ECO 521 Project

Kai Li

2022-05-12

```
#library(vtable)
library(ggplot2)
library(sampleSelection)
```

```
data <- read.csv("cps_March_2021.csv", header = TRUE)
data <- data[(data$REGION != 97) & !(data$METRO %in% c(0, 4, 9)) & (data$SEX != 9) &
             (data$RACE != 999) & !(data$MARST %in% c(7, 9)) & (data$POPSTAT == 1) &
             (data$FAMSIZE != 0) & (data$LABFORCE != 0) & (data$UHRSWORKLY != 999) &
             !(data$EDUC %in% c(0, 1, 999)) & (data$DIFFANY != 0) &
             (data$FTOTVAL != 9999999999) & !(data$INCWAGE %in% c(99999998, 99999999)),]
sum(is.na(data))
```

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

```
data$NORTHEAST <- ifelse(data$REGION %in% c(11, 12), 1, 0)
data$MIDWEST <- ifelse(data$REGION %in% c(21, 22), 1, 0)
data$SOUTH <- ifelse(data$REGION %in% c(31, 32, 33), 1, 0)
data$WEST <- ifelse(data$REGION %in% c(41, 42), 1, 0)

data$METRO <- ifelse(data$METRO == 1, 0, 1)

data <- data[(data$AGE > 15) & (data$AGE < 65),]

data$MALE <- ifelse(data$SEX == 1, 1, 0)

data$WHITE <- ifelse(data$RACE == 100, 1, 0)

data$MARRIED <- ifelse(data$MARST %in% c(1, 2), 1, 0)

data$CHLT5 <- ifelse(data$NCHLT5 != 0, 1, 0)

data$LABFORCE <- ifelse(data$LABFORCE == 1, 0, 1)

data$LESS_THAN_HIGH <- ifelse(data$EDUC %in% c(2, 10, 20, 30, 40, 50, 60, 71), 1, 0)
data$HIGH <- ifelse(data$EDUC == 73, 1, 0)
data$SOME_COLLEGE <- ifelse(data$EDUC == 81, 1, 0)
data$COLLEGE <- ifelse(data$EDUC %in% c(91, 92, 111), 1, 0)
data$GREATER_THAN_COLLEGE <- ifelse(data$EDUC %in% c(123, 124, 125), 1, 0)

data$DIFFANY <- ifelse(data$DIFFANY == 1, 0, 1)
```

```r
low_FTOTVAL <- quantile(data$FTOTVAL, probs=c(0.25, 0.75))[1]-IQR(data$FTOTVAL)
up_FTOTVAL <- quantile(data$FTOTVAL, probs=c(0.25, 0.75))[2]+IQR(data$FTOTVAL)
data <- data[(data$FTOTVAL > low_FTOTVAL) & (data$FTOTVAL < up_FTOTVAL),]

data$WAGE <- data$INCWAGE/data$WKSWORK1/data$UHRSWORKLY
low_WAGE <- quantile(data$WAGE, probs=c(0.25, 0.75))[1]-IQR(data$WAGE)
up_WAGE <- quantile(data$WAGE, probs=c(0.25, 0.75))[2]+IQR(data$WAGE)
data <- data[(data$WAGE > max(c(low_WAGE, 7.25))) & (data$WAGE < up_WAGE),]

data <- data[, !(names(data) %in% c("REGION", "SEX", "RACE", "MARST", "POPSTAT", "NCHLT5",
                                    "EDUC", "WKSWORK1", "UHRSWORKLY", "INCWAGE"))]
```

```r
str(data)
```

```
## 'data.frame':    46247 obs. of  20 variables:
##  $ METRO               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AGE                 : int  57 28 28 49 20 30 51 19 30 44 ...
##  $ FAMSIZE             : int  3 3 1 3 3 1 3 3 5 5 ...
##  $ LABFORCE            : num  1 1 1 1 1 1 1 0 1 1 ...
##  $ DIFFANY             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FTOTVAL             : int  27100 106123 111010 233467 233467 45046 140416 140416 115000 115000 ..
##  $ NORTHEAST           : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ MIDWEST             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SOUTH               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ WEST                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MALE                : num  1 1 1 0 1 1 0 0 0 1 ...
##  $ WHITE               : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ MARRIED             : num  1 0 0 0 0 0 1 0 1 1 ...
##  $ CHLT5               : num  0 0 0 0 0 0 0 0 1 1 ...
##  $ LESS_THAN_HIGH      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ HIGH                : num  1 0 1 0 1 0 0 1 0 1 ...
##  $ SOME_COLLEGE        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ COLLEGE             : num  0 1 0 1 0 1 1 0 1 0 ...
##  $ GREATER_THAN_COLLEGE: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ WAGE                : num  9.62 12.02 21.92 17.31 28.85 ...
```

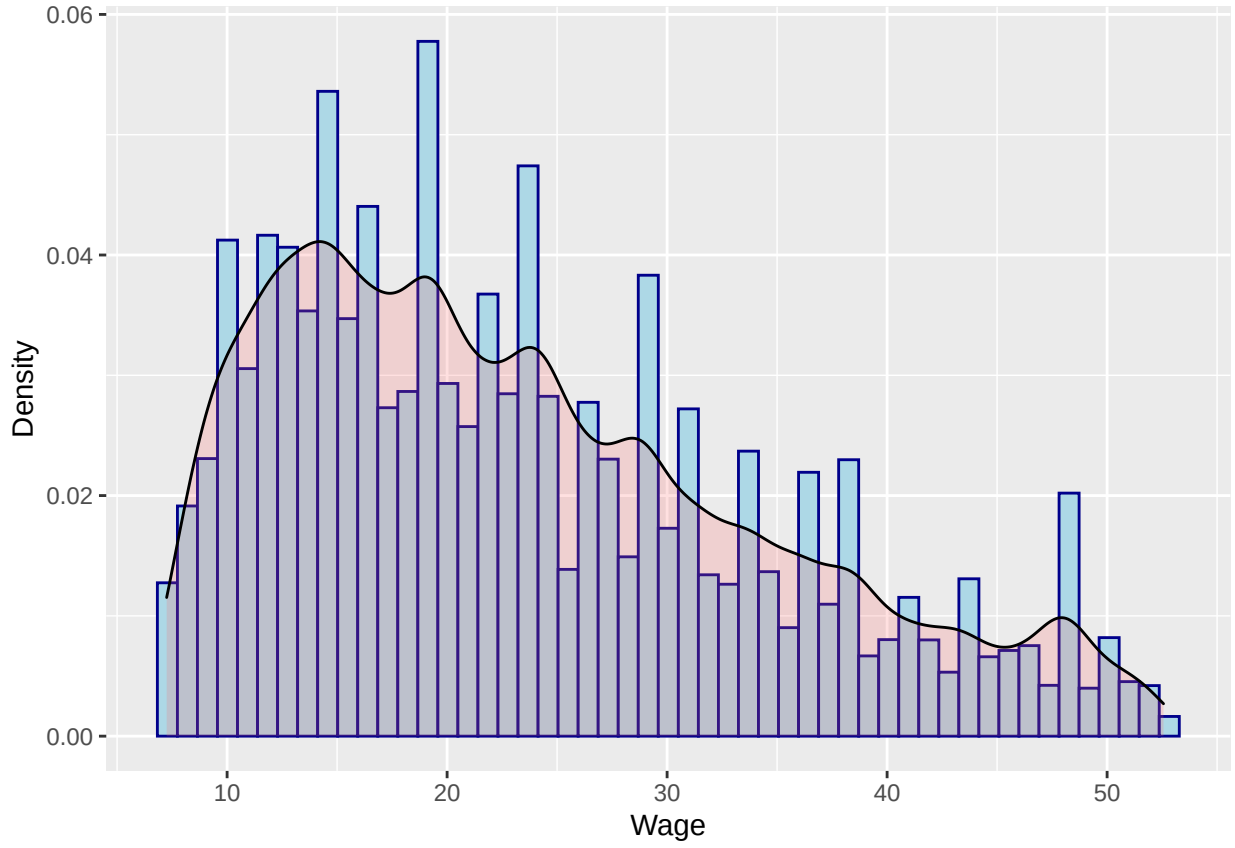
```r
summary(data)
```

```
##      METRO            AGE           FAMSIZE          LABFORCE
##  Min.   :0.000   Min.   :16.00   Min.   : 1.000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:30.00   1st Qu.: 2.000   1st Qu.:1.0000
##  Median :1.000   Median :40.00   Median : 3.000   Median :1.0000
##  Mean   :0.793   Mean   :40.22   Mean   : 3.111   Mean   :0.9464
##  3rd Qu.:1.000   3rd Qu.:50.00   3rd Qu.: 4.000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :64.00   Max.   :15.000   Max.   :1.0000
##     DIFFANY           FTOTVAL         NORTHEAST         MIDWEST
##  Min.   :0.00000   Min.   : -4996   Min.   :0.0000   Min.   :0.00
##  1st Qu.:0.00000   1st Qu.: 46901   1st Qu.:0.0000   1st Qu.:0.00
##  Median :0.00000   Median : 79456   Median :0.0000   Median :0.00
##  Mean   :0.03529   Mean   : 90000   Mean   :0.1492   Mean   :0.19
##  3rd Qu.:0.00000   3rd Qu.:123204   3rd Qu.:0.0000   3rd Qu.:0.00
##  Max.   :1.00000   Max.   :254002   Max.   :1.0000   Max.   :1.00
```

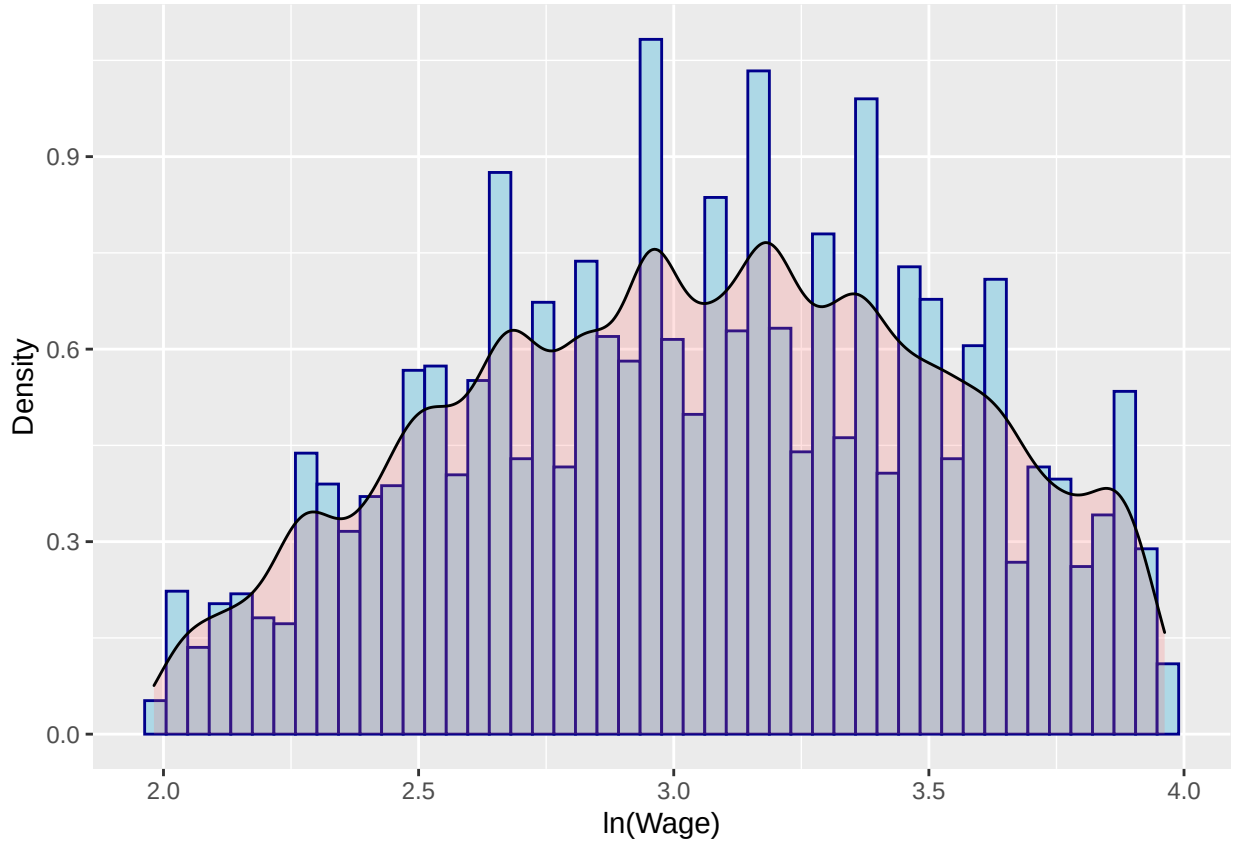
```
##      SOUTH             WEST              MALE             WHITE
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000
##  Median :0.000   Median :0.0000   Median :1.0000   Median :1.0000
##  Mean   :0.362   Mean   :0.2988   Mean   :0.5027   Mean   :0.7602
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     MARRIED           CHLT5        LESS_THAN_HIGH         HIGH
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.5216   Mean   :0.1414   Mean   :0.08422   Mean   :0.2797
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##   SOME_COLLEGE       COLLEGE      GREATER_THAN_COLLEGE      WAGE
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000       Min.   : 7.253
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000       1st Qu.:14.423
##  Median :0.0000   Median :0.000   Median :0.0000       Median :21.634
##  Mean   :0.1718   Mean   :0.351   Mean   :0.1133       Mean   :23.548
##  3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:0.0000       3rd Qu.:30.769
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000       Max.   :52.564
```

```
cols <- c("NORTHEAST", "MIDWEST", "SOUTH", "WEST", "METRO", "MALE", "WHITE", "MARRIED",
          "CHLT5", "LABFORCE", "LESS_THAN_HIGH", "HIGH", "SOME_COLLEGE", "COLLEGE",
          "GREATER_THAN_COLLEGE", "DIFFANY")
data[cols] <- lapply(data[cols], factor)
```

```
ggplot(data, aes(x = WAGE)) +
  geom_histogram(aes(y = ..density..), color = "darkblue", fill = "lightblue",
                 binwidth = 2*IQR(data$WAGE)/length(data$WAGE)^(1/3)) +
  geom_density(alpha = 0.2, fill = "#FF6666") + xlab("Wage") + ylab("Density")
```




```
ggplot(data, aes(x = log(WAGE))) +
  geom_histogram(aes(y = ..density..), color = "darkblue", fill = "lightblue",
                 binwidth = 2*IQR(log(data$WAGE))/length(data$WAGE)^(1/3)) +
  geom_density(alpha = 0.2, fill = "#FF6666") + xlab("ln(Wage)") + ylab("Density")
```




```
model <- selection(LABFORCE ~ I((AGE-mean(AGE))/100) + I(((AGE-mean(AGE))^2)/100) + MALE
                   + WHITE + MARRIED + FAMSIZE + CHLT5 + LESS_THAN_HIGH + HIGH + SOME_COLLEGE
                   + COLLEGE + DIFFANY + I(FTOTVAL/100000) + MALE:MARRIED,
                   log(WAGE) ~ NORTHEAST + SOUTH + MIDWEST + METRO + I((AGE-mean(AGE))/100)
                   + I(((AGE-mean(AGE))^2)/100) + MALE + WHITE + LESS_THAN_HIGH + HIGH
                   + SOME_COLLEGE + COLLEGE + MALE:WHITE, data = data, method = "2step")
summary(model)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 46247 observations (2481 censored and 43766 observed)
## 32 free parameters (df = 46216)
## Probit selection equation:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.052593   0.051601  39.778  < 2e-16 ***
## I((AGE - mean(AGE))/100)      0.740125   0.082625   8.958  < 2e-16 ***
## I(((AGE - mean(AGE))^2)/100) -0.144218   0.006291 -22.923  < 2e-16 ***
## MALE1                         0.014459   0.027396   0.528 0.597663
## WHITE1                        0.102930   0.023064   4.463 8.11e-06 ***
## MARRIED1                     -0.184979   0.030884  -5.989 2.12e-09 ***
## FAMSIZE                      -0.046188   0.006877  -6.716 1.89e-11 ***
## CHLT51                       -0.075652   0.032035  -2.362 0.018201 *
## LESS_THAN_HIGH1              -0.403295   0.052598  -7.667 1.79e-14 ***
## HIGH1                        -0.259946   0.045705  -5.687 1.30e-08 ***
```

```
## SOME_COLLEGE1                -0.379176   0.046810  -8.100 5.61e-16 ***
## COLLEGE1                     -0.167181   0.044565  -3.751 0.000176 ***
## DIFFANY1                     -0.616417   0.040424 -15.249  < 2e-16 ***
## I(FTOTVAL/1e+05)              0.246106   0.021959  11.208  < 2e-16 ***
## MALE1:MARRIED1                0.370757   0.041623   8.907  < 2e-16 ***
## Outcome equation:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.385589   0.017256 196.200  < 2e-16 ***
## NORTHEAST1                    0.012132   0.006712   1.808  0.07069 .
## SOUTH1                       -0.056649   0.004802 -11.797  < 2e-16 ***
## MIDWEST1                     -0.016206   0.005684  -2.851  0.00436 **
## METRO1                        0.077173   0.004676  16.505  < 2e-16 ***
## I((AGE - mean(AGE))/100)      0.241752   0.036124   6.692 2.22e-11 ***
## I(((AGE - mean(AGE))^2)/100)  0.027393   0.003448   7.944 2.01e-15 ***
## MALE1                         0.048669   0.010766   4.521 6.18e-06 ***
## WHITE1                       -0.025222   0.010644  -2.369  0.01782 *
## LESS_THAN_HIGH1              -0.390341   0.021325 -18.305  < 2e-16 ***
## HIGH1                        -0.325990   0.016275 -20.030  < 2e-16 ***
## SOME_COLLEGE1                -0.206446   0.017816 -11.588  < 2e-16 ***
## COLLEGE1                     -0.126635   0.015301  -8.276  < 2e-16 ***
## MALE1:WHITE1                  0.038950   0.008007   4.864 1.15e-06 ***
## Multiple R-Squared:0.2507,   Adjusted R-Squared:0.2504
##    Error terms:
##               Estimate Std. Error t value Pr(>|t|)
## invMillsRatio -2.13177    0.07097  -30.04   <2e-16 ***
## sigma          0.98539         NA      NA       NA
## rho           -2.16338         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```