_impt	32.137	32.230	−0.093	−0.008
C2R2MSCL_miss	0.014	0.014	0.000	−0.004

5d. Compute the within-stratum mean difference in the Grade 1 math score between students attending small classes and those attending large classes. Tabulate the stratum-specific treatment effects (see Hong (2015) Chapter 3 Table 3.2 on page 65 for an example). (The above can be done manually in Stata or R; alternatively, you may use `mmws.exe` to carry out these steps and save the stratified data.) Do you observe any systematic pattern in the estimated effect of class size on Grade 1 math score across the strata? (You may graph the results as illustrated in slide 50 of the Week 4 handouts.)

We present a table of the within-stratum differences in math outcomes for students in regular-sized and small classrooms below:

		Control			**Treated**		Mean
Stratum	*N*	Mean	SD	*N*	Mean	SD	Diff.
1	2111	56.652	16.329	561	58.142	16.647	1.490
2	2025	55.263	15.567	646	55.408	16.912	0.145
3	1875	58.717	15.192	796	56.918	15.396	−1.799
4	1791	56.218	13.929	880	56.831	14.454	0.613
5	1780	51.043	14.041	891	51.572	14.879	0.529

Now, we visualize these within-stratum mean differences:



Notice that the strata with the largest gaps in mean Grade 1 math outcomes between the treated and control groups are Strata 1 and 3. Notice also that the signs of these two differences are opposite: in Stratum 1, students in small classrooms average Grade 1 math outcomes that are roughly 1.5 points *higher* than students in regular-sized classrooms, while in Stratum 3, students in small classrooms average Grade 1 math outcomes that are roughly 1.8 points *lower* than students in regular-sized classrooms. It is noteworthy that for the three remaining strata, the mean differences are smaller in magnitude, but are positive (i.e., students in small classrooms had higher average scores than students in regular-sized classrooms). This may mean that while enrollment in small classrooms is beneficial for most students, it is actually detrimental for the learning of students who have a middle-of-the-road probability of being in one of those classrooms in the first place.

5e. Estimate the average effect of class size on Grade 1 math score by regressing the outcome on the treatment indicator and the stratum indicators. Also report the standard error, the hypothesis testing result, and the effect. What is your conclusion with regard to the class size effect?

Using our stratified data, we now regress Grade 1 math scores (`C4R2MSCL`) on treatment (small or regular class size). As in Questions 2 and 4, we cluster standard errors by school.

We estimate an ATE of approximately $\hat{\delta}_{ATE} = 0.150$. Intuitively, this means that, on average, enrollment in a small Grade 1 classroom caused a 0.150-unit *increase* in Grade 1 math scores compared to students enrolled in regular-sized Grade 1 classrooms. However, this estimate is *not* statistically significant. The standard error of our estimate of δ_{ATE} is approximately 0.528, which conveys considerable uncertainty. Indeed, assuming that Grade 1 math score conditional on Grade 1 class size is normally distributed (a common assumption in regression analysis), our standard error of 0.528 means that we can be ~95%

confident that the true change in Grade 1 math score caused by being in a small Grade 1 classroom rather than a regular-sized Grade 1 classroom is somewhere in the interval $(0.150-2\times 0.528, 0.150+2\times 0.528) \approx (-0.886, 1.187)$. The fact that this interval contains 0 means that we cannot be confident that there even *is* a nonzero relationship between Grade 1 classroom size and Grade 1 math score. This is also reflected in the model's hypothesis testing result, wherein our null hypothesis is that Grade 1 classroom size is not associated with Grade 1 math outcomes. The p-value on our estimate of δ_{ATE} is approximately 0.776, which means that there is a 0.776 probability of observing a disparity in Grade 1 math scores between students in regular-sized and small Grade 1 classrooms that is greater than the disparity we actually observed. In other words, the data we observed is not compelling enough to make us reject the null hypothesis.

We again estimate the effect size using Glass's delta statistic. Here, Glass's delta equals $0.010 \approx 0$, which tells us that the mean Grade 1 math score among students in small Grade 1 classrooms is barely more than zero standard deviations above he mean score among students in regular-sized Grade 1 classrooms. A 95% confidence interval for this effect size (calculated using Wald's method) is $(-0.06, 0.08)$. That this interval contains 0 means that at the $\alpha = 0.05$ significance level, Glass's delta is not significantly different from 0.

Factor	Coefficient Estimate	Standard Error	95% CI for Coef. Est.	p-value	Glass's Delta	95% CI for Glass's Delta
Treatment	0.150	0.528	$(-0.886, 1.187)$	0.776	0.010	$(-0.06, 0.08)$

In summary, we conclude that there is not statistical evidence that Grade 1 class size has any effect on Grade 1 math outcomes.

2.6 Question 6: Inverse-probability-of-treatment weighting (IPTW) for estimating the ATE

6a. Apply IPTW in Stata or R to estimate the ATE (e.g., use `teffects ipw` in Stata or `PSweight` in R).

We compute IPTW weights for estimating the ATE according to the following formulas:

- If student i is in a small Grade 1 classroom, her IPTW weight is $IPTW_i = \frac{\mathbb{P}(trt=1)}{\theta_i}$, and
- If student i is in a regular-sized Grade 1 classroom, her IPTW weight is $IPTW_i = \frac{1-\mathbb{P}(trt=1)}{1-\theta_i}$.

Some example calculations are shown below:

```r
# Compute IPTW weights for ATE
data_iptw = data_cs %>%
  mutate(weight_ate = ifelse(trt == 1,
                             mean(trt) / pscore,
                             (1 - mean(trt)) / (1 - pscore)))

#mean(data_iptw$trt) ## 0.283
head(data_iptw[c("CHILDID", "trt", "pscore", "logit", "weight_ate")]) %>%
  gt() %>%
  fmt_number(decimals = 3, columns = vars(pscore, logit, weight_ate))
```

CHILD IDENTIFICATION NUMBER	trt	pscore	logit	weight_ate
0001001C	1	0.308	−0.811	0.919
0001002C	1	0.257	−1.063	1.101
0001004C	1	0.311	−0.794	0.907
0001005C	0	0.234	−1.183	0.937
0001009C	0	0.271	−0.987	0.985
0001010C	1	0.333	−0.694	0.848

6b. Check balance in the logit propensity score after weighting. This can be done by simply using the logit propensity score as the outcome in a weighted analysis. Also check balance in the pretreatment covariates. In each case, compare the standardized difference before and after weighting (again, you may use `tebalance summarize` in Stata or `bal.tab()` from the `cobalt` package in R). Summarize the results in a table.

First, we check balance in logit propensity scores and pretreatment covariates *before* weighting, which are summarized in the output below. As shown, the balance is not great. The standardized difference in logit propensity scores is 0.247, while the variance ratio is 0.820, each of which are just barely within their respective desired thresholds of 0.25 and $[\frac{4}{5}, \frac{5}{4}]$. The balance in the covariates is not great, either. On the one hand, the maximum standardized difference in a pretreatment covariate is 0.129—for the Asian indicator variable—which is not awful. However, there are several pretreatment covariates with variance ratios below $\frac{4}{5}$, including the interactions of each of the Asian, Indigenous, and Other Race indicators with (imputed, if missing) teacher years of experience. This imbalance illustrates the need for a propensity-score based procedure such as IPTW weighting.

Balance Checks: Pre-Weighting

Variable	Type	Std. Diff	Var. Ratio
logit	Distance	0.247	0.820
race6_White	Binary	0.120	NA

race6_Asian	Binary	−0.129	NA
race6_Black	Binary	0.020	NA
race6_Hispanic	Binary	−0.064	NA
race6_Indigenous	Binary	−0.034	NA
race6_Other	Binary	−0.047	NA
B4YRSTC_impt	Contin.	−0.085	0.909
B4YRSTC_miss	Binary	−0.014	NA
I(B4YRSTC_impt^2)	Contin.	−0.099	0.852
C1RRSCAL_impt	Contin.	−0.068	0.957
C1RRSCAL_miss	Binary	−0.013	NA
C2RRSCAL_impt	Contin.	−0.071	1.022
C2RRSCAL_miss	Binary	−0.035	NA
C1R2MSCL_impt	Contin.	−0.053	0.994
C1R2MSCL_miss	Binary	0.025	NA
C2R2MSCL_impt	Contin.	−0.028	1.015
C2R2MSCL_miss	Binary	−0.027	NA
race6_White * B4YRSTC_impt	Contin.	0.019	0.917
race6_Asian * B4YRSTC_impt	Contin.	−0.102	0.577
race6_Black * B4YRSTC_impt	Contin.	−0.029	0.829
race6_Hispanic * B4YRSTC_impt	Contin.	−0.040	0.876
race6_Indigenous * B4YRSTC_impt	Contin.	−0.043	0.713
race6_Other * B4YRSTC_impt	Contin.	−0.037	0.743
race6_White * B4YRSTC_miss_0	Binary	0.124	NA
race6_Asian * B4YRSTC_miss_0	Binary	−0.128	NA
race6_Black * B4YRSTC_miss_0	Binary	0.019	NA
race6_Hispanic * B4YRSTC_miss_0	Binary	−0.067	NA
race6_Indigenous * B4YRSTC_miss_0	Binary	−0.034	NA
race6_Other * B4YRSTC_miss_0	Binary	−0.048	NA
race6_White * B4YRSTC_miss_1	Binary	−0.026	NA
race6_Asian * B4YRSTC_miss_1	Binary	−0.018	NA
race6_Black * B4YRSTC_miss_1	Binary	0.005	NA
race6_Hispanic * B4YRSTC_miss_1	Binary	0.016	NA
race6_Indigenous * B4YRSTC_miss_1	Binary	0.004	NA
race6_Other * B4YRSTC_miss_1	Binary	−0.001	NA
gender3_2	Binary	−0.037	NA

Now, we check balance in logit propensity scores and pretreatment covariates *after* weighting, again summarized in the output below. Balance has clearly improved. The standardized difference in logit propensity scores is now just $-0.005 \approx 0$, while the variance ratio is $1.037 \approx 1$. Balance in the pretreatment covariates has improved as well. The maximum standardized difference in a pretreatment covariate is now just 0.031—for gender—which is very close to 0. Meanwhile, the variance ratios for pretreatment covariates are all between 0.98—for the interaction of the Asian indicator with (imputed, if missing) teacher years of experience—and 1.165—for (imputed, if missing) kindergarten spring reading score—which are all within the desired $[\frac{4}{5}, \frac{5}{4}]$ range.

Balance Checks: Post-Weighting

Variable	Type	Std. Diff.	Var. Ratio
logit	Distance	−0.005	1.037
race6_White	Binary	−0.001	NA
race6_Asian	Binary	0.004	NA
race6_Black	Binary	0.000	NA
race6_Hispanic	Binary	−0.001	NA
race6_Indigenous	Binary	−0.001	NA
race6_Other	Binary	0.002	NA
B4YRSTC_impt	Contin.	0.000	0.996
B4YRSTC_miss	Binary	0.003	NA
I(B4YRSTC_impt^2)	Contin.	−0.001	0.988
C1RRSCAL_impt	Contin.	0.009	1.137
C1RRSCAL_miss	Binary	−0.007	NA
C2RRSCAL_impt	Contin.	0.008	1.165
C2RRSCAL_miss	Binary	0.002	NA
C1R2MSCL_impt	Contin.	0.007	1.135
C1R2MSCL_miss	Binary	−0.008	NA
C2R2MSCL_impt	Contin.	0.005	1.077
C2R2MSCL_miss	Binary	−0.001	NA
race6_White * B4YRSTC_impt	Contin.	−0.003	0.999
race6_Asian * B4YRSTC_impt	Contin.	0.003	0.980
race6_Black * B4YRSTC_impt	Contin.	0.001	1.013
race6_Hispanic * B4YRSTC_impt	Contin.	0.001	0.990
race6_Indigenous * B4YRSTC_impt	Contin.	0.000	1.049
race6_Other * B4YRSTC_impt	Contin.	0.002	1.016
race6_White * B4YRSTC_miss_0	Binary	−0.002	NA
race6_Asian * B4YRSTC_miss_0	Binary	0.004	NA
race6_Black * B4YRSTC_miss_0	Binary	0.000	NA
race6_Hispanic * B4YRSTC_miss_0	Binary	−0.001	NA
race6_Indigenous * B4YRSTC_miss_0	Binary	−0.001	NA
race6_Other * B4YRSTC_miss_0	Binary	0.002	NA
race6_White * B4YRSTC_miss_1	Binary	0.005	NA
race6_Asian * B4YRSTC_miss_1	Binary	−0.001	NA
race6_Black * B4YRSTC_miss_1	Binary	−0.001	NA
race6_Hispanic * B4YRSTC_miss_1	Binary	0.002	NA
race6_Indigenous * B4YRSTC_miss_1	Binary	−0.001	NA
race6_Other * B4YRSTC_miss_1	Binary	−0.002	NA
gender3_2	Binary	−0.031	NA

6c. Estimate the population average effect of class size on Grade 1 math achievement, report the standard error, the hypothesis testing result, and the effect

size. What is your conclusion with regard to the class size effect?

Using our weighted data, we now regress Grade 1 math scores (`C4R2MSCL`) on treatment (small or regular class size). As in Questions 2, 4, and 5, we cluster standard errors by school.

We estimate an ATE of approximately $\hat{\delta}_{ATE} = 0.319$. Intuitively, this means that, on average, enrollment in a small Grade 1 classroom caused a 0.319-unit *increase* in Grade 1 math scores compared to students enrolled in regular-sized Grade 1 classrooms. However, this estimate is *not* statistically significant. The standard error of our estimate of δ_{ATE} is approximately 0.535, which conveys considerable uncertainty. Indeed, assuming that Grade 1 math score conditional on Grade 1 class size is normally distributed (a common assumption in regression analysis), our standard error of 0.535 means that we can be ~95% confident that the true change in Grade 1 math score caused by being in a small Grade 1 classroom rather than a regular-sized Grade 1 classroom is somewhere in the interval $(0.309 - 2 \times 0.535, 0.309 + 2 \times 0.535) \approx (-0.731, 1.369)$. The fact that this interval contains 0 means that we cannot be confident that there even *is* a nonzero relationship between Grade 1 classroom size and Grade 1 math score. This is also reflected in the model's hypothesis testing result, wherein our null hypothesis is that Grade 1 classroom size is not associated with Grade 1 math outcomes. The p-value on our estimate of δ_{ATE} is approximately 0.551, which means that there is a 0.551 probability of observing a disparity in Grade 1 math scores between students in regular-sized and small Grade 1 classrooms that is greater than the disparity we actually observed. In other words, the data we observed is not compelling enough to make us reject the null hypothesis.

We again estimate the effect size using Glass's delta statistic. Here, Glass's delta equals $0.021 \approx 0$, which tells us that the mean Grade 1 math score among students in small Grade 1 classrooms is barely more than zero standard deviations above the mean score among students in regular-sized Grade 1 classrooms. A 95% confidence interval for this effect size (calculated using Wald's method) is $(-0.05, 0.09)$. That this interval contains 0 means that at the $\alpha = 0.05$ significance level, Glass's delta is not significantly different from 0.

Factor	Coefficient Estimate	Standard Error	95% CI for Coef. Est.	p-value	Glass's Delta	95% CI for Glass's Delta
Treatment	0.319	0.534	$(-0.731, 1.369)$	0.551	0.021	$(-0.047, 0.088)$

In summary, we conclude that there is not statistical evidence that Grade 1 class size has any effect on Grade 1 math outcomes.

2.7 Question 7: Marginal mean weighting through stratification (MMWS) for estimating the ATE

You may use the `mmws.exe` program to estimate the ATE.

(We use R code to perform this analysis as modeled in the Lab 6 code file.)

7a. After logistic regression analysis, use the data within the common support with caliper as your analytic sample. (You may use the same set of strata that you constructed in Question 5 for estimating the ATE.) Check balance in the logit propensity score and in each pretreatment covariate by computing the standardized difference and variance ratio before and after weighting. (The standardized differences after weighting are reported in the "Balance Checking Tables" in the output of the standalone mmws software.) Summarize the results in a table as you did in answering the previous questions.

First, we check balance in logit propensity scores and pretreatment covariates *before* weighting, which are summarized in the output below. Note that this table is identical to the pre-weighting balance check table shown in Question 6b, since we have not applied any weights yet. As discussed in Question 6b, the balance is not great. The standardized difference in logit propensity scores is 0.247, while the variance ratio is 0.820, each of which are just barely within their respective desired thresholds of 0.25 and $[\frac{4}{5}, \frac{5}{4}]$. The balance in the covariates is not great, either. On the one hand, the maximum standardized difference in a pretreatment covariate is 0.129—for the Asian indicator variable—which is not awful. However, there are several pretreatment covariates with variance ratios below $\frac{4}{5}$, including the interactions of each of the Asian, Indigenous, and Other Race indicators with (imputed, if missing) teacher years of experience. This imbalance illustrates the need for a propensity-score based procedure such as MMWS weighting.

Balance Checks: Pre-Weighting

Variable	Type	Std_Diff	Var_Ratio
logit	Distance	0.247	0.820
race6_White	Binary	0.120	NA
race6_Asian	Binary	−0.129	NA
race6_Black	Binary	0.020	NA
race6_Hispanic	Binary	−0.064	NA
race6_Indigenous	Binary	−0.034	NA
race6_Other	Binary	−0.047	NA
B4YRSTC_impt	Contin.	−0.085	0.909
B4YRSTC_miss	Binary	−0.014	NA
I(B4YRSTC_impt^2)	Contin.	−0.099	0.852
C1RRSCAL_impt	Contin.	−0.068	0.957
C1RRSCAL_miss	Binary	−0.013	NA
C2RRSCAL_impt	Contin.	−0.071	1.022
C2RRSCAL_miss	Binary	−0.035	NA
C1R2MSCL_impt	Contin.	−0.053	0.994
C1R2MSCL_miss	Binary	0.025	NA
C2R2MSCL_impt	Contin.	−0.028	1.015
C2R2MSCL_miss	Binary	−0.027	NA
race6_White * B4YRSTC_impt	Contin.	0.019	0.917
race6_Asian * B4YRSTC_impt	Contin.	−0.102	0.577

race6_Black * B4YRSTC_impt	Contin.	−0.029	0.829
race6_Hispanic * B4YRSTC_impt	Contin.	−0.040	0.876
race6_Indigenous * B4YRSTC_impt	Contin.	−0.043	0.713
race6_Other * B4YRSTC_impt	Contin.	−0.037	0.743
race6_White * B4YRSTC_miss_0	Binary	0.124	NA
race6_Asian * B4YRSTC_miss_0	Binary	−0.128	NA
race6_Black * B4YRSTC_miss_0	Binary	0.019	NA
race6_Hispanic * B4YRSTC_miss_0	Binary	−0.067	NA
race6_Indigenous * B4YRSTC_miss_0	Binary	−0.034	NA
race6_Other * B4YRSTC_miss_0	Binary	−0.048	NA
race6_White * B4YRSTC_miss_1	Binary	−0.026	NA
race6_Asian * B4YRSTC_miss_1	Binary	−0.018	NA
race6_Black * B4YRSTC_miss_1	Binary	0.005	NA
race6_Hispanic * B4YRSTC_miss_1	Binary	0.016	NA
race6_Indigenous * B4YRSTC_miss_1	Binary	0.004	NA
race6_Other * B4YRSTC_miss_1	Binary	−0.001	NA
gender3_2	Binary	−0.037	NA

Now, we check balance in logit propensity scores and pretreatment covariates *after* weighting, again summarized in the output below. Balance has clearly improved. The standardized difference in logit propensity scores is now just $0.021 \approx 0$, while the variance ratio is $0.955 \approx 1$. Balance in the pretreatment covariates has improved as well. The maximum standardized difference in a preatreatment covariate is now just 0.0431769—for the interaction of the Asian indicator with the missing teacher years of experience indicator—which is very close to 0. Meanwhile, the variance ratios for pretreatment covariates are all between 0.840687—again for the interaction of the Asian indicator with (imputed, if missing) teacher years of experience—and 1.1368624—for (imputed, if missing) kindergarten reading score, which are all within the desired $[\frac{4}{5}, \frac{5}{4}]$ range.

Balance Checks: Post-Weighting

Variable	Type	Std_Diff	Var_Ratio
logit	Distance	0.021	0.955
race6_White	Binary	0.005	NA
race6_Asian	Binary	−0.041	NA
race6_Black	Binary	−0.004	NA
race6_Hispanic	Binary	0.020	NA
race6_Indigenous	Binary	0.000	NA
race6_Other	Binary	0.004	NA
B4YRSTC_impt	Contin.	0.005	1.002
B4YRSTC_miss	Binary	0.005	NA
I(B4YRSTC_impt^2)	Contin.	0.005	0.998
C1RRSCAL_impt	Contin.	−0.007	1.104
C1RRSCAL_miss	Binary	0.005	NA

C2RRSCAL_impt	Contin.	−0.009	1.137
C2RRSCAL_miss	Binary	0.015	NA
C1R2MSCL_impt	Contin.	−0.008	1.113
C1R2MSCL_miss	Binary	0.000	NA
C2R2MSCL_impt	Contin.	−0.007	1.069
C2R2MSCL_miss	Binary	−0.005	NA
race6_White * B4YRSTC_impt	Contin.	0.007	1.015
race6_Asian * B4YRSTC_impt	Contin.	−0.029	0.841
race6_Black * B4YRSTC_impt	Contin.	0.000	1.009
race6_Hispanic * B4YRSTC_impt	Contin.	0.015	1.026
race6_Indigenous * B4YRSTC_impt	Contin.	−0.003	1.005
race6_Other * B4YRSTC_impt	Contin.	0.005	1.030
race6_White * B4YRSTC_miss_0	Binary	0.004	NA
race6_Asian * B4YRSTC_miss_0	Binary	−0.043	NA
race6_Black * B4YRSTC_miss_0	Binary	−0.003	NA
race6_Hispanic * B4YRSTC_miss_0	Binary	0.020	NA
race6_Indigenous * B4YRSTC_miss_0	Binary	0.000	NA
race6_Other * B4YRSTC_miss_0	Binary	0.005	NA
race6_White * B4YRSTC_miss_1	Binary	0.008	NA
race6_Asian * B4YRSTC_miss_1	Binary	0.001	NA
race6_Black * B4YRSTC_miss_1	Binary	−0.003	NA
race6_Hispanic * B4YRSTC_miss_1	Binary	0.003	NA
race6_Indigenous * B4YRSTC_miss_1	Binary	0.000	NA
race6_Other * B4YRSTC_miss_1	Binary	−0.002	NA
gender3_2	Binary	−0.034	NA

7b. Estimate the average effect of class size on Grade 1 math achievement. Also report the standard error, the hypothesis testing result, and the effect size. What is your conclusion with regard to the class size effect? How does the result compare with what you obtained in Question 5?

Using our weighted data, we now regress Grade 1 math scores (`C4R2MSCL`) on treatment (small or regular class size). As in questions 2, 4, 5, and 6, we cluster standard errors by school.

We estimate an ATE of approximately $\hat{\delta}_{ATE} = 0.196$. Intuitively, this means that, on average, enrollment in a small Grade 1 classroom caused a 0.196-unit *increase* in Grade 1 math scores compared to students enrolled in regular-sized Grade 1 classrooms. However, this estimate is *not* statistically significant. The standard error of our estimate of δ_{ATE} is approximately 0.534, which conveys considerable uncertainty. Indeed, assuming that Grade 1 math score conditional on Grade 1 class size is normally distributed (a common assumption in regression analysis), our standard error of 0.534 means that we an be ~95% confident that the true change in Grade 1 math score caused by being in a small Grade 1 classroom rather than a regular-sized Grade 1 classroom is somewhere in the interval $(0.196 - 2 \times 0.534, 0.196 + 2 \times 0.534) \approx (-0.854, 1.246)$. The fact that this interval contains 0 means that we cannot be confident that there even *is* a nonzero relationship between Grade 1 classroom size and Grade 1 math score. This is also reflected in the model's hypothesis testing result, wherein our null hypothesis is that Grade 1 classroom size is not associated with Grade 1 math outcomes. The p-value on our estimate of δ_{ATE} is approximately 0.714, which means that there is a 0.714 probability of observing a disparity in Grade 1 math scores between students in regular-sized and small Grade 1 classrooms that is greater than the disparity we actually observed. In other words, the data we observed is not compelling enough to make us reject the null hypothesis.

We again estimate the effect size using Glass's delta statistic. Here, Glass's delta equals $0.013 \approx 0$, which tells us that the mean Grade 1 math score among students in small Grade 1 classrooms is barely more than zero standard deviations above the mean score among students in regular-sized Grade 1 classrooms. A 95% confidence interval for this effect size (calculated using Wald's method) is $(-0.06, 0.09)$. That this interval contains 0 means that at the $\alpha = 0.05$ significance level, Glass's delta is not significantly different from 0.

Factor	Coefficient Estimate	Standard Error	95% CI for Coef. Est.	p-value	Glass's Delta	95% CI for Glass's Delta
Treatment	0.196	0.534	$(-0.854, 1.246)$	0.714	0.013	$(-0.056, 0.089)$

In summary, we conclude that there is not statistical evidence that Grade 1 class size has any effect on Grade 1 math outcomes.

2.8 Question 8: Identification assumption

8a. What is the key identification assumption required for the above results to have causal validity? State your assumption both in symbols and in words. Explain each notation used when you express the identification assumption in symbols.

The key identification assumption required for all of our above results to have causal validity is that in each case, $\delta_{PF} = \delta$, where $\delta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ and $\delta_{PF} = \mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]$.

- Here, δ is the causal effect of being in a small classroom rather than a regular-sized classroom on Grade 1 math outcome. In Question 4, we considered δ to be the ATT, while in Questions 5–7, we considered δ to be the ATE. $\mathbb{E}[Y(1)]$ is the population average potential Grade 1 math score associated with being in a small Grade 1 classroom, while $\mathbb{E}[Y(0)]$ is the population average potential Grade 1 math score associated with being in a regular-sized Grade 1 classroom. Since $\delta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$, δ is the population average difference in Grade 1 math scores associated with being in a small Grade 1 classroom relative to a regular-sized Grade 1 classroom.

- Meanwhile, δ_{PF} is the *prima facie* causal effect of being in a small classroom rather than a regular-sized classroom on Grade 1 math outcome. That is, δ_{PF} is the difference between the average Grade 1 math score among students who were enrolled in small Grade 1 classrooms and the average Grade 1 math score among students who were enrolled in regular-sized Grade 1 classrooms. $\mathbb{E}[Y|Z=1]$ is the average Grade 1 math score among students in the study who were enrolled in small Grade 1 classrooms, while $\mathbb{E}[Y|Z=0]$ is the average Grade 1 math score among students in the study who were enrolled in regular-sized Grade 1 classrooms.

The assumption that $\delta_{PF}=\delta$ is equivalent to the *ignorability* assumption that each student's Grade 1 class size Z (where $Z=0$ denotes a regular-sized class and $Z=1$ denotes a small class) is independent of (1) their potential Grade 1 math score if assigned to a regular-sized Grade 1 classroom, $Y(0)$, and (2) their potential Grade 1 math score if assigned to a small Grade 1 classroom, $Y(1)$. Symbolically, we assume that $Z \perp\!\!\!\perp Y(0)$ and $Z \perp\!\!\!\perp Y(1)$.

Intuitively, we assume that each student's potential Grade 1 math scores—whether they are in a small or regular-sized classroom—have no bearing on the size of the classroom in which they are enrolled. Similarly, the size of the classroom in which each student is enrolled must have no bearing on their potential Grade 1 math scores in either class size. For example, it must *not* be the case that the students who perform best in small classrooms are disproportionately likely to be enrolled in small classrooms. We instead would assume that even if some students perform best in small classrooms, they are just as likely as any other student to be enrolled in a regular-sized classroom or a small-classroom.

8b. Please think of one unmeasured covariate that might potentially confound the class size effect. Explain why the omission of such a covariate from the analysis might have potentially biased your result.

One key covariate that was not measured (or at least not included in the dataset) in this study is parental income/wealth. Parental income/wealth could very easily confound our estimates of the class size effect.

Firstly, wealthier parents can generally afford to send their children to better-resourced school districts with small average class sizes—whether by living in a wealthy community with well-funded public schools, or by sending their children to private schools. Poorer families are much more likely to live in inexpensive communities where public schools are overcrowded, under-resourced, and consequently have large average class sizes. As such, parental income/wealth is likely correlated with a child's Grade 1 class size.

Moreover, parental income/wealth likely affects Grade 1 math outcomes through pathways *other than* class size. For example, wealthier parents may be able to afford tutoring services for their child that poorer parents cannot, giving wealthier children a leg up in their studies. Even more simply, children of poorer parents may experience food or housing insecurity that interferes with their learning, setting their progress behind wealthier students who have reliable meals and housing. (Some of our analyses may have partially addressed this problem by clustering standard errors by school, since students attending the same school likely live in the same community, and students living in the same community likely have parents with comparable income levels. But without explicit parental income/wealth data, this is certainly not a perfect fix.) Thus, parental income/wealth is also likely correlated with a child's Grade 1 math achievement.

By omitting this likely confounder from our analyses, our estimates of the class size effect could very well be biased. In particular, we have posited that "treatment" (i.e., being in a small classroom) and parental income/wealth are positively associated: the wealthier a child's parents, the more likely they are to be treated, on average. We have also posited that Grade 1 math outcomes and parental income/wealth are positively associated: the wealthier a child's parents, the higher their math score, on average. So, our estimates of the effect of enrollment in a small Grade 1 classroom on math outcomes may have "absorbed" the positive effect of parental wealth on math outcomes that is disproportionately enjoyed by students in small classrooms. This would introduce a positive bias in our estimates; in other words, we likely have overestimated the benefits of enrollment in a small Grade 1 classroom on math outcomes. Since many of our analyses found that there was not a significant, nonzero effect of Grade 1 class size on math outcomes, accounting for this confounder and removing the positive bias may even mean that being in a small Grade 1 classroom has a negative effect on math outcomes.

8c. What is the purpose of a sensitivity analysis in causal inference in general? In the context of the current study, under what conditions would you consider an analytic result to be sensitive to potential bias associated with an omitted confounder?

In any causal analysis, we assume—as described in Part (a)—that the *prima facie* effect we have estimated properly identifies the true causal effect we are interested in. However, there is no way to test or guarantee that the identification assumption made in our original analysis holds true—that's why it's an assumption! Sensitivity analysis involves checking whether our conclusions are robust/immune to potential violations of the identification assumptions, since it may very well be the case that some of our identification assumptions *have* been violated. If our conclusions after performing sensitivity analyses are (qualitatively) the same as our original conclusion, that lends a lot of support to the validity of the original conclusion, since it means that our conclusions are valid even if our assumptions were not. On the other hand, if minor violations of identification assumptions yield conclusions that are very different from those of the original analysis, it renders the initial conclusions highly suspicious. After all, if a violation of the identification assumptions invalidates a causal result, and there is no guarantee that the identification assumptions are valid, then there is no guarantee that the causal result is valid, either.

In our current study of the effects of classroom size on Grade 1 math outcomes, we might consider using a weighting-based approach to sensitivity analysis. In particular, assuming parental income/wealth is an omitted confounder—as discussed in Part (b)—we would compute a new set of weights that take into account the probabilities of being in a small or regular-sized classroom *conditional on* particular values of parental income/wealth (as well as the other pretreatment covariates we were already conditioning on). We would then compare the results of our analysis using the "original" weights (that do not account for parental income/wealth) and our analysis using the "new" weights (that do account for parental income/wealth). If using the new weights gives a completely different conclusion than using the original weights did, then our causal estimate would be considered *sensitive* to the bias associated with omitting parental income/wealth as a confounder, and our original conclusions would be called into question.

For example, in Question 6, we found using IPTW weights that Grade 1 class size did not have a significant, nonzero effect on Grade 1 math outcomes. If we were to redo this analysis with recalculated weights that accounted for parental income/wealth and found that either (a) being in a small Grade 1 classroom caused an *increase* in Grade 1 math scores compared to being in a regular-sized Grade 1 classroom, or (b) being in a small Grade 1 classroom caused a *decrease* in Grade 1 math scores compared to being in a regular-sized Grade 1 classroom, our conclusion would be qualitatively different than our original conclusion. This would make our original conclusion of no class size effect *sensitive* to the omission of parental income/wealth data, calling into question the validity of this "no effect" conclusion.