ECON 21110 - Applied Microeconometrics - Assignment 3

Jack Ogle in collaboration with Eva Haque, Matt Lohrs, and Jack Knickrehm

Problem 1

(a)

Banerjee, Cole, Duflo, and Liden are trying to understand two opposing ideas about education. (1) According to the Millennium Development Goals put forth by the United Nations, primary education should be a universal right. (2) However, the quality of education currently offered to the poor is of bad quality. These authors are trying to understand why the children of poor countries are receiving inferior education. They conduct a field experiment where they measure the effect of the Balsakhi Program and CAL the (Computer Assisted Learning) Program on the learning levels of elementary school children in India. The students learning levels are measured by mathematical and verbal standardized tests. The students take the tests pre and post treatment.

(b)

To ensure a balanced sample, assignment was stratified by language, pretest score, and gender.

We stratify the pretest scores and gender like they did in the study. We can accomplish this by regressing the balsakhi on the male and bal variable. Which gives us the following result. We see that the pre_tot coefficient is -0.0002 close to zero, and are statistically significant. We can also see that the male coefficient is close to 0 and not statistically significant. We can conclude that the treatment and control groups are balanced because the effect of gender and pretest scores have very little effect on whether a student is placed in the treatment or control group. Yes, we should cluster standard errors and they are labeled in the table.

Table 1: Regression Results (b)

	Dependent variable:	
	bal	
	Without Clustered Std. Errors	With Clustered Std. Errors
	(1)	(2)
pre_tot	−0.0002	−0.0002
	(0.0002)	(0.0009)
male	0.018*	0.018*
	(0.010)	(0.0431)
Constant	0.485***	0.485***
	(0.010)	(0.0524)
Observations	9,745	9,745
R^2	0.0004	0.0004
Adjusted R^2	0.0002	0.0002
Residual Std. Error (df = 9742)	0.500	0.500
F Statistic (df = 2; 9742)	1.934	1.934
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses	

(c)

Table 2: Regression Results (c)

	Dependent variable:			
	bal			
	Pretest Treatment	Pretest Comparision	Posttest Treatment	Posttest Comparision
	(1)	(2)	(3)	(4)
Math	−0.00523	0.000519	0.3987	0.19556
Verbal	0.00823	−0.00816	0.8558	0.6976

Our results match the paper and so the papers interpretation is verified and the Balsakhi program is successful. In Vadodara, in the first year, the difference in posttest between treatment and comparison groups was 0.203 standard deviations for math and 0.158 standard deviations for verbal. This is a pretty substantial improvement.

(d)

Table 3: Regression Results (d)

	Dependent variable:	
	(post_tot - pre_tot)	
	Year 3	Year 4
pre_tot	-0.236^{***}	-0.177^{***}
	(0.036)	(0.024)
bal	3.216^{***}	3.928^{***}
	(1.79)	(1.632)
Constant	15.251^{***}	14.625^{***}
	(0.621)	(0.764)
Observations	2,644	2,750
R^2	0.077	0.062
Adjusted R^2	0.077	0.062
Residual Std. Error	17.168 (df = 2641)	16.544 (df = 2747)
F Statistic	110.858^{***} (df = 2; 2641)	91.154^{***} (df = 2; 2747)
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses	

The estimated treatment effect is represented by the coefficient of bal. For students in grade 3 (column 1), treatment seems to have improved their total score by 3.216 standard deviations, which is slightly less than the amount by which students in grade 4 benefited from treatment. Treated grade 4 students saw their total scores increase by 3.28 standard deviations.

(e)

Table 4:

	Dependent variable: Total Score (Year 2 - Year 1)
pre-test score	-0.407^{***}
	(0.017)
bal	0.244^{***}
	(0.035)
Constant	0.954^{***}
	(0.025)
Observations	3,145
R^2	0.159
Adjusted R^2	0.158
Residual Std. Error	0.974 (df = 3142)
F Statistic	296.024^{***} (df = 2; 3142)
Note:	*p<0.1; **p<0.05; ***p<0.01

Estimate the two-year balsakhi treatment. We have clustered our standard errors. We can see that there is a 0.244 standard deviation increase in receiving the treatment balsakhi from the pretest (year 1 to year 2).

(f) The main threats to internal validity threats are selection bias, attrition, and hawthorne effect. Selection bias might be possible because we are comparing the students who are already behind to the students who are on track. The students in the treatment group are already behind those students who are not so the gains they make might be for a variety of reasons that have little to do with the treatment. We need randomization here to fix the selection bias.

Attrition will also affect the study because the students who left school or dropped out and might have benefited alot from the treatment will not be measured the next year in the posttest. This would skew our results to say that the treatment had less of an effect on the education level than it actually did.

Hawthorne Effect had an impact on the study because the students who participated in the study knew that they were participating in a study and therefore might feel pressure from the researchers to act or study better or worse. Essentially they might think they are helping the researcher by answering a question this way or another because they are in a study. This might affect the results because the students aren't behaving genuinely.

The main threat to external validity are that the schooling in the cities in th paper are not a perfect representation of India as a whole. India is a very large country and a very diverse one at that therefore conclusions about how they decided who to put in the treatment group and control group might vary alot region to region. For example who is considered to be behind would differ in New Delhi compared to Mumbai.

(g)

Table 5: Regression Results (g)

	Dependent variable:	
	treated	
	Non-Robust Std. Errors	Robust Std. Errors
std	−0.006	−0.006
	(0.009)	(0.0086)
male	0.001	0.001
	(0.009)	(0.0086)
numstud	0.0002	0.0002
	(0.0002)	(0.00016)
pre_totnorm	-0.013^{**}	-0.013^{**}
	(0.006)	(0.0062)
income	0.002^{***}	0.002^{***}
	(0.0002)	(0.0001)
Constant	0.280^{***}	0.280^{***}
	(0.039)	(0.0392)
Observations	12,415	12,415
R^2	0.028	0.028
Adjusted R^2	0.028	0.028
Residual Std. Error (df = 12409)	0.483	0.483
F Statistic (df = 5; 12409)	71.411^{***}	71.411^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses	

From these results I would argue that the pre-treatment covariates are balanced between the treatment and control group. From the regression we see that each coefficient is close to 0 meaning that they have little effect on whether a student is in the treatment or control group.

(h)

Table 6: Regression Results (h)

	Dependent variable:
	Finalscore
treated	0.298***
	(0.003)
male	−0.001
	(0.003)
income	0.00001
	(0.0001)
std	0.0005
	(0.003)
pre_totnorm	0.998***
	(0.002)
Constant	−0.002
	(0.014)
Observations	12,415
R^2	0.972
Adjusted R^2	0.972
Residual Std. Error	0.173 (df = 12409)
F Statistic	87,588.000*** (df = 5; 12409)
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses

On average if a student is in the treatment group there is a 0.298 standard deviation increase in Final Score. Yes we would need to control for other pre-treatment variables such as amount of time that the parents spent with their children and variables that would improve and effect development outcomes and thus effect the pre test scores. This would improve consistency of the treatment variable because it would make MLR.4 more credible in reducing the amount of variables in the error term that are correlated with the independent variables.

(i)

We can improve the precision of the estimator by performing a BP test to test for heteroskedasticity. Then we can account for the heteroskedasticity by using robust standard errors to make our results more meaningful.

The results of our BP test yield the following results: BP = 1.6069, df = 5, p-value = 0.9004

Because the p-value is greater than 0.05 we cannot reject our null hypothesis of homoskedasticity. We can conclude that the model is not heteroskedastic and we don't need use Robust Standard Errors to make our model more meaningful and precise. However, this is one way to make the model more precise to view these standard error look to table 5.

(j)

Table 7: Regression Results (j)

	Dependent variable:		
	Final Scores	High Income	Low Income
treated	0.298***	0.284	−0.459
	(0.003)	(0.578)	(0.493)
pre_totnorm	0.999***	17.766***	14.862***
	(0.002)	(0.257)	(0.403)
Constant	0.0005	174.508***	125.123***
	(0.002)	(0.510)	(0.405)
Observations	12,415	6,104	6,311
R^2	0.972	0.439	0.177
Adjusted R^2	0.972	0.438	0.177
Residual Std. Error	0.173 (df = 12412)	20.627 (df = 6101)	19.585 (df = 6308)
F Statistic	219,019.700*** (df = 2; 12412)	2,383.763*** (df = 2; 6101)	680.495*** (df = 2; 6308)
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses		

Table 8: Regression Results (j)

	Dependent variable:		
	High Students	Low Students	Male
treated	0.521	−0.193	0.001
	(0.498)	(0.311)	(0.009)
pre_totnorm	1.260***	−0.080	−0.022***
	(0.248)	(0.150)	(0.004)
Constant	84.584***	43.289***	0.499***
	(0.386)	(0.239)	(0.007)
Observations	5,876	6,539	12,415
R^2	0.005	0.0001	0.002
Adjusted R^2	0.005	-0.0002	0.002
Residual Std. Error	18.555 (df = 5873)	12.258 (df = 6536)	0.500 (df = 12412)
F Statistic	14.280*** (df = 2; 5873)	0.376 (df = 2; 6536)	12.061*** (df = 2; 12412)
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses		

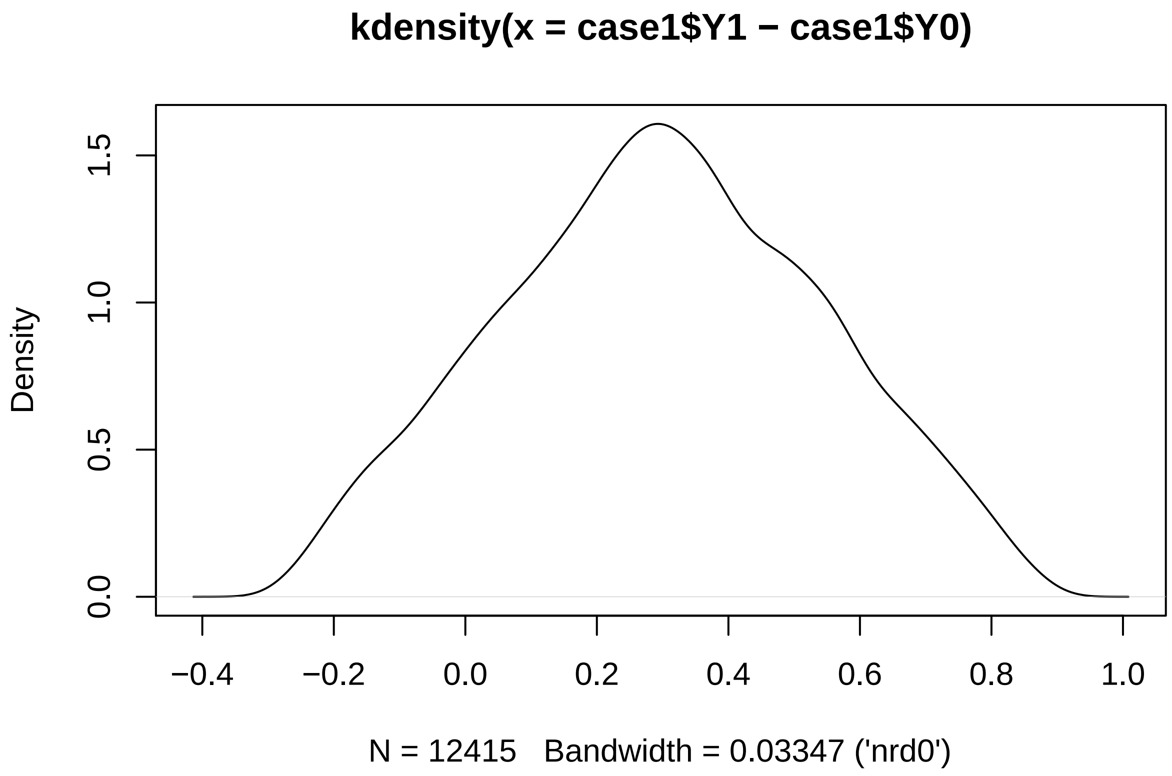
I conducted the subgroup test by first dividing up the income. The high income group has an income that is greater than the median income. The low income group has an income that is less than the mean. Then I divided by class size. The high student number group has a student number that above the mean and the low student number group has one below the mean. I also analyzed the sub groups final score and male.

The results I achieved are interesting. I found that There are pretty big differences in the treatment effect coefficient for high income versus low income from 0.284 to -0.459. The effect number of student per class also yielded interesting results there was as significant difference in the treatment coefficient for high students and low students. From 0.521 to -0.193. There was very little difference for gender.

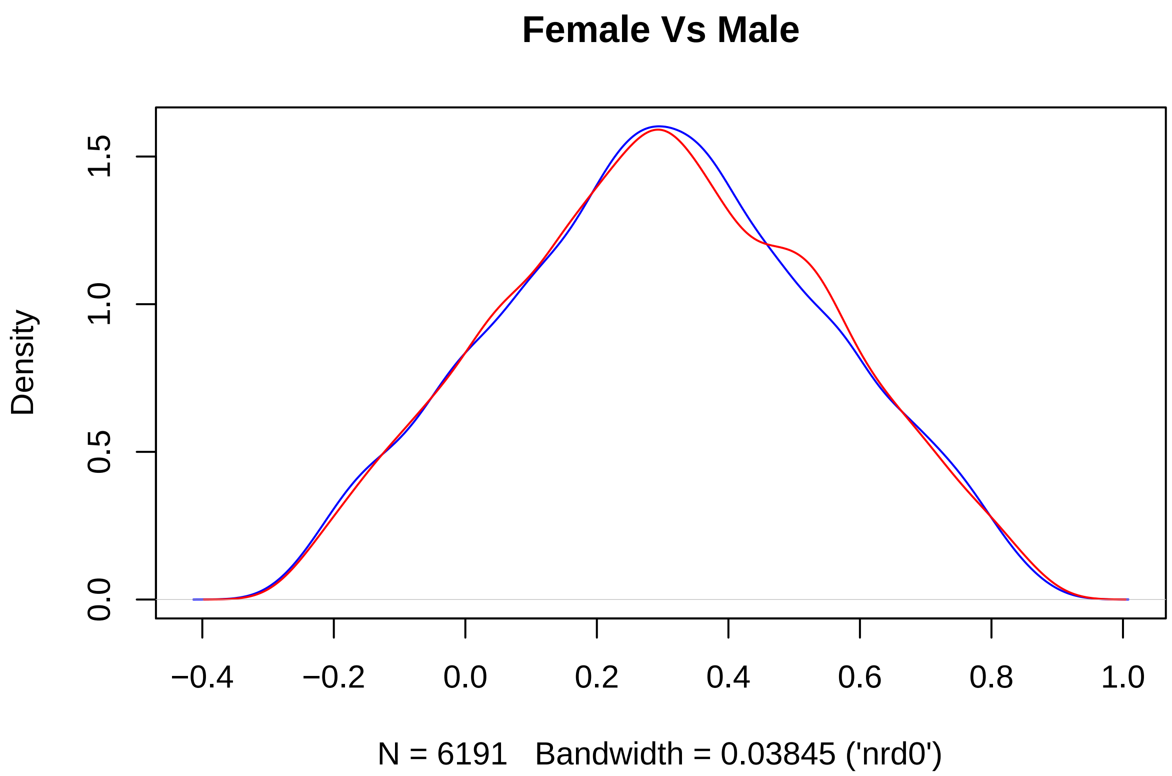
(k)

```
library(kdensity)
plot(kdensity(case1$Y1-case1$Y0))
```

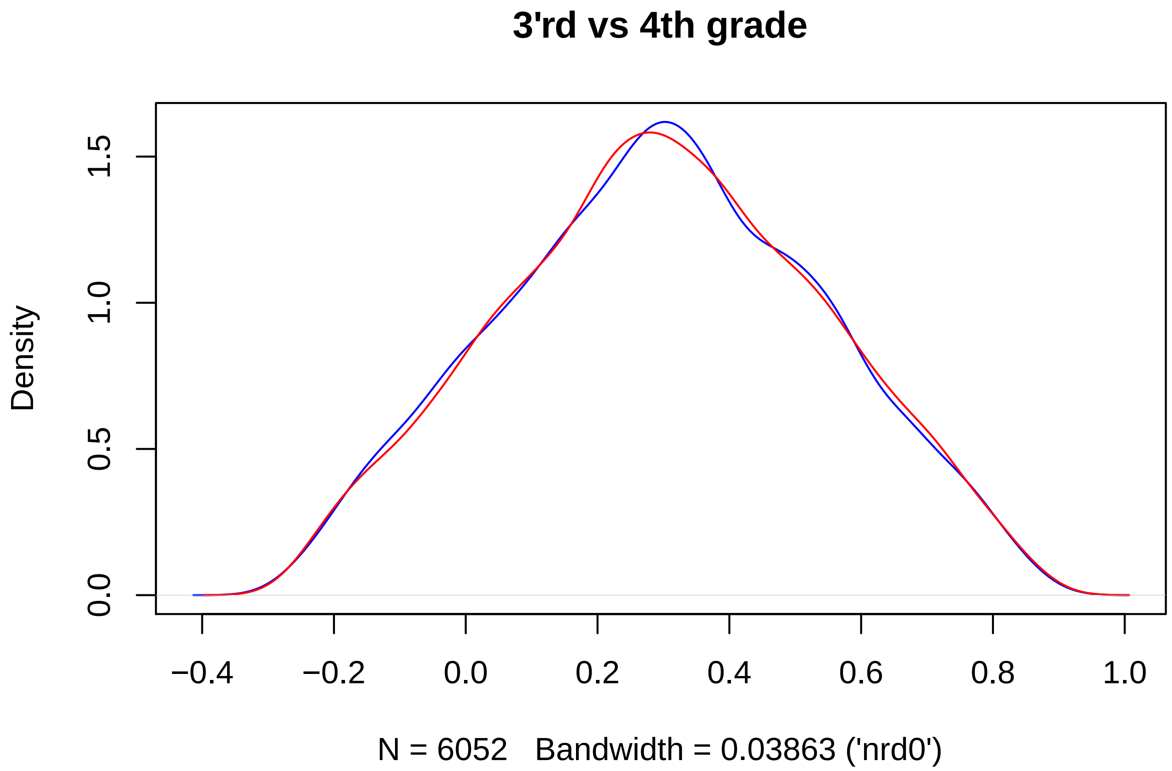
```
## Warning: namespace 'actuar' is not available and has been replaced
## by .GlobalEnv when processing object '<unknown>'
```



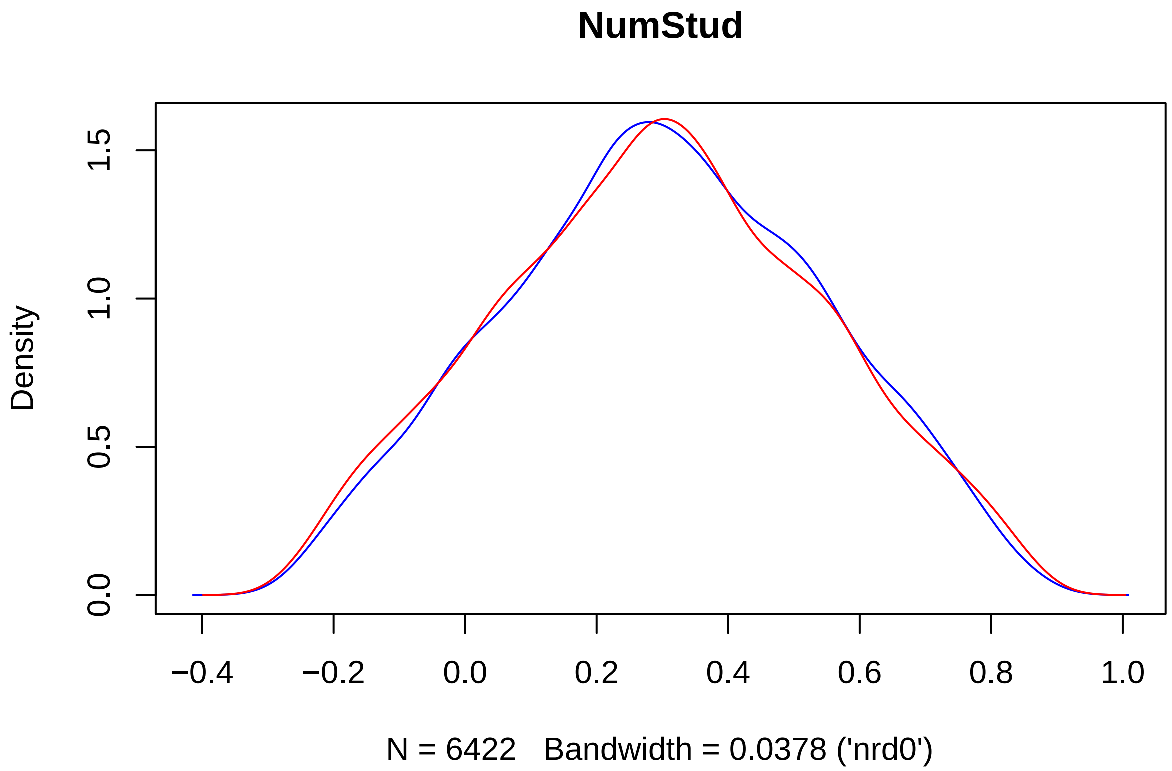
```
plot(kdensity(case1$Y1[case1$male==1]-case1$Y0[case1$male==1]),col="blue",main="Female Vs Male")
lines(kdensity(case1$Y1[case1$male==0]-case1$Y0[case1$male==0]),col="red")
```



```
plot(kdensity(case1$Y1[case1$std==3]-case1$Y0[case1$std==3]),col="blue",main="3’rd vs 4th grade")
lines(kdensity(case1$Y1[case1$std==4]-case1$Y0[case1$std==4]),col="red")
```



```
plot(kdensity(case1$Y1[case1$numstud>=60]-case1$Y0[case1$numstud>=60]),col="blue",main="NumStud")
lines(kdensity(case1$Y1[case1$numstud<60]-case1$Y0[case1$numstud<60]),col="red")
```


NumStud
Density
0.0
0.5
1.0
1.5
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
N = 6422   Bandwidth = 0.0378 ('nrd0')

(l)

Table 9: Regression Results (l)

	Dependent variable:	
	treated	
	Without Robust Std. Errors	With Robust Std. Errors
	(1)	(2)
std	−0.006	−0.006
	(0.009)	(8.987e-03)
male	−0.001	−0.001
	(0.009)	(8.987e-03)
numstud	0.00002	0.00002
	(0.0002)	(1.728e-04)
pre_totnorm	−0.003	−0.003
	(0.007)	(6.514e-03)
income	−0.00001	−0.00001
	(0.0002)	(1.559e-04)
Constant	0.520***	0.520***
	(0.041)	(4.057e-02)
Observations	12,415	12,415
R^2	0.0001	0.0001
Adjusted R^2	−0.0003	−0.0003
Residual Std. Error (df = 12409)	0.500	0.500
F Statistic (df = 5; 12409)	0.238	0.238
Note:		*p<0.1; **p<0.05; ***p<0.01

From our results it looks like there is balance between the treatment group and the control. This is because most of the coefficients are very close to 0. However, we know that few of the results are statistically significant so we cannot be sure that the groups are balanced.

(m)

Table 10: Regression Results (l)

	Dependent variable:
	Finalscore
treated	0.047^{***}
	(0.004)
pre_totnorm	0.999^{***}
	(0.002)
Constant	-0.001
	(0.003)
Observations	12,415
R^2	0.942
Adjusted R^2	0.942
Residual Std. Error	0.249 (df = 12412)
F Statistic	$101{,}181.400^{***}$ (df = 2; 12412)
Note:	*p<0.1; **p<0.05; ***p<0.01

The model we regress comes from regressing our final outcome score on the treatment and then the pretest score. We need MLR.4 which is the zero conditional mean to hold for consistency and causality to be considered. I think that MLR.4 is not credible. There are many different variables that are in the error term and are correlated with the independent variables. For example, how much the student studies indepentely of the study, how much help the child has when studying outside the school, the mental and physical health of the student. We would want to control for these variables to make a causal conclusion.

(n)

We can improve the precision of the estimator by performing a BP test to test for heteroskedasticity. Then we can account for the heteroskedasticity by using robust standard errors to make our results more meaningful.

The results of our BP test yield the following results: BP = 2294, df = 2, p-value < 2.2e-16

Because the p-value is greater than 0.05 we can reject our null hypothesis of homoskedasticity. We can conclude that the model is heteroskedastic and we need use Robust Standard Errors to make our model more meaningful and precise. However, this is one way to make the model more precise to view these standard error look to the table.

(o)

Table 11: Regression Results (n)

	Dependent variable:
	Finalscore
treated	0.047***
	(0.0044686)
pre_totnorm	0.999***
	(0.0022304)
Constant	$-$0.001
	(0.0022013)
Observations	12,415
R^2	0.942
Adjusted R^2	0.942
Residual Std. Error	0.249 (df = 12412)
F Statistic	101,181.400*** (df = 2; 12412)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 12: Regression Results (o)

	Dependent variable:		
	Final Scores	High Income	Low Income
	(1)	(2)	(3)
treated	0.047***	$-$0.774	0.303
	(0.004)	(0.528)	(0.493)
pre_totnorm	0.999***	17.761***	14.862***
	(0.002)	(0.257)	(0.403)
Constant	$-$0.001	175.097***	124.743***
	(0.003)	(0.403)	(0.404)
Observations	12,415	6,104	6,311
R^2	0.942	0.439	0.177
Adjusted R^2	0.942	0.439	0.177
Residual Std. Error	0.249 (df = 12412)	20.623 (df = 6101)	19.586 (df = 6308)
F Statistic	101,181.400*** (df = 2; 12412)	2,385.462*** (df = 2; 6101)	680.199*** (df = 2; 6308)

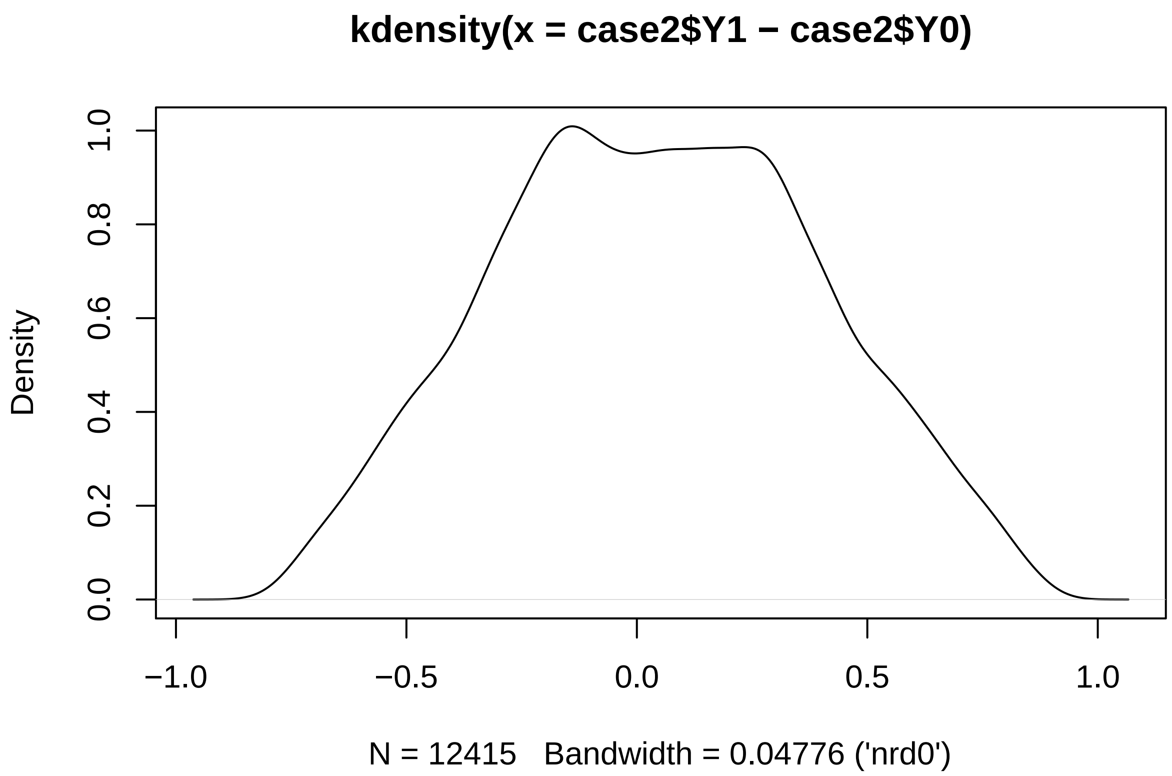
Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses

Table 13: Regression Results (o)

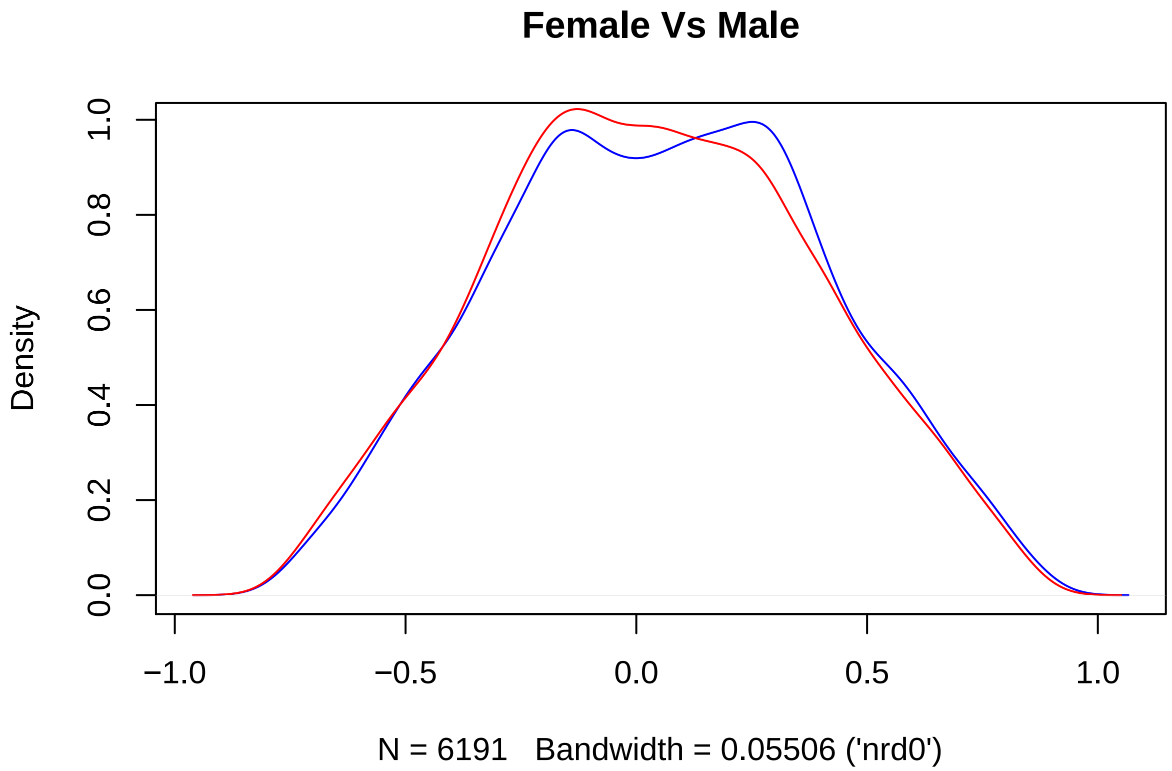
	Dependent variable:		
	High Students	Low Students	Male
	(1)	(2)	(3)
treated	−0.338	0.018	−0.001
	(0.484)	(0.303)	(0.009)
pre_totnorm	1.289***	−0.090	−0.022***
	(0.246)	(0.149)	(0.004)
Constant	85.067***	43.166***	0.500***
	(0.343)	(0.214)	(0.006)
Observations	5,876	6,539	12,415
R^2	0.005	0.0001	0.002
Adjusted R^2	0.004	-0.0002	0.002
Residual Std. Error	18.556 (df = 5873)	12.259 (df = 6536)	0.500 (df = 12412)
F Statistic	13.975*** (df = 2; 5873)	0.183 (df = 2; 6536)	12.068*** (df = 2; 12412)
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses		

(p)

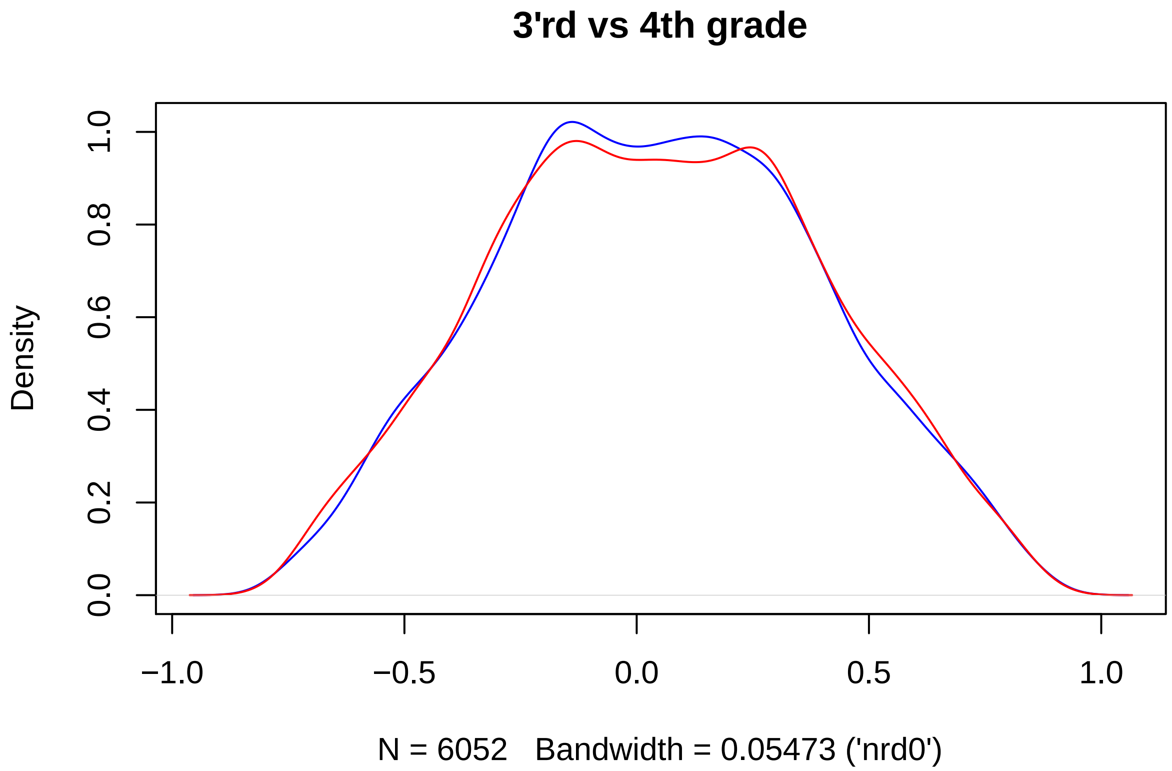
```
plot(kdensity(case2$Y1-case2$Y0))
```



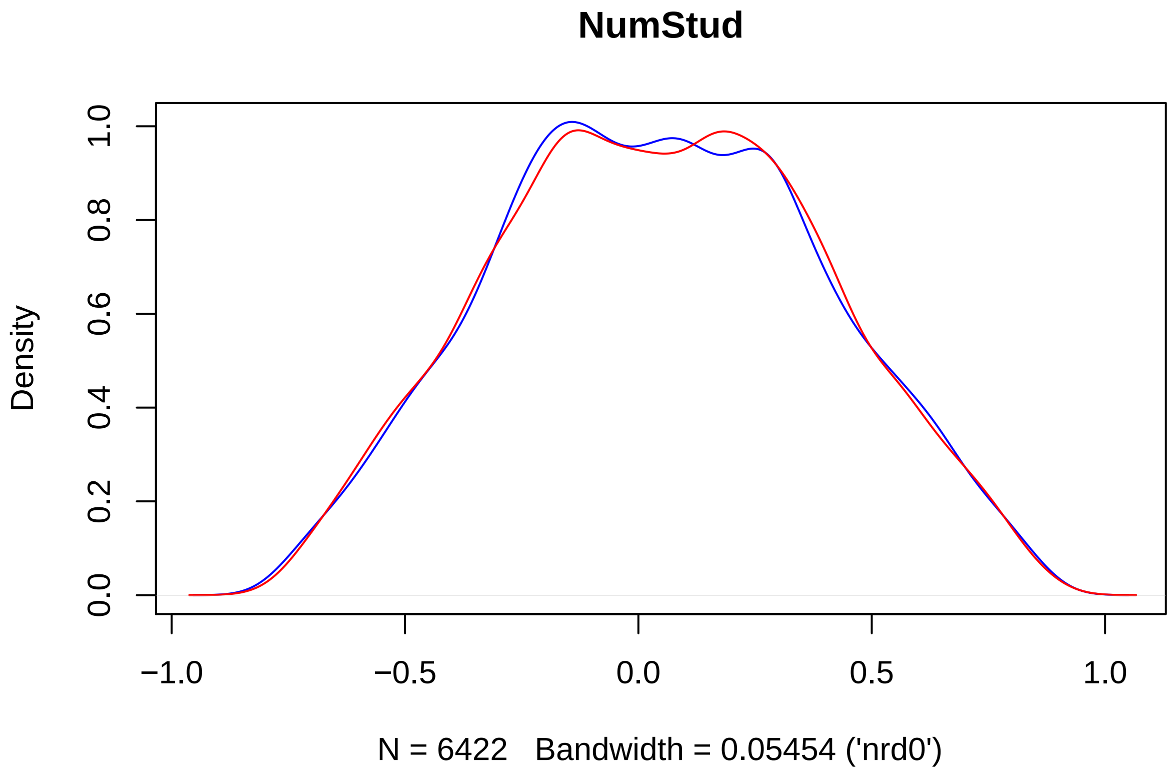
```
plot(kdensity(case2$Y1[case2$male==1]-case2$Y0[case2$male==1]),col="blue",main="Female Vs Male")
lines(kdensity(case2$Y1[case2$male==0]-case2$Y0[case2$male==0]),col="red")
```



```
plot(kdensity(case2$Y1[case2$std==3]-case2$Y0[case2$std==3]),col="blue",main="3'rd vs 4th grade")
lines(kdensity(case2$Y1[case2$std==4]-case2$Y0[case2$std==4]),col="red")
```

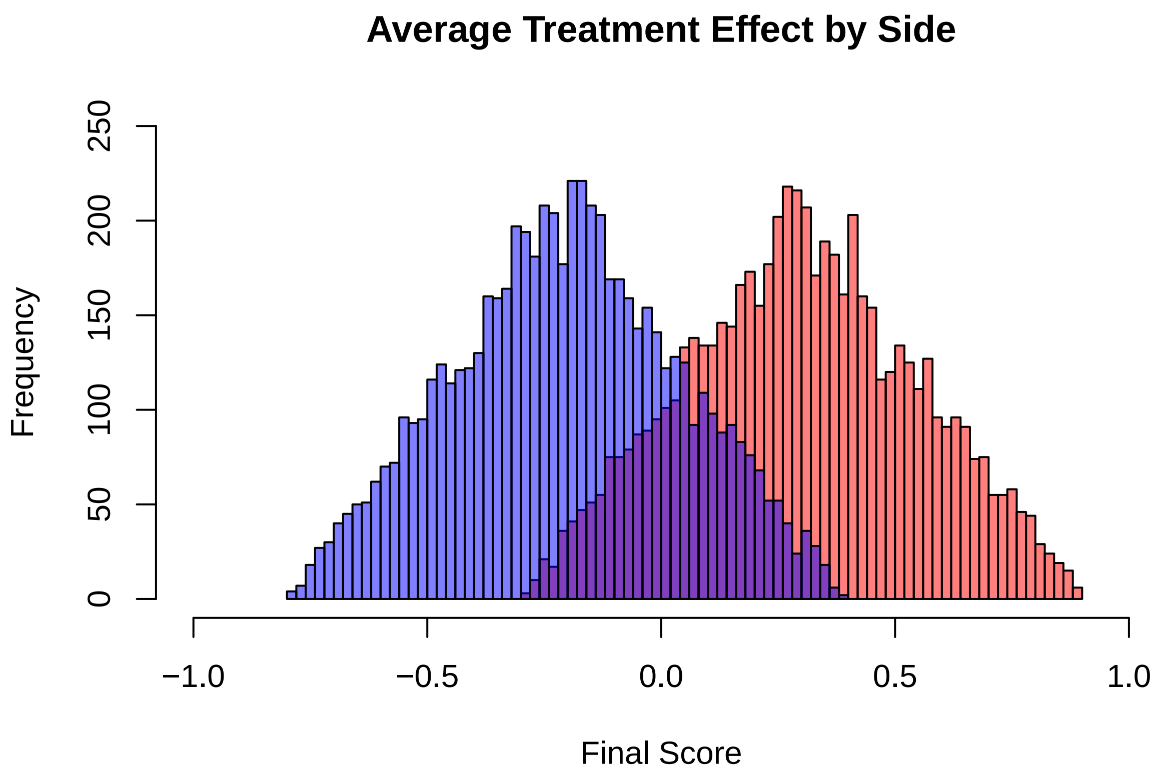


```
plot(kdensity(case2$Y1[case2$numstud>=60]-case2$Y0[case2$numstud>=60]),col="blue",main="NumStud")
lines(kdensity(case2$Y1[case2$numstud<60]-case2$Y0[case2$numstud<60]),col="red")
```



(q)

```r
hist(case2$Y1[case2$side == 1] - case2$Y0[case2$side == 1], main = "Average Treatment Effect by Side",
    col = rgb(1, 0, 0, .5),
    xlab = "Final Score",
    ylab = "Frequency",
    xlim = c(-1, 1),
    ylim = c(0, 250),
    breaks = 50)
hist(case2$Y1[case2$side == 0] - case2$Y0[case2$side == 0],
    col = rgb(0, 0, 1, .5),
    breaks = 50,
    add = TRUE)
```



The red is a presence of side and red is where side is not present.

Table 14: Regression Results (q)

	Dependent variable:
	Finalscore
std	-0.005
	(0.004)
male	-0.0004
	(0.004)
numstud	-0.0001
	(0.0001)
pre_totnorm	0.998^{***}
	(0.003)
income	0.00001
	(0.0001)
treated	0.049^{***}
	(0.004)
side	0.252^{***}
	(0.004)
Constant	-0.106^{***}
	(0.018)
Observations	12,415
R^2	0.957
Adjusted R^2	0.957
Residual Std. Error	0.214 (df = 12407)
F Statistic	$39,490.520^{***}$ (df = 7; 12407)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ Standard errors in parentheses

As we can see in the model above, the addition of side greatly increase the R-squared value and adjusted R-squared value. Because we did not see a significant change in our coefficeints and we saw this increase of fit to our regression model we would want to keep side included in our model.

(r)

The average treament effect does not mean much. As we see from the graph in a the effect of the treatment is not homogeneous with respect to side. If side is present, than the average effect of the treatment is shown to be significantly greater, because students, on average, achieve much higher final scores. Compare this to when side is not present, where the average effect of the treatment is shown to be significantly less, because students, on average, achieve much lower final scores. If side were = to wealth it would follow these previous results. If students have more resources in general then they are more likely to be a good student because they can purchase tutors and better equipment and have more time to study and not work.

(s)

Table 15: Regression Results (s)

	Dependent variable:
	Finalscore
std	−0.005
	(0.004)
male	−0.0004
	(0.004)
numstud	−0.0001
	(0.0001)
pre_totnorm	0.998***
	(0.003)
income	0.00001
	(0.0001)
treated	0.049***
	(0.004)
side	0.252***
	(0.004)
Constant	−0.106***
	(0.018)
Observations	12,415
R^2	0.957
Adjusted R^2	0.957
Residual Std. Error	0.214 (df = 12407)
F Statistic	39,490.520*** (df = 7; 12407)
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses

We can conclude that the two groups are balanced because all of the coefficients from this regression are close to 0 and pretest treated and side are all statistically significant values.

Randomization is not successful. In the regression on assigned treatment, the coefficient for normalized pre test score was not obtained with statistical significance. Although the assignment to treatment may have been random, randomization was not successful. This was due to the significant estimate for the coefficient of the normalized pre-test score variable.

(t)

Table 16: Regression Results (t)

	Dependent variable:
	(TreatmentGroup - treated)
std	-0.008
	(0.006)
male	0.007
	(0.006)
numstud	0.0002
	(0.0001)
pre_totnorm	0.014^{***}
	(0.005)
income	-0.002^{***}
	(0.0001)
Constant	0.490^{***}
	(0.029)
Observations	12,415
R^2	0.049
Adjusted R^2	0.049
Residual Std. Error	0.352 (df = 12409)
F Statistic	129.166^{***} (df = 5; 12409)
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses

Our coefficients here represent the likelihood or probability that someone will dropout of school and in effect the treatment or control group. We can see that a higher pretest scores lead to a higher likelihood that the student will drop out. Conversely higher income means that the student is less likely to drop out.

(u)

When we find the mean final score for the students who were treated we get 0.4668035 and when we get the mean final score for no assigned treatment group we get 0.02937351. We get ITT = 0.4668035 - 0.02937351 = 0.43743. Knowing the ITT coefficient is 0.43743 tells us and other policy makers that the average effect of assigning treatment is 0.43743.

(v)

Table 17: Regression Results (v)

	Dependent variable:	
	Finalscore	
	(1)	(2)
treated	0.516^{***}	
	(0.019)	
pre_totnorm		1.015^{***}
		(0.002)
Constant	-0.049^{***}	0.103^{***}
	(0.011)	(0.002)
Observations	12,415	12,415
R^2	0.055	0.954
Adjusted R^2	0.055	0.954
Residual Std. Error (df = 12413)	1.015	0.224
F Statistic (df = 1; 12413)	729.238^{***}	$258,009.200^{***}$
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses	

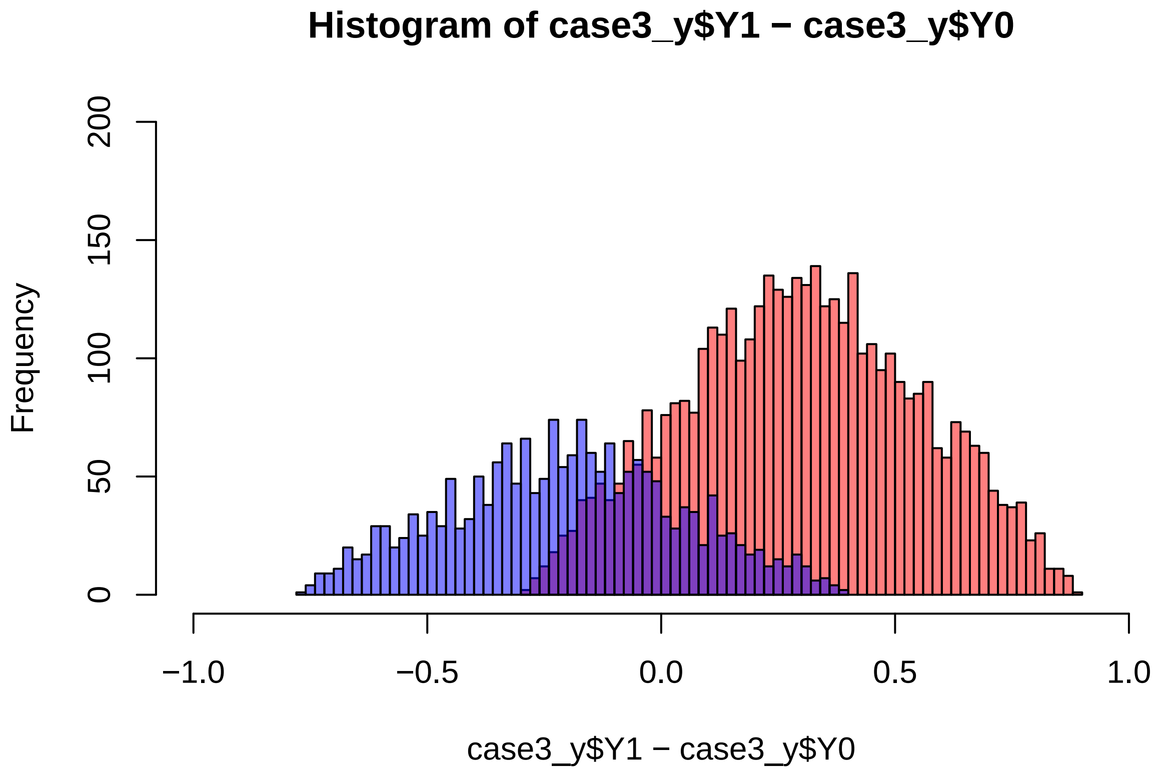
No our estimated treatment effect is biased we already know that there is attrition students are dropping out of the program. There is attrition bias. Additionally, MLR.4 is not very credible here either. In order for the zero conditional mean assumption to hold there must be no variables in the error term that are correlated with the independent variables. To prove this we only need to perform a simple regression of Final Scores on pretest normalized scores to see that they are correlated. If this is the case then there must be variales in the error term that are correlated with pretest scores.

(y)

```r
case3_y <- subset(case3, TreatmentGroup == 1)
case3_y <- subset(case3_y, treated == 1)

hist(case3_y$Y1 - case3_y$Y0,
    col = rgb(1, 0, 0, .5),
    xlim = c(-1, 1),
    ylim = c(0, 200),
    breaks = 50)

case3_y <- subset(case3, TreatmentGroup == 1)
case3_y <- subset(case3_y, treated == 0)
hist(case3_y$Y1 - case3_y$Y0,
    col = rgb(0, 0, 1, .5),
    breaks = 50,
    add = TRUE)
```



The blue are the students who dropped out the red are the students who were eventually treated. Those that were eventually treated by the treatment had better educational outcomes than those who dropped out. These results are not consistent with rationality or rational individuals. For example all rational students want to do well in school and improve there scores on standardized tests so that they can improve their knowledge and get a better higher paying job when they are older. However, children and students are not all equally rational and that is why we see attrition bias. Some students drop out which is consistent with the attrition bias we saw in t.