1. Linear expansions of the matching estimators

The bias-corrected estimator is given by

$$\hat{\tau^{mbc}} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{Y}_i(1) - \hat{Y}_i(0)\} - \frac{1}{n}\sum_{i=1}^{n}\left\{(2Z_i - 1)\frac{1}{M}\sum_{j\in J_i}\{\hat{\mu}_{1-Z_i}(X_i) - \hat{\mu}_{1-Z_i}(X_k)\}\right\} \tag{1}$$

The right hand side is given by

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + (2Z_i - 1)(1 + \frac{K_i}{M})\{Y_i - \hat{\mu}_{Z_i}(X_i)\}\right\} \tag{2}$$

where K_i is the times that unit i is used as a match and M is the number of units that are used for matching.
Note that

$$\sum_{i=1}^{n}\{\hat{Y}_i(1) - \hat{Y}_i(0)\} = \sum_{i=1}^{n}(2Z_i - 1)\left\{Y_i - \frac{1}{M}\sum_{k\in J_i}Y_k\right\} \tag{3}$$

The key to the proof is the following equation.

$$\sum_{i=1}^{n}\sum_{k\in J_i}f(X_k) = \sum_{i=1}^{n}K_i f(X_i) \tag{4}$$

Using (3) and (4), the bias-corrected estimator can be further shown as

$$\begin{aligned}
\hat{\tau^{mbc}} &= \frac{1}{n}\sum_{i=1}^{n}\{\hat{Y}_i(1) - \hat{Y}_i(0)\} - \frac{1}{n}\sum_{i=1}^{n}\left\{(2Z_i - 1)\frac{1}{M}\sum_{j\in J_i}\{\hat{\mu}_{1-Z_i}(X_i) - \hat{\mu}_{1-Z_i}(X_k)\}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}(2Z_i - 1)\left\{Y_i - \hat{\mu}_{1-Z_i}(X_i) - \frac{K_i}{M}(Y_i - \hat{\mu}_{1-Z_i}(X_i))\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}(2Z_i - 1)(Y_i - \hat{\mu}_{1-Z_i}(X_i))(1 - \frac{K_i}{M})
\end{aligned} \tag{5}$$

Subtract (2) by (5)

$$\begin{aligned}
(*) =& \frac{1}{n}\sum_{i=1}^{n}\left\{(2Z_i - 1)\left[(1 - \frac{K_i}{M})(Y_i - \hat{\mu}_{1-Z_i}(X_i) - (1 + \frac{K_i}{M})Y_i + (1 + \frac{K_i}{M})\hat{\mu}_{Z_i}(X_i)\right]\right\} \\
=& \frac{1}{n}\sum_{i=1}^{n}\left\{(2Z_i - 1)\frac{K_i}{M}\left[(2Y_i - \hat{\mu}_{Z_i}(X_i) - \hat{\mu}_{1-Z_i}(X_i) + \hat{\mu}_{Z_i}(X_i) - \hat{\mu}_{1-Z_i}(X_i)\right] - \hat{\mu}_1(X_i) + \hat{\mu}_0(X_i)\right\}
\end{aligned} \tag{6}$$

Note that

$$(2Z_i - 1)(\hat{\mu}_{Z_i}(X_i) - \hat{\mu}_{1-Z_i}(X_i)) = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \tag{7}$$

$$2\bar{Y} = \bar{\mu}_1(X) + \bar{\mu}_0(X) \tag{8}$$

Hence

$$\frac{1}{n}\sum_{i=1}^{n}(2Z_i - 1)\frac{K_i}{M}[(2Y_i - \hat{\mu}_{Z_i}(X_i) - \hat{\mu}_{1-Z_i}(X_i)] = 0 \tag{9}$$

$$\frac{1}{n}\sum_{i=1}^{n}\{(2Z_i - 1)[\hat{\mu}_{Z_i}(X_i) - \hat{\mu}_{1-Z_i}(X_i)] - \hat{\mu}_1(X_i) + \hat{\mu}_0(X_i)\} = 0 \tag{10}$$

The Bias Corrected Estimator on the Treated

Similarly,

$$\begin{aligned}
\hat{\tau}_T^{mbc} &= \frac{1}{n_1}\sum_{i=1}^{n}\left\{Z_i(Y_i - Y_i(0)) - Z_i\frac{1}{M}\sum_{k\in J_i}(\hat{\mu}_0(X_i) - \hat{\mu}_0(X_k))\right\} \\
&= \frac{1}{n_1}\sum_{i=1}^{n}\left\{Z_i[Y_i - \frac{1}{M}\sum_{k\in J_i}(Y_k + \hat{\mu}_0(X_i) - \hat{\mu}_0(X_k))]\right\} \\
&= \frac{1}{n_1}\sum_{i=1}^{n}Z_i\left\{Y_i - \hat{\mu}_0(X_i) + \frac{K_i}{M}(Y_i - \hat{\mu}_0(X_i))\right\}
\end{aligned} \tag{11}$$

$$\begin{aligned}
\frac{1}{n_1}\sum_{i=1}^{n}\hat{\psi}_{T,i} &= \frac{1}{n_1}\sum_{i=1}^{n}\left\{Z_iY_i - Z_i\hat{\mu}_0(X_i) - (1 - Z_i)\frac{K_i}{M}(Y_i - \hat{\mu}_0(X_i))\right\} \\
&= \frac{1}{n_1}\sum_{i=1}^{n}Z_i\left\{Y_i - \hat{\mu}_0(X_i) + \frac{K_i}{M}(Y_i - \hat{\mu}_0(X_i))\right\} \\
&= \hat{\tau}_T^{mbc}
\end{aligned} \tag{12}$$

Note that $(1 - Z_i) = -Z_i$

Problem 2

In a bivariate linear regression,

$$\begin{aligned}
\hat{\tau}_{unadj} &= \frac{cov(Z, Y)}{var(Z)} \\
&= \frac{cov(aX + bU + \epsilon_z, \tau(aX + bU + \epsilon_z) + cU + \epsilon_Y)}{var(aX + bU + \epsilon_z)} \\
&= \frac{\tau var(Z) + bcvar(U)}{var(Z)}
\end{aligned} \tag{13}$$

The cross-products of most terms becomes zero as we assumed i.i.d. draws of (X, U, ϵ_z, ϵ_Y) Note that $var(U) = 1$, $var(Z) = a^2var(X) + b^2var(U) + var(\epsilon_z) = a^2 + b^2 + 1$

$$\tau_{unadj} = \tau + \frac{bc}{a^2 + b^2 + 1} \tag{14}$$

In the multivariate regression, $\begin{cases} cov(X_i, \epsilon_i) = 0 \\ cov(Z_i, \epsilon_i) = 0 \end{cases}$

$$\begin{cases} cov(Z, Y) - \tau_{adj}var(Z) - cov(X, Z)\alpha = 0 \\ cov(X, Y) - \tau_{adj}cov(X, Z) - var(X)\alpha = 0 \end{cases} \tag{15}$$

In the specific linear system,

$$\tau a^2 + \tau b^2 + \tau + bc - \tau_{adj}(a^2 + b^2 + 1) - a\alpha = 0 \tag{16}$$

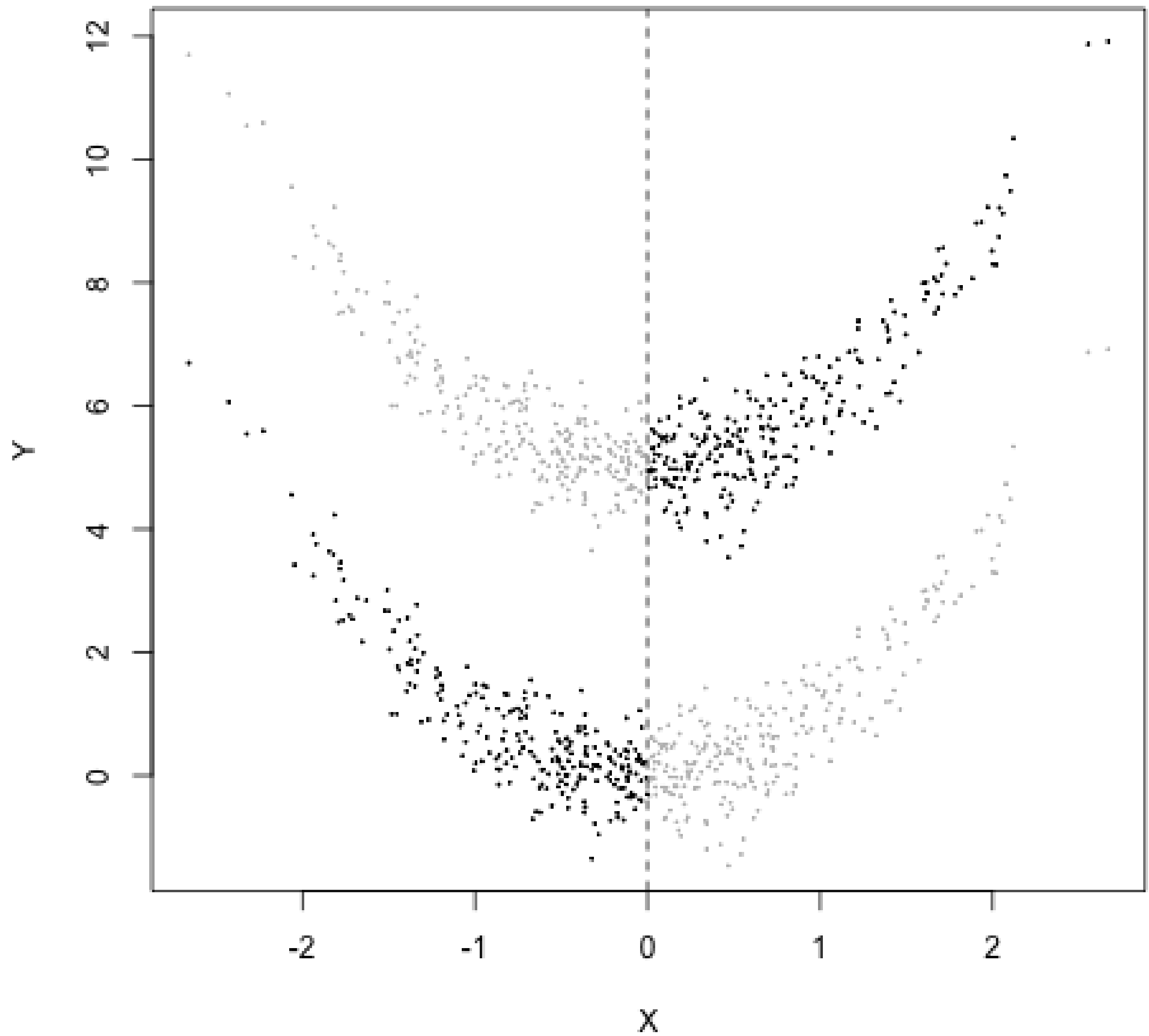
$$\tau a^2 - \tau_{adj}a^2 - a\alpha = 0 \tag{17}$$

Subtract (14) by (15)

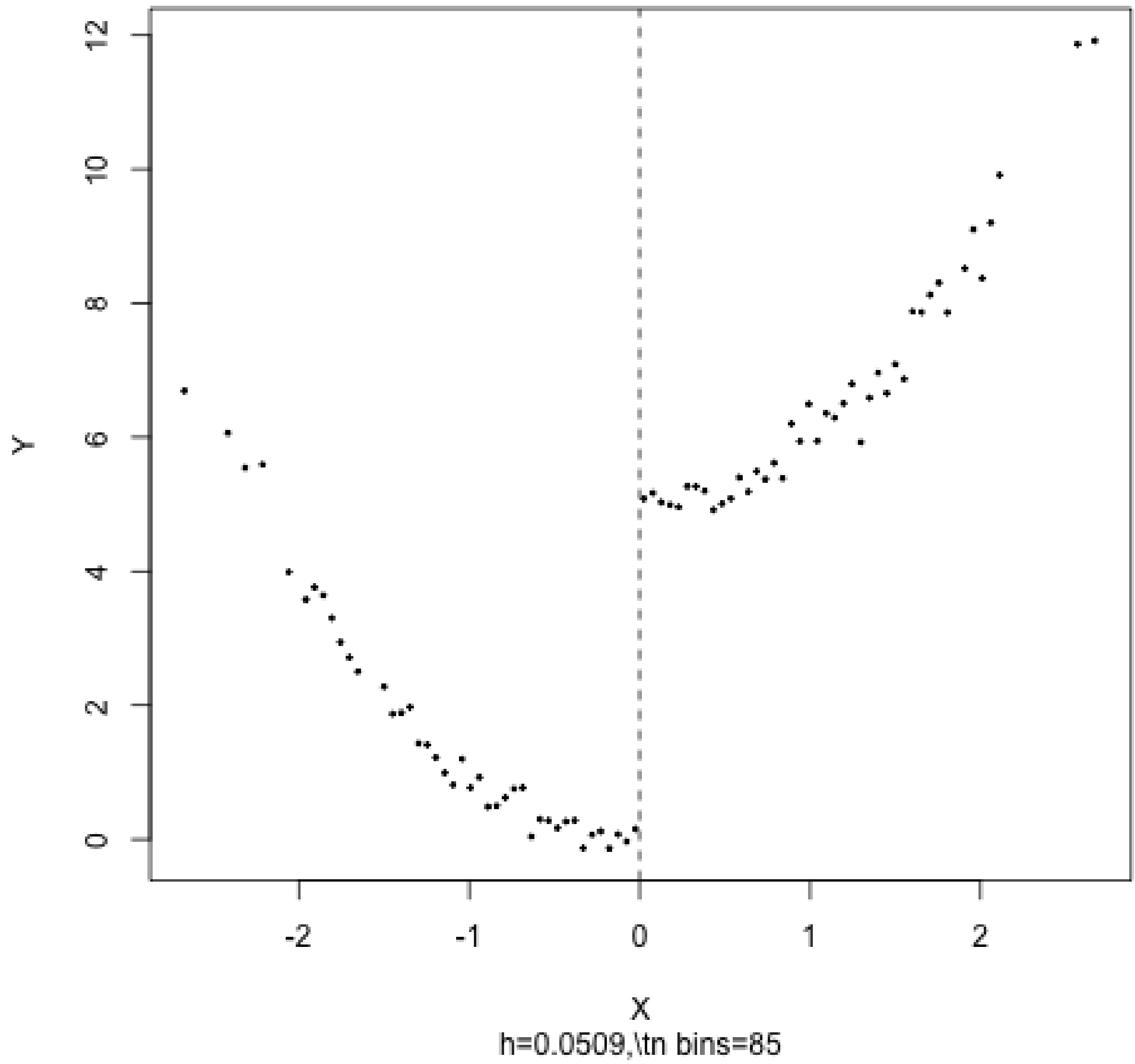
$$\tau_{adj} = \tau + \frac{bc}{b^2 + 1} \tag{18}$$

Problem 4

```r
## RDD numerical examples
set.seed(1000)
n    = 500
x    = rnorm(n)
y0   = x^2 + rnorm(n, 0, 0.5)
y1   = y0 + 5
z    = (x>=0)
y    = z*y1 + (1-z)*y0
plot(y0 ~ x, col = "grey", pch = 19, cex = 0.1,
     ylim = c(min(y), max(y)),
     xlab = "X", ylab = "Y")
points(y1 ~ x, col = "grey", pch = 19, cex = 0.1)
points(y ~ x, col = "black", pch = 19, cex = 0.1)
abline(v = 0, lty = 2)
```




```r
plot(rdd_data(x=x, y=y,cutpoint=0),
     xlab = "X", ylab = "Y", cex = 0.3)
```



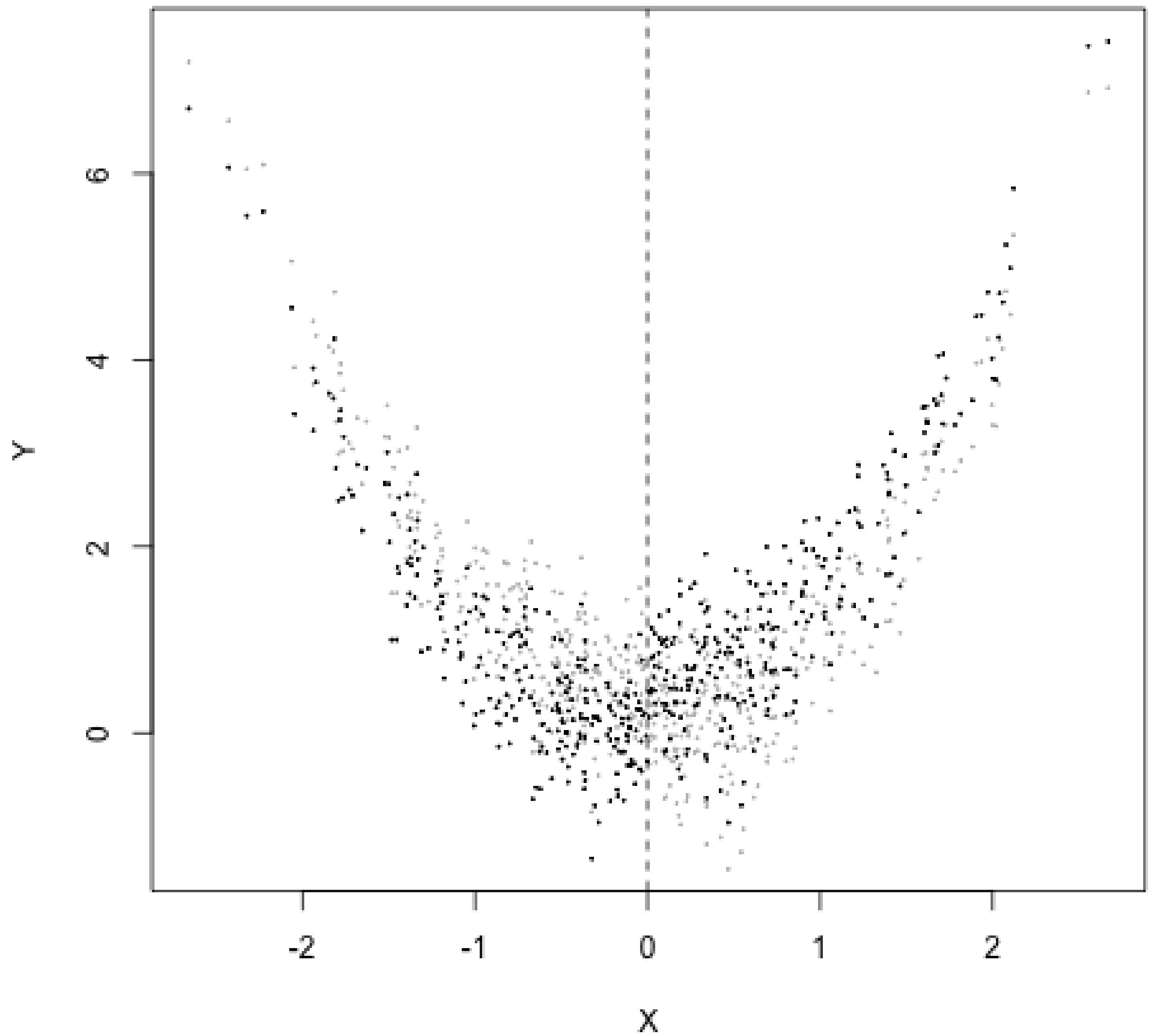

```r
# The point estimate, variance estimate, and confidence interval for the causal effect
RDDest = rdrobust(y, x)
# Generally they're providing us with a good estimate of the causal effect (5)
cbind(RDDest$coef, RDDest$ci)
```

```
##                    Coeff CI Lower CI Upper
## Conventional    5.054218 4.848778 5.259658
## Bias-Corrected 5.049684 4.844244 5.255123
## Robust          5.049684 4.804010 5.295357
```

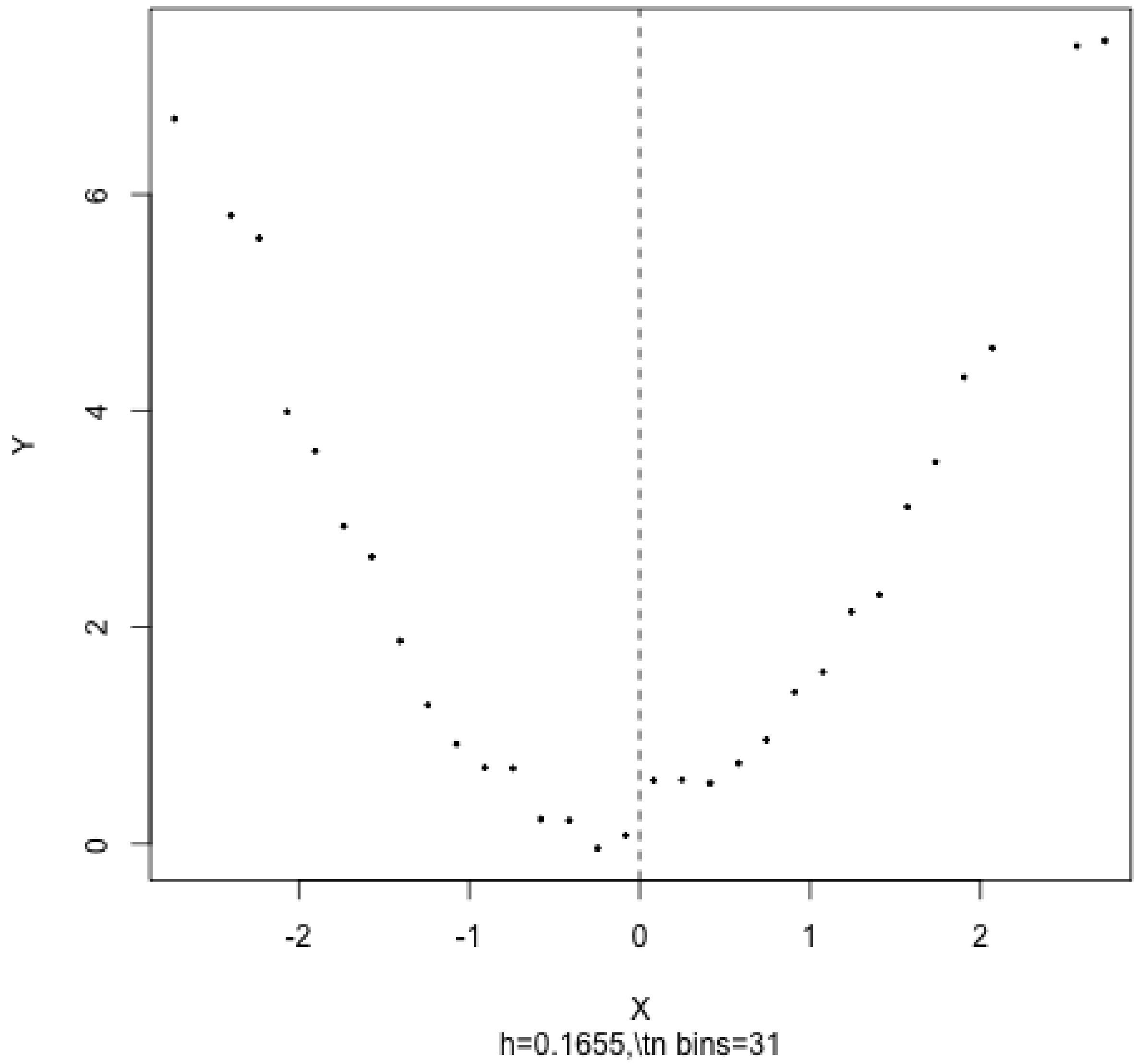
```r
# Lin's estimator
Greg = lm(y ~ z + x + z*x)
cbind(coef(Greg)[2], confint(Greg, 'zTRUE'))
```

```
##                   2.5 \%   97.5 \%
## zTRUE 4.97995 4.790766 5.169134
```

```r
# Shrink the real effect
set.seed(1000)
n   = 500
x   = rnorm(n)
y0  = x^2 + rnorm(n, 0, 0.5)
y1  = y0 + 0.5
z   = (x>=0)
y   = z*y1 + (1-z)*y0
plot(y0 ~ x, col = "grey", pch = 19, cex = 0.1,
     ylim = c(min(y), max(y)),
     xlab = "X", ylab = "Y")
points(y1 ~ x, col = "grey", pch = 19, cex = 0.1)
points(y ~ x, col = "black", pch = 19, cex = 0.1)
abline(v = 0, lty = 2)
```




```
plot(rdd_data(x=x, y=y,cutpoint=0),
     xlab = "X", ylab = "Y", cex = 0.3)
```



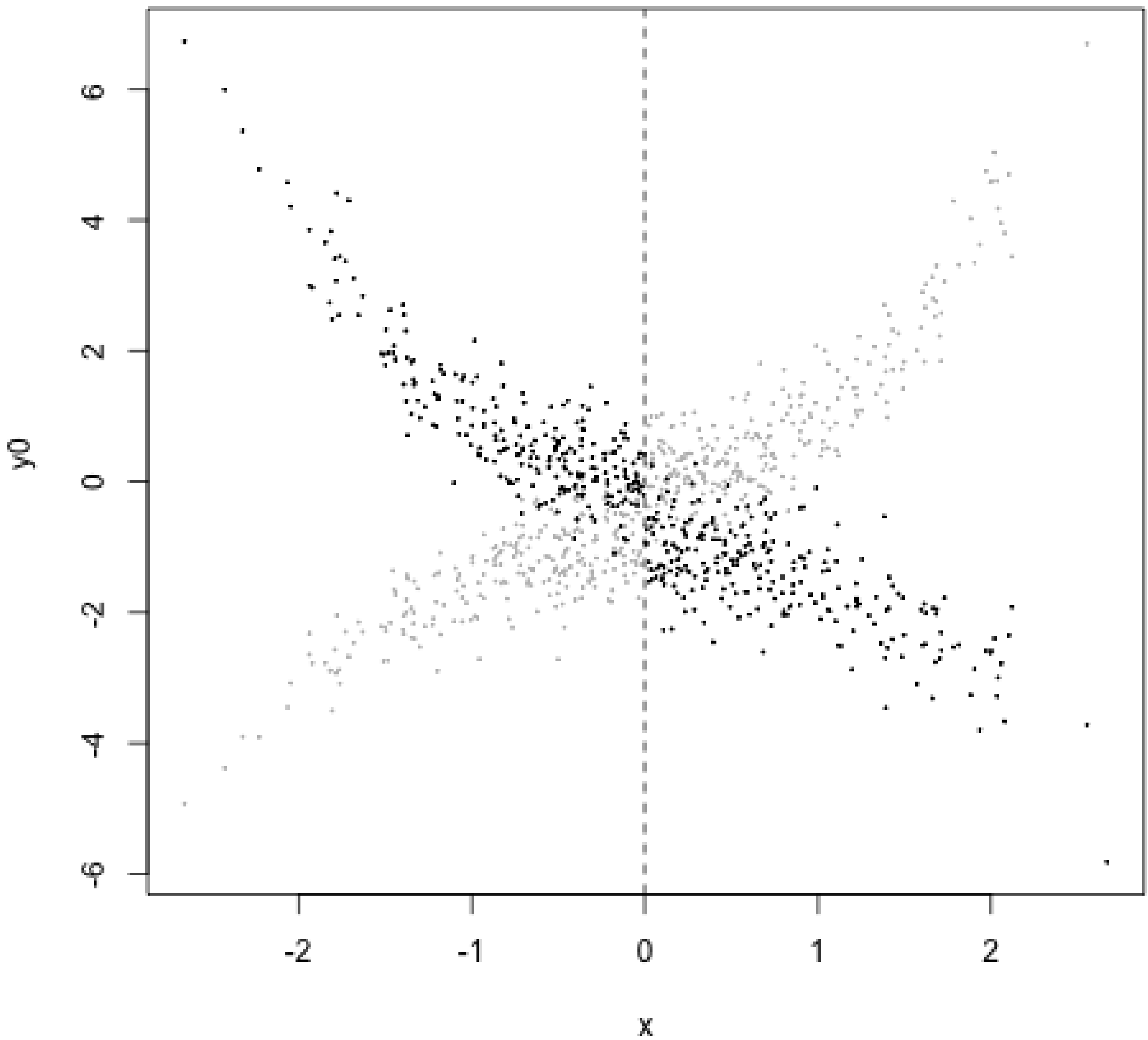

```r
# The point estimate, variance estimate, and confidence interval for the causal effect
RDDest = rdrobust(y, x)
# Stil, that's a pretty good estimate.
cbind(RDDest$coef, RDDest$ci)
```

```
##                     Coeff  CI Lower  CI Upper
## Conventional    0.5542181 0.3487782 0.7596579
## Bias-Corrected 0.5496835 0.3442436 0.7551234
## Robust          0.5496835 0.3040096 0.7953574
```

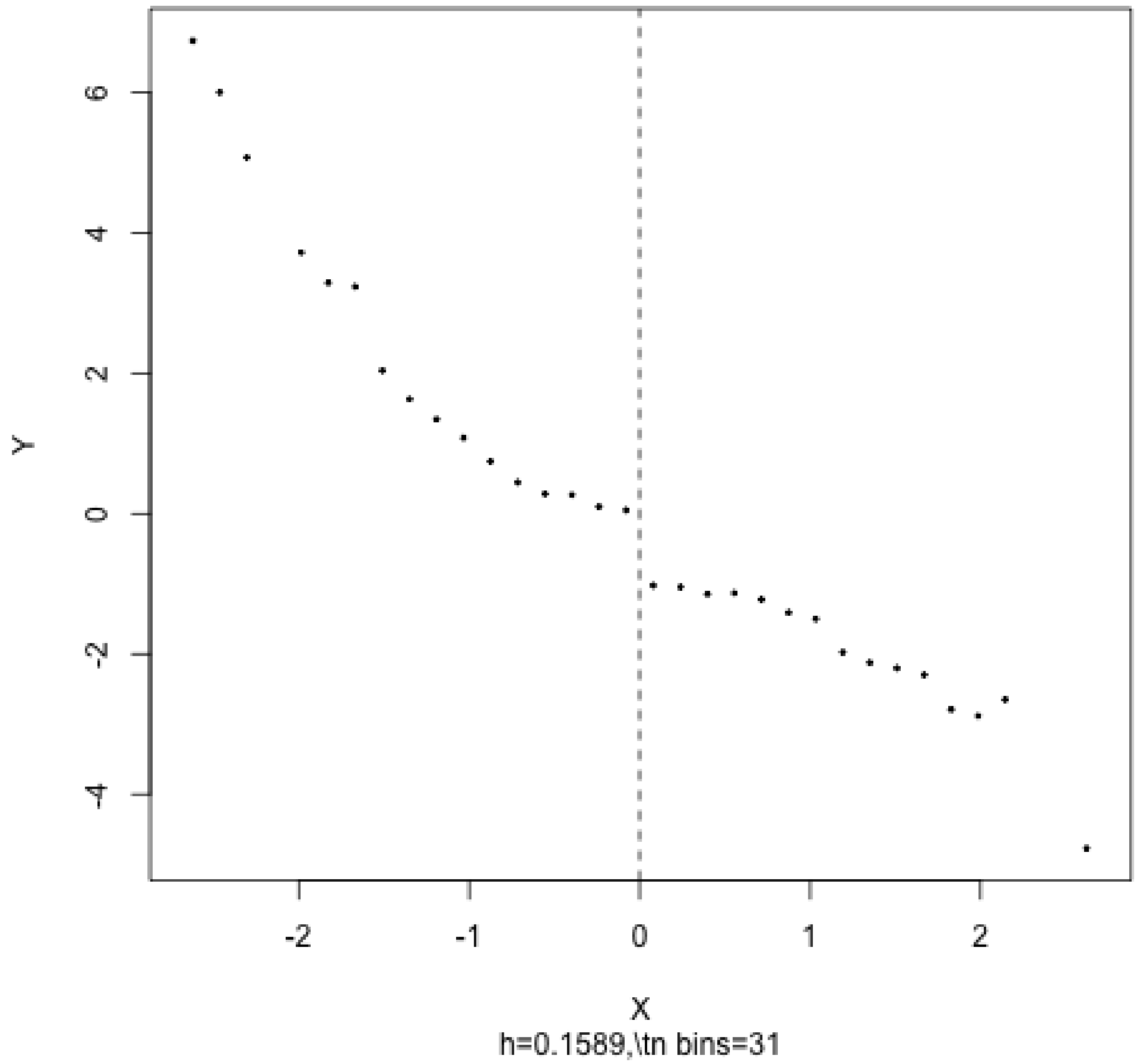
```
# Lin's estimator (Note that the advantage of RDD is that it does not require unconfou
Greg = lm(y ~ z + x + z*x)
cbind(coef(Greg)[2], confint(Greg, 'zTRUE'))
```

```
##                     2.5 \%    97.5 \%
## zTRUE 0.4799505 0.2907665 0.6691345
```

```
y0  = x^2 + rnorm(n, 0, 0.5)
y1  = -1-0.5*x^2 + rnorm(n, 0, 0.5)
z   = (x>=0)
y   = z*y1 + (1-z)*y0
plot(y0 ~ x, col = "grey", pch = 19, cex = 0.1,
     ylim = c(min(y), max(y)))
points(y1 ~ x, col = "grey", pch = 19, cex = 0.1)
points(y ~ x, col = "black", pch = 19, cex = 0.1)
abline(v = 0, lty = 2)
```




```
plot(rdd_data(x=x, y=y,cutpoint=0),
     xlab = "X", ylab = "Y", cex = 0.3)
```




```r
# RDD is still robust assuming non-linear trend
RDDest = rdrobust(y, x)
cbind(RDDest$coef, RDDest$ci)
```

```
##                      Coeff  CI Lower   CI Upper
## Conventional    -0.9798502 -1.224561 -0.7351395
## Bias-Corrected -1.0308944 -1.275605 -0.7861837
## Robust          -1.0308944 -1.305329 -0.7564594
```

```r
# Lin's estimator loses it power
Greg = lm(y ~ z + x + z*x)
cbind(coef(Greg)[2], confint(Greg, 'zTRUE'))
```

```
##                          2.5 \%      97.5 \%
## zTRUE -0.1962193 -0.3689827 -0.02345585
```

Problem 5

```r
data <- matrix(c(0,0,1,1,1,1,
                 0,0,0,0,1,1,
                 0,1,0,1,0,1,
                 74,11514,34,2385,12,9663), nrow = 6,ncol = 4)
# Estimating ITT
ITT <- (data[5,4] + data[6,4])/(data[5,4] + data[6,4] + data[3,4] + data[4,4])
# Estimating the Local Average Treatment Effect
LATE <- (data[4,4] + data[6,4]) / sum(data[c(3,4,5,6),4]) - data[2,4] / sum(data[c(1,2),
tau <- LATE / ITT
print(paste("The Average Causal Effect (Estimated by IV) is ", tau, sep = ""))
```

```
## [1] "The Average Causal Effect (Estimated by IV) is 0.0032280386285733"
```

```r
IV_Wald = function(Z, D, Y)
{
       tau_D = mean(D[Z==1]) - mean(D[Z==0])
       tau_Y = mean(Y[Z==1]) - mean(Y[Z==0])
       CACE  = tau_Y/tau_D

       return(list(tau_D = tau_D, tau_Y = tau_Y,
                   CACE  = CACE))
}

## IV se via the delta method
IV_Wald_delta = function(Z, D, Y)
{
       est         = IV_Wald(Z, D, Y)
       AdjustedY   = Y - D*est$CACE
       VarAdj      = var(AdjustedY[Z==1])/sum(Z) +
                          var(AdjustedY[Z==0])/sum(1 - Z)
       return(sqrt(VarAdj)/abs(est$tau_D))
```

```
}

##IV se via the bootstrap
IV_Wald_bootstrap = function(Z, D, Y, n.boot = 200)
{
       CACEboot  = replicate(n.boot,
                    {
                    bindex = sample(1:length(Z), replace = TRUE)
                    IV_Wald(Z[bindex], D[bindex], Y[bindex])$CACE
                    })

       return(sd(CACEboot))
}

## covariate adjustment in IV analysis
IV_Lin = function(Z, D, Y, X)
{
  X     = scale(as.matrix(X))
  tau_D = lm(D ~ Z + X + Z*X)$coef[2]
  tau_Y = lm(Y ~ Z + X + Z*X)$coef[2]
  names(tau_D) = NULL
  names(tau_Y) = NULL
  CACE  = tau_Y/tau_D

  return(list(tau_D = tau_D, tau_Y = tau_Y,
              CACE  = CACE))
}

## IV_adj se via the delta method
IV_Lin_delta = function(Z, D, Y, X)
{
  X      = scale(as.matrix(X))
  est    = IV_Lin(Z, D, Y, X)

  betaY1 = lm(Y ~ X, subset = (Z == 1))$coef[-1]
  betaY0 = lm(Y ~ X, subset = (Z == 0))$coef[-1]
  betaD1 = lm(D ~ X, subset = (Z == 1))$coef[-1]
  betaD0 = lm(D ~ X, subset = (Z == 0))$coef[-1]

  AdjustedY1    = Y - X%*%betaY1 -
                    (D - X%*%betaD1)*est$CACE
```

```
  AdjustedY0   = Y - X%*%betaY0 -
                 (D - X%*%betaD0)*est$CACE
  VarAdj       = var(AdjustedY1[Z==1])/sum(Z) +
                 var(AdjustedY0[Z==0])/sum(1 - Z)

  return(sqrt(VarAdj)/abs(est$tau_D))
}

##IV_adj se via the bootstrap
IV_Lin_bootstrap = function(Z, D, Y, X, n.boot = 200)
{
  X         = scale(as.matrix(X))
  CACEboot  = replicate(n.boot,
                        {
                          bindex = sample(1:length(Z), replace = TRUE)
                          IV_Lin(Z[bindex], D[bindex], Y[bindex], X[bindex])$CACE
                        })

  return(sqrt(var(CACEboot)))
}
```

Problem 6

```
data <- read.table("fludata.txt")
# Without Covariates
data %>% filter(assign == 1) %>% summarise(mean(receive))
```

```
##   mean(receive)
## 1     0.3077446
```

```
ITT <- data %>% filter(assign == 1) %>% summarise(mean(receive)) - data %>% filter(assi
LATE <- data %>% filter(assign == 1) %>% summarise(mean(outcome)) - data %>% filter(ass
tau_IV <- LATE / ITT
# That's identical to the result by the function in class
IV_Wald(data$assign, data$receive, data$outcome)$CACE
```

```
## [1] -0.1245575
```

```
print(paste("The Average Causal Effect (Estimated by 2sls, without covariates) is ", tau
```

```
## [1] "The Average Causal Effect (Estimated by 2sls, without covariates) is -0.12455748
```

```r
# This process can also be understood as a two-stage least square, which will offer id
# Stage 1
model1 <- lm(receive ~ data$assign, data = data)$fitted.values
# Stage 2
model <- lm(data$outcome ~  model1)
model$coefficients[2]
```

```
##     model1
## -0.1245575
```

```r
# With Covariates (Lin)
IV_Lin(data$assign, data$receive, data$outcome, data[,c(-1,-2,-3)])$CACE
```

```
## [1] -0.125214
```

```r
# Variance Estimation
IV_Lin_delta(data$assign, data$receive, data$outcome, data[,c(-1,-2,-3)])
```

```
## [1] 0.08844344
```

```r
# CI
print(paste("CI: [", IV_Lin(data$assign, data$receive, data$outcome, data[,c(-1,-2,-3)])
```

```
## [1] "CI: [-0.298563111659852,0.0481351769249073]"
```

```r
# # Stage 1
# model1 <- lm(receive ~ .-outcome, data = data)
# # Stage 2
# model <- lm(data$outcome ~ data$age+ data$copd+data$dm+data$heartd+data$race+data$ra
# print(paste("The Average Causal Effect (Estimated by IV, with covariates) is ", mode
```

```r
data <-  read.table("karolinska.txt", header = T)
Y <- as.numeric(data$YearsSurvivingAfterDiagnosis)
X <- as.numeric(data$HighVolTreatHosp)
IV <- as.numeric(data$HighVolDiagHosp)
covariates <- matrix(c(data$FromRuralArea, data$Male, data$AgeAtDiagnosis), ncol = 3)
# Lin's estimator (Unbiased)
IV_Lin(IV, X, Y, covariates)$CACE
```

```
## [1] 0.1849535
```

```r
# Variance estimation
IV_Lin_delta(IV, X, Y, covariates)^2
```

```
## [1] 0.04253986
```

```r
# The result is insignificant
print(paste("CI: [", IV_Lin(IV, X, Y, covariates)$CACE - 1.96*IV_Lin_delta(IV, X, Y, co
```

```
## [1] "CI: [-0.21930025888364,0.589207292815465]"
```

```r
# With Covariates (2SLS)
# Stage 1
model1 <- lm(X~IV+covariates + data$YearOfDiagnosis)
# Stage 2
model <- lm(Y~model1$fitted.values + covariates + data$YearOfDiagnosis)
print(paste("The Average Causal Effect (Estimated by 2SLS, with covariates) is ", model$
```

```
## [1] "The Average Causal Effect (Estimated by 2SLS, with covariates) is 0.174521858685
```

```r
# Note that 2SLS is in itself biased with even larger std error 0.24
temp <- summary(model)
temp$coefficients[2,2]
```

```
## [1] 0.2357129
```