ECON883_HW1

Yutong Shao

2023-01-24

Introduction

1. Paper: Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American economic review.*
2. Abstract of the paper:

   > To the extent that students benefit from high-achieving peers, tracking will help strong ?students and hurt weak ones. However, all students may benefit if tracking allows teachers to better tailor their instruction level. Lower-achieving pupils are particularly likely to benefit from tracking when teachers have incentives to teach to the top of the distribution. We propose a simple model nesting these effects and test its implications in a randomized tracking experiment conducted with 121 primary schools in Kenya. While the direct effect of high-achieving peers is positive, tracking benefited lower-achieving pupils indirectly by allowing teachers to teach to their level.

3. Data source: https://www.openicpsr.org/openicpsr/project/112446/version/V1/view
4. This paper used regression discontinuity (RD) design to test whether students at the median are better off being assigned to the top section (measured by percentile of student).
5. Model:

$$y_{ij} = \delta B_{ij} + \lambda_1 P_{ij} + \lambda_1 P_{ij}^2 + \lambda_1 P_{ij}^3 + X_{ij}\beta + \epsilon_{ij}$$

   where P_{ij} is the percentile of the child on the baseline distribution in his or her school, y_{ij} is the standardized test score, B_{ij} are control variables.

6. Variables:
   1. y = standardized student's score
   2. x = students' percentile in previous exams
   3. control variables:
      - gender: dummy
      - age
      - teacher: dummy, whether the student was taught by civil-service teacher.

```r
# remove variables from work space
rm(list = ls())

set.seed(2023)  # for reproducability

# LIBRARIES
library(haven)  # for data loading
library(dplyr)  # for easy data shaping
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)  # for plotting
library(binsreg)  # for binscatter
library(rdrobust)
library(sandwich)

# kernel mean and local linear regression
source("/Users/shaoyutong/Library/Mobile Documents/com~apple~CloudDocs/ECON883/HW/HW1/plot_funs.R")
source("/Users/shaoyutong/Library/Mobile Documents/com~apple~CloudDocs/ECON883/HW/HW1/loc_lin.R")

path_figure_save <- "/Users/shaoyutong/Library/Mobile Documents/com~apple~CloudDocs/ECON883/HW/HW1/figu
```

Summary of major data set

`Student_test_data.dta` is the main data set in wide format (one observation per student).

It includes baseline characteristics of the students, their test scores at both the endline (fall 2006) and long-term follow-up (fall 2007) tests, and the "treatment" dummies – whether the school was sampled for "Tracking", whether the student was assigned to the Contract Teacher, etc.

I simplify the data set by extracting variables of interest, including:

- y = standardized student's score
- x = students' percentile in previous exams
- control variables: gender, age, teacher

The scatter plot are as follows.

```r
df1 <- read_dta('/Users/shaoyutong/Library/Mobile Documents/com~apple~CloudDocs/ECON883/HW/HW1/data/112
df_pres <- read_dta('/Users/shaoyutong/Library/Mobile Documents/com~apple~CloudDocs/ECON883/HW/HW1/data


# simplify data set
df_simp <- data.frame(SD_std_mark = df1$SDstream_std_mark,
                      MEAN_std_mark = df1$MEANstream_std_mark,
                      x = df1$percentile,
                      xx = df1$realpercentile,
                      total_score = df1$totalscore,
                      y = df1$std_mark,
                      gender = df1$girl,
                      age = df1$agetest,
                      teacher = df1$etpteacher
                      )
dat_use <- na.omit(df_simp) # remove missing values

head(dat_use)
```

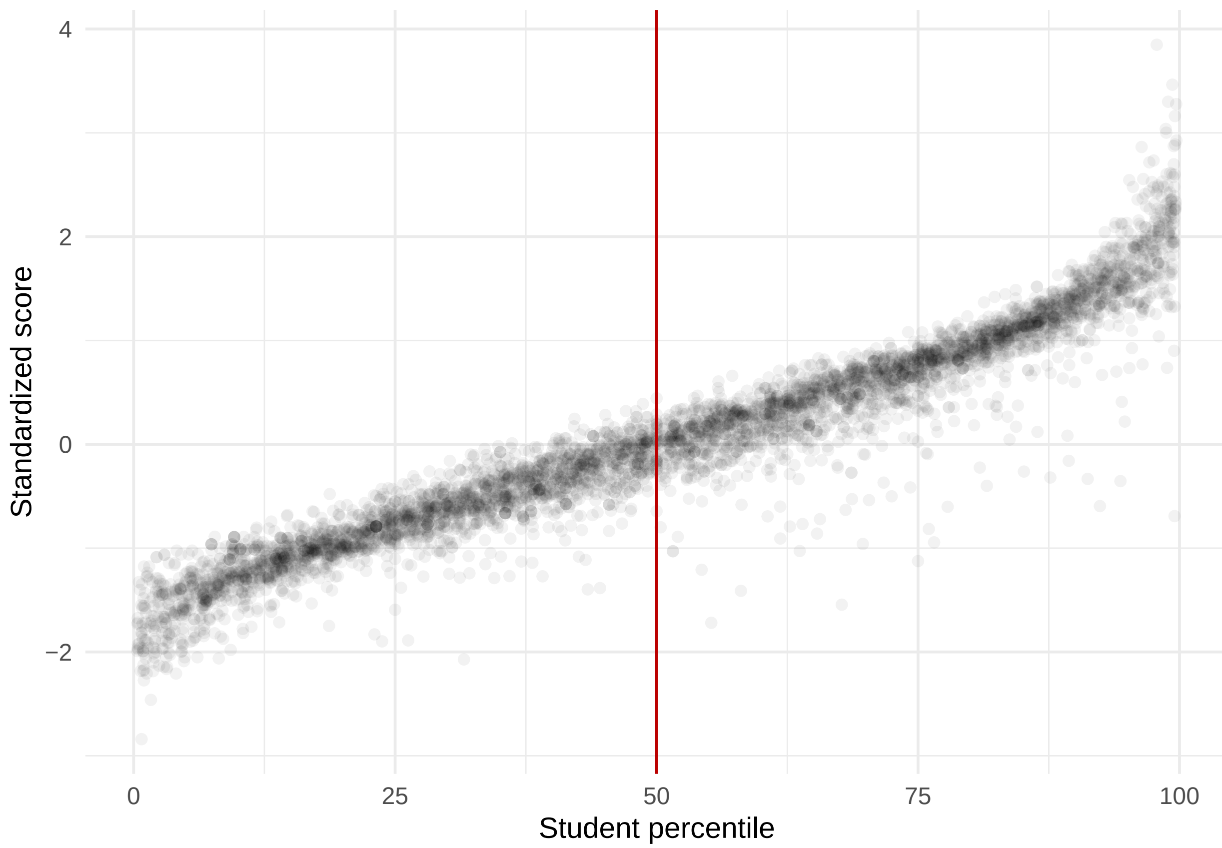
```
##   SD_std_mark MEAN_std_mark         x xx total_score         y gender age
## 3   0.6692453    -0.8192898  3.205132  4   2.9000001 -2.145025      1   8
## 4   0.6692453    -0.8192898  4.487181  5   0.2357143 -1.879484      0  14
## 5   0.6692453    -0.8192898  5.769229  6   9.8571424 -1.809605      0  11
## 6   0.6692453    -0.8192898  8.333338  9   6.5999999 -1.194667      0  10
## 7   0.6692453    -0.8192898 10.897434 11   5.1999998 -1.362377      0  10
## 8   0.6692453    -0.8192898 12.179488 13   9.7428570 -1.124788      1   9
##   teacher
## 3       0
## 4       0
## 5       0
## 6       0
## 7       0
## 8       0
```

```r
summary(dat_use)
```

```
##   SD_std_mark     MEAN_std_mark             x                 xx
##  Min.   :0.1541   Min.   :-0.906560   Min.   : 0.3546   Min.   :  1.00
##  1st Qu.:0.5428   1st Qu.:-0.741663   1st Qu.:27.1792   1st Qu.: 28.00
##  Median :0.7965   Median : 0.005767   Median :51.8182   Median : 52.00
##  Mean   :0.7597   Mean   : 0.006725   Mean   :51.3407   Mean   : 51.83
##  3rd Qu.:0.9880   3rd Qu.: 0.737004   3rd Qu.:75.7501   3rd Qu.: 76.00
##  Max.   :1.1925   Max.   : 0.889537   Max.   :99.7059   Max.   :100.00
##   total_score           y                gender              age
##  Min.   : 0.000   Min.   :-2.84003   Min.   :0.0000   Min.   : 5.000
##  1st Qu.: 5.857   1st Qu.:-0.73795   1st Qu.:0.0000   1st Qu.: 8.000
##  Median :11.286   Median :-0.01431   Median :0.0000   Median : 9.000
##  Mean   :13.068   Mean   : 0.03059   Mean   :0.4913   Mean   : 9.318
##  3rd Qu.:18.861   3rd Qu.: 0.76442   3rd Qu.:1.0000   3rd Qu.:10.000
##  Max.   :42.729   Max.   : 3.84869   Max.   :1.0000   Max.   :19.000
##     teacher
##  Min.   :0.0000
```

```
##   1st Qu.:0.0000
##   Median :1.0000
##   Mean   :0.5026
##   3rd Qu.:1.0000
##   Max.   :1.0000
```

```r
# plot(x=df_simp$x, y = df_simp$y,
#      main = 'Standardized score vs student percentile',
#      xlab = 'percentile',
#      ylab = 'standardized score')
# x = df_simp$x
# y = df_simp$std_score
ggplot(mapping=aes(x=x,y=y)) +
    geom_point(data=dat_use, alpha= 0.05) +
    geom_vline(aes(xintercept=50), colour="#BB0000") +
    xlab('Student percentile') +
    ylab('Standardized score') +
    theme_mp()
```

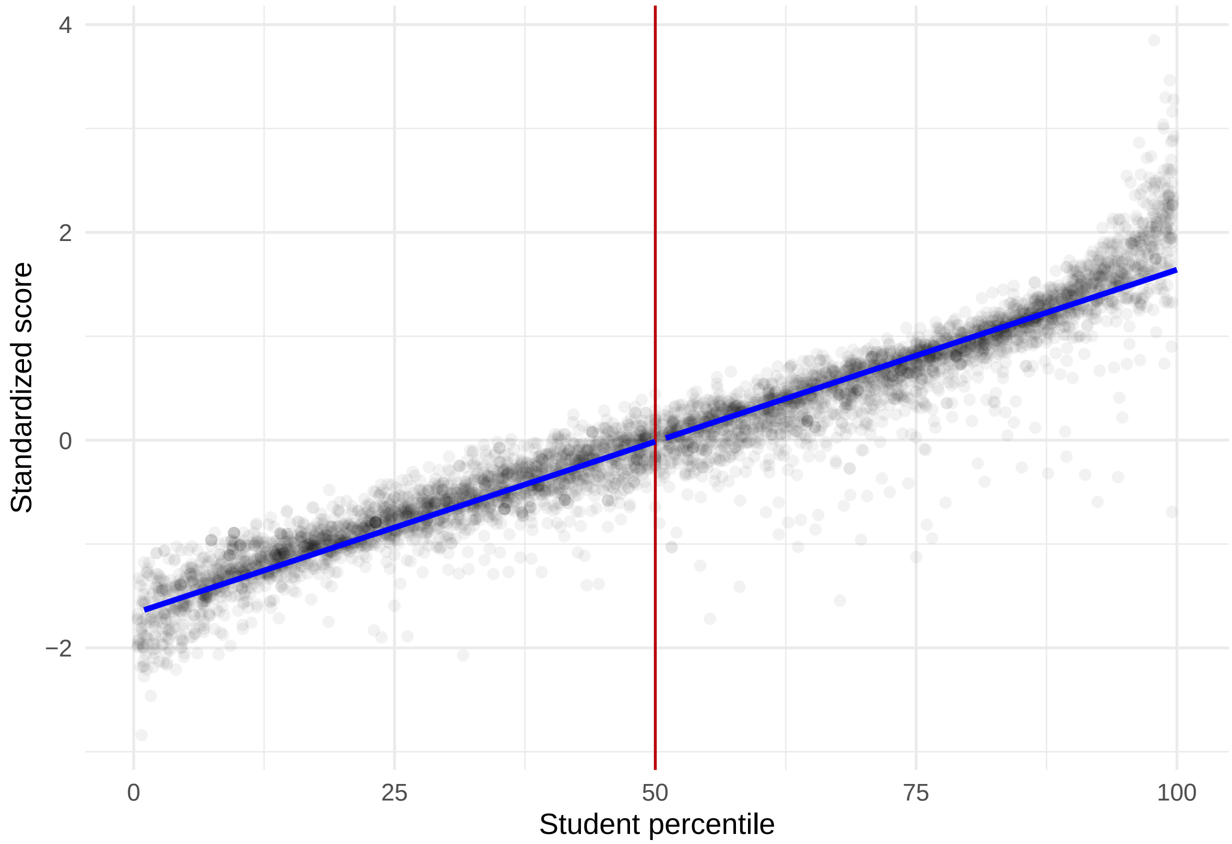


Single covariate

Linear regression

I first run a basic linear regression on each side of the discontinuity. Note that the author used polynomial form, but I implemented a basic one for illustration.

```
group1 <- subset(dat_use, x <= 50)
group2 <- subset(dat_use, x > 50)
lm_1 <- lm(formula = y ~ x, data=dat_use)
lm_2 <- lm(formula = y ~ x, data=dat_use)

# mean of first group
x_curve1 <- sort(unique(group1$xx))

y_curve1 <- predict(lm_1, newdata = data.frame(x=x_curve1))
dat_curve1 <- data.frame(y=y_curve1, x=x_curve1) # population mean curve (blue line)
# dat_curve1
# mean of second group
x_curve2 <- sort(unique(group2$xx))

y_curve2 <- predict(lm_2, newdata = data.frame(x=x_curve2))
dat_curve2 <- data.frame(y=y_curve2, x=x_curve2)


ggplot(mapping=aes(x=x,y=y)) +
    geom_point(data=dat_use, alpha= 0.05) +
    geom_vline(aes(xintercept=50), colour="#BB0000") +
    geom_line(data = dat_curve1, color='blue', linewidth=1.) +
    geom_line(data = dat_curve2, color='blue', linewidth=1.) +
    xlab('Student percentile') +
    ylab('Standardized score') +
    theme_mp()
```



Kernel Regression

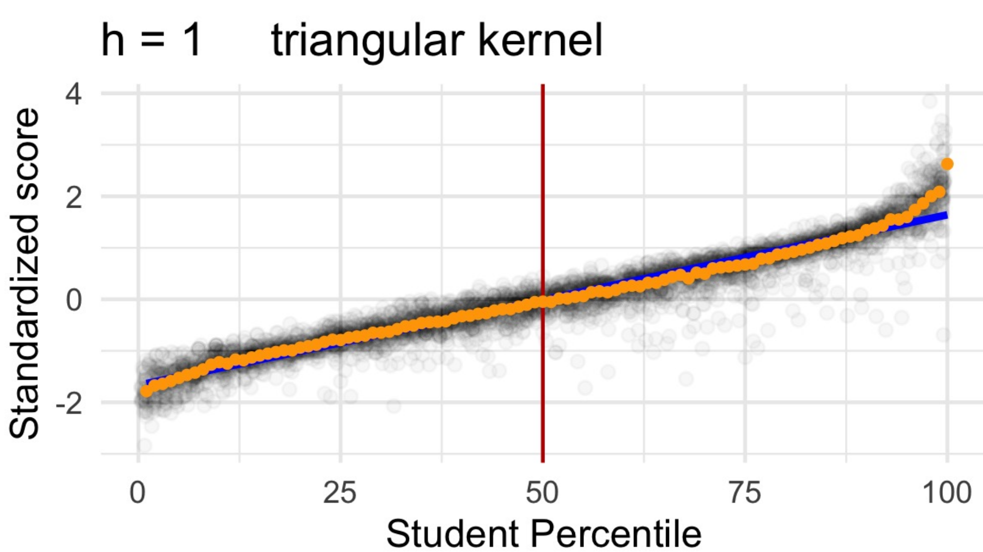
I then applied kernel regression using triangular kernel with bandwidth from 1 to 10. And plotted regression results for 4 of them.

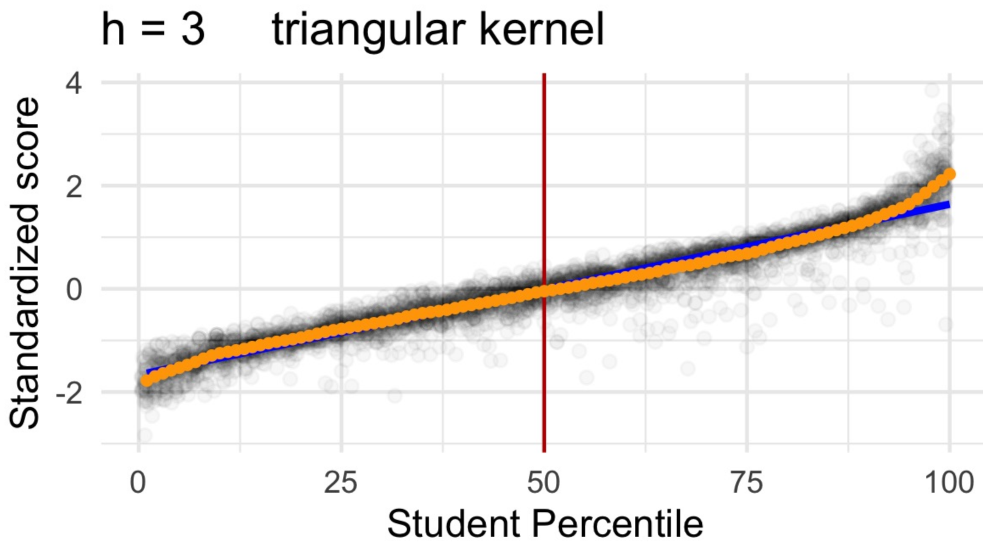
```r
xx <- sort(unique(dat_use$xx))
# xx
# xx <- seq(1, 100, 2)

yy <- kernel_mean(y=dat_use$y, x=dat_use$x, xx=xx, hs=1:10)

# yy
# create plots to illustrate
for (h in c(1,3,5,10)) {
  ggplot(mapping=aes(x=x,y=y)) +
    geom_point(data=dat_use, alpha=0.03) +
    geom_vline(aes(xintercept=50), colour="#BB0000") +
    geom_line(data = dat_curve1, color='blue', linewidth=1.) +
    geom_line(data = dat_curve2, color='blue', linewidth=1.) +
    geom_point(data=data.frame(x=xx,y=yy[,h]), color="orange", size=1) +
    xlab('Student Percentile') +
    ylab('Standardized score') +
    ggtitle(paste("h","=",h,"   ","triangular kernel")) +
    theme_mp()
```

```
  saveplot(paste0(path_figure_save,"triangular_h",h))
}
```

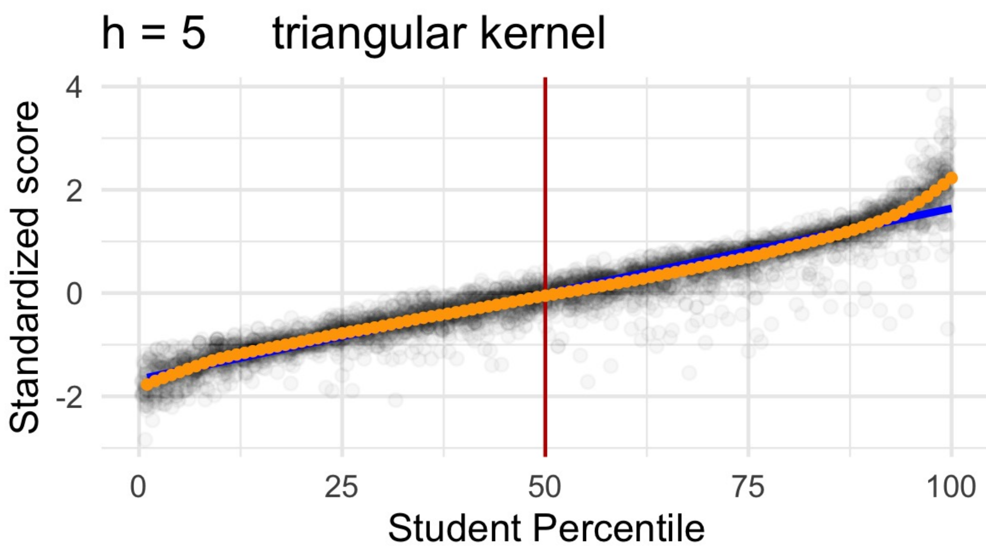
Local Linear Regression

```
kern <- "triangular"
xx <- sort(unique(dat_use$xx))  # evaluation points
reg_loclin <- loclin_reg(y=dat_use$y, x=dat_use$x, xx=xx, hs=1:10, kernel=kern)

yy <- reg_loclin$yy
# yy
# create plots to illustrate
saveplot <- function(filename,
                     plot=last_plot(),
                     width=4, height=2.25, units="in") {
  ggsave(filename=paste0(filename,".jpg"), plot=plot, width=width, height=height, units=units)
}


for (h in c(1,3,5,10)) {
  ggplot(mapping=aes(x=x,y=y)) +
    geom_point(data=dat_use, alpha=0.03) +
    geom_vline(aes(xintercept=50), colour="#BB0000") +
    geom_line(data = dat_curve1, color='blue', linewidth=1.) +
    geom_line(data = dat_curve2, color='blue', linewidth=1.) +
    geom_point(data=data.frame(x=xx,y=yy[,h]), color="orange", size=1) +
    xlab('Student Percentile') +
    ylab('Standardized score') +
    ggtitle(paste("h","=",h,"   ","triangular kernel")) +
    theme_mp()
  saveplot(paste0(path_figure_save,"loclin_",kern,"_h",h))
}
```

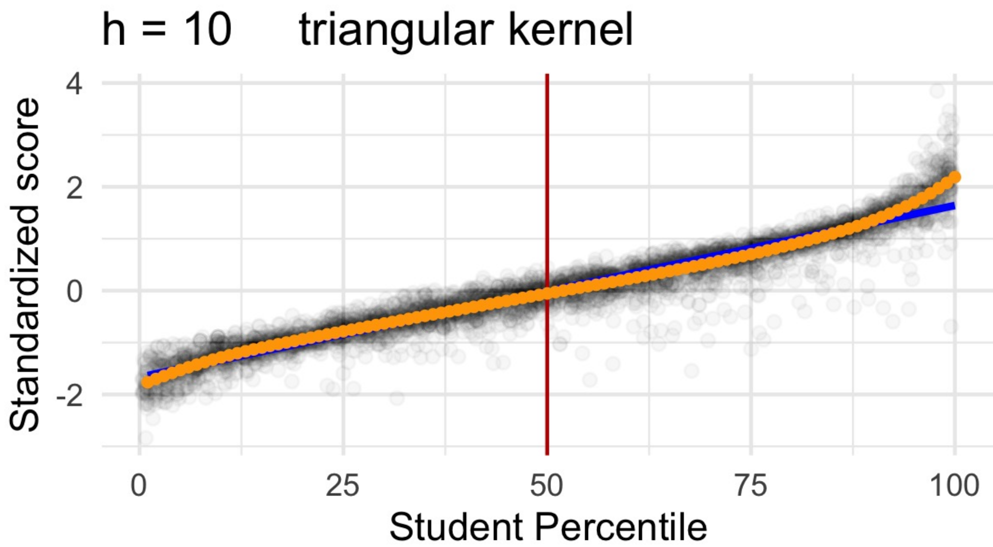




h = 3   triangular kernel
Standardized score
4
2
0
-2
0
25
50
75
100
Student Percentile



h = 5   triangular kernel
Standardized score
4
2
0
-2
0
25
50
75
100
Student Percentile

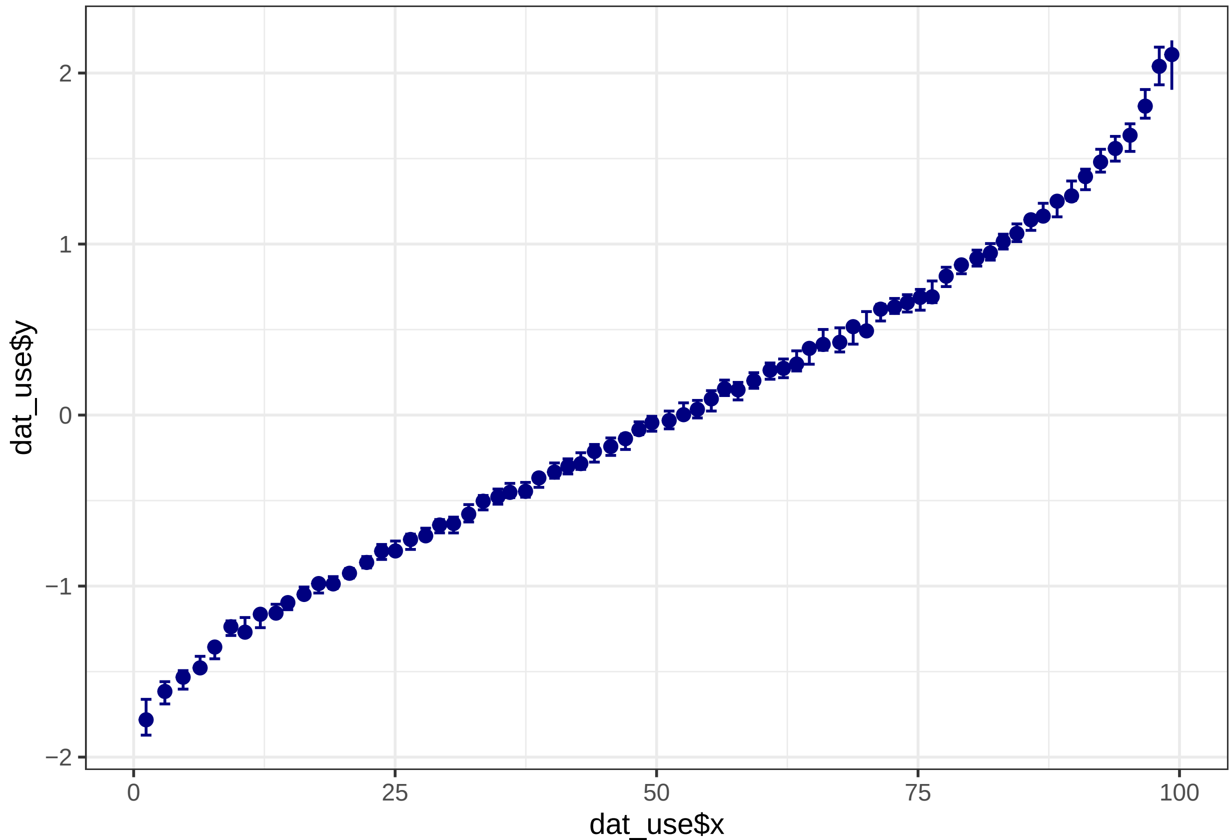


Binscatter Regression

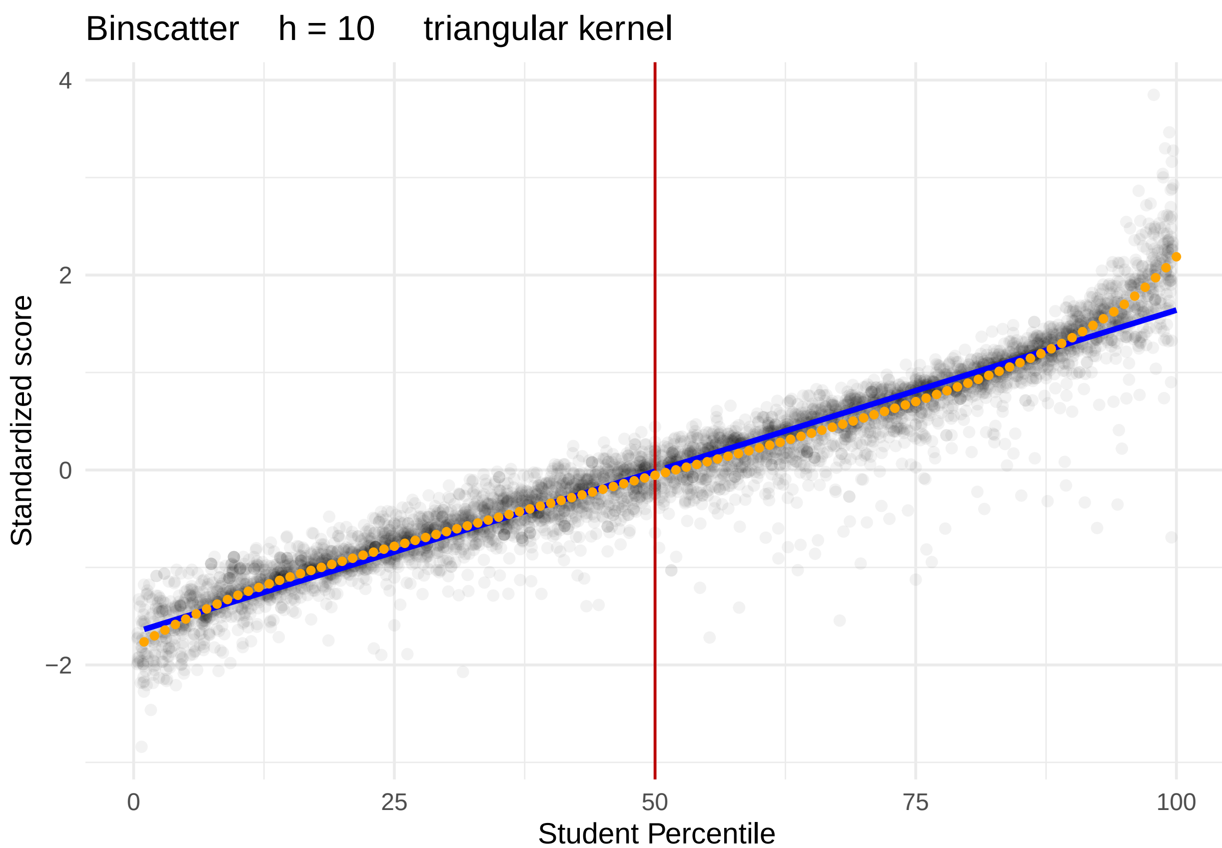
Next I implemented binscatter regression using `R` package. Results are as follows.

```r
# binscatter
breg <- binsreg(dat_use$y,dat_use$x,ci=c(3,3))
```

```
## Warning in binsreg(dat_use$y, dat_use$x, ci = c(3, 3)): To speed up computation,
## bin/degree selection uses a subsample of roughly max(5,000, 0.01n) observations
## if the sample size n>5,000. To use the full sample, set randcut=1.
```




```r
# ci=c(3,3) is what the authors recommend to calculate standard errors
# can also play with many other options to get local linear, etc.

# for ggplot2 plot, grab data underlying figure:
df_breg <- data.frame(x=breg$data.plot$`Group Full Sample`$data.dots$x,
                      y=breg$data.plot$`Group Full Sample`$data.dots$fit#,
                      # ymin=breg$data.plot$`Group Full Sample`$data.ci$ci.l,
                      # ymax=breg$data.plot$`Group Full Sample`$data.ci$ci.r
                      )
ggplot(data=df_breg, aes(x=x,y=y)) +
  geom_point(data=dat_use, alpha=0.05) +
  # geom_errorbar(aes(ymin=min(y),ymax=max(y)), color='blue') +
  # geom_line(data=dat_curve, color="blue", linewidth=1.5) +
  # geom_point(color='orange',size=2) +
  geom_vline(aes(xintercept=50), colour="#BB0000") +
  geom_line(data = dat_curve1, color='blue', linewidth=1.) +
  geom_line(data = dat_curve2, color='blue', linewidth=1.) +
  geom_point(data=data.frame(x=xx,y=yy[,h]), color="orange", size=1) +
  xlab('Student Percentile') +
  ylab('Standardized score') +
  ggtitle(paste('Binscatter','  ',"h","=",h,"   ","triangular kernel")) +
  theme_mp()
```




```
saveplot(paste0(path_figure_save,"binscatter"))
```

Controlling other variables

Incorrect regression: partialing out covariates before binscatter regression

```
source("/Users/shaoyutong/Library/Mobile Documents/com~apple~CloudDocs/ECON883/HW/HW1/plot_funs.r")

nbins <- 20
colnames(dat_use)
```

```
## [1] "SD_std_mark"   "MEAN_std_mark" "x"             "xx"
## [5] "total_score"   "y"             "gender"        "age"
## [9] "teacher"
```

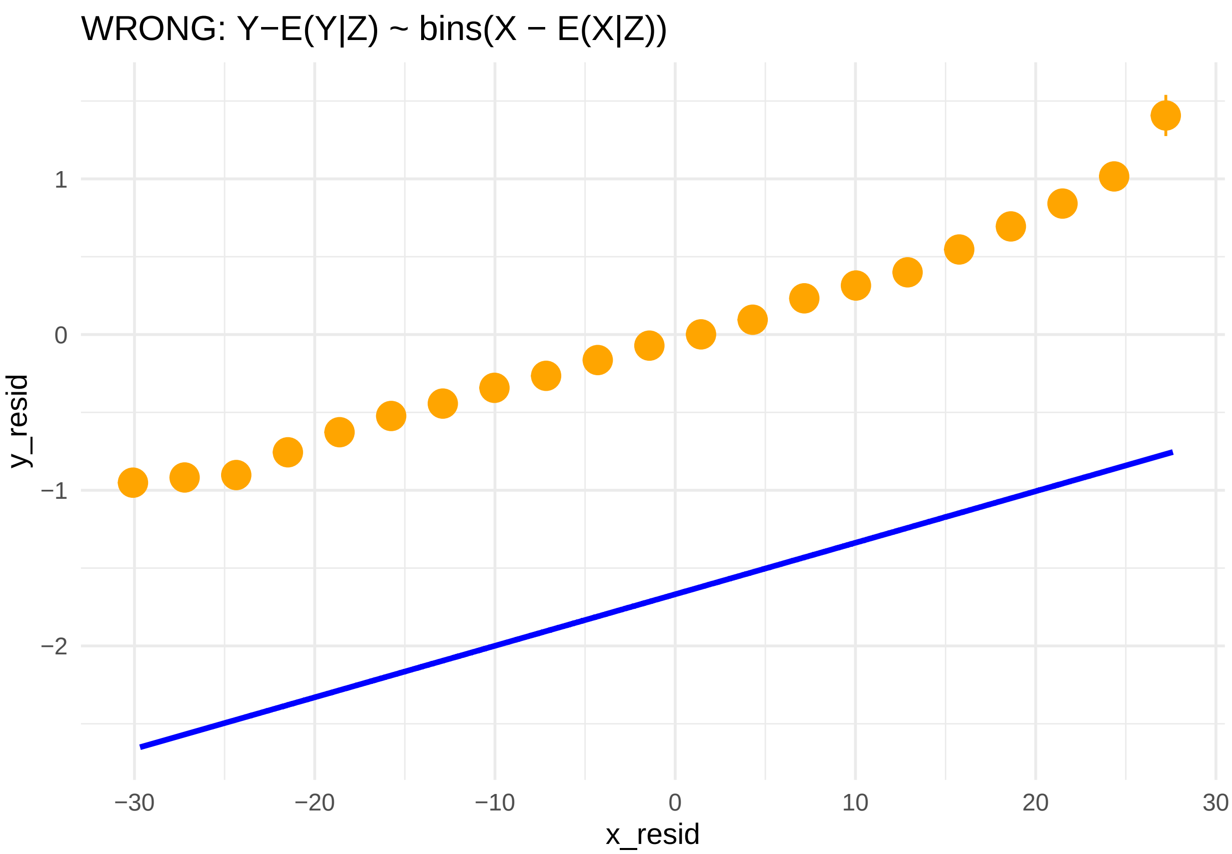
```
# group 1
x1 <- group1$x
y1 <- group1$y
w_gp1 <- group1[,c('gender','age','teacher')]
names(w_gp1) <- c('w1', 'w2', 'w3')
# incorrect: residualize x and y w.r.t. w, then use binscatter on residuals
x_resid1 <- lm(x1 ~ w1 + w2 + w3, data=w_gp1)$residuals
```

```r
y_resid1 <- lm(y1 ~ w1 + w2 + w3, data=w_gp1)$residuals
true1 <- predict(lm_1, newdata = data.frame(x=x_resid1))

df_gp1 <- data.frame(x=x1, y=y1, x_resid=x_resid1, y_resid=y_resid1,
                     true=true1)

# group 2
x2 <- group2$x
y2 <- group2$y
w_gp2 <- group2[,c('gender','age','teacher')]
names(w_gp2) <- c('w1', 'w2', 'w3')
# incorrect: residualize x and y w.r.t. w, then use binscatter on residuals
x_resid2 <- lm(x2 ~ w1 + w2 + w3, data=w_gp2)$residuals
y_resid2 <- lm(y2 ~ w1 + w2 + w3, data=w_gp2)$residuals
true2 <- predict(lm_2, newdata = data.frame(x=x_resid2))
df_gp2 <- data.frame(x=x2, y=y2, x_resid=x_resid2, y_resid=y_resid2,
                     true=true2)

df_all <- rbind(df_gp1, df_gp2)
colnames(df_all)
```

```
## [1] "x"       "y"       "x_resid" "y_resid" "true"
```

```r
# df1 <- data.frame(y1=y1, x1=x1, x_resid=x_resid, y_resid=y_resid)
ggplot(data=df_all, aes(x=x_resid, y=y_resid)) +
  # geom_line(aes(x=dat_curve1$x, y=dat_curve1$y), color='blue', linewidth=1.) +
  # geom_line(aes(x=dat_curve2$x, y=dat_curve2$y), color='blue', linewidth=1.) +
  geom_line(aes(y=true), size=1, color='blue') +
  # geom_line(data = df_gp2, size=1, color='blue') +
  stat_summary_bin(fun.data = mean_se, bins=nbins, size= 0.4, color='orange') +
  stat_summary_bin(fun='mean', bins=nbins,
                   color='orange', size=1) +
  ggtitle('WRONG: Y-E(Y|Z) ~ bins(X - E(X|Z))') +
  theme_mp()
```

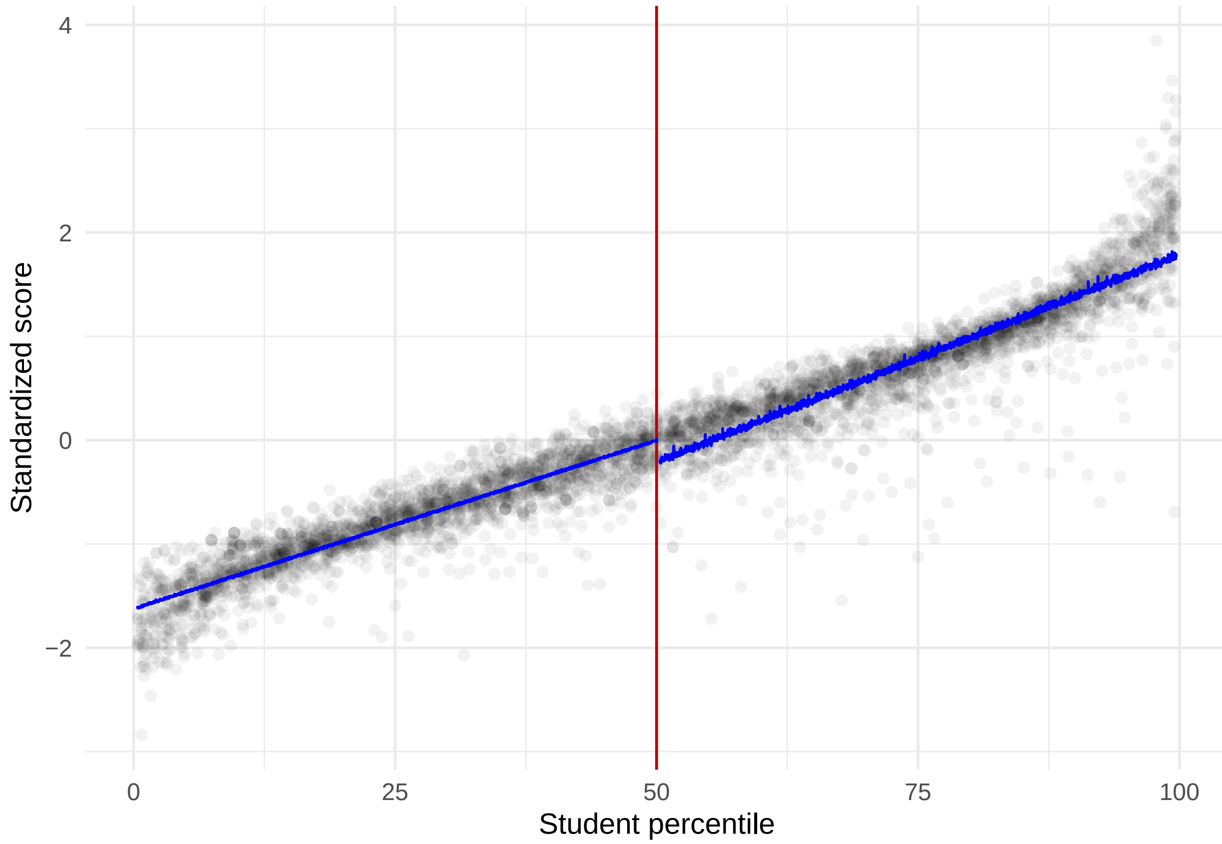
```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.

## Warning: Removed 1 rows containing missing values ('geom_segment()').

## Warning: Removed 21 rows containing missing values ('geom_segment()').
```



Correct regression

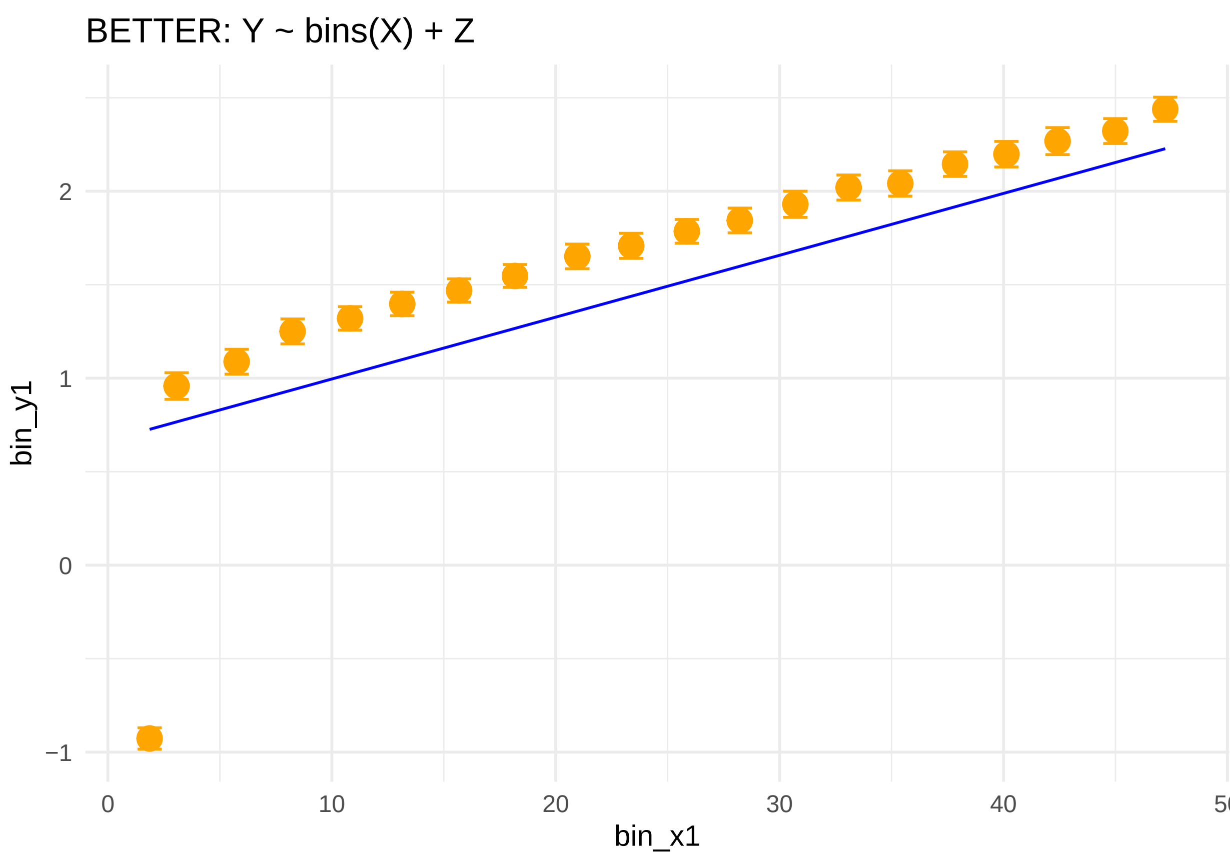
```r
# colnames(dat_use)
# group 1
# group 1
lm_ctrl1 <- lm(y1 ~ x1 + gender + age + teacher, data=group1)

# colnames(group1)
# group 2
lm_ctrl2 <- lm(y2 ~ x2 + gender + age + teacher, data=group2)

y_ctrl_curve1 <- predict(lm_ctrl1, newdata = data.frame(x1=group1$x,
                                                       gender=group1$gender,
                                                       age=group1$age,
                                                       teacher=group1$teacher))

y_ctrl_curve2 <- predict(lm_ctrl2, newdata = data.frame(x2=group2$x,
                                                       gender=group2$gender,
                                                       age=group2$age,
                                                       teacher=group2$teacher))

ctrl_curve1 <- data.frame(x=group1$x, y=y_ctrl_curve1)
ctrl_curve2 <- data.frame(x=group2$x, y=y_ctrl_curve2)
```

```r
ggplot(mapping=aes(x=x,y=y)) +
    geom_point(data=dat_use, alpha= 0.05) +
    geom_vline(aes(xintercept=50), colour="#BB0000") +
    geom_line(data = ctrl_curve1, color='blue') +
    geom_line(data = ctrl_curve2, color='blue') +
    xlab('Student percentile') +
    ylab('Standardized score') +
    theme_mp()
```




```r
xbin1 <- cut(x1, breaks=quantile(x1,probs=seq(0,1,length.out=nbins+1)),
             labels=FALSE)
xbin1[is.na(xbin1)] <- 1

# xbin1

df_bin1 <- data.frame(y=y1, x1=x1,
                      w11=group1$gender,
                      w22=group1$age - mean(group1$age),
                      w33=group1$teacher,
                      xbin1=as.factor(xbin1))
reg_bins1 <- lm(y~1+xbin1+w11+w22+w33, data=df_bin1,x=TRUE)
reg_bins_xval1 <- lm(x1~1+xbin1, data=df_bin1)
df2_1 <- data.frame(bin_x1 = reg_bins_xval1$coefficients[1:nbins],
                    bin_y1 = reg_bins1$coefficients[1:nbins] - mean(y1),
                    bin_y_se1 = sqrt(diag(vcovHC(reg_bins1)))[1:nbins])
```

```r
ggplot(data=df2_1, aes(x=bin_x1, y=bin_y1)) +
  geom_line(
    data =
      data.frame(
    x = seq(min(df2_1$bin_x1), max(df2_1$bin_x1), length.out=100),
    y = predict(lm_1, newdata = data.frame(
      x = seq(min(df2_1$bin_x1), max(df2_1$bin_x1), length.out=100)
    )) - 3*mean(y1)
    ),
    aes(x=x, y=y), color='blue') +
  # geom_line(data = ctrl_curve2, color='blue') +
  geom_errorbar(aes(ymin=bin_y1-1.96*bin_y_se1,ymax=bin_y1+1.96*bin_y_se1),
              color='orange') +
  geom_point(color='orange',size=4) +
  ggtitle('BETTER: Y ~ bins(X) + Z') +
  theme_mp()
```




```r
saveplot(paste0(path_figure_save,"sim_residualize_better"))
```

Reference

1. Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American economic review*, *101*(5), 1739-74.