Applied Micro PSET 3

Matthew Zhao working with Cedric Elkouh

2023-01-16

```r
library(tidyverse)
library(haven)
library(stargazer)
library(xtable)
library(lmtest)
library(sandwich)
library(broom)
library(whitestrap)
options(scipen=999)
```

Question 1

```r
df1 <- read_stata('data/Y1.dta')
```

a)

One of the key motivators that the authors cite is that in developing nations, there has been a large push in making primary school universal and specifically getting students enrolled in primary school. However, the literature reveals that while we are able to get students to attend primary school, this has had little impact on actual learning, with students typically many grade levels behind where they should be in terms of core competencies e.g. literacy and numeracy.

A survey of past studies reveals that additional resources e.g. smaller classroom size or more textbooks, has little effect for the average student but some effect for students who were above average. As a result, the authors in this paper look to seek if supporting weaker students with remedial education, i.e. providing education that is catered towards their actual level of knowledge, and more broadly if resources which target/adjust to a student's actual level of understanding, would produce gains in literacy and numeracy.

b)

```r
# cleaning and standardizing data
year1df <- df1 %>%
  select(!c(starts_with('mid'),contains('norm'))) %>%
  filter(!(is.na(pre_verb) | is.na(pre_math) | is.na(post_verb) | is.na(post_math)))

y1_mean_pre_verb_ctrl <- mean((year1df %>% filter(bal == 0))$pre_verb)
y1_sd_pre_verb_ctrl <- sd((year1df %>% filter(bal == 0))$pre_verb)
```

```
y1_mean_pre_math_ctrl <- mean((year1df %>% filter(bal == 0))$pre_math)
y1_sd_pre_math_ctrl <- sd((year1df %>% filter(bal == 0))$pre_math)
y1_mean_pre_tot_ctrl <- mean((year1df %>% filter(bal == 0))$pre_tot)
y1_sd_pre_tot_ctrl <- sd((year1df %>% filter(bal == 0))$pre_tot)

y1.cleaned <- year1df %>%
  mutate(pre_verbnorm = (pre_verb - y1_mean_pre_verb_ctrl)/y1_sd_pre_verb_ctrl,
         pre_mathnorm = (pre_math - y1_mean_pre_math_ctrl)/y1_sd_pre_math_ctrl,
         pre_totnorm = (pre_tot - y1_mean_pre_tot_ctrl)/y1_sd_pre_tot_ctrl,
         post_verbnorm = (post_verb - y1_mean_pre_verb_ctrl)/y1_sd_pre_verb_ctrl,
         post_mathnorm = (post_math - y1_mean_pre_math_ctrl)/y1_sd_pre_math_ctrl,
         post_totnorm = (post_tot - y1_mean_pre_tot_ctrl)/y1_sd_pre_tot_ctrl
         )
```

```
# regression balance test
bal_test1_mlr <- lm(bal ~ pre_totnorm + male, data = y1.cleaned)
bal_test1_pretotnorm <- lm(bal ~ pre_totnorm, data = y1.cleaned)
bal_test1_male <- lm(bal ~ male, data = y1.cleaned)
models <- list(bal_test1_pretotnorm,bal_test1_male,bal_test1_mlr)
stargazer(models,type='text',digits=3,
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 1 - Balance Test')
```

```
##
## Table 1 - Balance Test
## ===========================================================================
##                                   Dependent variable:
##                    -------------------------------------------------------
##                                          bal
##                         (1)               (2)               (3)
## ---------------------------------------------------------------------------
## pre_totnorm           -0.005                               -0.005
##                       (0.006)                              (0.006)
##
## male                                     0.010              0.010
##                                         (0.011)            (0.011)
##
## Constant              0.485***          0.480***           0.480***
##                       (0.006)           (0.008)            (0.008)
##
## ---------------------------------------------------------------------------
## Observations          8,065             8,065              8,065
## F Statistic  0.873 (df = 1; 8063) 0.831 (df = 1; 8063) 0.823 (df = 2; 8062)
## ===========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

We find that the coefficients are insignificantly different from 0 with the implication that the groups are likely balanced. Standard errors for this analysis should be clustered since treatment is randomized at the school level, hence the SEs should be clustered at the school level.

c)

```r
# table II reproduction
t2_means <- y1.cleaned %>%
  select(c(schoolid,bal,contains('norm'))) %>%
  pivot_longer(cols = c(contains('norm')),names_to = 'score_cat',values_to='score') %>%
  mutate(bal_cat = ifelse(bal == 1,'Treatment','Comparison')) %>%
  group_by(bal_cat,score_cat) %>%
  summarise(means = mean(score),.groups='keep') %>%
  pivot_wider(names_from = bal_cat, values_from = means) %>%
  filter(!str_detect(score_cat,'tot'))

t2_math_pre <- lm(pre_mathnorm ~ bal, data = y1.cleaned)
t2_verb_pre <- lm(pre_verbnorm ~ bal, data = y1.cleaned)
t2_math_post <- lm(post_mathnorm ~ bal, data = y1.cleaned)
t2_verb_post <- lm(post_verbnorm ~ bal, data = y1.cleaned)

t2_math_pre_robust_clustered <- tidy(coeftest(t2_math_pre,
                                     vcov = vcovCL,
                                     type = "HC1",
                                     cluster = ~schoolid))
t2_verb_pre_robust_clustered <- tidy(coeftest(t2_verb_pre,
                                     vcov = vcovCL,
                                     type = "HC1",
                                     cluster = ~schoolid))
t2_math_post_robust_clustered <- tidy(coeftest(t2_math_post,
                                     vcov = vcovCL,
                                     type = "HC1",
                                     cluster = ~schoolid))
t2_verb_post_robust_clustered <- tidy(coeftest(t2_verb_post,
                                     vcov = vcovCL,
                                     type = "HC1",
                                     cluster = ~schoolid))

#rows <- c('t2_math_pre','t2_verb_pre','t2_math_post','t2_verb_post')
t2_regs <- bind_rows(t2_math_post_robust_clustered,t2_verb_post_robust_clustered, t2_math_pre_robust_clu
tableII <- bind_cols(t2_means,t2_regs) %>% mutate(across(where(is.numeric), round, 4))
tableII
```

```
## # A tibble: 4 x 5
## # Groups:   score_cat [4]
##   score_cat     Comparison Treatment estimate std.error
##   <chr>              <dbl>     <dbl>    <dbl>     <dbl>
## 1 post_mathnorm      0.125     0.271    0.146    0.0843
## 2 post_verbnorm      0.608     0.711    0.103    0.0797
## 3 pre_mathnorm       0        -0.0447  -0.0447   0.0916
## 4 pre_verbnorm       0         0.0089   0.0089   0.0742
```

I was able to get similar estimates as Table II of the paper. The estimated treatment effect is 0.146 standard deviations for math and 0.103 standard deviations for language. These seem fairly small, and given the size of the robust clustered standard errors, we cannot say that the treatment has a positive and statistically significant effect on student scores.

d)

```r
t3_math <- lm((post_mathnorm - pre_mathnorm) ~ bal + pre_totnorm, data=y1.cleaned)
t3_verb <- lm((post_verbnorm - pre_verbnorm) ~ bal + pre_totnorm, data=y1.cleaned)
t3_tot <- lm((post_totnorm - pre_totnorm) ~ bal + pre_totnorm, data=y1.cleaned)

t3_math_robust_clustered <- coeftest(t3_math,
                                     vcov = vcovCL,
                                     type = "HC1",
                                     cluster = ~schoolid)
t3_verb_robust_clustered <- coeftest(t3_verb,
                                     vcov = vcovCL,
                                     type = "HC1",
                                     cluster = ~schoolid)
t3_tot_robust_clustered <- coeftest(t3_tot,
                                    vcov = vcovCL,
                                    type = "HC1",
                                    cluster = ~schoolid)
models <- list(t3_math_robust_clustered,
               t3_verb_robust_clustered,
               t3_tot_robust_clustered)
stargazer(models,type='text',digits=3,
          dep.var.labels=c('Math','Language','Total'),
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 3 - Table III replication: Vadodara Y1 pooled')
```

```
##
## Table 3 - Table III replication: Vadodara Y1 pooled
## ========================================
##                  Dependent variable:
##             ----------------------------
##                         Math
##              (1)       (2)       (3)
## ----------------------------------------
## bal         0.186***  0.092*   0.150***
##             (0.043)   (0.047)  (0.041)
##
## pre_totnorm -0.214*** -0.129*** -0.184***
##             (0.017)   (0.017)  (0.017)
##
## Constant    0.125***  0.608*** 0.368***
##             (0.035)   (0.038)  (0.035)
##
## ========================================
## ========================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```

Compared with the previous simple comparison of means, we find that there is a positive and statistically significant effect of treatment across outcome categories, with all being significant at the 10% level and math and total score being significant at the 1% level. Specifically, we see that controlling for normalized total pretest score improves the precision of estimates of the treatment effect.

e)

```
df2 <- read_stata('data/Y2.dta')

df_bothyr <- left_join(df1 %>%
                    select(!contains('post')),
                df2 %>%
                    select(!contains('pre')),
                suffix=c("",".y"), by='studentid',) %>%
  select(!c(starts_with('mid'),contains('norm'),ends_with('.y'))) %>%
  filter(!(is.na(pre_verb) | is.na(pre_math) | is.na(post_verb) | is.na(post_math)))

bothyr_mean_pre_verb_ctrl <- mean((df_bothyr %>% filter(bal == 0))$pre_verb)
bothyr_sd_pre_verb_ctrl <- sd((df_bothyr %>% filter(bal == 0))$pre_verb)
bothyr_mean_pre_math_ctrl <- mean((df_bothyr %>% filter(bal == 0))$pre_math)
bothyr_sd_pre_math_ctrl <- sd((df_bothyr %>% filter(bal == 0))$pre_math)
bothyr_mean_pre_tot_ctrl <- mean((df_bothyr %>% filter(bal == 0))$pre_tot)
bothyr_sd_pre_tot_ctrl <- sd((df_bothyr %>% filter(bal == 0))$pre_tot)

bothyr.cleaned <- df_bothyr %>%
  mutate(pre_verbnorm = (pre_verb - bothyr_mean_pre_verb_ctrl)/bothyr_sd_pre_verb_ctrl,
         pre_mathnorm = (pre_math - bothyr_mean_pre_math_ctrl)/bothyr_sd_pre_math_ctrl,
         pre_totnorm = (pre_tot - bothyr_mean_pre_tot_ctrl)/bothyr_sd_pre_tot_ctrl,
         post_verbnorm = (post_verb - bothyr_mean_pre_verb_ctrl)/bothyr_sd_pre_verb_ctrl,
         post_mathnorm = (post_math - bothyr_mean_pre_math_ctrl)/bothyr_sd_pre_math_ctrl,
         post_totnorm = (post_tot - bothyr_mean_pre_tot_ctrl)/bothyr_sd_pre_tot_ctrl
         )

bothyr_t3_math <- lm((post_mathnorm - pre_mathnorm) ~ bal + pre_totnorm, data=bothyr.cleaned)
bothyr_t3_verb <- lm((post_verbnorm - pre_verbnorm) ~ bal + pre_totnorm, data=bothyr.cleaned)
bothyr_t3_tot <- lm((post_totnorm - pre_totnorm) ~ bal + pre_totnorm, data=bothyr.cleaned)

bothyr_t3_math_robust_clustered <- coeftest(bothyr_t3_math,
                                    vcov = vcovCL,
                                    type = "HC1",
                                    cluster = ~schoolid)
bothyr_t3_verb_robust_clustered <- coeftest(bothyr_t3_verb,
                                    vcov = vcovCL,
                                    type = "HC1",
                                    cluster = ~schoolid)
bothyr_t3_tot_robust_clustered <- coeftest(bothyr_t3_tot,
                                    vcov = vcovCL,
                                    type = "HC1",
                                    cluster = ~schoolid)
models <- list(bothyr_t3_math_robust_clustered,
               bothyr_t3_verb_robust_clustered,
               bothyr_t3_tot_robust_clustered)
stargazer(models,type='text',digits=3,
          dep.var.labels=c('Math','Language','Total'),
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 3 - Table III replication: Vadodara Two-year pooled')
```

##

```
## Table 3 - Table III replication: Vadodara Two-year pooled
## ==========================================
##                      Dependent variable:
##              -----------------------------
##                            Math
##                (1)        (2)        (3)
## ------------------------------------------
## bal          0.321***   0.157*    0.260***
##              (0.109)    (0.088)    (0.095)
##
## pre_totnorm -0.401*** -0.293*** -0.372***
##              (0.030)    (0.031)    (0.028)
##
## Constant     1.134***   1.329***   1.300***
##              (0.073)    (0.061)    (0.065)
##
## ==========================================
## ==========================================
## Note:           *p<0.1; **p<0.05; ***p<0.01
```

We find similar two-year effects as the paper. Our coefficients are larger now at similar significance levels to part d. Our estimated two-year treatment effect is larger than the one-year treatment effect which makes sense given that the study was designed so that students who received treatment in year 1 in grade 3 would continue to do so in year 2 in grade 4.

f)

Internal Validity: could be issues with selection e.g. randomized at school level but could be the case that certain areas are more wealthy and this could result in schools being skewed leading to biased effects?

Another more likely one is partial compliance since we see that in Mumbai not all schools assigned for treatment actually got it since there were administrative issues. Given the reason for not getting treatment, it could be the case that these schools are systematically lower quality than the schools that did receive treatment. This could have potentially biased the estimated effect upwards if students in these schools were more likely to get less benefit from the program.

There are a larger number of threats to external validity. In particular, since the experiment was conducted in large cities in India, there are doubts as to whether this could generalize to local regions where finding balsakhi may be more difficult. This could be due to administrative burden like in Mumbai or because there are fewer educated young women in less populated areas.

Furthermore, we do not know if this program could be used in other developing nations to support learning, since it relies on the presence of balsakhi or at least some substitute to keep the program going.

g)

```
case1 <- read_stata('data/case1.dta')
```

```
# regression balance test
bal_case1_mlr <- lm(treated ~ pre_totnorm + male + income + numstud, data = case1)
bal_case1_pretotnorm <- lm(treated ~ pre_totnorm, data = case1)
bal_case1_male <- lm(treated ~ male, data = case1)
```

```
bal_case1_inc <- lm(treated ~ income, data = case1)
bal_case1_numstud <- lm(treated ~ numstud, data = case1)

models <- list(bal_case1_numstud,bal_case1_inc,bal_case1_male)
stargazer(models,type='text',digits=3,column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 4 - Case 1 Balance Test')
```

```
##
## Table 4 - Case 1 Balance Test
## =========================================================
##                               Dependent variable:
##                       ----------------------------------
##                                      treated
##                           (1)        (2)        (3)
## ---------------------------------------------------------
## numstud                  0.0001
##                         (0.0002)
##
## income                              0.002***
##                                    (0.0001)
##
## male                                          -0.004
##                                               (0.009)
##
## Constant                0.590***   0.307***   0.601***
##                         (0.012)    (0.016)    (0.006)
##
## ---------------------------------------------------------
## Observations            12,415     12,415     12,415
## F Statistic (df = 1; 12413)  0.765  350.630***  0.214
## =========================================================
## Note:                         *p<0.1; **p<0.05; ***p<0.01
```

```
models <- list(bal_case1_pretotnorm,bal_case1_mlr)
stargazer(models,type='text',digits=3,column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 4 (cont) - Case 1 Balance Test')
```

```
##
## Table 4 (cont) - Case 1 Balance Test
## ================================================================
##                            Dependent variable:
##              --------------------------------------------------
##                                   treated
##                        (1)                        (2)
## ----------------------------------------------------------------
## pre_totnorm          0.053***                   -0.012**
##                      (0.004)                     (0.006)
##
## male                                             0.0005
##                                                  (0.009)
##
```

7

```
## income                                          0.002***
##                                                 (0.0002)
##
## numstud                                         0.0002
##                                                 (0.0002)
##
## Constant                    0.597***            0.260***
##                             (0.004)             (0.026)
##
## ----------------------------------------------------------------
## Observations             12,415                  12,415
## F Statistic  147.074*** (df = 1; 12413) 89.149*** (df = 4; 12410)
## ================================================================
## Note:                                 *p<0.1; **p<0.05; ***p<0.01
```

Here, we identify a potential issue with randomization where it appears that the treatment may not have been perfectly randomized over income, resulting in a significant coefficient and F-stat. Normalized total pre-test score also seems to have a similar relationship which could be explained by the design of the program but should be noted. The other variables appear to be balanced.

h)

```
case1_base <- lm((Finalscore - pre_totnorm) ~ treated, data=case1)
case1_ov <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + income, data=case1)
models <- list(case1_base,case1_ov)
stargazer(models,type='text',digits=3,
          column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 5 - Case 1 Estimate of Causal Effect')
```

```
##
## Table 5 - Case 1 Estimate of Causal Effect
## ==========================================================================
##                                 Dependent variable:
##                    ------------------------------------------------------
##                              (Finalscore - pre_totnorm)
##                          (1)                         (2)
## --------------------------------------------------------------------------
## treated                0.298***                    0.298***
##                        (0.003)                     (0.003)
##
## pre_totnorm                                        -0.002
##                                                    (0.002)
##
## income                                             0.00001
##                                                    (0.0001)
##
## Constant               0.001                       -0.001
##                        (0.002)                     (0.008)
##
## --------------------------------------------------------------------------
## Observations           12,415                      12,415
```

```
## F Statistic  8,821.418*** (df = 1; 12413) 2,940.522*** (df = 3; 12411)
## =======================================================================
## Note:                                              *p<0.1; **p<0.05; ***p<0.01
```

I estimate the causal effect of treatment on score to be around 0.3 standard deviations. I show that controlling for other pre-treatment variables noted in the balance test does not change the effect from the baseline regression. Since the estimate of the causal effect is robust to these variables, I believe that this estimate is the causal effect of interest to the extent that we have shown randomization and can assume the treatment is uncorrelated with unobservables from that. We are also told there is no attrition.

i)

```r
case1_base <- lm((Finalscore - pre_totnorm) ~ treated, data=case1)
case1_ov <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + income, data=case1)
case1_ctrl <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + income + male + std + numstud, da

# checking for heteroskedasticity
bptest(case1_base)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  case1_base
## BP = 0.0039828, df = 1, p-value = 0.9497
```

```r
white_test(case1_base)
```

```
## White's test results
##
## Null hypothesis: Homoskedasticity of the residuals
## Alternative hypothesis: Heteroskedasticity of the residuals
## Test Statistic: 0
## P-value: 0.998011
```

```r
models <- list(case1_base,case1_ov,case1_ctrl)
stargazer(models,type='text',digits=3,
          column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 6 - Case 1 Alternative Specifications')
```

```
##
## Table 6 - Case 1 Alternative Specifications
## ==================================================================================================
##                                         Dependent variable:
##                     ------------------------------------------------------------------------------
##                                      (Finalscore - pre_totnorm)
##                             (1)                      (2)                          (3)
## ------------------------------------------------------------------------------------------
## treated                  0.298***                 0.298***                     0.298***
##                          (0.003)                  (0.003)                      (0.003)
```

9

```
## 
## pre_totnorm                                                    -0.002                        -0.002
##                                                                (0.002)                       (0.002)
## 
## income                                                         0.00001                       0.00001
##                                                               (0.0001)                      (0.0001)
## 
## male                                                                                         -0.001
##                                                                                              (0.003)
## 
## std                                                                                           0.001
##                                                                                              (0.003)
## 
## numstud                                                                                     -0.00001
##                                                                                             (0.0001)
## 
## Constant                      0.001                            -0.001                        -0.001
##                              (0.002)                           (0.008)                       (0.014)
## 
## ----------------------------------------------------------------------------------------------------
## Observations                 12,415                            12,415                         12,415
## F Statistic  8,821.418*** (df = 1; 12413) 2,940.522*** (df = 3; 12411) 1,469.965*** (df = 6; 12408)
## ====================================================================================================
## Note:                                                                    *p<0.1; **p<0.05; ***p<0.01
```

We do not see any heteroskedasticity based on the BP and White tests. Furthermore, we cannot improve
the precision of our estimate with additional controls.

j)

```
case1_hte_inc <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + male + income*treated, data=ca
case1_hte_male <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + income + male*treated, data=ca
case1_hte_full <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + income*treated + male*treated


models <- list(case1_hte_inc,case1_hte_male,case1_hte_full)
stargazer(models,type='text',digits=3,
          column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 6 (cont) - Case 1 Alternative Specifications with HTE')
```

```
## 
## Table 6 (cont) - Case 1 Alternative Specifications with HTE
## ===================================================================================================
##                                             Dependent variable:
##              ------------------------------------------------------------------------------------
##                                          (Finalscore - pre_totnorm)
##                         (1)                             (2)                             (3)
## ---------------------------------------------------------------------------------------------------
## treated              0.301***                         0.297***                        0.299***
##                      (0.012)                          (0.005)                         (0.012)
## 
```

```
## pre_totnorm              -0.002                  -0.002                   -0.002
##                          (0.002)                 (0.002)                  (0.002)
##
## male                     -0.001                  -0.003                   -0.003
##                          (0.003)                 (0.005)                  (0.005)
##
## treated:male                                      0.002                    0.002
##                                                  (0.006)                  (0.006)
##
## income                    0.00002                 0.00001                  0.00002
##                          (0.0001)                (0.0001)                 (0.0001)
##
## treated:income           -0.00002                                        -0.00001
##                          (0.0001)                                        (0.0001)
##
## Constant                 -0.001                   0.001                   -0.001
##                          (0.011)                 (0.009)                  (0.011)
##
## --------------------------------------------------------------------------------
## Observations             12,415                  12,415                   12,415
## F Statistic   1,764.088*** (df = 5; 12409) 1,764.126*** (df = 5; 12409) 1,469.997*** (df = 6; 12408)
## ================================================================================
## Note:                                                    *p<0.1; **p<0.05; ***p<0.0
```

Using interactions between treatment and various controls, we test for potential heterogeneous treatment
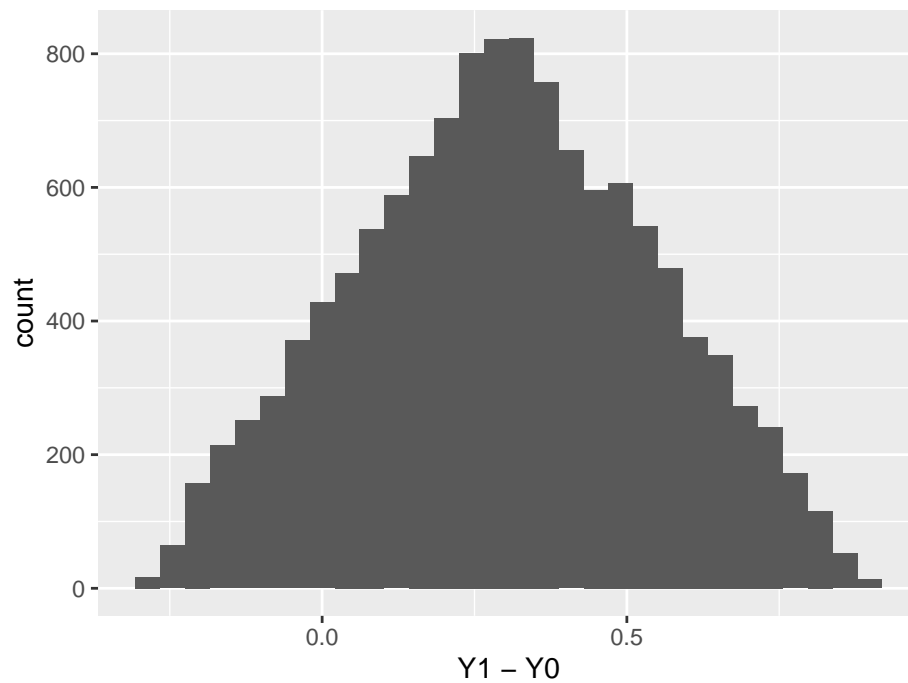effects and find little evidence that this is the case.

k)

```
# true effect
mean(case1$Y1 - case1$Y0)
```
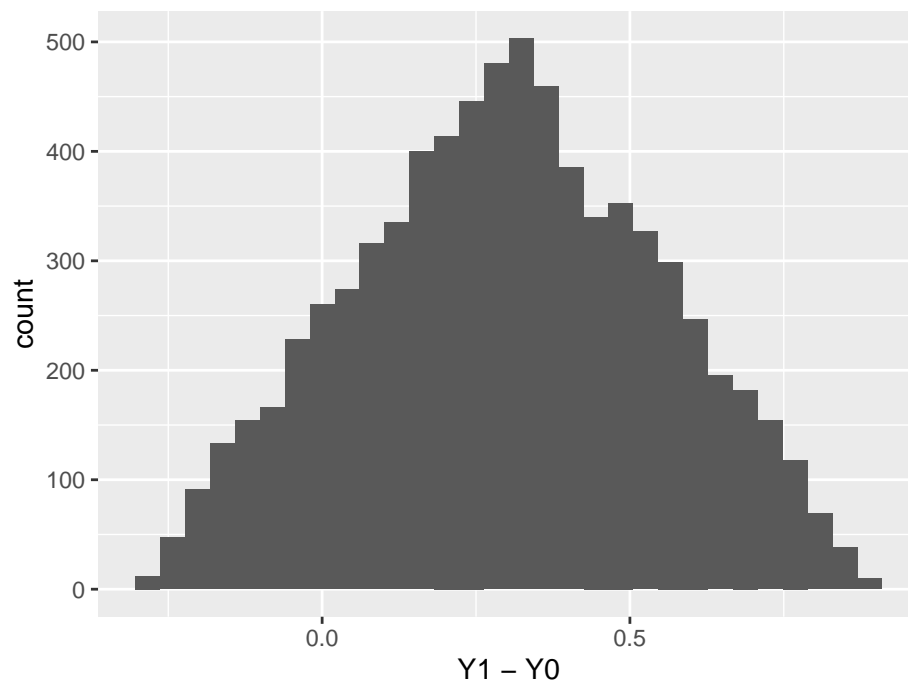
```
## [1] 0.2990752
```

```
ggplot(data = case1, mapping=aes(x=Y1 - Y0)) +
  geom_histogram(bins=30) +
  ggtitle('Treatment effect for whole sample')
```
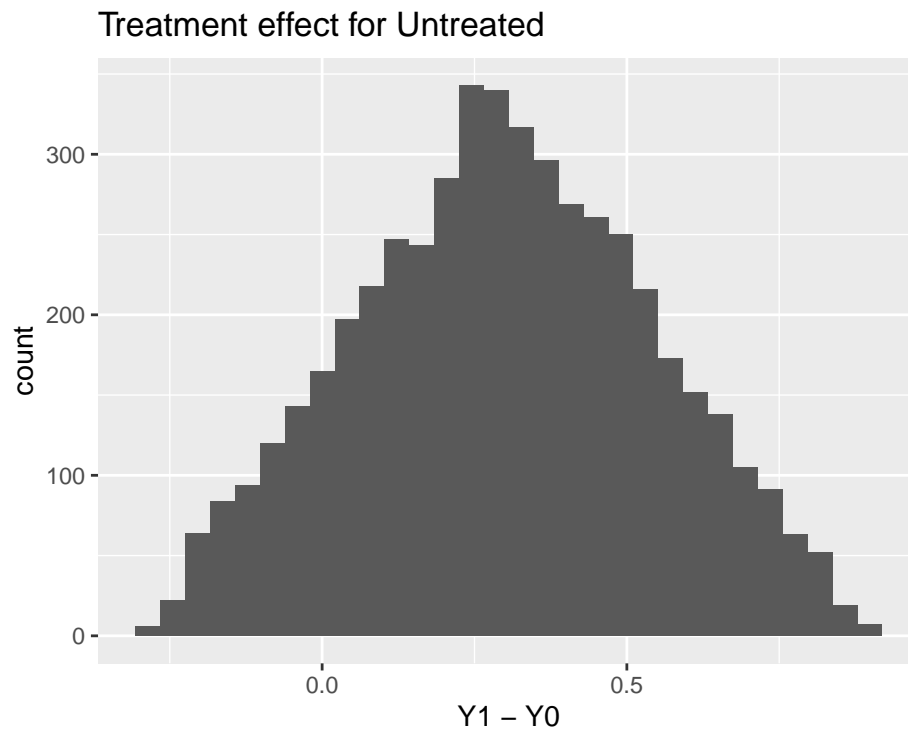
Treatment effect for whole sample


```
ggplot(data = case1 %>% filter(treated == 1), mapping=aes(x=Y1 - Y0)) +
  geom_histogram(bins=30) +
  ggtitle('Treatment effect for Treated')
```

Treatment effect for Treated

```
ggplot(data = case1 %>% filter(treated == 0), mapping=aes(x=Y1 - Y0)) +
  geom_histogram(bins=30) +
  ggtitle('Treatment effect for Untreated')
```

Treatment effect for Untreated


```
case1 %>%
  mutate(trt_effect = cut(Y1-Y0,10)) %>%
  group_by(trt_effect) %>%
  count()
```

```
## # A tibble: 10 x 2
## # Groups:   trt_effect [10]
##    trt_effect           n
##    <fct>            <int>
##  1 (-0.296,-0.177]    266
##  2 (-0.177,-0.0581]   750
##  3 (-0.0581,0.0604]  1240
##  4 (0.0604,0.179]    1701
##  5 (0.179,0.297]     2242
##  6 (0.297,0.416]     2204
##  7 (0.416,0.534]     1702
##  8 (0.534,0.653]     1260
##  9 (0.653,0.771]      763
## 10 (0.771,0.891]      287
```

We see that the treatment effect is roughly normally distributed across the sample. Furthermore, we see little evidence of sorting or any selection in relation to the treatment. Since our subgroup analysis in part j revealed little difference across observable groups, we do not further investigate these variables.

l)

```
case2 <- read_stata('data/case2.dta')
```

```
# regression balance test
bal_case2_mlr <- lm(treated ~ pre_totnorm + male + income + numstud + side, data = case2)
bal_case2_pretotnorm <- lm(treated ~ pre_totnorm, data = case2)
bal_case2_side <- lm(treated ~ side, data = case2)
bal_case2_male <- lm(treated ~ male, data = case2)
bal_case2_inc <- lm(treated ~ income, data = case2)
bal_case2_numstud <- lm(treated ~ numstud, data = case2)

models <- list(bal_case2_numstud,bal_case2_inc,bal_case2_male)
stargazer(models,type='text',digits=3,column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 7 - Case 2 Balance Test')
```

```
##
## Table 7 - Case 2 Balance Test
## ============================================================
##                              Dependent variable:
##                     ----------------------------------------
##                                     treated
##                        (1)          (2)          (3)
## ------------------------------------------------------------
## numstud               0.00003
##                      (0.0002)
##
## income                            -0.0001
##                                   (0.0001)
##
## male                                         -0.001
##                                              (0.009)
##
## Constant             0.497***    0.510***    0.499***
##                      (0.012)     (0.017)     (0.006)
##
## ------------------------------------------------------------
## Observations          12,415       12,415      12,415
## F Statistic (df = 1; 12413)  0.024   0.487       0.010
## ============================================================
## Note:                       *p<0.1; **p<0.05; ***p<0.01
```

```
models <- list(bal_case2_pretotnorm,bal_case2_side,bal_case2_mlr)
stargazer(models,type='text',digits=3,column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 7 (cont) - Case 2 Balance Test')
```

```
##
## Table 7 (cont) - Case 2 Balance Test
## ==========================================================================
##                                   Dependent variable:
```

14

```
##                  ----------------------------------------------------------------
##                                         treated
##                      (1)                 (2)                 (3)
##                  ----------------------------------------------------------------
## pre_totnorm        -0.004                                   -0.003
##                    (0.004)                                  (0.007)
##
## male                                                        -0.001
##                                                             (0.009)
##
## income                                                      -0.00001
##                                                             (0.0002)
##
## numstud                                                     0.00002
##                                                             (0.0002)
##
## side                                     -0.008             -0.008
##                                          (0.009)            (0.009)
##
## Constant            0.499***            0.503***            0.504***
##                    (0.004)              (0.006)            (0.027)
##
##                  ----------------------------------------------------------------
## Observations        12,415              12,415              12,415
## F Statistic  0.758 (df = 1; 12413) 0.793 (df = 1; 12413) 0.316 (df = 5; 12409)
##                  ================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Since the F-stats are all small and insignificant, we can conclude that the groups are balanced.

m)

```
case2_base <- lm((Finalscore - pre_totnorm) ~ treated, data=case2)
case2_unobs <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + income, data=case2)
case2_ov <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + income + side, data=case2)

stargazer(list(case2_base,case2_unobs,case2_ov),type='text',digits=4,
        column.sep.width = "-15pt",
        omit.stat=c('rsq','adj.rsq',"ser"),
        title='Table 8 - Case 2 Estimate of Causal Effect')
```

```
##
## Table 8 - Case 2 Estimate of Causal Effect
## ============================================================================================
##                                          Dependent variable:
##                  --------------------------------------------------------------------------
##                                        (Finalscore - pre_totnorm)
##                      (1)                     (2)                         (3)
##                  --------------------------------------------------------------------------
## treated            0.0467***               0.0467***                   0.0487***
##                    (0.0045)                (0.0045)                    (0.0038)
##
```

15

```
## pre_totnorm                                        -0.0016                    -0.0020
##                                                     (0.0032)                   (0.0028)
##
## income                                              0.00002                    0.00001
##                                                     (0.0001)                   (0.0001)
##
## side                                                                           0.2518***
##                                                                                (0.0038)
##
## Constant                      -0.0009               -0.0044                    -0.1285***
##                               (0.0032)              (0.0120)                   (0.0105)
##
## ------------------------------------------------------------------------------------------
## Observations                   12,415                12,415                     12,415
## F Statistic   109.4783*** (df = 1; 12413) 36.5659*** (df = 3; 12411) 1,107.4050*** (df = 4; 12410)
## ==========================================================================================
## Note:                                                        *p<0.1; **p<0.05; ***p<0.01
```

```r
# checking for heteroskedasticity
print('Base Case:')
```

```
## [1] "Base Case:"
```

```r
bptest(case2_base)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  case2_base
## BP = 2293, df = 1, p-value < 0.00000000000000022
```

```r
white_test(case2_base)
```

```
## White's test results
##
## Null hypothesis: Homoskedasticity of the residuals
## Alternative hypothesis: Heteroskedasticity of the residuals
## Test Statistic: 2292.99
## P-value: 0
```

```r
print('With Controls:')
```

```
## [1] "With Controls:"
```

```r
bptest(case2_unobs)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  case2_unobs
## BP = 2295.3, df = 3, p-value < 0.00000000000000022
```

```
white_test(case2_unobs)
```

```
## White's test results
##
## Null hypothesis: Homoskedasticity of the residuals
## Alternative hypothesis: Heteroskedasticity of the residuals
## Test Statistic: 2288.78
## P-value: 0
```

```
print('With Controls + side:')
```

```
## [1] "With Controls + side:"
```

```
bptest(case2_ov)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  case2_ov
## BP = 2.6669, df = 4, p-value = 0.615
```

```
white_test(case2_ov)
```

```
## White's test results
##
## Null hypothesis: Homoskedasticity of the residuals
## Alternative hypothesis: Heteroskedasticity of the residuals
## Test Statistic: 3.12
## P-value: 0.209972
```

```
case2_base_robust <- coeftest(case2_base, vcov = vcovHC(case2_base, type = 'HC1'))
case2_unobs_robust <- coeftest(case2_unobs, vcov = vcovHC(case2_unobs, type = 'HC1'))

round(case2_base_robust,4)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value          Pr(>|t|)
## (Intercept)  -0.0009     0.0022   -0.42            0.6745
## treated       0.0467     0.0045   10.45 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
round(case2_unobs_robust,4)
```

```
##
## t test of coefficients:
##
```

```
##             Estimate Std. Error t value           Pr(>|t|)
## (Intercept)  -0.0044     0.0118 -0.3703             0.7112
## treated       0.0467     0.0045 10.4444 <0.0000000000000002 ***
## pre_totnorm  -0.0016     0.0033 -0.4748             0.6350
## income        0.0000     0.0001  0.3001             0.7641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After examining the inital regression results, I testd for heteroskedasticity using the BP and White tests. I find that the tests strongly reject the null of homoskedasticity for both the baseline model and the model with observed controls As a result, I re-estimate the models with robust standard errors, finding negligible change.

I estimate the causal effect of treatment on score to be around 0.0467 standard deviations. After using robust standard errors, the baseline model should no longer exhibit heteroskedasticity, with the standard errors and estimate of treatment effect not changing. Furthermore, I believe that the estimate of β_D is consistent here because side is correlated with the outcome (Finalscore-pre_totnorm) but is uncorrelated with treated.

n)

```
case2_base <- lm((Finalscore - pre_totnorm) ~ treated, data=case2)
case2_ov <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + income + side, data=case2)
case2_ctrl <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + income + male + std + numstud + s:
```

```
models <- list(case2_base,case2_ov,case2_ctrl)
stargazer(models,type='text',digits=4,
          column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 9 - Case 2 Alternative Specifications')
```

```
##
## Table 9 - Case 2 Alternative Specifications
## ==============================================================================================
##                                         Dependent variable:
##                  ---------------------------------------------------------------------------
##                                       (Finalscore - pre_totnorm)
##                         (1)                      (2)                      (3)
## --------------------------------------------------------------------------------------------
## treated              0.0467***                0.0487***                0.0487***
##                       (0.0045)                 (0.0038)                 (0.0038)
##
## pre_totnorm                                    -0.0020                  -0.0021
##                                                (0.0028)                 (0.0028)
##
## income                                         0.00001                  0.00001
##                                                (0.0001)                 (0.0001)
##
## male                                                                    -0.0004
##                                                                         (0.0039)
##
## std                                                                     -0.0047
##                                                                         (0.0039)
```

```
## 
## numstud                                                             -0.0001
##                                                                     (0.0001)
## 
## side                                            0.2518***            0.2518***
##                                                 (0.0038)             (0.0038)
## 
## Constant              -0.0009                  -0.1285***           -0.1064***
##                       (0.0032)                 (0.0105)             (0.0176)
## 
## ----------------------------------------------------------------------------------
## Observations          12,415                   12,415               12,415
## F Statistic  109.4783*** (df = 1; 12413) 1,107.4050*** (df = 4; 12410) 633.2490*** (df = 7; 12407)
## ==================================================================================
## Note:                                                    *p<0.1; **p<0.05; ***p<0.01
```

We can improve the precision of our estimate of the treatment effect by adding in side since it uniquely accounts for a portion of the variation in our outcome since it is correlated with Y but not D or any other controls.

o)

```
case2_hte_inc <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + male + side + income*treated, 
case2_hte_side <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + income + side*treated + male, 
case2_hte_full <- lm((Finalscore - pre_totnorm) ~ treated + pre_totnorm + income*treated + male*treated


models <- list(case2_hte_inc,case2_hte_side,case2_hte_full)
stargazer(models,type='text',digits=3,
          column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 9 (cont) - Case 2 Alternative Specifications with HTE')
```

```
## 
## Table 9 (cont) - Case 2 Alternative Specifications with HTE
## ==================================================================================
##                                        Dependent variable:
##                        -----------------------------------------------------------
##                                       (Finalscore - pre_totnorm)
##                           (1)                   (2)                   (3)
## ----------------------------------------------------------------------------------
## treated                  0.028*               -0.202***            -0.220***
##                         (0.014)               (0.004)              (0.012)
## 
## pre_totnorm             -0.002                -0.001               -0.001
##                         (0.003)               (0.002)              (0.002)
## 
## male                    -0.0004               -0.002               -0.004
##                         (0.004)               (0.003)              (0.004)
## 
## treated:side                                  0.506***             0.506***
##                                               (0.006)              (0.006)
```

19

```
## 
## side                   0.252***              -0.001                   -0.0005
##                         (0.004)              (0.004)                   (0.004)
## 
## income                 -0.0001               0.00001                  -0.00004
##                        (0.0001)              (0.0001)                  (0.0001)
## 
## treated:income         0.0001                                          0.0001
##                        (0.0001)                                       (0.0001)
## 
## treated:male                                                           0.006
##                                                                       (0.006)
## 
## Constant               -0.118***             -0.001                    0.008
##                         (0.013)              (0.009)                   (0.011)
## 
## ------------------------------------------------------------------------------------------------
## Observations            12,415                12,415                   12,415
## F Statistic   738.655*** (df = 6; 12408) 2,238.204*** (df = 6; 12408) 1,679.069*** (df = 8; 12406)
## ================================================================================================
## Note:                                                      *p<0.1; **p<0.05; ***p<0.01
```
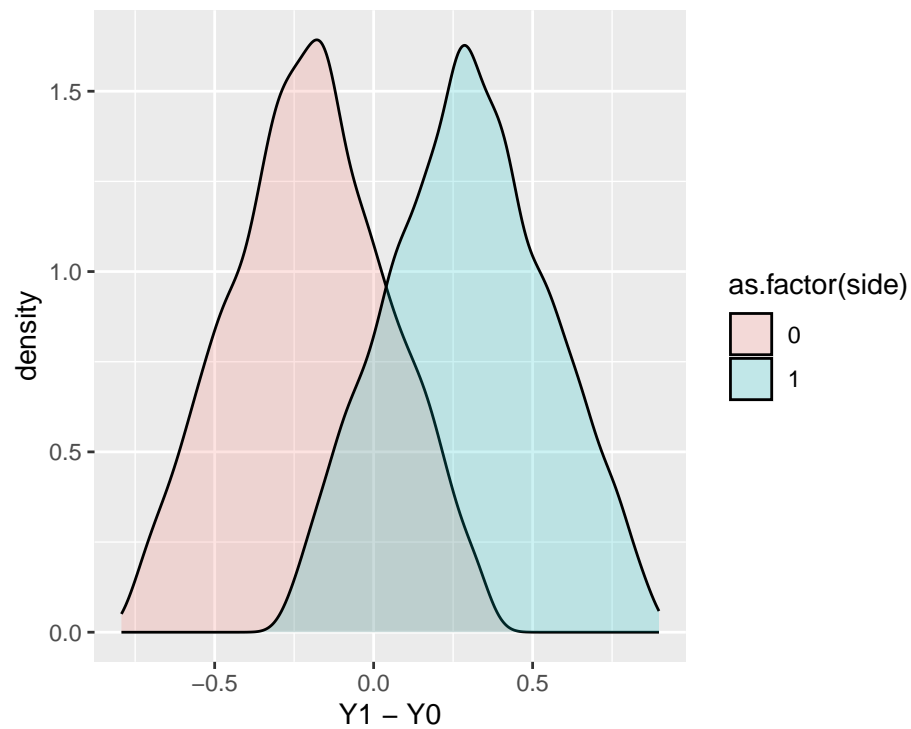
We see strong evidence that the treatment effect is heterogeneous for the side group based on the fact that the interaction term is highly significant, with side $==$ 1 resulting in greater treatment effect.
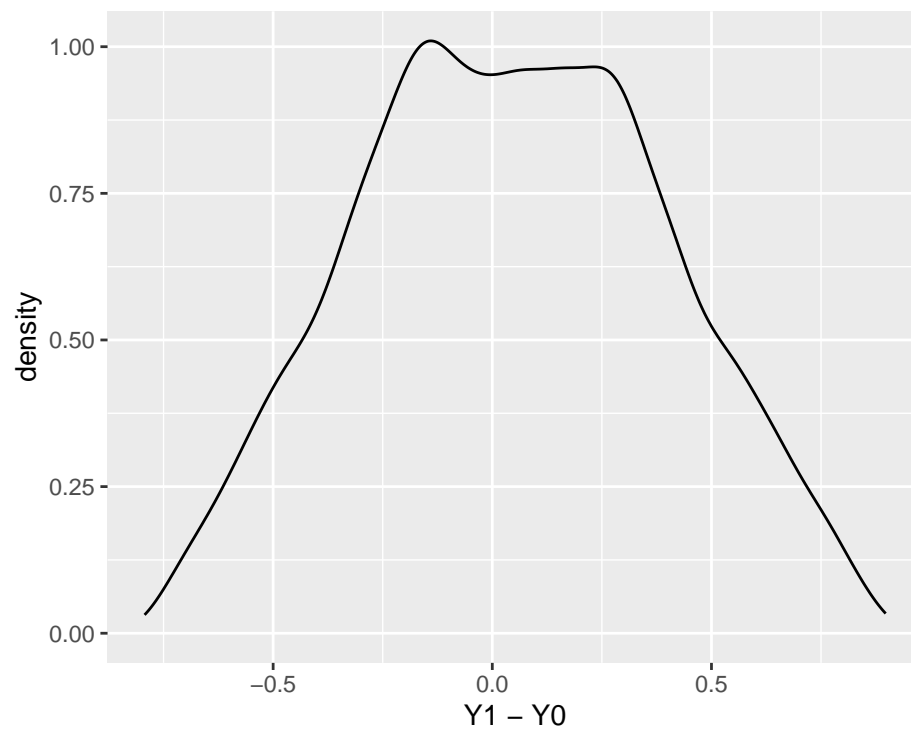
p)

```
# true effect
mean(case2$Y1 - case2$Y0)
```

```
## [1] 0.04638325
```

```
ggplot(data=case2, mapping=aes(x=Y1-Y0)) +
  geom_density(mapping=aes(fill=as.factor(side)),alpha = 0.2)
```

```
ggplot(data=case2, mapping=aes(x=Y1-Y0)) +
  geom_density(alpha = 0.2)
```

```
case2 %>%
  group_by(side) %>%
  summarise(score_diff = mean(Finalscore-pre_totnorm),.groups = 'keep')
```

```
## # A tibble: 2 x 2
## # Groups:   side [2]
##    side score_diff
##   <dbl>      <dbl>
## 1     0     -0.102
## 2     1      0.149
```

Our estimate of the treatment effect is slightly higher than the actual treatment effect. This could be due to sorting, where those who will benefit more from the treatment e.g. those with side == 1 tend to be in the treatment group more often. Since side is unobserved, it is difficult to account for when stratifying a sample for randomization and poses selection concerns.

q)

```
ggplot(data=case2, mapping=aes(x=Y1-Y0,fill = as.factor(side))) +
  geom_density(alpha = 0.2)
```
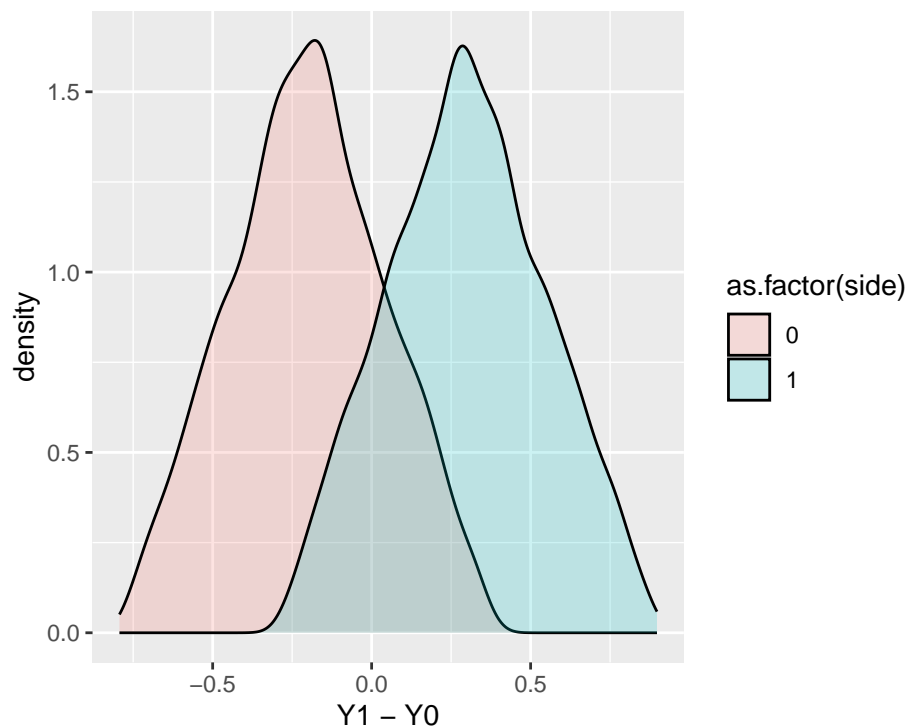

The actual treatment effect is higher for those with side == 1 than side == 0. As a result, our estimated causal effect of treatment in part m is slightly off because while the distribution of side == 1 and side == 0 is roughly even across treated and untreated groups, the effect of treatment on those with side == 1 is more positive than the effect of treatment on those with side == 0 is negative. As a result, there is a slight upwards bias on the estimated effect in m since the two groups which split the sample have a non-insignificant difference in conditional average treatment effect.

r)

The average effect for the treated is not meaningful in this case because there are two distinctly separate groups which have different CATT that are not only different magnitude but more importantly in different directions. As a result, we care significantly more about the conditional average treatment effect, especially if side is an indicator for a threshold of total family wealth e.g. if side == 1 represents below poverty line we would want to specifically target those who would actually benefit and avoid treating those who would be negatively affected.

s)

```
case3 <- read_stata('data/case3.dta')
```

```
# regression balance test
tgbal_case3_mlr <- lm(TreatmentGroup ~ pre_totnorm + male + income + numstud, data = case3)
tgbal_case3_pretotnorm <- lm(TreatmentGroup ~ pre_totnorm, data = case3)
tgbal_case3_male <- lm(TreatmentGroup ~ male, data = case3)
tgbal_case3_inc <- lm(TreatmentGroup ~ income, data = case3)
tgbal_case3_numstud <- lm(TreatmentGroup ~ numstud, data = case3)

models <- list(tgbal_case3_numstud,tgbal_case3_inc,tgbal_case3_male)
stargazer(models,type='text',digits=3,column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 10 - Case 3 Treatment Group Balance Test')
```

```
##
## Table 10 - Case 3 Treatment Group Balance Test
## ============================================================
##                              Dependent variable:
##                     ----------------------------------
##                              TreatmentGroup
##                         (1)        (2)        (3)
## ------------------------------------------------------------
## numstud               0.0002
##                       (0.0002)
##
## income                           0.00004
##                                 (0.0001)
##
## male                                        0.002
##                                            (0.009)
##
## Constant             0.489***   0.496***   0.501***
##                       (0.012)    (0.017)    (0.006)
##
## ------------------------------------------------------------
## Observations          12,415     12,415     12,415
## F Statistic (df = 1; 12413)  1.454      0.164      0.051
## ============================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

```
models <- list(tgbal_case3_pretotnorm,tgbal_case3_mlr)
stargazer(models,type='text',digits=3,column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 10 (cont) - Case 3 Treatment Group Balance Test')
```

```
##
## Table 10 (cont) - Case 3 Treatment Group Balance Test
## ===========================================================
##                          Dependent variable:
##              ---------------------------------------------
##                           TreatmentGroup
##                       (1)                    (2)
## -----------------------------------------------------------
## pre_totnorm          -0.002                 -0.006
##                      (0.004)                (0.007)
##
## male                                         0.002
##                                             (0.009)
##
## income                                       0.0002
##                                            (0.0002)
##
## numstud                                      0.0002
##                                            (0.0002)
##
## Constant             0.502***               0.465***
##                      (0.004)                (0.027)
##
## -----------------------------------------------------------
## Observations          12,415                 12,415
## F Statistic  0.135 (df = 1; 12413) 0.643 (df = 4; 12410)
## ===========================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

```
# regression balance test
trtbal_case3_mlr <- lm(treated ~ pre_totnorm + male + income + numstud, data = case3)
trtbal_case3_pretotnorm <- lm(treated ~ pre_totnorm, data = case3)
trtbal_case3_male <- lm(treated ~ male, data = case3)
trtbal_case3_inc <- lm(treated ~ income, data = case3)
trtbal_case3_inc_pretotnorm <- lm(treated ~ pre_totnorm + income, data = case3)
trtbal_case3_numstud <- lm(treated ~ numstud, data = case3)

models <- list(trtbal_case3_numstud,trtbal_case3_inc,trtbal_case3_male)
stargazer(models,type='text',digits=3,column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 11 - Case 3 Treated Balance Test')
```

```
##
## Table 11 - Case 3 Treated Balance Test
## ===========================================================
##                              Dependent variable:
##                    -----------------------------------------
##                                    treated
```

```
##                                  (1)         (2)         (3)
## --------------------------------------------------------------
## numstud                        -0.00002
##                                (0.0002)
##
## income                                     0.002***
##                                            (0.0001)
##
## male                                                   -0.009
##                                                        (0.009)
##
## Constant                       0.349***    0.056***   0.353***
##                                (0.011)     (0.016)    (0.006)
##
## --------------------------------------------------------------
## Observations                    12,415      12,415     12,415
## F Statistic (df = 1; 12413)     0.015     372.148***    1.016
## ==============================================================
## Note:                             *p<0.1; **p<0.05; ***p<0.01
```

```
models <- list(trtbal_case3_pretotnorm,trtbal_case3_inc_pretotnorm,trtbal_case3_mlr)
stargazer(models,type='text',digits=3,column.sep.width = "-15pt",
        omit.stat=c('rsq','adj.rsq',"ser"),
        title='Table 11 (cont) - Case 3 Treated Balance Test')
```

```
##
## Table 11 (cont) - Case 3 Treated Balance Test
## ===============================================================================================
##                                              Dependent variable:
##                -------------------------------------------------------------------------------
##                                                   treated
##                        (1)                         (2)                         (3)
## -----------------------------------------------------------------------------------------------
## pre_totnorm          0.049***                   -0.020***                   -0.020***
##                      (0.004)                     (0.006)                     (0.006)
##
## male                                                                         -0.005
##                                                                              (0.008)
##
## income                                           0.002***                    0.002***
##                                                  (0.0001)                    (0.0001)
##
## numstud                                                                      0.0001
##                                                                              (0.0002)
##
## Constant             0.347***                     0.004                       0.003
##                      (0.004)                     (0.022)                     (0.025)
##
## -----------------------------------------------------------------------------------------------
## Observations          12,415                      12,415                      12,415
## F Statistic   135.683*** (df = 1; 12413) 191.665*** (df = 2; 12412) 95.918*** (df = 4; 12410)
## ===============================================================================================
## Note:                                                            *p<0.1; **p<0.05; ***p<0.01
```

While randomization was successful and the treatment and control groups are initially balanced, the eventual groups that did and did not receive treatment are not balanced across income and pre-treatment test score.

t)

```
case3.t <- case3 %>%
  mutate(dropout = TreatmentGroup - treated)

dropout_case3_mlr <- lm(dropout ~ pre_totnorm + male + income + numstud, data = case3.t)
dropout_case3_inc <- lm(dropout ~ income, data = case3.t)
dropout_case3_pretotnorm <- lm(dropout ~ pre_totnorm, data = case3.t)

models <- list(dropout_case3_pretotnorm,dropout_case3_inc,dropout_case3_mlr)
stargazer(models,type='text',digits=3,column.sep.width = "-15pt",
        omit.stat=c('rsq','adj.rsq',"ser"),
        title='Table 12 - Case 3 Determinants of Dropout')
```

```
##
## Table 12 - Case 3 Determinants of Dropout
## =================================================================================
##                                      Dependent variable:
##               -------------------------------------------------------------------
##                                            dropout
##                      (1)                    (2)                    (3)
## -------------------------------------------------------------------------------
## pre_totnorm        -0.051***                                     0.014***
##                     (0.003)                                      (0.005)
##
## male                                                              0.007
##                                                                  (0.006)
##
## income                                     -0.002***            -0.002***
##                                             (0.0001)             (0.0001)
##
## numstud                                                          0.0002
##                                                                  (0.0001)
##
## Constant            0.155***               0.439***              0.462***
##                     (0.003)                (0.012)               (0.019)
##
## -------------------------------------------------------------------------------
## Observations        12,415                 12,415                12,415
## F Statistic  254.661*** (df = 1; 12413) 631.895*** (df = 1; 12413) 161.051*** (df = 4; 12410)
## =================================================================================
## Note:                                           *p<0.1; **p<0.05; ***p<0.01
```

The primary determinants of treatment group dropout are income and normalized pre-treatment total score.

u)

```
trtgrp_case3 <- case3 %>% filter(TreatmentGroup == 1)
ctrlgrp_case3 <- case3 %>% filter(TreatmentGroup == 0)
mean(trtgrp_case3$Finalscore - trtgrp_case3$pre_totnorm) -
  mean(ctrlgrp_case3$Finalscore - ctrlgrp_case3$pre_totnorm)
```

```
## [1] 0.2077142
```

Here we are estimating the effect of being assigned to the treatment group on the difference in normalized total score. The magnitude from a policy point of view is meaningful since 0.2 standard deviations is a fairly large improvement. The ITT effect itself is not meaningful from a policy perspective since it is not the effect of the actual treatment because we have differential attrition across groups, potentially resulting in selection.

v)

```
case3_base <- lm(Finalscore-pre_totnorm ~ TreatmentGroup, data=case3)

stargazer(case3_base,type='text',digits=3,column.sep.width = "-15pt",
          omit.stat=c('rsq','adj.rsq',"ser"),
          title='Table 13 - Case 3 Baseline ITT')
```

```
##
## Table 13 - Case 3 Baseline ITT
## =========================================
##                       Dependent variable:
##                    ----------------------------
##                        Finalscore - pre_totnorm
## -----------------------------------------
## TreatmentGroup              0.208***
##                             (0.004)
##
## Constant                    -0.001
##                             (0.003)
##
## -----------------------------------------
## Observations                12,415
## F Statistic      3,392.362*** (df = 1; 12413)
## =========================================
## Note:            *p<0.1; **p<0.05; ***p<0.01
```

We do not get an unbiased estimate of the causal effect of treatment on the outcome because attrition was differential across different observable groups. As a result, we can no longer assume randomization and that there was selection at play where people with different observables/unobservables chose to not be treated despite being assigned to the treatment group. As a result, we have OVB in our regression.

w)

```r
library(ivreg)

case3.tsls <- case3 %>%
  mutate(Y = Finalscore-pre_totnorm,
         D = treated,
         Z = TreatmentGroup)

case3_firststage <- lm(D ~ Z, data = case3.tsls)
case3_reduced <- lm(Y ~ Z, data = case3.tsls)
case3_tsls <- ivreg(formula = Y ~ D|Z, data = case3.tsls)

stargazer(list(case3_firststage,case3_reduced,case3_tsls),
          type='text',digits=3,column.sep.width = "-15pt",
          dep.var.labels=c('First Stage','Reduced Form','TSLS Estimate'),
          title='Table 14 - Case 3 TSLS')
```

```
##
## Table 14 - Case 3 TSLS
## ===============================================================
##                                  Dependent variable:
##                        ----------------------------------------
##                        First Stage       Reduced Form
##                            OLS          OLS       instrumental
##                                                     variable
##                            (1)          (2)           (3)
## -------------------------------------------------------------
## Z                        0.693***     0.208***
##                          (0.006)      (0.004)
##
## D                                                   0.300***
##                                                     (0.005)
##
## Constant                 -0.000       -0.001        -0.001
##                          (0.004)      (0.003)       (0.002)
##
## -------------------------------------------------------------
## Observations             12,415       12,415        12,415
## R2                       0.529        0.215         0.398
## Adjusted R2              0.529        0.215         0.398
## Residual Std. Error (df = 12413)   0.327    0.199    0.174
## F Statistic (df = 1; 12413)  13,961.060*** 3,392.362***
## ===============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

This is an unbiased estimate of the treatment effect. This part of the population relates to compliers, that this, individuals who were assigned to the treatment group and eventually were treated. LATE is higher because in ITT our treatment effect includes people who were assigned to the treatment group and did not receive treatment. This means that since our treatment effect is positive, these noncompliers lower our estimated treatment effect since we are counting these individuals who likely have a lower outcome since they never got treated as being treated. Hence the LATE is higher because it only looks at those who actually got treated.
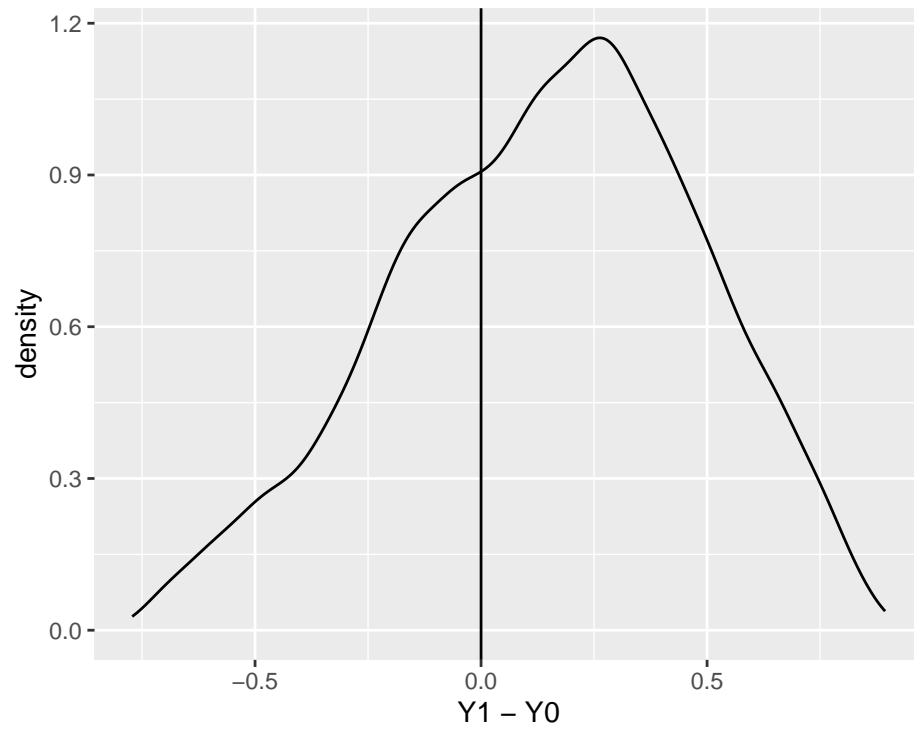
```
x)
```

```
# true effect
mean((case3 %>% filter(treated==1))$Y1 - (case3 %>% filter(treated==1))$Y0)
```

```
## [1] 0.297617
```

```
ggplot(data=case3, mapping=aes(x=Y1-Y0)) +
  geom_density(mapping=aes(fill=cut(income,4)),alpha = 0.2) +
  geom_vline(xintercept=0)
```
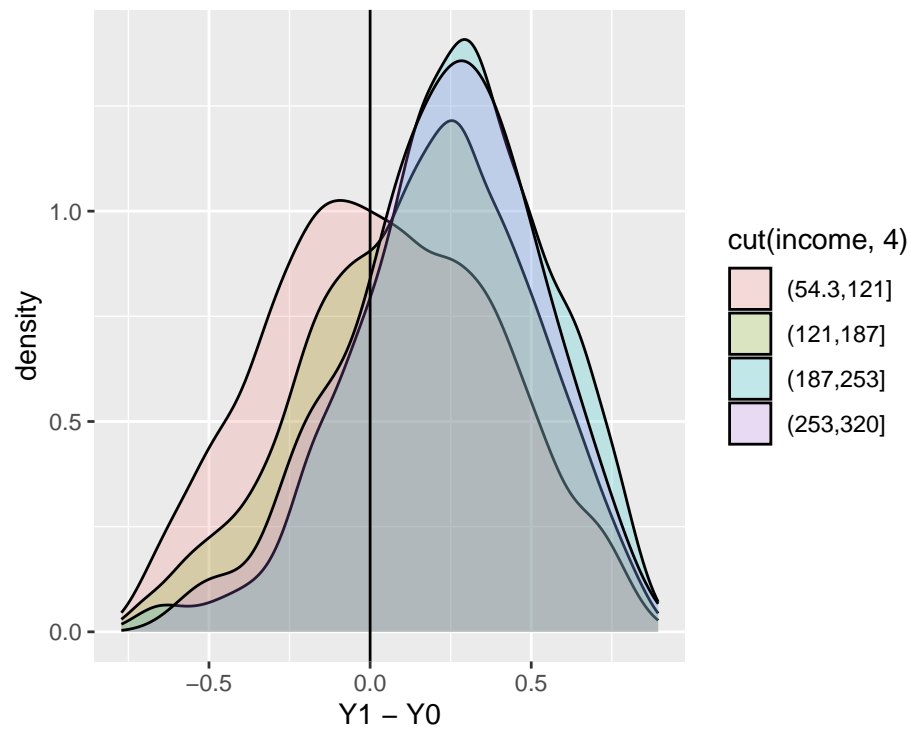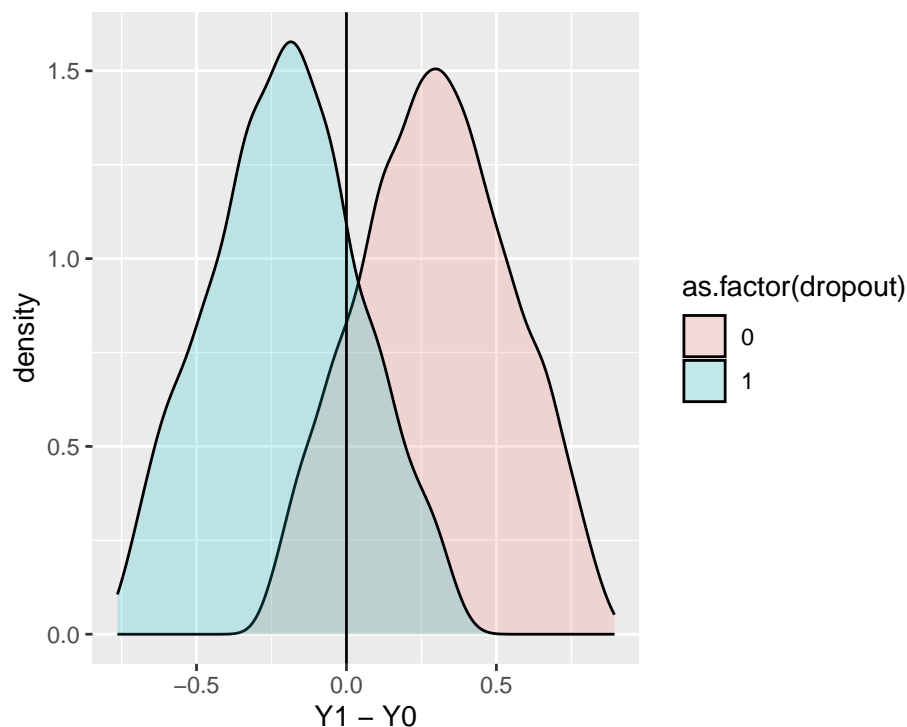


```
ggplot(data=case3, mapping=aes(x=Y1-Y0)) +
  geom_density(alpha = 0.2) +
  geom_vline(xintercept=0)
```

y)

```
ggplot(data=case3.t %>% filter(TreatmentGroup == 1), mapping=aes(x=Y1-Y0)) +
  geom_density(mapping=aes(fill=as.factor(dropout)),alpha = 0.2) +
  geom_vline(xintercept=0)
```

It appears that people who were likely to have a negative treatment effect chose to dropout and not receive the treatment. This behavior is logical since if an individual has some idea that they may not benefit from the treatment they naturally should attempt to not receive the treatment. However, this does not explain those who would receive a positive treatment effect dropping out. This is most likely because people do not exactly know their true treatment effect (i.e. imperfect information about treatment) and at the margin may choose to drop out.

z)

We have seen that treatment effects can vary across both observed and unobserved dimensions of the data and that it is important to check for heterogeneous treatment effects as well as heteroskedasticity in your results. Additionally, we see that compliance is extremely important for the validity of RCT results, given that differential attrition can be due to selection by individuals on dimensions that are unobservable for the researcher. While RCT provides a gold standard for causal analyses and relies on relatively few dubious/difficult to validate assumptions, we must be careful in designing our experiment and subsequently analyzing its results to ensure that the necessary conditions hold to get the causal treatment effect of interest.

Problem 2

I will be working on replicating the paper "Is the price elasticity of demand for coal in China increasing?" Burke and Liao (2015) with Cedric Elkouh.

a)

How much did a change in coal prices affect coal consumption in China? Y is consumption of coal and X is an output price index for the mining and washing of coal deflated by the industrial producer price index (Burke and Liao 2015).

b)

The ideal randomized experiment would be to have suppliers randomly set prices and see how much coal is demanded to trace out a demand curve for each period to get the elasticity of demand.

c)

Naturally it would be impossible to force coal sellers to randomize their prices and potentially some general equilibrium effects and other potential violations of RCT.

d)

We would want to add individual level contract information as a control variable so that we could account for different coal sellers and their involvement in setting prices via contracts. This could credibly reduce the selection bias towards zero since we account for the matching in the coal seller-buyer market.