Optimal Bandwidths and Balance of the Samples

Falco J. Bargagli Stoffi

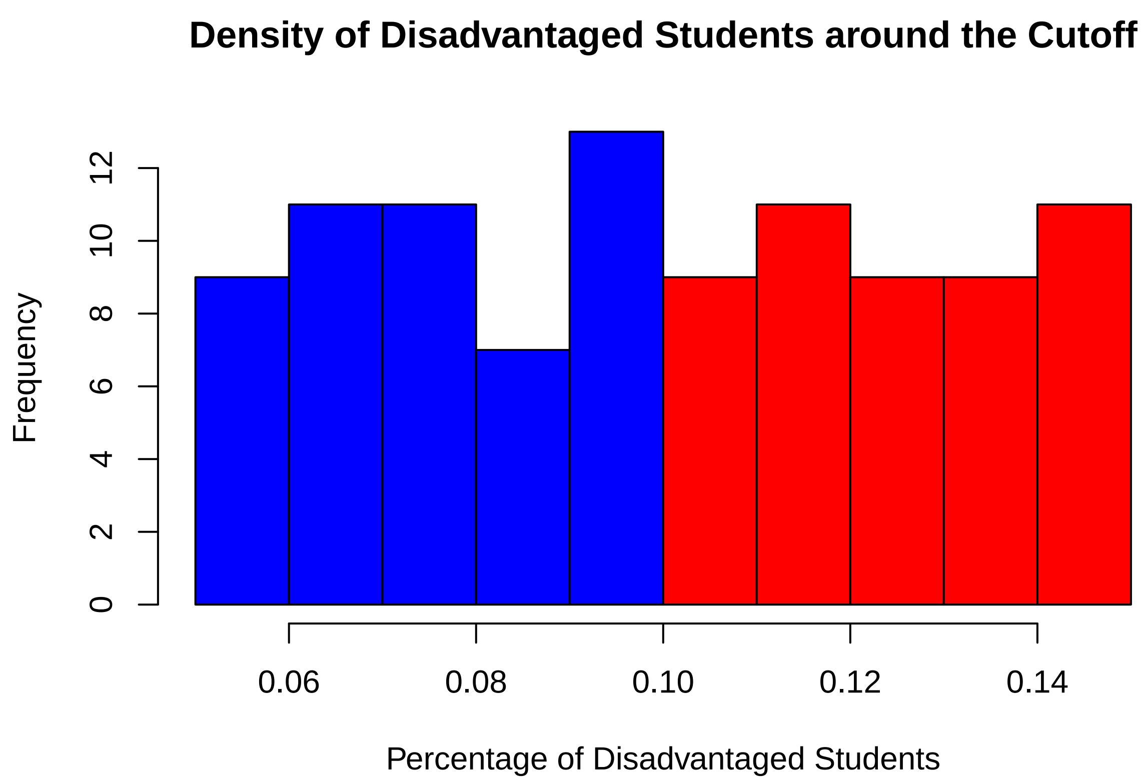
Cutoff Manipulation Check

Before estimating the optimal bandwidth it is important to check wheter or not there is manipulation of the cutoff. We can test the manipulation in two ways:

1. by looking at the distribution of the density of disadvantaged students around the cutoff in order to visualize if there is any sign of manipulation;
2. by using the McCrary manipulation test.

Both these diagnostics are performed aggregating the students at school level.

```
setwd("G:\\Il mio Drive\\Causal Tree IV\\Honest Causal Tree\\new_data")
school_data <- read_dta("first_stage_2011.dta")
h <-hist(school_data$GOKpercentage[which(school_data$GOKpercentage>0.05 &
                                          school_data$GOKpercentage<0.15)],
         breaks=10,
         plot = FALSE)
cuts <- cut(h$breaks, c(-Inf,.099999,.1000000,Inf))
plot(h, col=c("blue","red","red")[cuts],
     xlab = "Percentage of Disadvantaged Students",
     main="Density of Disadvantaged Students around the Cutoff")
```



```r
summary(rddensity(school_data$GOKpercentage, c = 0.10, h = 0.035))
```

```
##
## RD Manipulation Test using local polynomial density estimation.
##
## Number of obs =       649
## Model =               unrestricted
## Kernel =              triangular
## BW method =           mannual
## VCE method =          jackknife
##
## Cutoff c = 0.1        Left of c           Right of c
## Number of obs         54                  595
## Eff. Number of obs    39                  35
## Order est. (p)        2                   2
## Order bias (q)        3                   3
## BW est. (h)           0.035               0.035
##
## Method                T                   P > |T|
## Robust                -0.7497             0.4534
```

Bandwidth Estimation

$\hat{b}_{CCT,p}$ and $\hat{h}_{CCT,p,q}$ are the estimated bandwidths used to construct the fuzzy RD point estimator and the RD bias-correction, respectively. The latter function implements the bias-corrected robust (to "large" bandwidth choices) inference procedure proposed by Calonico, Cattaneo, and Titiunik (2014a, CCT hereafter). The subscript p specifies the order of the local-polynomial used to construct the point-estimator (i.e., $p = 1$ is for the local-linear fuzzy RD estimator; $p = 2$ is the local-quadratic fuzzy kink RD estimator) and q specifies the order of the local-polynomial used to construct the bias-correction, which is built using a possibly different bandwidth (default is $q = 2$ (local quadratic regression). For more references on the package see "rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs" (Calonico, Cattaneo, Titiunik, 2015).

The bandwidth are constructed for the tree different response variables: *certificate* (whether or not the student got an A certificate grade) and *progress_school* (whether or not the student progressed school).

Student Level Analysis of Optimal Bandiwidth

Let's now see what happens when we take as a unit level of out analysis the students. We load the new data and construct the new *GOKschool_up* variable.

```r
setwd("G:\\Il mio Drive\\Causal Tree IV\\Honest Causal Tree\\new_data")
#students_data_2011 <- read_dta("first_stage_students_trimmed_2011.dta")
students_data_2011 <- read_dta("first_stage_students_2011.dta")
table(students_data_2011$GOKschool, students_data_2011$D)
```

```
##
##          0      1
##   0  13084  16405
##   1      0 106193
```

```r
students_data_2011$eligible_dummy <-  students_data_2011$D
```

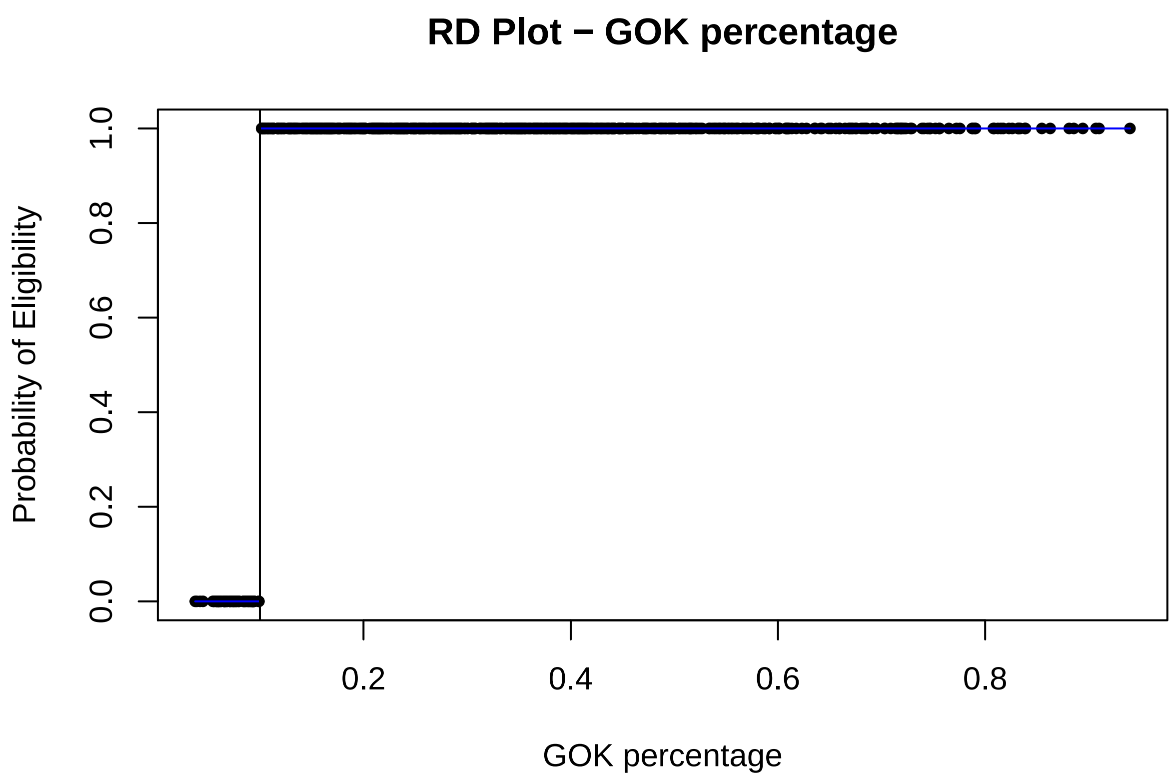
Again, we first compute the probabilities of being eligible with a logit model where the regressor is simply the percentage of disadvantaged students (*GOKpercentage*):

$$p(eligible_i = 1|\theta, GOK_{\%,i}) = \frac{1}{1 + e^{\theta GOK_{\%,i}}} \tag{1}$$

where the index i refers to the students.

```r
logit<-glm(eligible_dummy ~ GOKpercentage,
           data = students_data_2011, family = binomial(link = "logit"))
summary(logit)
probs<-predict(logit, students_data_2011, type="response")
```

We plot the probability of elegibility v. *GOKpercentage*. The discontinuity in this case is "clear-cut".




```
## Call: rdplot
##
## Number of Obs.                 135682
## Kernel                        Uniform
##
## Number of Obs.                  13084          122598
## Eff. Number of Obs.             13084          122598
## Order poly. fit (p)                 4               4
## BW poly. fit (h)                0.062           0.840
## Number of bins scale                1               1
```

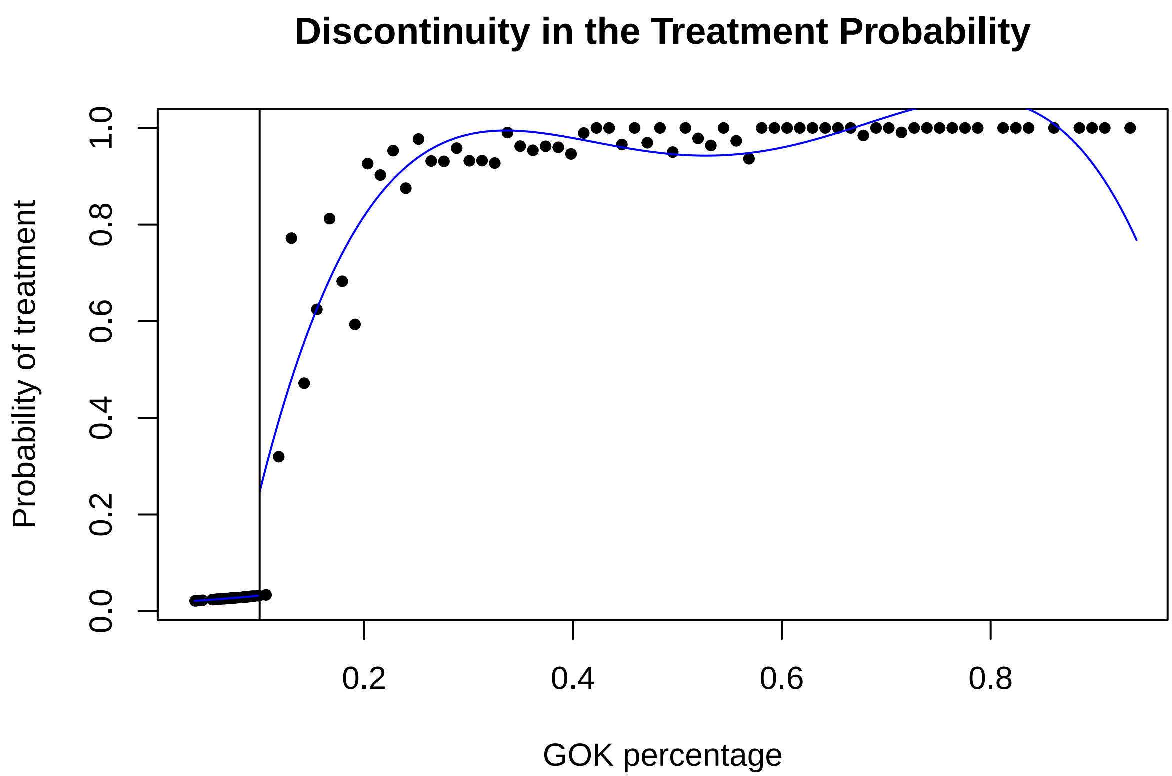
However, it is more meaningful to construct the probability of being treated (namely, being eligible and having more than six hours of extra classes) using again a logit model where now the regressors are the percentage of disadvantaged students (*GOKpercentage*) and a dummy variable at school level for the implementation of the six hours of classes (*GOKschool_up*):

$$p(eligible_i = 1|\Theta, GOK_{\%,i}, GOKschool\_up_i) = \frac{1}{1 + e^{\theta_1 GOK_{\%,i} + \theta_2 GOKschool\_up_i}} \tag{2}$$

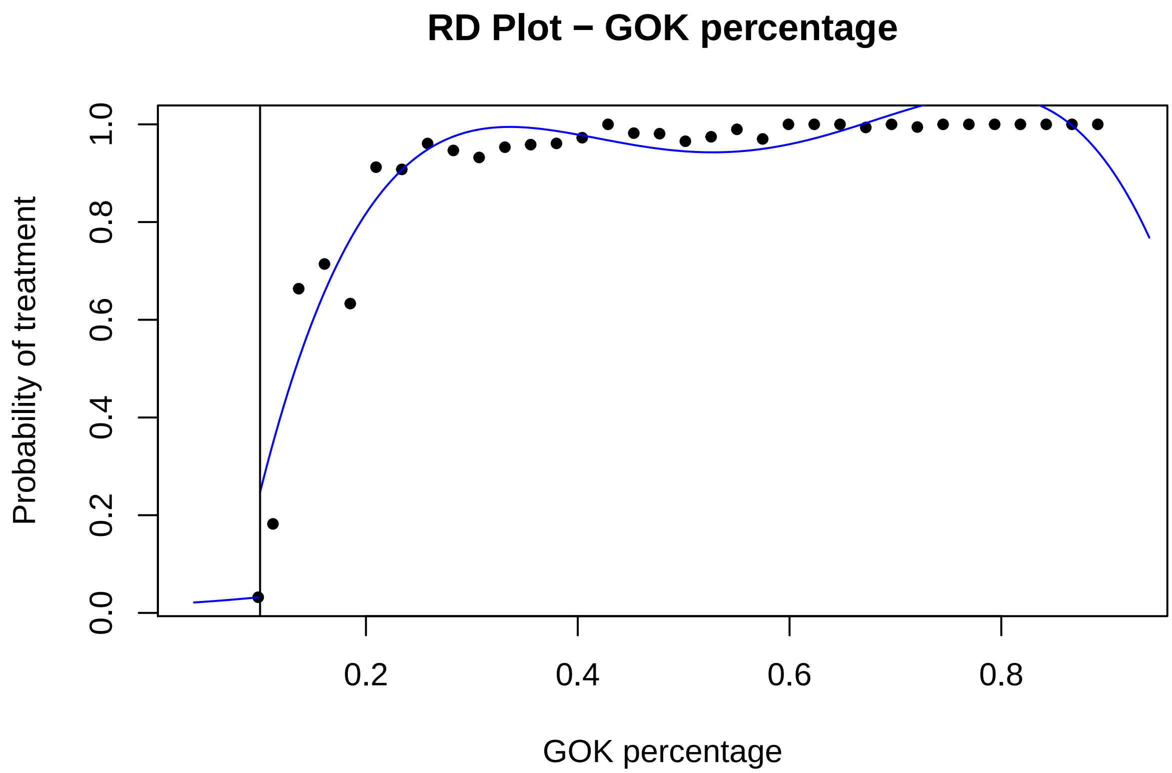
where the index i refers to the school.

```
logit<-glm(GOKschool ~ GOKpercentage + GOKschool_up,
           data = students_data_2011, family = binomial(link = "logit"))
summary(logit)
probs<-predict(logit, students_data_2011, type="response")
```

In this case the discontinuity is less “clear-cut”, as we can see from the following plots. The probability of being treated decreases as the observations approach the threshold. However, the discontinuity is still present.




```
## Call: rdplot
##
## Number of Obs.                 135682
## Kernel                        Uniform
##
## Number of Obs.                  13084          122598
## Eff. Number of Obs.             13084          122598
## Order poly. fit (p)                 4               4
## BW poly. fit (h)                0.062           0.840
## Number of bins scale                0               0
```




```
## Call: rdplot
##
## Number of Obs.                 135682
## Kernel                        Uniform
##
## Number of Obs.                 13084         122598
## Eff. Number of Obs.            13084         122598
## Order poly. fit (p)                4              4
## BW poly. fit (h)               0.062          0.840
## Number of bins scale               0              0
```

Certificate

Since from the previous plots there seems to be some form of non linearity we decide to build the optimal bandwidths using a second degree polinomial ($p = 2$). The optimal bandwidths for school levels, when the response variable is *certificate*, are $\hat{b}_{CCT,2} = 0.024$ and $\hat{h}_{CCT,2,2} = 0.035$.

```
rdrobust(y = students_data_2011$certificate, x = students_data_2011$GOKpercentage,
         c = 0.10,
         fuzzy = students_data_2011$GOKschool_up  == 1, p = 2,    all = TRUE)
```

```
## Call: rdrobust
##
## Number of Obs.                 135682
## BW type                         mserd
## Kernel                     Triangular
```

```
## VCE method                          NN
##
## Number of Obs.                   13084      122598
## Eff. Number of Obs.               5931        6530
## Order est. (p)                      2           2
## Order bias  (p)                     3           3
## BW est. (h)                      0.024       0.024
## BW bias (b)                      0.035       0.035
## rho (h/b)                        0.670       0.670
```

Progress School

The optimal bandwidths for school levels, when the response variable is *Progress School*, are $\hat{b}_{CCT,2} = 0.026$ and $\hat{h}_{CCT,2,2} = 0.037$.

```r
rdrobust(y = students_data_2011$progress_school, x = students_data_2011$GOKpercentage,
         c = 0.10, p = 2,
         fuzzy = students_data_2011$GOKschool_up, all = TRUE)
```

```
## Call: rdrobust
##
## Number of Obs.                  135682
## BW type                          mserd
## Kernel                      Triangular
## VCE method                          NN
##
## Number of Obs.                   13084      122598
## Eff. Number of Obs.               6928        6530
## Order est. (p)                      2           2
## Order bias  (p)                     3           3
## BW est. (h)                      0.026       0.026
## BW bias (b)                      0.037       0.037
## rho (h/b)                        0.713       0.713
```

The optimal bandwidth for all the different outcome variables are from 0.24 to 0.26 for $\hat{b}_{CCT,p}$ and from 0.35 to 0.37 for $\hat{h}_{CCT,p,q}$.

Balance in the Samples

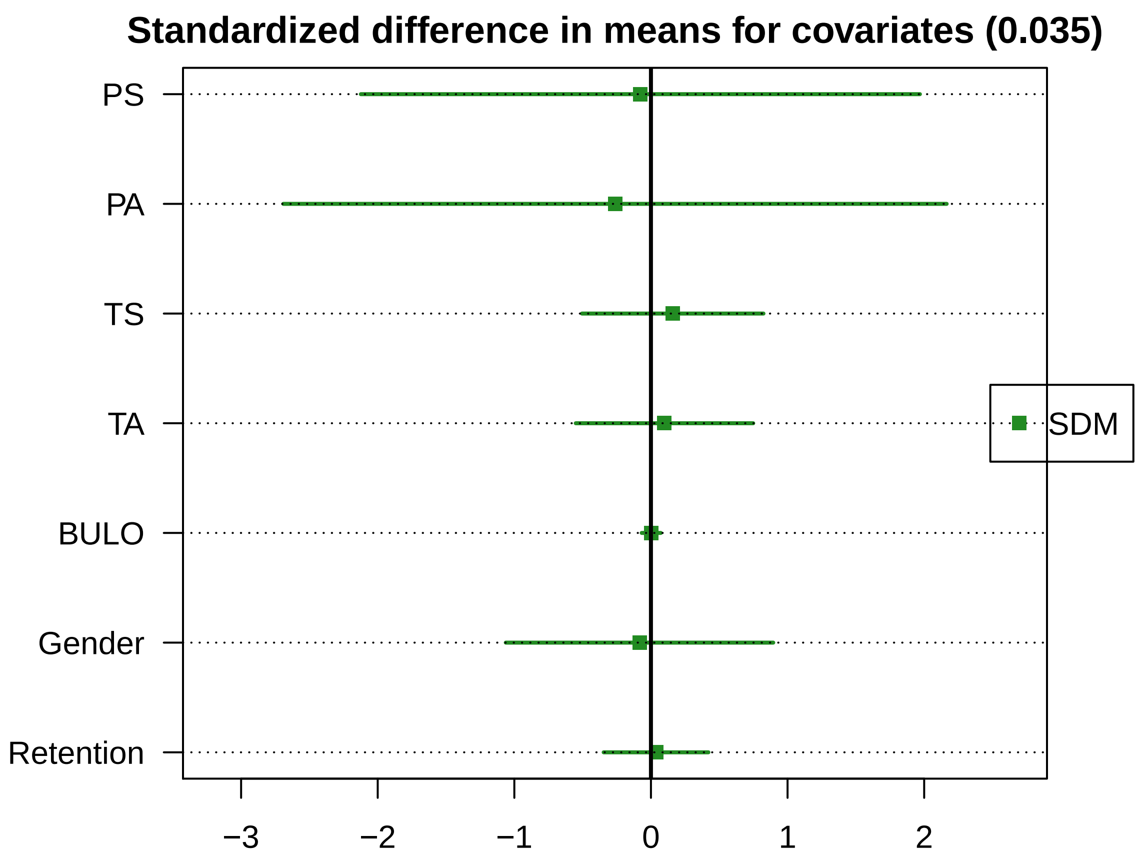
Bandwidth 0.035

Let's now check the balance of the samples of treated and control units when the bandwidth is 0.03. The analysis refers to students as unit level.

```r
myvars <- c("llncode", "schooljaar", "school", "GOKpercentage", "eligible_dummy",
            "GOKschool", "progress_school", "certificate",
            "primary_retention", "man", "BULO", "leerkracht_age",
            "leerkracht_seniority", "directie_age",
            "directie_seniority")

students_data_2011 <- students_data_2011[myvars]
head(students_data_2011)
table(students_data_2011$GOKschool, students_data_2011$eligible_dummy)
```

```
students_data_randomized_03_2011 <-
  students_data_2011[which(students_data_2011$GOKpercentage >= .065
                          & students_data_2011$GOKpercentage <= .135),]
```

```
students_data_randomized_03_2011_variables <-
students_data_randomized_03_2011[, 9:ncol(students_data_randomized_03_2011)]

mean.full <- as.data.frame(apply(students_data_randomized_03_2011_variables, 2, mean))
sd.full <- as.data.frame(apply(students_data_randomized_03_2011_variables, 2, sd))

mean.treated<-as.data.frame(apply(students_data_randomized_03_2011_variables[
  which(students_data_randomized_03_2011$eligible_dummy==1),], 2, mean))
mean.control<-as.data.frame(apply(students_data_randomized_03_2011_variables[
  which(students_data_randomized_03_2011$eligible_dummy==0),], 2, mean))
sd.treated<-as.data.frame(apply(students_data_randomized_03_2011_variables[
  which(students_data_randomized_03_2011$eligible_dummy==1),], 2, sd))
sd.control<-as.data.frame(apply(students_data_randomized_03_2011_variables[
  which(students_data_randomized_03_2011$eligible_dummy==0),], 2, sd))


# Standardized difference in means
diff.means <- mean.treated[,1] - mean.control[,1]
standard.diff.means <-  (mean.treated[,1] - mean.control[,1])/
                         sqrt((sd.treated^2 + sd.control^2)/2)

# Check this t-test
t =  mean.treated[,1] - mean.control[,1]/
  sqrt((sd.control^2/length(which(students_data_randomized_03_2011$eligible_dummy == 0))) +
         (sd.treated^2/length(which(students_data_randomized_03_2011$eligible_dummy == 1))))

# 99% CI (t-student distribution)
x0 <- standard.diff.means - 1.96 * sqrt((sd.treated^2 + sd.control^2)/2)
#check that dimension is correct
x1 <- standard.diff.means + 1.96 * sqrt((sd.treated^2 + sd.control^2)/2)
#check that dimension is correct


rownames(standard.diff.means)<-c("Retention", "Gender", "BULO", "TA", "TS", "PA", "PS")
```



We now check whetere or not the means for the covariates in the groups of units assigned to the control and to the treatment are significatively different.

First, we perform the analysis sampling 50 students from each school.

```
schools <-
students_data_randomized_03_2011$school[which(
!duplicated(students_data_randomized_03_2011$school))]

sample_students <- as.data.frame(matrix(data = NA, nrow = 50*length(schools),
                                        ncol = ncol(students_data_randomized_03_2011)))
colnames(sample_students) <- colnames(students_data_randomized_03_2011)

for (j in (0:(length(schools)-1))){
  set.seed(j  + 123)
  sample_students[(1+(j*50)):(50+(j*50)),]  <-
    students_data_randomized_03_2011[which(
    students_data_randomized_03_2011$school %in%
    schools[j+1]),][sample(1:nrow(students_data_randomized_03_2011[which(
    students_data_randomized_03_2011$school %in% schools[j+1]),]),
    50, replace = FALSE ),]
}

length(which(is.na(sample_students)))

sample_students <- round(sample_students[, -(1:4)], 0)
sample_students_variables <- sample_students[, -(1:4)]
```

```r
mean.full <- as.data.frame(apply(sample_students_variables, 2, mean))
sd.full <- as.data.frame(apply(sample_students_variables, 2, sd))


mean.treated<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==1),], 2, mean))
mean.control<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==0),], 2, mean))
sd.treated<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==1),], 2, sd))
sd.control<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==0),], 2, sd))

# Sampled Data
pvalue <- matrix(NA, ncol=1, nrow = 7)

#Primary retentions
pvalue[1,]<-t.test(
  sample_students$primary_retention[which(sample_students$eligible_dummy==1)],
  sample_students$primary_retention[which(sample_students$eligible_dummy==0)])$p.value

#BULO
pvalue[3,]<-t.test(
  sample_students$BULO[which(sample_students$eligible_dummy==1)],
  sample_students$BULO[which(sample_students$eligible_dummy==0)])$p.value

#Man
pvalue[2,]<-t.test(
  sample_students$man[which(sample_students$eligible_dummy==1)],
  sample_students$man[which(sample_students$eligible_dummy==0)])$p.value

#leerkracht_age
pvalue[4,]<-t.test(
  sample_students$leerkracht_age[which(sample_students$eligible_dummy==1)],
  sample_students$leerkracht_age[which(sample_students$eligible_dummy==0)])$p.value

#leerkracht_seniority
pvalue[5,]<-t.test(
  sample_students$leerkracht_seniority[which(sample_students$eligible_dummy==1)],
  sample_students$leerkracht_seniority[which(sample_students$eligible_dummy==0)])$p.value

#directie_age
pvalue[6,]<-t.test(
  sample_students$directie_age[which(sample_students$eligible_dummy==1)],
  sample_students$directie_age[which(sample_students$eligible_dummy==0)])$p.value

#directie_seniority
pvalue[7,]<-t.test(
  sample_students$directie_seniority[which(sample_students$eligible_dummy==1)],
  sample_students$directie_seniority[which(sample_students$eligible_dummy==0)])$p.value

pvalue <- as.data.frame(pvalue)
```

```
table <- as.data.frame(cbind(mean.control, sd.control, mean.treated, sd.treated,
                             mean.full, sd.full, pvalue))

rownames(table)<-c("Retention", "Gender", "BULO", "TA", "TS", "PA", "PS")

xtable(table, type = "latex", file = "filename.tex", digits=c(3,3,3,3,3,3,3,3))
```

Second, we perform the same analysis but sampling a number of students according to the size of the smallest school (62 students).

```
schools <-
students_data_randomized_03_2011$school[which(
!duplicated(students_data_randomized_03_2011$school))]


sample_students <- as.data.frame(matrix(data = NA, nrow = 62*length(schools),
                                        ncol = ncol(students_data_randomized_03_2011)))
colnames(sample_students) <- colnames(students_data_randomized_03_2011)

for (j in (0:(length(schools)-1))){
  set.seed(j  + 123)
  sample_students[(1+(j*62)):(62+(j*62)),]  <-
  students_data_randomized_03_2011[which(
  students_data_randomized_03_2011$school %in%
  schools[j+1]),][sample(1:nrow(students_data_randomized_03_2011[which(
  students_data_randomized_03_2011$school %in% schools[j+1]),]),
  62, replace = FALSE ),]
}

length(which(is.na(sample_students)))

sample_students <- round(sample_students[, -(1:4)], 0)
sample_students_variables <- sample_students[, -(1:4)]

mean.full <- as.data.frame(apply(sample_students_variables, 2, mean))
sd.full <- as.data.frame(apply(sample_students_variables, 2, sd))


mean.treated<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==1),], 2, mean))
mean.control<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==0),], 2, mean))
sd.treated<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==1),], 2, sd))
sd.control<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==0),], 2, sd))

# Sampled Data
pvalue <- matrix(NA, ncol=1, nrow = 7)

#Primary retentions
pvalue[1,]<-t.test(
  sample_students$primary_retention[which(sample_students$eligible_dummy==1)],
  sample_students$primary_retention[which(sample_students$eligible_dummy==0)])$p.value
```

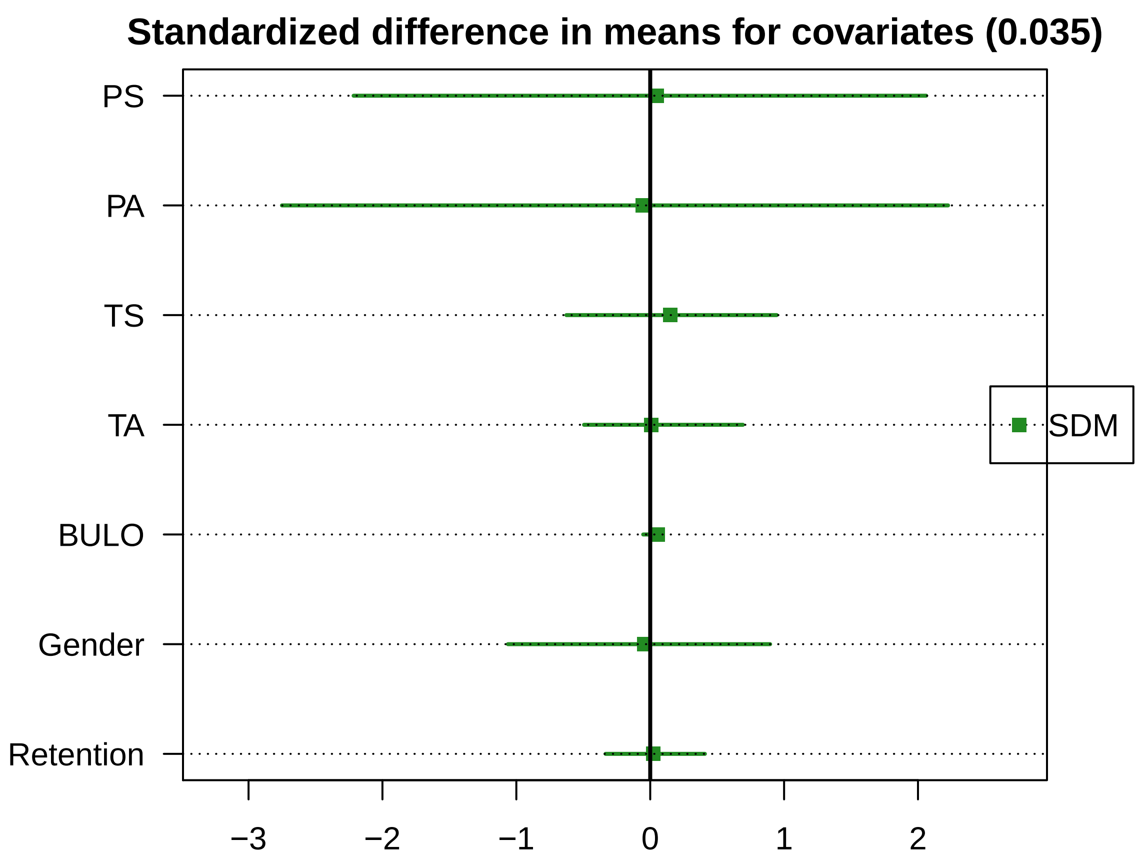
```
#BULO
pvalue[3,]<-t.test(
  sample_students$BULO[which(sample_students$eligible_dummy==1)],
  sample_students$BULO[which(sample_students$eligible_dummy==0)])$p.value

#Man
pvalue[2,]<-t.test(
  sample_students$man[which(sample_students$eligible_dummy==1)],
  sample_students$man[which(sample_students$eligible_dummy==0)])$p.value

#leerkracht_age
pvalue[4,]<-t.test(
  sample_students$leerkracht_age[which(sample_students$eligible_dummy==1)],
  sample_students$leerkracht_age[which(sample_students$eligible_dummy==0)])$p.value

#leerkracht_seniority
pvalue[5,]<-t.test(
  sample_students$leerkracht_seniority[which(sample_students$eligible_dummy==1)],
  sample_students$leerkracht_seniority[which(sample_students$eligible_dummy==0)])$p.value

#directie_age
pvalue[6,]<-t.test(
  sample_students$directie_age[which(sample_students$eligible_dummy==1)],
  sample_students$directie_age[which(sample_students$eligible_dummy==0)])$p.value

#directie_seniority
pvalue[7,]<-t.test(
  sample_students$directie_seniority[which(sample_students$eligible_dummy==1)],
  sample_students$directie_seniority[which(sample_students$eligible_dummy==0)])$p.value

pvalue <- as.data.frame(pvalue)

table <- as.data.frame(cbind(mean.control, sd.control, mean.treated, sd.treated,
                             mean.full, sd.full, pvalue))

rownames(table)<-c("Retention", "Gender", "BULO", "TA", "TS", "PA", "PS")

xtable(table, type = "latex", file = "filename.tex", digits=c(3,3,3,3,3,3,3,3))
```

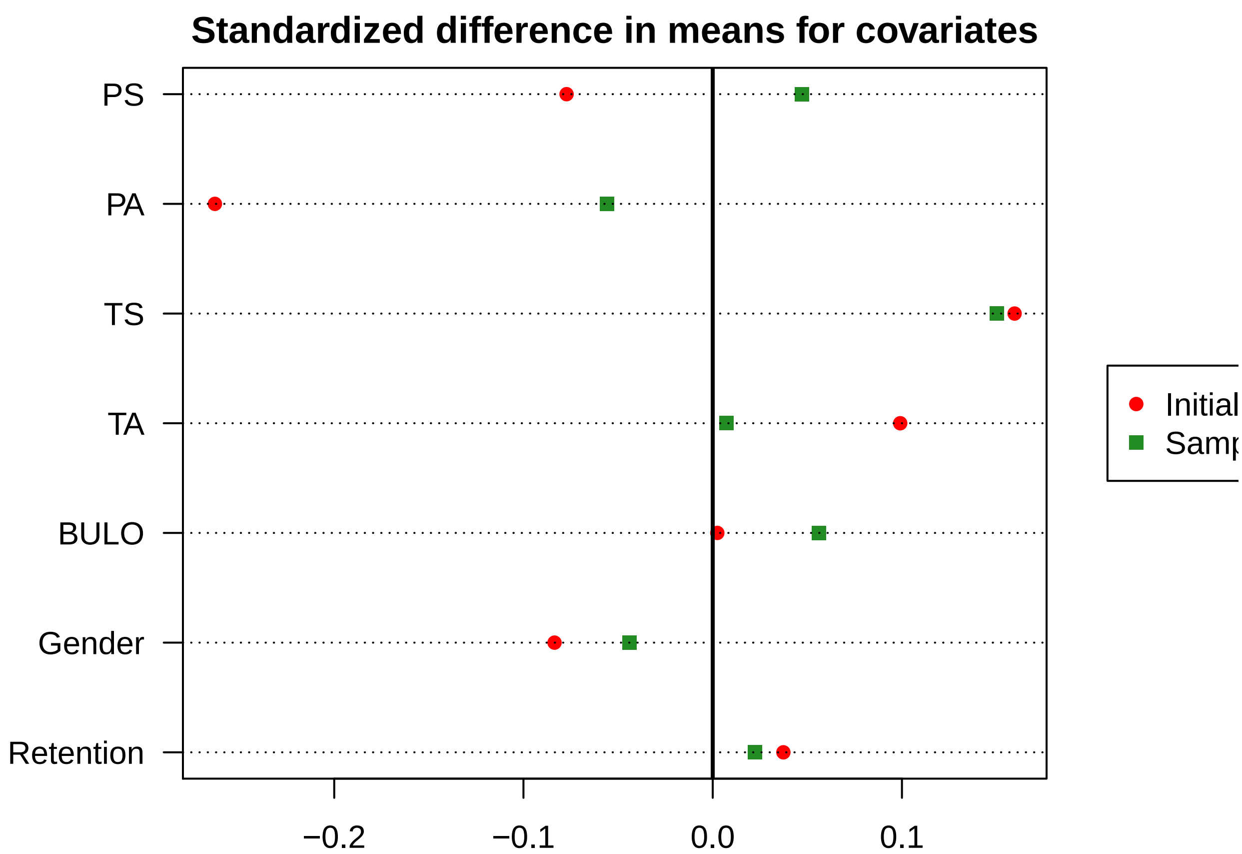
Moreover, we plot the standardized difference in means in the latter sample and the improvement in the fit between the overall population of units in the bandwidth and the sampled units.

```
mean.full <- as.data.frame(apply(sample_students, 2, mean))
sd.full <- as.data.frame(apply(sample_students, 2, sd))


mean.treated<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==1),], 2, mean))
mean.control<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==0),], 2, mean))
sd.treated<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==1),], 2, sd))
sd.control<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==0),], 2, sd))
```

```r
# Standardized difference in means
diff.means <- mean.treated[,1] - mean.control[,1]
standard.diff.means.sample <-  (mean.treated[,1] - mean.control[,1])/
                                sqrt((sd.treated^2 + sd.control^2)/2)

# Check this t-test
t =  mean.treated[,1] - mean.control[,1]/
  sqrt((sd.control^2/length(which(sample_students$eligible_dummy == 0))) +
         (sd.treated^2/length(which(sample_students$eligible_dummy == 1))))

# 99% CI (t-student distribution)
x0 <- standard.diff.means - 1.96 * sqrt((sd.treated^2 + sd.control^2)/2)
#check that dimension is correct
x1 <- standard.diff.means + 1.96 * sqrt((sd.treated^2 + sd.control^2)/2)
#check that dimension is correct


rownames(standard.diff.means.sample)<-c("Retention", "Gender", "BULO", "TA", "TS", "PA", "PS")
```

Two variables result to be unbalanced with respect to the standadized difference in means: *teacher female* and *teacher diploma*. There are more female teachers in the sample of treated units and less teachers with the diploma.



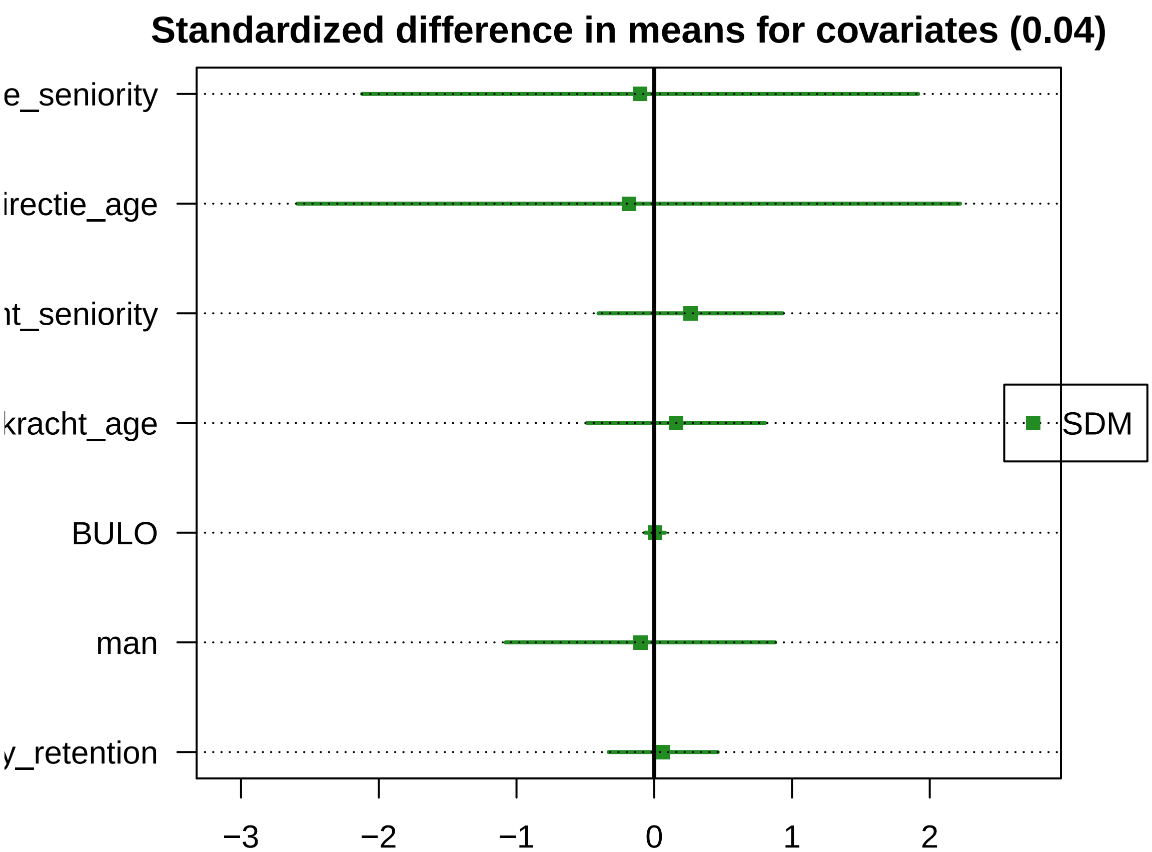


Bandwidth 0.04

Let's now check the balance of the samples of treated and control units when the bandwidth is 0.04. The analysis refers to students as unit level.

```
students_data_randomized_03_2011 <-
students_data_2011[which(students_data_2011$GOKpercentage >= .063
                         & students_data_2011$GOKpercentage <= .137),]


students_data_randomized_03_2011_variables <-
students_data_randomized_03_2011[, 9:ncol(students_data_randomized_03_2011)]

mean.treated<-as.data.frame(apply(students_data_randomized_03_2011_variables[
  which(students_data_randomized_03_2011$eligible_dummy==1),], 2, mean))
mean.control<-as.data.frame(apply(students_data_randomized_03_2011_variables[
  which(students_data_randomized_03_2011$eligible_dummy==0),], 2, mean))
sd.treated<-as.data.frame(apply(students_data_randomized_03_2011_variables[
  which(students_data_randomized_03_2011$eligible_dummy==1),], 2, sd))
sd.control<-as.data.frame(apply(students_data_randomized_03_2011_variables[
  which(students_data_randomized_03_2011$eligible_dummy==0),], 2, sd))


# Standardized difference in means
diff.means <- mean.treated[,1] - mean.control[,1]
```

```
standard.diff.means <-  (mean.treated[,1] - mean.control[,1])/
                         sqrt((sd.treated^2 + sd.control^2)/2)

# Check this t-test
t =  mean.treated[,1] - mean.control[,1]/
  sqrt((sd.control^2/length(which(students_data_randomized_03_2011$eligible_dummy == 0))) +
      (sd.treated^2/length(which(students_data_randomized_03_2011$eligible_dummy == 1))))

# 99% CI (t-student distribution)
x0 <- standard.diff.means - 1.96 * sqrt((sd.treated^2 + sd.control^2)/2)
#check that dimension is correct
x1 <- standard.diff.means + 1.96 * sqrt((sd.treated^2 + sd.control^2)/2)
#check that dimension is correct


rownames(standard.diff.means)<-c(names(students_data_randomized_03_2011_variables))
```



We now check whetere or not the means for the covariates in the groups of units assigned to the control and to the treatment are significatively different.

First, we perform the analysis sampling 50 students from each school.

```
schools <-
students_data_randomized_03_2011$school[which(
!duplicated(students_data_randomized_03_2011$school))]

sample_students <- as.data.frame(matrix(data = NA, nrow = 50*length(schools),
```

```
                                            ncol = ncol(students_data_randomized_03_2011)))
colnames(sample_students) <- colnames(students_data_randomized_03_2011)

for (j in (0:(length(schools)-1))){
  set.seed(j  + 123)
  sample_students[(1+(j*50)):(50+(j*50)),]  <-
  students_data_randomized_03_2011[which(
    students_data_randomized_03_2011$school %in%
    schools[j+1]),][sample(1:nrow(students_data_randomized_03_2011[which(
    students_data_randomized_03_2011$school %in% schools[j+1]),]),
    50, replace = FALSE ),]
}

length(which(is.na(sample_students)))

sample_students <- round(sample_students[, -(1:4)], 0)
sample_students_variables <- sample_students[, -(1:4)]

mean.full <- as.data.frame(apply(sample_students_variables, 2, mean))
sd.full <- as.data.frame(apply(sample_students_variables, 2, sd))


mean.treated<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==1),], 2, mean))
mean.control<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==0),], 2, mean))
sd.treated<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==1),], 2, sd))
sd.control<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==0),], 2, sd))

# Sampled Data
pvalue <- matrix(NA, ncol=1, nrow = 7)

#Primary retentions
pvalue[1,]<-t.test(
  sample_students$primary_retention[which(sample_students$eligible_dummy==1)],
  sample_students$primary_retention[which(sample_students$eligible_dummy==0)])$p.value

#BULO
pvalue[3,]<-t.test(
  sample_students$BULO[which(sample_students$eligible_dummy==1)],
  sample_students$BULO[which(sample_students$eligible_dummy==0)])$p.value

#Man
pvalue[2,]<-t.test(
  sample_students$man[which(sample_students$eligible_dummy==1)],
  sample_students$man[which(sample_students$eligible_dummy==0)])$p.value

#leerkracht_age
pvalue[4,]<-t.test(
  sample_students$leerkracht_age[which(sample_students$eligible_dummy==1)],
  sample_students$leerkracht_age[which(sample_students$eligible_dummy==0)])$p.value
```

```
#leerkracht_seniority
pvalue[5,]<-t.test(
  sample_students$leerkracht_seniority[which(sample_students$eligible_dummy==1)],
  sample_students$leerkracht_seniority[which(sample_students$eligible_dummy==0)])$p.value

#directie_age
pvalue[6,]<-t.test(
  sample_students$directie_age[which(sample_students$eligible_dummy==1)],
  sample_students$directie_age[which(sample_students$eligible_dummy==0)])$p.value

#directie_seniority
pvalue[7,]<-t.test(
  sample_students$directie_seniority[which(sample_students$eligible_dummy==1)],
  sample_students$directie_seniority[which(sample_students$eligible_dummy==0)])$p.value

pvalue <- as.data.frame(pvalue)

table <- as.data.frame(cbind(mean.control, sd.control, mean.treated, sd.treated,
                             mean.full, sd.full, pvalue))

rownames(table)<-c("Retention", "Gender", "BULO", "TA", "TS", "PA", "PS")

xtable(table, type = "latex", file = "filename.tex", digits=c(3,3,3,3,3,3,3,3))
```

Second, we perform the same analysis but sampling a number of students according to the size of the smallest school (62 students).

```
schools <-
students_data_randomized_03_2011$school[which(
!duplicated(students_data_randomized_03_2011$school))]


sample_students <- as.data.frame(matrix(data = NA, nrow = 62*length(schools),
                                        ncol = ncol(students_data_randomized_03_2011)))
colnames(sample_students) <- colnames(students_data_randomized_03_2011)

for (j in (0:(length(schools)-1))){
  set.seed(j  + 123)
  sample_students[(1+(j*62)):(62+(j*62)),]  <-
  students_data_randomized_03_2011[which(
    students_data_randomized_03_2011$school %in%
    schools[j+1]),][sample(1:nrow(students_data_randomized_03_2011[which(
    students_data_randomized_03_2011$school %in% schools[j+1]),]),
    62, replace = FALSE ),]
}

length(which(is.na(sample_students)))

sample_students <- round(sample_students[, -(1:4)], 0)
sample_students_variables <- sample_students[, -(1:4)]

mean.full <- as.data.frame(apply(sample_students_variables, 2, mean))
sd.full <- as.data.frame(apply(sample_students_variables, 2, sd))
```

```r
mean.treated<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==1),], 2, mean))
mean.control<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==0),], 2, mean))
sd.treated<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==1),], 2, sd))
sd.control<-as.data.frame(apply(sample_students_variables[
  which(sample_students$eligible_dummy==0),], 2, sd))

# Sampled Data
pvalue <- matrix(NA, ncol=1, nrow = 7)

#Primary retentions
pvalue[1,]<-t.test(
  sample_students$primary_retention[which(sample_students$eligible_dummy==1)],
  sample_students$primary_retention[which(sample_students$eligible_dummy==0)])$p.value

#BULO
pvalue[3,]<-t.test(
  sample_students$BULO[which(sample_students$eligible_dummy==1)],
  sample_students$BULO[which(sample_students$eligible_dummy==0)])$p.value

#Man
pvalue[2,]<-t.test(
  sample_students$man[which(sample_students$eligible_dummy==1)],
  sample_students$man[which(sample_students$eligible_dummy==0)])$p.value

#leerkracht_age
pvalue[4,]<-t.test(
  sample_students$leerkracht_age[which(sample_students$eligible_dummy==1)],
  sample_students$leerkracht_age[which(sample_students$eligible_dummy==0)])$p.value

#leerkracht_seniority
pvalue[5,]<-t.test(
  sample_students$leerkracht_seniority[which(sample_students$eligible_dummy==1)],
  sample_students$leerkracht_seniority[which(sample_students$eligible_dummy==0)])$p.value

#directie_age
pvalue[6,]<-t.test(
  sample_students$directie_age[which(sample_students$eligible_dummy==1)],
  sample_students$directie_age[which(sample_students$eligible_dummy==0)])$p.value

#directie_seniority
pvalue[7,]<-t.test(
  sample_students$directie_seniority[which(sample_students$eligible_dummy==1)],
  sample_students$directie_seniority[which(sample_students$eligible_dummy==0)])$p.value

pvalue <- as.data.frame(pvalue)

table <- as.data.frame(cbind(mean.control, sd.control, mean.treated, sd.treated,
                             mean.full, sd.full, pvalue))
rownames(table)<-c("Retention", "Gender", "BULO", "TA", "TS", "PA", "PS")
```

```r
xtable(table, type = "latex", file = "filename.tex", digits=c(3,3,3,3,3,3,3,3))
```