Assignment 6

Kaicheng Luo

2019/11/28

Problem 1

An example of proper IV data is

```
data <- data.frame("Z" = c(1,1,0,0), "D" = c(1,1,0,0), "Y" = c(1,1,0,0))
data
```

```
##   Z D Y
## 1 1 1 1
## 2 1 1 1
## 3 0 0 0
## 4 0 0 0
```

```
Q <- cbind(data$D * data$Y, data$D*(1 - data$Y), data$Y*(1-data$D), data$D+data$Y - data$D * data$Y)
mean(Q[1:2,1]) - mean(Q[3:4,1])
```

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

```
mean(Q[1:2,2]) - mean(Q[3:4,2])
```

```
## [1] 0
```

```
mean(Q[1:2,3]) - mean(Q[3:4,3])
```

```
## [1] 0
```

```
mean(Q[1:2,4]) - mean(Q[3:4,4])
```

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

An example of improper IV data is

```
data <- data.frame("Z" = c(1,1,0,0), "D" = c(1,1,1,0), "Y" = c(1,1,0,0))
data
```

```
##   Z D Y
## 1 1 1 1
## 2 1 1 1
## 3 0 1 0
## 4 0 0 0
```

```
Q <- cbind(data$D * data$Y, data$D*(1 - data$Y), data$Y*(1-data$D), data$D+data$Y - data$D * data$Y)
mean(Q[1:2,1]) - mean(Q[3:4,1])
```

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

```
mean(Q[1:2,2]) - mean(Q[3:4,2])
```

```
## [1] -0.5
```

```
mean(Q[1:2,3]) - mean(Q[3:4,3])
```

```
## [1] 0
```

```
mean(Q[1:2,4]) - mean(Q[3:4,4])
```

```
## [1] 0.5
```

Problem 3

Some sketch about our data: there are many tricky properties in our data. For missing values, we interpolate the income and work length data as zero for those unemployed. This helps us avoid dropping so many NA values in our observations. The total observations dropped is 354 out of 10454 (already dropping those with no response).

```
xdat <- read.csv("X.csv", header = T, sep = ",")
ydat <- read.csv(file="Y.csv", header = TRUE, sep = ",")
dat <- bind_cols(xdat, ydat)
names(dat)
```

```
##  [1] "const"     "wgt"       "female"    "age"       "haschld"   "educ"
##  [7] "educ.m"    "educ.f"    "currjob"   "mosinjob"  "yr_work1"  "earn_yr"
## [13] "white"     "partnered" "evarrst"   "p_inc"     "hh_inc"    "R.educ.m"
## [19] "R.educ.f"  "R.hh_inc"  "I.educ.m"  "I.educ.f"  "I.hh_inc"  "ID"
## [25] "wgt1"      "assign"    "treat"     "r52"       "work52"    "y52"
## [31] "h52"       "w52"       "r130"      "work130"   "y130"      "h130"
## [37] "w130"      "r208"      "work208"   "y208"      "h208"      "w208"
## [43] "TOTHRSW"   "EARNY4"
```

```
dat <- dat %>%
  filter(r52 == 1 & r130 == 1 & r208 == 1) %>%
  mutate(w52 = ifelse(is.na(w52), 0, w52)) %>%
  mutate(w130 = ifelse(is.na(w130), 0, w130)) %>%
  mutate(w208 = ifelse(is.na(w208), 0, w208))
sum(is.na(dat))
```

```
## [1] 354
```

```r
dat <- na.omit(dat)
# 1. Without Covariates
Z <- dat %>%
  select(assign)
D <- dat %>%
  select(treat)
Y <- dat[,39:44]
X <- dat[,2:23]
```

```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))
}
```

```r
# IV wald estimation without covariates
for (i in 1:6)
{
  est = IV_Wald(Z, D, Y[,i])
  estVar = IV_Wald_delta(Z, D, Y[,i])
  print(paste("The (Wald) point estimate of the treatment effect for ", colnames(Y)[i] ," is ", est$CAC
}
```

```
## [1] "The (Wald) point estimate of the treatment effect for work208 is 0.0505451407972023 and the con
## [1] "The (Wald) point estimate of the treatment effect for y208 is 29.3200533263004 and the confiden
## [1] "The (Wald) point estimate of the treatment effect for h208 is 2.25222199510049 and the confiden
## [1] "The (Wald) point estimate of the treatment effect for w208 is 0.580465717577646 and the confide
## [1] "The (Wald) point estimate of the treatment effect for TOTHRSW is -1.24813761309078 and the conf
## [1] "The (Wald) point estimate of the treatment effect for EARNY4 is 19.0338473074643 and the confid
```

```r
IV_Lin = function(Z, D, Y, X)
{
  X = as.matrix(X)
  D = as.matrix(D)
  Y = as.matrix(Y)
  Z = as.matrix(Z)
  tau_D = lm(D ~ Z + X + Z*X)$coef[2]
  tau_Y = lm(Y ~ Z + X + Z*X)$coef[2]
```

```r
  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 = as.matrix(X)
  D = as.matrix(D)
  Y = as.matrix(Y)
  Z = as.matrix(Z)
  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))
}
```

```r
# 2. With Covariates
# IV wald estimation without covariates
for (i in 1:6)
{
  est = IV_Lin(Z, D, Y[,i], X)
  estVar = IV_Lin_delta(Z, D, Y[,i], X)
  print(paste("The (covariate adjusted) point estimate of the treatment effect for ", colnames(Y)[i] ,"
}
```

```
## [1] "The (covariate adjusted) point estimate of the treatment effect for work208 is 0.19753431533633
## [1] "The (covariate adjusted) point estimate of the treatment effect for y208 is 51.3147887798107 an
## [1] "The (covariate adjusted) point estimate of the treatment effect for h208 is 4.96361358500102 an
## [1] "The (covariate adjusted) point estimate of the treatment effect for w208 is 2.75778311527301 an
## [1] "The (covariate adjusted) point estimate of the treatment effect for TOTHRSW is 14.4808177850419
## [1] "The (covariate adjusted) point estimate of the treatment effect for EARNY4 is -13.9791072184366
```

Problem 4

The threshold we chose here is 12 years, which is the number of years of education before college. Due to the fact that we are investigating a particular instrumental variable which is the distance to college. It seems unnatural to choose any threshold below that. We'll discuss about the stability of the estimation using 16/18 years as thresholds, implying the treatment effect of attending grad school.

```
card <- read.csv("card.csv")
Y <- card$lwage
Z <- card$nearc2
D <- ifelse(card$educ > 12, 1, 0)
```

```
est = IV_Wald(Z, D, Y)
estVar = IV_Wald_delta(Z, D, Y)
print(paste("The (Wald) point estimate of the treatment effect for logwage is ", est$CACE))
```

```
## [1] "The (Wald) point estimate of the treatment effect for logwage is  1.83417131676141"
```

```
print(paste(" and the confidence interval is [", est$CACE - 1.96*estVar,",",est$CACE + 1.96*estVar,"]",
```

```
## [1] " and the confidence interval is [0.435784069917934,3.23255856360489]"
```

```
# That's significantly different from zero.
```

```
# With covariates (Note that we're somehow arbitrarily dropping all the observations with na values. Th
# Note that we dropped all the data with south indicator 1 here, leading to NA values in the regression
card <- na.omit(card)
Y <- card$lwage
Z <- card$nearc2
D <- ifelse(card$educ > 12, 1, 0)
X <- card %>% select(-id, -nearc2, -nearc4, -educ, -lwage, -south66, -reg669)
```

```
est = IV_Lin(Z, D, Y, X)
estVar = IV_Lin_delta(Z, D, Y, X)
print(paste("The (covariate adjusted) point estimate of the treatment effect for ", colnames(Y)[i] ," i
```

```
## [1] "The (covariate adjusted) point estimate of the treatment effect for  is 0.827472400711612 and th
```

```
# That's significantly different from zero.
```

```
# Stability of our analysis.
# 1/ if we use the 4 mile indicator instead of 2 mile
card <- read.csv("card.csv")
Y <- card$lwage
Z <- card$nearc4
D <- ifelse(card$educ > 12, 1, 0)
est = IV_Wald(Z, D, Y)
estVar = IV_Wald_delta(Z, D, Y)
print(paste("The (Wald) point estimate of the treatment effect for logwage is ", est$CACE, " and the co
```

```
## [1] "The (Wald) point estimate of the treatment effect for logwage is 1.27867156323074 and the confi
```

```
# That's still significantly different from zero.
```

```
card <- na.omit(card)
Y <- card$lwage
Z <- card$nearc4
D <- ifelse(card$educ > 12, 1, 0)
X <- card %>% select(-id, -nearc2, -nearc4, -educ, -lwage, -south66, -reg669)
est = IV_Lin(Z, D, Y, X)
estVar = IV_Lin_delta(Z, D, Y, X)
print(paste("The (covariate adjusted) point estimate of the treatment effect for ", colnames(Y)[i] ," i
```

```
## [1] "The (covariate adjusted) point estimate of the treatment effect for  is 0.22876331457567 and the
```

```
# That's still significantly different from zero.
```

```
# 2/ If we use a different threshold for the treatment indicator
card <- read.csv("card.csv")
Y <- card$lwage
Z <- card$nearc4
D <- ifelse(card$educ > 16, 1, 0)
est = IV_Wald(Z, D, Y)
estVar = IV_Wald_delta(Z, D, Y)
print(paste("The (Wald) point estimate of the treatment effect for logwage is ", est$CACE, " and the co
```

```
## [1] "The (Wald) point estimate of the treatment effect for logwage is 3.19790703936761 and the confi
```

```
# The treatment effect is even stronger when the treatment indicator is whether a person attends grad s
```

```
card <- na.omit(card)
Y <- card$lwage
Z <- card$nearc4
D <- ifelse(card$educ > 16, 1, 0)
X <- card %>% select(-id, -nearc2, -nearc4, -educ, -lwage, -south66, -reg669)
est = IV_Lin(Z, D, Y, X)
estVar = IV_Lin_delta(Z, D, Y, X)
print(paste("The (covariate adjusted) point estimate of the treatment effect for ", colnames(Y)[i] ," i
```

```
## [1] "The (covariate adjusted) point estimate of the treatment effect for  is -0.231781645767038 and t
```

```
# However, we observed a flip of sign when we adjust our results with covariates. That indicates the ke
```

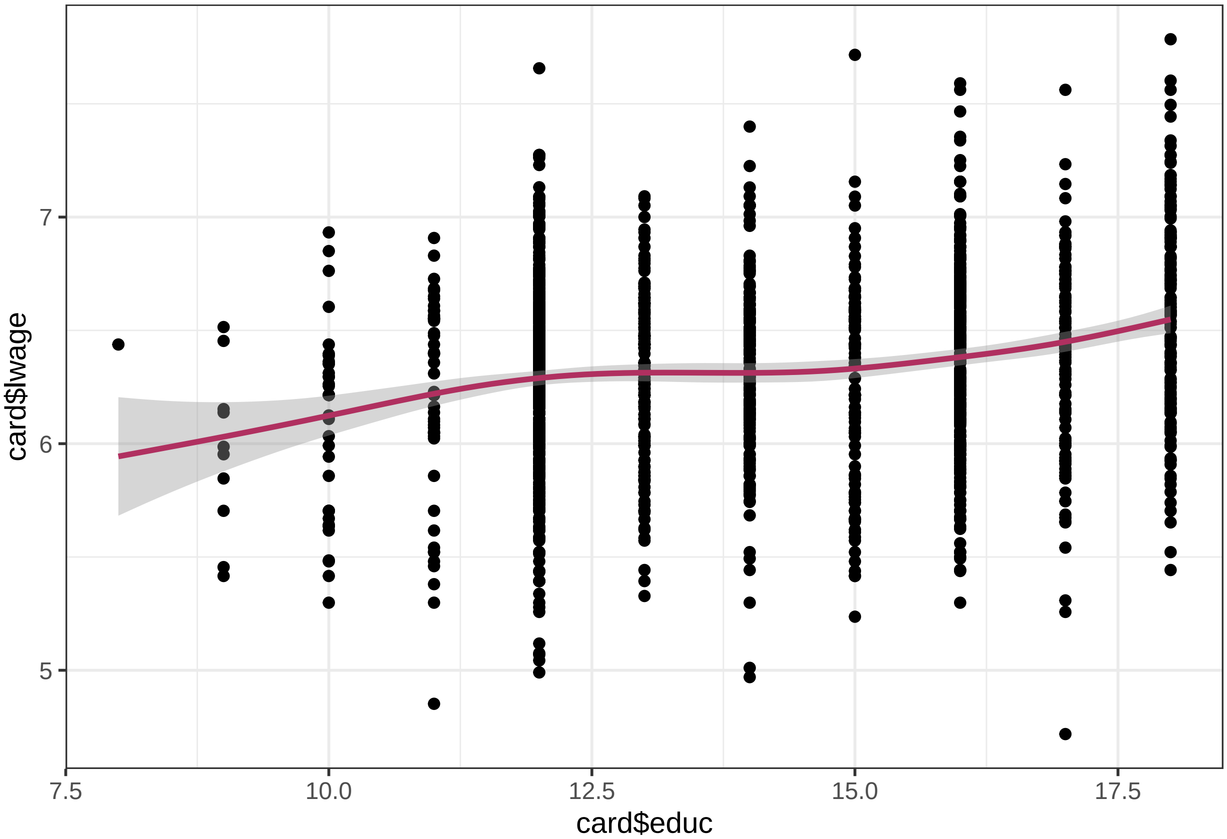
```
# 2SLS
TSLS <- function(Z, D, Y, X=0){
  X = as.matrix(X)
  D = as.matrix(D)
  Y = as.matrix(Y)
  Z = as.matrix(Z)
  Dhat    = lm(D ~ Z + X)$fitted.values
  tslsreg = lm(Y ~ Dhat + X)
  LATE <- coef(tslsreg)[2]
```

```
  res.correct       = Y - cbind(1, D, X)%*%coef(tslsreg)
  tslsreg$residuals = as.vector(res.correct)
  stderr <- sqrt(hccm(tslsreg, type = "hc0")[2, 2])
  return(list("LATE" = LATE, "se" = stderr))
}
print(paste("The point estimate of the causal effect is ", TSLS(Z, D, Y, X)$LATE, " and the CI is [", TS
```

```
## [1] "The point estimate of the causal effect is 1.61775198587005 and the CI is [-9.22563773240543,12
```

```
# That's no longer significantly different from zero!
```

```
ggplot() +
  geom_point(aes(x = card$educ, y = card$lwage)) + theme_bw() +
  geom_smooth(aes(x = card$educ, y = card$lwage), color = "maroon")
```




```
# There's no clear indicator that the influence of education is non-linear
```

```
data <- read.csv("EF.csv")
# Baseline: Complete randomized experiment
Z = data[,1]
Y = data[,5] - 0.75*data[,4] - 0.25*data[,3]
# Note that D is not randomized. The difference in means between the treatment and control group is
mean(Y[Z == 1]) - mean(Y[Z == 0])
```

```
## [1] -24.836
```

```r
# with standard error
sqrt(sd(Y[Z == 1])^2 / length(Y[Z == 1]) + sd(Y[Z == 0])^2 / length(Y[Z == 0]))
```

```
## [1] 2.655732
```

```r
# The drug did have a "positive" (in the sense of decreasing cholesterol level) effect.
```

```r
# However, there could be non-compliance. Namely, those assigned to the treatment group do not neccesar
# A possible solution is to define the "Compliance Level" as C3 - C4. Those who listens to our advice i
Z = data[,1]
D = data[,2]
Y = data[,5]
X <- data[,3] - data[,4]
# 1. As a covariate, use Lin's regression adjustment
model <- lm(Y~ Z:D + X + Z:D:X)
model$coef["Z:D"]
```

```
##        Z:D
## -0.4693973
```

```r
# The std error estimation is
sqrt(hccm(model, type = "hc0")[2, 2])
```

```
## [1] 0.1426412
```

```r
print(paste("One percent of increase in the effective drug intake leads to a decrease of 0.47 in the ch
```

```
## [1] "One percent of increase in the effective drug intake leads to a decrease of 0.47 in the cholest
```

```r
# 2. As an instrumental variable. Then we have to focus on those who receives the real treatment.
Z = data[,3] - data[,4]
D = data[,1] * data[,2]
Y = data[,5]
TSLS_simple <- function(Z, D, Y){
  D = as.matrix(D)
  Y = as.matrix(Y)
  Z = as.matrix(Z)
  Dhat    = lm(D ~ Z)$fitted.values
  tslsreg = lm(Y ~ Dhat)
  LATE <- coef(tslsreg)[2]
  res.correct       = Y - cbind(1, D)%*%coef(tslsreg)
  tslsreg$residuals = as.vector(res.correct)
  stderr <- sqrt(hccm(tslsreg, type = "hc0")[2, 2])
  return(list("LATE" = LATE, "se" = stderr))
}
print(paste("The point estimate of the average causal effect is ", TSLS_simple(Z, D, Y)$LATE, " and the
```

```
## [1] "The point estimate of the average causal effect is -1.88023867357047 and the CI is [-3.89355825
```

```
# Under 95% significance level, we do not reject the null hypothesis. For those who receives the real t
```