HW4

Wanlin Ji

5/9/2018

Problem 1

a. $\hat{\tau_{OLS}} = \frac{\sum_{i=1}^{I} D_i \times Y_i}{\sum_{i=1}^{I} Y_i^2}$, where I is number of total observations.

b. $\hat{\tau_{OLS}} = \frac{Cov(Y,D)}{Var(Y)}$

c.

d. $\tau_{OLS} = \sum_{i=1}^{N} P_i \tau_{Xi}$ Where N is the number of blocking group.

e. It's different from the OLS one. The weighted one is unbiased and the OLS one is not because there are some random effects and unobserved effects having impact on the OLS result.

Problem 2

a.

```
library(Matching)
```

```
## Loading required package: MASS
```

```
## ##
## ##  Matching (Version 4.9-2, Build Date: 2015-12-25)
## ##  See http://sekhon.berkeley.edu/matching for additional documentation.
## ##  Please cite software as:
## ##   Jasjeet S. Sekhon. 2011. ``Multivariate and Propensity Score Matching
## ##   Software with Automated Balance Optimization: The Matching package for R.''
## ##   Journal of Statistical Software, 42(7): 1-52.
## ##
```

```
library(foreign)

data <- read.csv("~/Desktop/CausalInf/Homeworks/HW4/dataQ2.csv")
head(data)
```

```
##   D X           Y
## 1 0 0  0.01013758
## 2 1 0  9.44139306
## 3 1 0  9.57398593
## 4 0 0 -0.23094179
## 5 1 0 10.32727881
## 6 1 0 10.10340707
```

```
lm1 <- lm(data$Y ~ data$D)
summary(lm1)
```

```
##
## Call:
## lm(formula = data$Y ~ data$D)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.355  -0.735  -0.311   1.428  12.077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7072     0.5242   1.349   0.1776
## data$D       -2.2204     1.1811  -1.880   0.0604 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.86 on 998 degrees of freedom
## Multiple R-squared:  0.003529,   Adjusted R-squared:  0.00253
## F-statistic: 3.534 on 1 and 998 DF,  p-value: 0.0604
```

```
lm2 <- lm(data$Y ~ data$D + data$X)
summary(lm2)
```

```
##
## Call:
## lm(formula = data$Y ~ data$D + data$X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -89.967  -3.639  -3.215   6.793  11.196
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.611      0.620   5.824 7.74e-09 ***
## data$D        -4.243      1.171  -3.625 0.000304 ***
## data$X        -8.269      1.013  -8.162 9.89e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.39 on 997 degrees of freedom
## Multiple R-squared:  0.06594,    Adjusted R-squared:  0.06407
## F-statistic: 35.19 on 2 and 997 DF,  p-value: 1.706e-15
```

```
data0 <- data[data$X==0,]
data1 <- data[data$X==1,]

dif0 <- mean(data0$Y[data0$D==1])-mean(data0$Y[data0$D==0])
dif1 <- mean(data1$Y[data1$D==1])-mean(data1$Y[data1$D==0])
ATE <- dif0*(length(data0)/length(data))+dif1*(length(data1)/length(data))
ATE
```

```
## [1] -90.12302
```

(i) Our estimate is -2.2204 with se as 1.1811
(ii) Our estimate is -4.243 with se as 1.171
(iii) Our estimate is -90.12302

```
Match(data$Y,data$D,data$X, ties= TRUE, estimand = "ATE")
Match(data$Y,data$D,data$X, ties= FALSE, estimand = "ATE")
```

(iv) Our estimate is -23.36667 with se as 3.464732
(v) Our estimate is -23.37438

b. They are different. ATE is resembles the partial regression coefficient in a linear model. The coefficient varies by different model as well as different covariate.

c. Conditions should be D is independent with X. In the methods above, (iii) and (iv) give the unbiased estimate of the ATE.

d. ATE=-100.12456*(1000/1333) - 10.00154*(333/1333)=-77.61071

Problem 3

a.

```
library(dplyr)
```

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

```
## The following object is masked from 'package:MASS':
##
##     select
```

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

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

```
sim1<- read.csv("~/Desktop/CausalInf/Homeworks/HW4/simdata1.csv")
head(sim1)
```

```
##   D          X          Y
## 1 1 -12.203239  188.68025
## 2 1 -22.669386  992.70995
## 3 1   6.847158   75.32212
## 4 1  -7.379723   48.28052
## 5 0  33.307333 -572.23089
## 6 0  -1.629455   -6.24319
```

```
sim11 <- filter(sim1, sim1$D == 1)
sim10 <- sim1[sim1$D==0,]
summary(sim11$X)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -27.700  -4.865   1.668   1.877   8.805  32.940
```

```
summary(sim10$X)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -33.35000  -6.76300  -0.05347  -0.21900   6.92900  33.31000
```

```
summary(sim1$X)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -33.3500  -5.8780   0.7345   0.8941   8.0800  33.3100
```

```
MatchBalance(sim1$D~sim1$X)
```

```
##
```

```
## ***** (V1) sim1$X *****
## before matching:
## mean treatment........ 1.8772
## mean control.......... -0.21898
## std mean diff......... 21.101
##
## mean raw eQQ diff..... 2.1013
## med  raw eQQ diff..... 2.001
## max  raw eQQ diff..... 6.3195
##
## mean eCDF diff........ 0.05669
## med  eCDF diff........ 0.060611
## max  eCDF diff........ 0.094363
##
## var ratio (Tr/Co)..... 0.94415
## T-test p-value........ 0.0010793
## KS Bootstrap p-value.. 0.024
## KS Naive p-value...... 0.023709
## KS Statistic.......... 0.094363
```

b.

(1)

```
lm31=lm(sim1$Y~sim1$D)
summary(lm31)
```

```
##
## Call:
## lm(formula = sim1$Y ~ sim1$D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -682.82  -86.73  -45.82   15.66 2688.67
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  110.587      9.456  11.695   <2e-16 ***
## sim1$D       -16.062     12.977  -1.238    0.216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 204.8 on 998 degrees of freedom
## Multiple R-squared:  0.001533,   Adjusted R-squared:  0.0005322
## F-statistic: 1.532 on 1 and 998 DF,  p-value: 0.2161
```

(2)

```
lm32=lm(sim1$Y~sim1$D+sim1$X)
summary(lm32)
```

```
##
## Call:
## lm(formula = sim1$Y ~ sim1$D + sim1$X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -367.29 -118.08  -46.05   77.70 2376.89
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 108.5263     8.3875  12.939   <2e-16 ***
## sim1$D        3.6665    11.5710   0.317    0.751
## sim1$X       -9.4115     0.5709 -16.486   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 181.6 on 997 degrees of freedom
## Multiple R-squared:  0.2154, Adjusted R-squared:  0.2138
## F-statistic: 136.9 on 2 and 997 DF,  p-value: < 2.2e-16
```

(3)

```
lm33=lm(sim1$Y~sim1$D+sim1$X+sim1$X^2+sim1$X^3)
summary(lm33)
```

```
##
## Call:
## lm(formula = sim1$Y ~ sim1$D + sim1$X + sim1$X^2 + sim1$X^3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -367.29 -118.08  -46.05   77.70 2376.89
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 108.5263     8.3875  12.939   <2e-16 ***
## sim1$D        3.6665    11.5710   0.317    0.751
## sim1$X       -9.4115     0.5709 -16.486   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 181.6 on 997 degrees of freedom
## Multiple R-squared:  0.2154, Adjusted R-squared:  0.2138
## F-statistic: 136.9 on 2 and 997 DF,  p-value: < 2.2e-16
```

(4)

```
Match(sim1$Y,sim1$D,sim1$X)
```

$orig.nobs [1] 1000

$orig.wnobs [1] 1000

$orig.treated.nobs [1] 531

$nobs [1] 1000

$wnobs [1] 531

Notice:(1)The more variables are added into a linear model, the larger R square it would be. (2) Matching and regression with controls for covariate can yeilds unbiased estimate.(3) Adding Square term and cubic term of covariate into regression model is not a good idea.

c.

```r
sim2<- read.csv("~/Desktop/CausalInf/Homeworks/HW4/simdata2.csv")
head(sim2)
```

```
##   D         X          Y
## 1 1 11.230823 121.376224
## 2 0 -2.604471  -5.382584
## 3 1 16.122747 140.927670
## 4 0  4.381380  37.235998
## 5 0  1.646444  11.230539
## 6 1 25.862801 -56.256116
```

```r
sim21=sim2[sim2$D==1,]
sim20=sim2[sim2$D==0,]
summary(sim21$X)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -9.54   13.83   20.06   19.99   26.17   53.14
```

```r
summary(sim20$X)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -29.9800  -7.0330  -0.1055  -0.1861   6.5330  28.0800
```

```r
summary(sim2$X)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -29.9800  -0.8914   9.7580   9.5200  20.0200  53.1400
```

```r
lm34 <- lm(sim2$Y ~ sim2$D)
summary(lm34)
```

```
##
## Call:
## lm(formula = sim2$Y ~ sim2$D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4341.2   -94.5     1.8   160.9  1988.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   108.15      13.44   8.046 2.42e-15 ***
## sim2$D       -171.52      19.38  -8.850  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 306.2 on 998 degrees of freedom
## Multiple R-squared:  0.07276,    Adjusted R-squared:  0.07183
## F-statistic: 78.32 on 1 and 998 DF,  p-value: < 2.2e-16
```

```r
lm35=lm(sim2$Y~sim2$D+sim2$X)
summary(lm35)
```

```
##
## Call:
## lm(formula = sim2$Y ~ sim2$D + sim2$X)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3729.1  -135.8    29.3   152.1  1438.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104.7158    10.8963   9.610   <2e-16 ***
## sim2$D      201.0845    22.6429   8.881   <2e-16 ***
## sim2$X      -18.4652     0.8081 -22.850   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 248.2 on 997 degrees of freedom
## Multiple R-squared:  0.3915, Adjusted R-squared:  0.3902
## F-statistic: 320.7 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
lm36=lm(sim2$Y~sim2$D+sim2$X+sim2$X^2+sim2$X^3)
summary(lm36)
```

```
##
## Call:
## lm(formula = sim2$Y ~ sim2$D + sim2$X + sim2$X^2 + sim2$X^3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3729.1  -135.8    29.3   152.1  1438.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104.7158    10.8963   9.610   <2e-16 ***
## sim2$D      201.0845    22.6429   8.881   <2e-16 ***
## sim2$X      -18.4652     0.8081 -22.850   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 248.2 on 997 degrees of freedom
## Multiple R-squared:  0.3915, Adjusted R-squared:  0.3902
## F-statistic: 320.7 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
rr=Match(sim2$Y,sim2$D,sim2$X)
summary(rr)
```

```
##
## Estimate...  -79.733
## AI SE......  55.203
## T-stat.....  -1.4444
## p.val......  0.14864
##
## Original number of observations..............  1000
## Original number of treated obs...............  481
## Matched number of observations...............  481
## Matched number of observations  (unweighted).  611
```

```
MatchBalance(sim2$D~sim2$X,match.out = rr)
```

```
##
```

```
## ***** (V1) sim2$X *****
##                          Before Matching        After Matching
## mean treatment........     19.993             19.993
## mean control..........   -0.18612             19.005
## std mean diff.........     215.57             10.549
##
## mean raw eQQ diff.....     20.232            0.89534
## med  raw eQQ diff.....     20.125           0.073802
## max  raw eQQ diff.....     25.066             25.066
##
## mean eCDF diff........    0.42999           0.015327
## med  eCDF diff........    0.47196          0.0032733
## max  eCDF diff........    0.71204            0.15385
##
## var ratio (Tr/Co).....    0.86809             1.4151
## T-test p-value........ < 2.22e-16         7.9048e-14
## KS Bootstrap p-value.. < 2.22e-16         < 2.22e-16
## KS Naive p-value...... < 2.22e-16         1.0482e-06
## KS Statistic..........    0.71204            0.15385
```

Differences:(2)and(3)method,i.e., regression with covariate can not give the stable result to balence different value in D on X ,but matching can. This mathcing method did balance treatment and control groups on x, because the standard mean difference has decreased after matching.

Problem 4

1.

```
library(foreign)
ck=read.dta('~/Desktop/CausalInf/Homeworks/HW4/card_krueger.dta')
cknj=ck[ck$nj==1,]
ckpa=ck[ck$pa==1,]
t.test(cknj$bk,ckpa$bk,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  cknj$bk and ckpa$bk
## t = -0.5152, df = 116.88, p-value = 0.6074
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1557951  0.0914714
## sample estimates:
## mean of x mean of y
## 0.4108761 0.4430380
```

```
t.test(cknj$kfc,ckpa$kfc,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  cknj$kfc and ckpa$kfc
## t = 1.1557, df = 128.98, p-value = 0.25
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03812126  0.14519993
## sample estimates:
```

```
## mean of x mean of y
## 0.2054381 0.1518987
```

```
t.test(cknj$roys,ckpa$roys,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  cknj$roys and ckpa$roys
## t = 0.62288, df = 122.03, p-value = 0.5345
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07088651  0.13597504
## sample estimates:
## mean of x mean of y
## 0.2477341 0.2151899
```

```
t.test(cknj$wendys,ckpa$wendys,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  cknj$wendys and ckpa$wendys
## t = -1.1176, df = 107.87, p-value = 0.2662
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1495624  0.0417189
## sample estimates:
## mean of x mean of y
## 0.1359517 0.1898734
```

```
t.test(cknj$co_owned,ckpa$co_owned,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  cknj$co_owned and ckpa$co_owned
## t = -0.2169, df = 116.95, p-value = 0.8287
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1321120  0.1060307
## sample estimates:
## mean of x mean of y
## 0.3413897 0.3544304
```

```
t.test(ckpa$emptot,ckpa$emptot2,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  ckpa$emptot and ckpa$emptot2
## t = 1.3142, df = 135.86, p-value = 0.191
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.093086  5.424254
## sample estimates:
## mean of x mean of y
```

```
##  23.33117  21.16558
```

```
t.test(ckpa$wage_st,ckpa$wage_st2,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  ckpa$wage_st and ckpa$wage_st2
## t = 0.21637, df = 143.96, p-value = 0.829
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1030482  0.1283818
## sample estimates:
## mean of x mean of y
##  4.630132  4.617465
```

```
t.test(ckpa$pmeal,ckpa$pmeal2,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  ckpa$pmeal and ckpa$pmeal2
## t = 0.16356, df = 144.99, p-value = 0.8703
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1745544  0.2060518
## sample estimates:
## mean of x mean of y
##  3.042368  3.026620
```

```
t.test(ckpa$pmeal,ckpa$pmeal2,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  ckpa$pmeal and ckpa$pmeal2
## t = 0.16356, df = 144.99, p-value = 0.8703
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1745544  0.2060518
## sample estimates:
## mean of x mean of y
##  3.042368  3.026620
```

```
t.test(ckpa$hrsopen,ckpa$hrsopen2,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  ckpa$hrsopen and ckpa$hrsopen2
## t = -0.2758, df = 154.99, p-value = 0.7831
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0491182  0.7920588
## sample estimates:
## mean of x mean of y
##  14.52532  14.65385
```

```r
t.test(ckpa$hrsopen,ckpa$hrsopen2,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  ckpa$hrsopen and ckpa$hrsopen2
## t = -0.2758, df = 154.99, p-value = 0.7831
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0491182  0.7920588
## sample estimates:
## mean of x mean of y
##  14.52532  14.65385
```

```r
t.test(cknj$emptot,cknj$emptot2,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  cknj$emptot and cknj$emptot2
## t = -0.80843, df = 637.55, p-value = 0.4191
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.0163300  0.8402872
## sample estimates:
## mean of x mean of y
##  20.43941  21.02743
```

```r
t.test(cknj$wage_st,cknj$wage_st2,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  cknj$wage_st and cknj$wage_st2
## t = -22.97, df = 368.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5088421 -0.4285889
## sample estimates:
## mean of x mean of y
##  4.612134  5.080849
```

```r
t.test(cknj$pmeal,cknj$pmeal2,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  cknj$pmeal and cknj$pmeal2
## t = -1.2349, df = 613.98, p-value = 0.2173
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.16498200  0.03759601
## sample estimates:
## mean of x mean of y
##  3.351061  3.414754
```

```
t.test(cknj$pmeal,cknj$pmeal2,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  cknj$pmeal and cknj$pmeal2
## t = -1.2349, df = 613.98, p-value = 0.2173
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.16498200  0.03759601
## sample estimates:
## mean of x mean of y
##  3.351061  3.414754
```

```
t.test(cknj$hrsopen,cknj$hrsopen2,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  cknj$hrsopen and cknj$hrsopen2
## t = -0.0062802, df = 649.94, p-value = 0.995
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4243697  0.4216639
## sample estimates:
## mean of x mean of y
##  14.41843  14.41978
```

```
t.test(cknj$hrsopen,cknj$hrsopen2,paired=F)
```

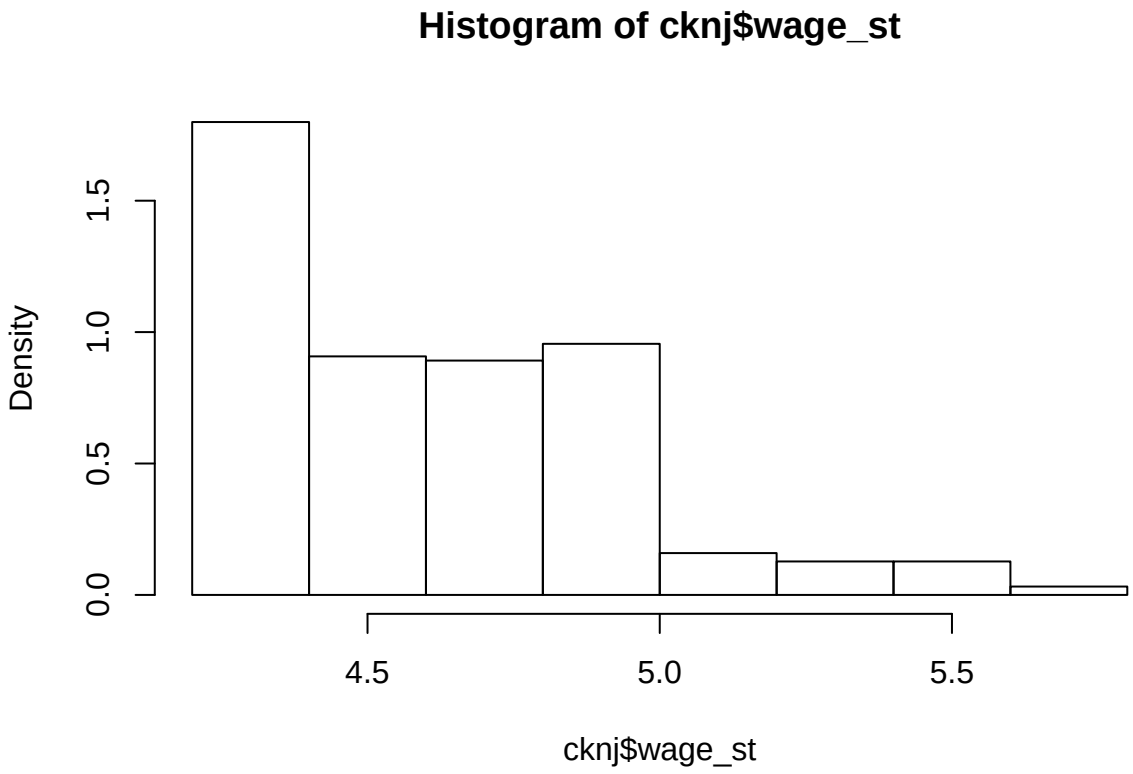
```
##
##  Welch Two Sample t-test
##
## data:  cknj$hrsopen and cknj$hrsopen2
## t = -0.0062802, df = 649.94, p-value = 0.995
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4243697  0.4216639
## sample estimates:
## mean of x mean of y
##  14.41843  14.41978
```

2.

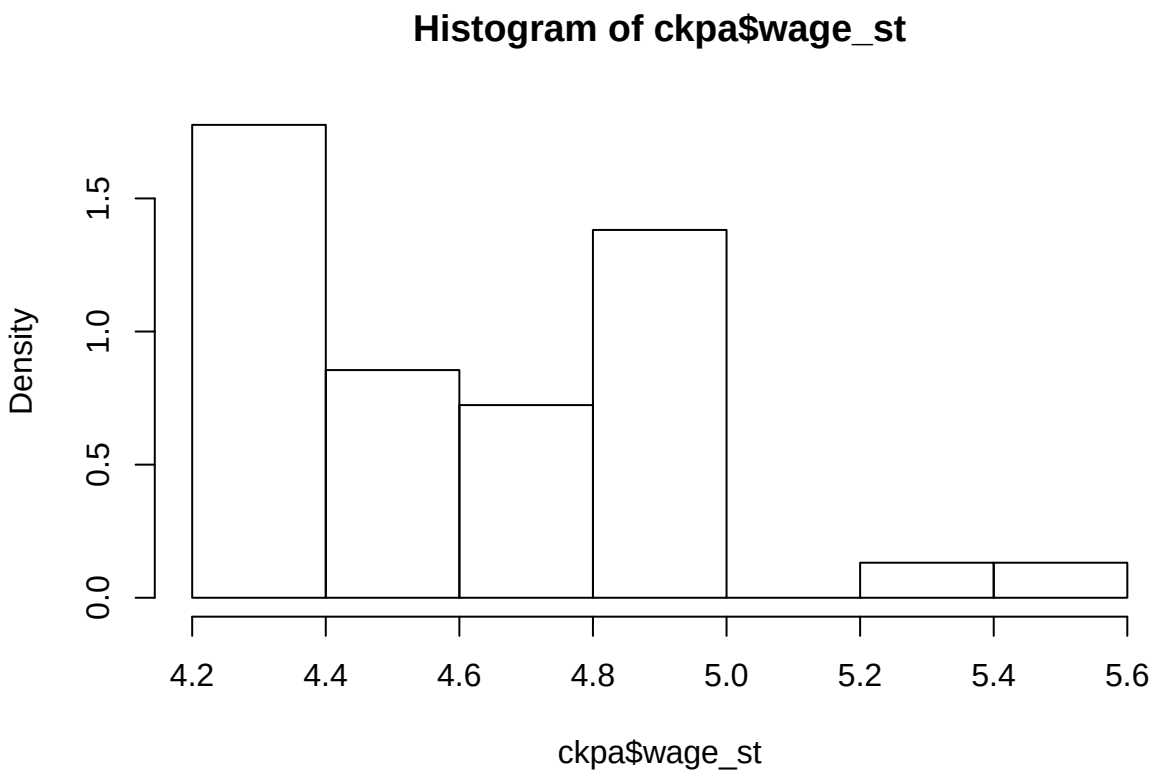
There were no significant difference before and after the changing of minimum wages.

3.

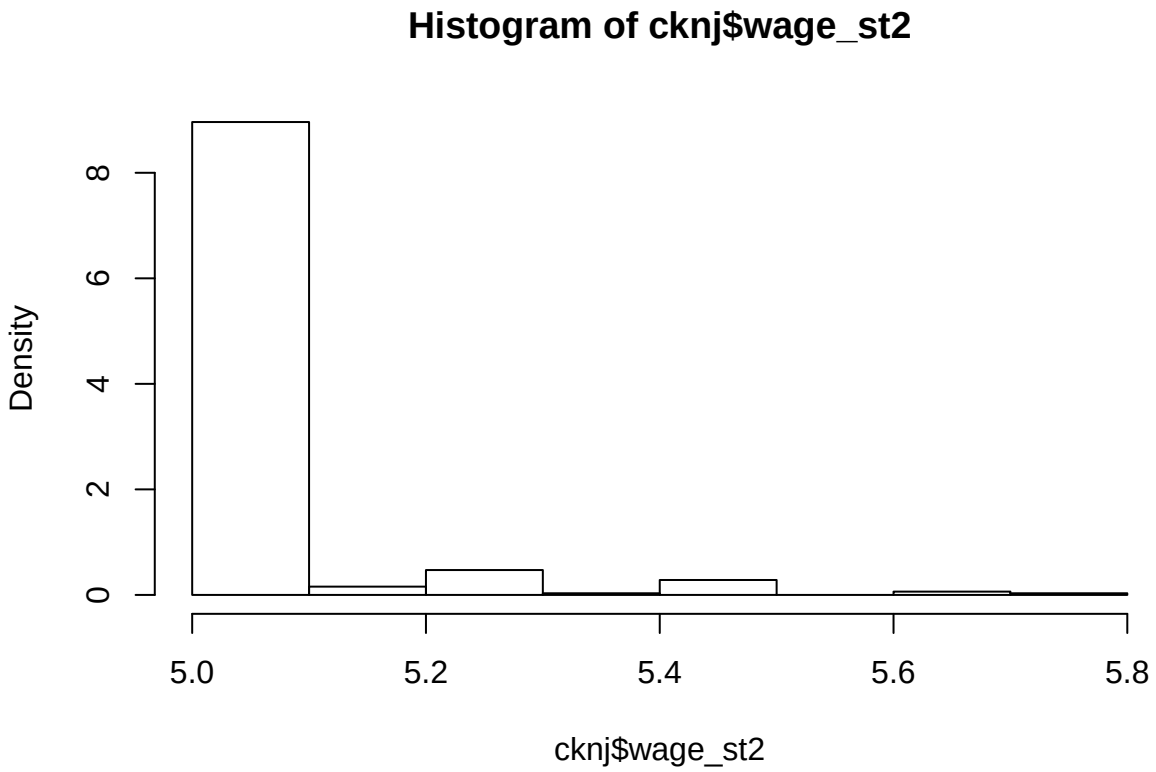
```
hist(cknj$wage_st,freq=F)
```



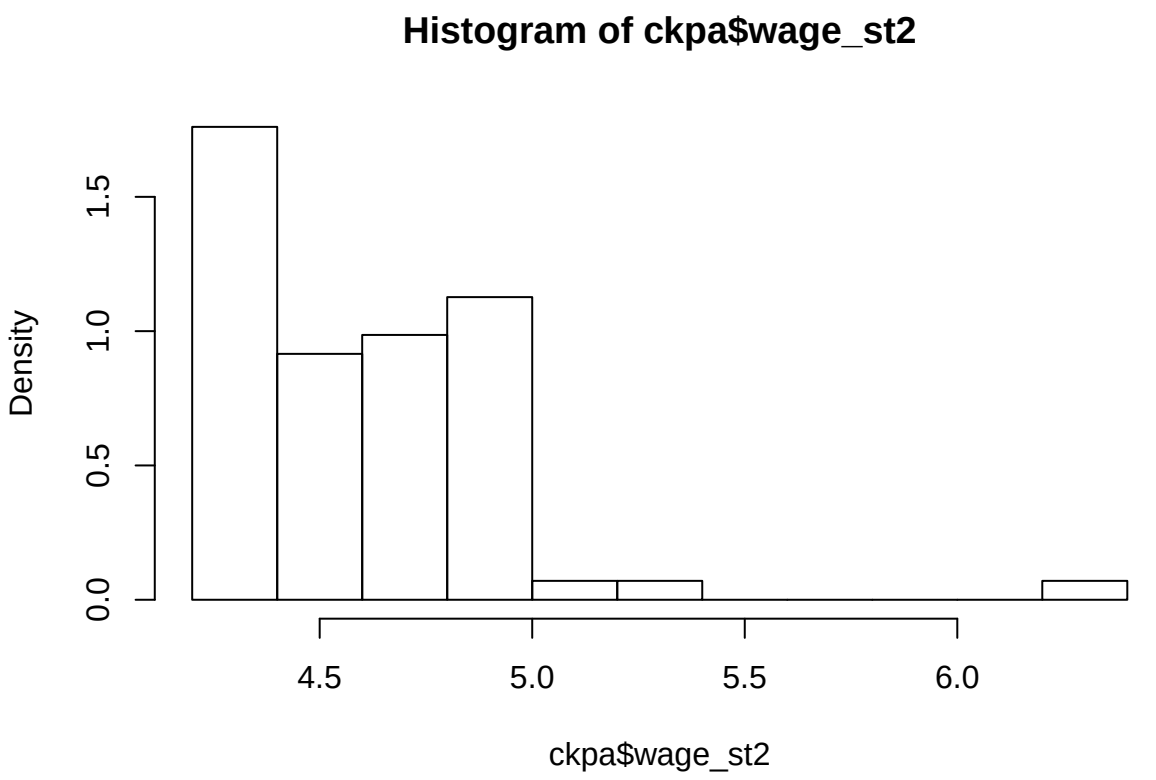

```
hist(ckpa$wage_st,freq=F)
```




```
hist(cknj$wage_st2,freq=F)
```




```
hist(ckpa$wage_st2,freq=F)
```



After adjusting the minimum wage, most of the companies fixed their wage at the minimum wage.

4.

Variable Creation

```
empc=ck$emptot2-ck$emptot
ckpa2<-data.frame(ckpa,iwg=0)
cknj2<-data.frame(cknj,iwg=0)
for(i in 1:331){
  if(is.na(cknj$wage_st[i])=="TRUE"){
    cknj2$iwg[i]=0
  }else
  if(5.05-cknj$wage_st[i]>0){
    cknj2$iwg[i]=5.05-cknj$wage_st[i]
  }else{cknj2$iwg[i]=0
  }
}
ck22=rbind(ckpa2,cknj2)
ck2<-data.frame(ck22,empc)
```

Regressions

```
lm41 = lm(empc ~ nj, data = ck2)
summary(lm41)
```

```
##
## Call:
## lm(formula = empc ~ nj, data = ck2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.217  -3.967   0.533   4.533  33.533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.283      1.036  -2.205   0.0280 *
## nj             2.750      1.154   2.382   0.0177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.968 on 382 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.01464,    Adjusted R-squared:  0.01206
## F-statistic: 5.675 on 1 and 382 DF,  p-value: 0.01769
```

```
lm42=lm(empc~nj+bk+roys+wendys,data = ck2)
summary(lm42)
```

```
##
## Call:
## lm(formula = empc ~ nj + bk + roys + wendys, data = ck2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.734  -3.861   0.564   4.277  33.167
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.4142     1.4032  -1.008   0.3142
```

```
## nj            2.7757     1.1539   2.405   0.0166 *
## bk           -0.3518     1.2321  -0.286   0.7754
## roys         -2.7478     1.3647  -2.013   0.0448 *
## wendys       -0.5282     1.6002  -0.330   0.7415
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.938 on 379 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.02875,    Adjusted R-squared:  0.0185
## F-statistic: 2.804 on 4 and 379 DF,  p-value: 0.02564
```

```
lm43=lm(empc~iwg,data = ck2)
summary(lm43)
```

```
##
## Call:
## lm(formula = empc ~ iwg, data = ck2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.377  -3.912   0.484   4.263  35.123
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.123      0.674  -1.666   0.0965 .
## iwg            2.984      1.401   2.130   0.0338 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.981 on 382 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.01174,    Adjusted R-squared:  0.009151
## F-statistic: 4.537 on 1 and 382 DF,  p-value: 0.0338
```

```
lm44=lm(empc~iwg+bk+roys+wendys,data = ck2)
summary(lm44)
```

```
##
## Call:
## lm(formula = empc ~ iwg + bk + roys + wendys, data = ck2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.825  -3.982   0.357   4.181  34.467
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1994     1.1590  -0.172   0.8635
## iwg           2.8366     1.4239   1.992   0.0471 *
## bk           -0.4753     1.2334  -0.385   0.7002
## roys         -2.6083     1.3720  -1.901   0.0580 .
## wendys       -0.2677     1.6222  -0.165   0.8690
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 8.96 on 379 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.02414,    Adjusted R-squared:  0.01384
## F-statistic: 2.344 on 4 and 379 DF,  p-value: 0.05436
```

```
lm45=lm(empc~iwg+bk+roys+wendys+southj+centralj+pa1+pa2,data = ck2)
summary(lm45)
```

```
##
## Call:
## lm(formula = empc ~ iwg + bk + roys + wendys + southj + centralj +
##     pa1 + pa2, data = ck2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.451  -3.894   0.368   4.164  33.070
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.02156    1.39327   0.733   0.4639
## iwg          1.44253    1.68456   0.856   0.3924
## bk          -0.39013    1.24300  -0.314   0.7538
## roys        -2.95114    1.38210  -2.135   0.0334 *
## wendys      -0.09189    1.62859  -0.056   0.9550
## southj      -0.16853    1.19422  -0.141   0.8878
## centralj    -1.23923    1.37014  -0.904   0.3663
## pa1         -4.58032    1.84114  -2.488   0.0133 *
## pa2         -0.68044    1.71890  -0.396   0.6924
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.927 on 375 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.04147,    Adjusted R-squared:  0.02102
## F-statistic: 2.028 on 8 and 375 DF,  p-value: 0.04225
```

Above all, Initial Wage Gap is the most significant variable to interpret the change of wage, regardless of other variables.

5.

```
empcnj=cknj$emptot2-cknj$emptot
empcpa=ckpa$emptot2-ckpa$emptot
t.test(empcnj,empcpa)
```

```
##
##  Welch Two Sample t-test
##
## data:  empcnj and empcpa
## t = 2.0487, df = 96.884, p-value = 0.0432
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.08578438 5.41421563
## sample estimates:
##  mean of x  mean of y
```

```
##  0.4666667 -2.2833333
```

The t value is just a little larger then critical value. So, to statistical level, the result is credible. However, there should be more variable that could have impact in the change of wage does not includes in this data frame. Thus, there may be a omitted variable bias of this method.