#ECON613 A4 Mengzhi Chen

#==========

```
#Set working directory
setwd("/Users/halcyonchan/Desktop/Econ613/A4/Data")
library(data.table)
library(tidyverse)
#Read the file
dat = fread("dat_A4.csv")
```

#Exercise1==========

#1.1==========

```
#1.1==========
#Create additional variable for the age of the agent ”age”,
#total work experience measured in years ”work exp”.
#Hint: ”CV WKSWK JOB DLI.01” denotes the number of weeks a person ever worked at JOB 01.

#Create variable "age" by the following formula
dat[,'age'] = 2019 - dat$KEY_BDATE_Y_1997

#Replace NA in variables "CV WKSWK JOB DLI" with 0
dat[is.na(dat$CV_WKSWK_JOB_DLI.01_2019),'CV_WKSWK_JOB_DLI.01_2019'] = 0
dat[is.na(dat$CV_WKSWK_JOB_DLI.02_2019),'CV_WKSWK_JOB_DLI.02_2019'] = 0
dat[is.na(dat$CV_WKSWK_JOB_DLI.03_2019),'CV_WKSWK_JOB_DLI.03_2019'] = 0
dat[is.na(dat$CV_WKSWK_JOB_DLI.04_2019),'CV_WKSWK_JOB_DLI.04_2019'] = 0
dat[is.na(dat$CV_WKSWK_JOB_DLI.05_2019),'CV_WKSWK_JOB_DLI.05_2019'] = 0
dat[is.na(dat$CV_WKSWK_JOB_DLI.06_2019),'CV_WKSWK_JOB_DLI.06_2019'] = 0
dat[is.na(dat$CV_WKSWK_JOB_DLI.07_2019),'CV_WKSWK_JOB_DLI.07_2019'] = 0
dat[is.na(dat$CV_WKSWK_JOB_DLI.08_2019),'CV_WKSWK_JOB_DLI.08_2019'] = 0
dat[is.na(dat$CV_WKSWK_JOB_DLI.09_2019),'CV_WKSWK_JOB_DLI.09_2019'] = 0
dat[is.na(dat$CV_WKSWK_JOB_DLI.10_2019),'CV_WKSWK_JOB_DLI.10_2019'] = 0
dat[is.na(dat$CV_WKSWK_JOB_DLI.11_2019),'CV_WKSWK_JOB_DLI.11_2019'] = 0
#Sum the weeks for each job, divide by 52, and round up to the nearest number of year
dat[,'work_exp'] = round((dat$CV_WKSWK_JOB_DLI.01_2019 + dat$CV_WKSWK_JOB_DLI.02_2019 + dat$CV_WKSWK_JOB_DLI.03_2019
                        + dat$CV_WKSWK_JOB_DLI.04_2019 + dat$CV_WKSWK_JOB_DLI.05_2019+ dat$CV_WKSWK_JOB_DLI.06_2019
                        + dat$CV_WKSWK_JOB_DLI.07_2019 + dat$CV_WKSWK_JOB_DLI.08_2019+ dat$CV_WKSWK_JOB_DLI.09_2019
                        + dat$CV_WKSWK_JOB_DLI.10_2019 + dat$CV_WKSWK_JOB_DLI.11_2019)/52)
view(dat)
```

19	CV_WKSWK_JOB_DLI.09_2019	CV_WKSWK_JOB_DLI.10_2019	CV_WKSWK_JOB_DLI.11_2019	YSCH.3113_2019	YINC_1700_2019	age	work_exp
0	0	0	0	NA	NA	38	0
0	0	0	0	3	100000	37	12
0	0	0	0	5	59000	36	2
0	0	0	0	3	27000	38	2
0	0	0	0	3	100000	37	13
0	0	0	0	3	17000	37	2
0	0	0	0	1	NA	36	2
0	0	0	0	5	60000	38	4
0	0	0	0	6	90000	37	3
0	0	0	0	6	100000	35	5
0	0	0	0	5	38000	37	12
0	0	0	0	2	72000	38	15
0	0	0	0	2	24000	35	0
0	0	0	0	NA	NA	39	0
0	0	0	0	4	50000	36	10
0	0	0	0	5	NA	37	3

Showing 1 to 16 of 8,984 entries, 32 total columns

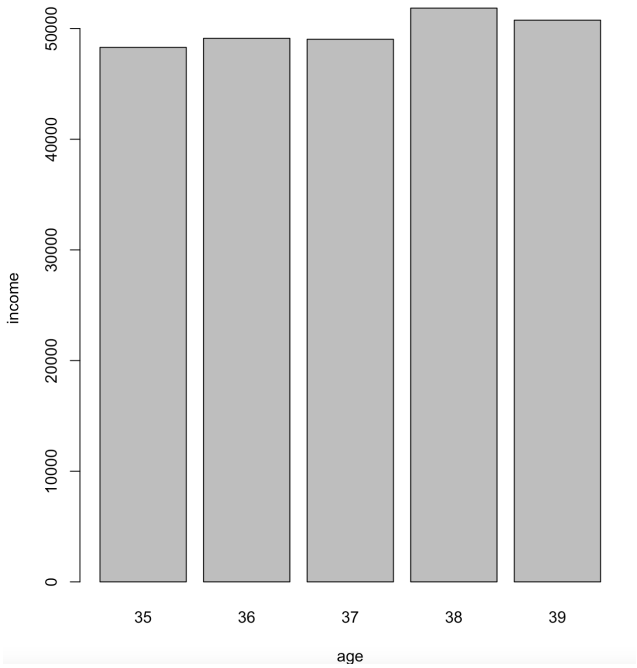
#1.2==========

```
#1.2================================================
#Create additional education variables indicating total years of schooling from all variables
#related to education (eg, ”BIOLOGICAL FATHERS HIGHEST GRADE COMPLETED”) in our dataset.

#Replace 95 ungraded with 0
dat[which(dat$CV_HGC_BIO_DAD_1997 == 95),'CV_HGC_BIO_DAD_1997'] = 0
dat[which(dat$CV_HGC_BIO_MOM_1997 == 95),'CV_HGC_BIO_MOM_1997'] = 0
dat[which(dat$CV_HGC_RES_DAD_1997 == 95),'CV_HGC_RES_DAD_1997'] = 0
dat[which(dat$CV_HGC_RES_MOM_1997 == 95),'CV_HGC_RES_MOM_1997'] = 0
#Switch "YSCH.3113_2019" into years of schooling
dat[which(dat$YSCH.3113_2019 == 1),'YSCH.3113_2019_year'] = 0
dat[which(dat$YSCH.3113_2019 == 2),'YSCH.3113_2019_year'] = 4
dat[which(dat$YSCH.3113_2019 == 3),'YSCH.3113_2019_year'] = 12
dat[which(dat$YSCH.3113_2019 == 4),'YSCH.3113_2019_year'] = 14
dat[which(dat$YSCH.3113_2019 == 5),'YSCH.3113_2019_year'] = 16
dat[which(dat$YSCH.3113_2019 == 6),'YSCH.3113_2019_year'] = 18
dat[which(dat$YSCH.3113_2019 == 7),'YSCH.3113_2019_year'] = 23
dat[which(dat$YSCH.3113_2019 == 8),'YSCH.3113_2019_year'] = 22
view(dat)
```

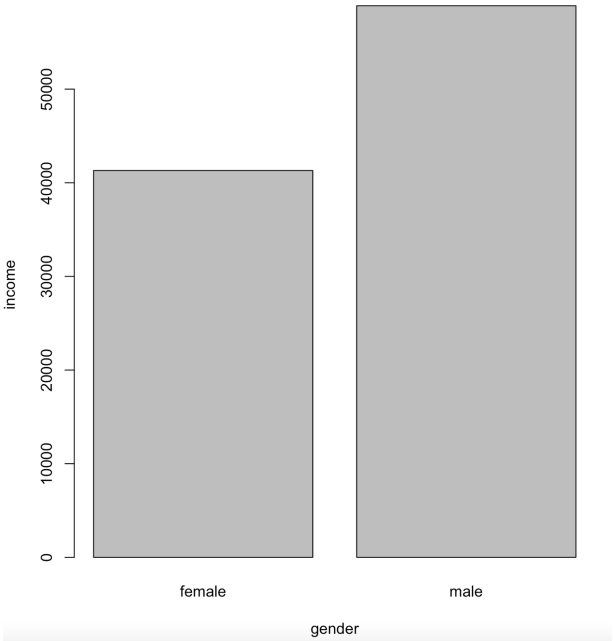
I.09_2019	CV_WKSWK_JOB_DLI.10_2019	CV_WKSWK_JOB_DLI.11_2019	YSCH.3113_2019	YINC_1700_2019	age	work_exp	YSCH.3113_2019_year
0	0	0	*NA*	*NA*	38	0	*NA*
0	0	0	3	100000	37	12	12
0	0	0	5	59000	36	2	16
0	0	0	3	27000	38	2	12
0	0	0	3	100000	37	13	12
0	0	0	3	17000	37	2	12
0	0	0	1	*NA*	36	2	0
0	0	0	5	60000	38	4	16
0	0	0	6	90000	37	3	18
0	0	0	6	100000	35	5	18
0	0	0	5	38000	37	12	16
0	0	0	2	72000	38	15	4
0	0	0	2	24000	35	0	4
0	0	0	*NA*	*NA*	39	0	*NA*
0	0	0	4	50000	36	10	14
0	0	0	5	*NA*	37	3	16

Showing 1 to 16 of 8,984 entries, 33 total columns

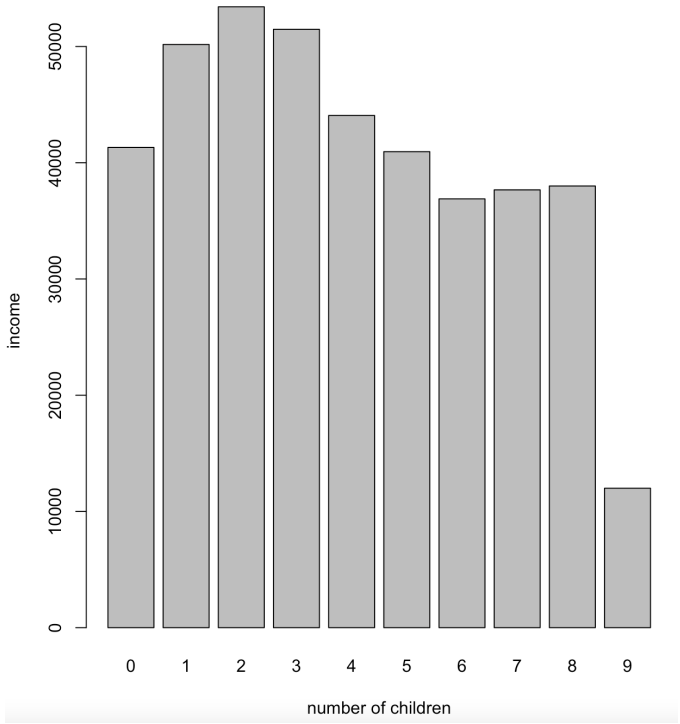
```
#1.3================================================
#Provide the following visualizations.
# - Plot the income data (where income is positive) by i) age groups, ii) gender groups and iii) number of children

#Delete the rows where income data is NA
delete = which(is.na(dat$YINC_1700_2019) == TRUE)
dat1 = dat[-delete,]
#Delete the rows where income data is 0
dat_plot = dat1[-which(dat1$YINC_1700_2019 == 0),]

#plot the income data by age groups
dat_plot_age = aggregate(dat_plot$YINC_1700_2019,by=list(dat_plot$age),FUN=mean)
barplot(dat_plot_age$x,names.arg = dat_plot_age$Group.1,xlab = 'age',ylab = 'income')
```




```r
#plot the income data by gender groups
dat_plot[which(dat_plot$KEY_SEX_1997 == 1),'gender'] = 'male'
dat_plot[which(dat_plot$KEY_SEX_1997 == 2),'gender'] = 'female'
dat_plot_gender = aggregate(dat_plot$YINC_1700_2019,by=list(dat_plot$gender),FUN=mean)
barplot(dat_plot_gender$x,names.arg = dat_plot_gender$Group.1,xlab = 'gender',ylab = 'income')
```




```r
#plot the income data by number of children
delete1 = which(is.na(dat_plot$CV_BIO_CHILD_HH_U18_2019) == TRUE)
dat_plot = dat_plot[-delete1,]
dat_plot_child = aggregate(dat_plot$YINC_1700_2019,by=list(dat_plot$CV_BIO_CHILD_HH_U18_2019),FUN=mean)
barplot(dat_plot_child$x,names.arg = dat_plot_child$Group.1,xlab = 'number of children',ylab = 'income')
```



#1.3==========

```
# - Table the share of "0" in the income data by i) age groups, ii) gender groups,
#iii) number of children and marital status

#Table the share of "0" in the income data by age groups
dat_share_age = dat %>% group_by(age) %>% mutate(share = length(which(YINC_1700_2019 == 0))
                                                 /length(which(YINC_1700_2019 >= 0)))
dat_share_age = unique(dat_share_age[,c('age','share')])
#Sort the age column
dat_share_age[order(dat_share_age[,1]),]
```

```
# Groups:   age [5]
    age   share
  <dbl>   <dbl>
1    35 0.00929
2    36 0.00630
3    37 0.00542
4    38 0.00896
5    39 0.00299
```

```
#Table the share of "0" in the income data by gender groups
dat[which(dat$KEY_SEX_1997 == 1),'gender'] = 'male'
dat[which(dat$KEY_SEX_1997 == 2),'gender'] = 'female'
dat_share_gender = dat %>% group_by(gender) %>% mutate(share = length(which(YINC_1700_2019 == 0))
                                                       /length(which(YINC_1700_2019 >= 0)))
dat_share_gender = unique(dat_share_gender[,c('gender','share')])
view(dat_share_gender)
```

	gender	share
1	female	0.005742726
2	male	0.007500000

```
#Table the share of "0" in the income data by number of children and marital status
delete2 = which(is.na(dat$CV_BIO_CHILD_HH_U18_2019) == TRUE)
dat_share_children_marital = dat[-delete2,]
delete3 = which(is.na(dat_share_children_marital$CV_MARSTAT_COLLAPSED_2019) == TRUE)
dat_share_children_marital = dat_share_children_marital[-delete3,]
#Create a new variable "children_marital" to represent the combination of children and marital
dat_share_children_marital['children_marital'] = paste(dat_share_children_marital$CV_BIO_CHILD_HH_U18_2019,
                                                       dat_share_children_marital$CV_MARSTAT_COLLAPSED_2019)
#Compute the share grouped by "children_marital"
dat_share_children_marital = dat_share_children_marital %>% group_by(children_marital) %>%
  mutate(share = length(which(YINC_1700_2019 == 0))/length(which(YINC_1700_2019 >= 0)))
dat_share_children_marital = unique(dat_share_children_marital[,c('children_marital','share')])
#Split the first column into two columns
library(stringr)
split = str_split_fixed(dat_share_children_marital$children_marital," ",2)
#The two columns correspond to number of children and marital
dat_share_children_marital['number of children'] = split[,1]
dat_share_children_marital['marital'] = split[,2]
view(dat_share_children_marital)
```

	children_marital	share	number of children	marital
1	3 1	0.004597701	3	1
2	1 0	0.011080332	1	0
3	2 1	0.008456660	2	1
4	1 3	0.000000000	1	3
5	1 1	0.008156607	1	1
6	0 0	0.000000000	0	0
7	3 0	0.017699115	3	0
8	2 0	0.000000000	2	0
9	2 3	0.000000000	2	3
10	0 3	0.006802721	0	3
11	1 4	0.000000000	1	4
12	4 0	0.000000000	4	0
13	3 3	0.000000000	3	3

Showing 1 to 13 of 35 entries, 4 total columns

```
# - interpret the visualizations from above
```

From the graph above, the income will increase with the age and males' average income is larger than females' average income. There seems no direct relationship between income and number of children. Form the tables, the age group of 35 reports largest share of "0" in income data, while age group of 39 reports less share of "0". Male group reports larger share of "0" in income than female. According to number of children and marital, there are 9 groups report "0" in income. In particular, the individual whose marital is separated and has 3 children report the largest share of "0" in income data.

share	number of children	marital
0.142857143	3	2
0.136363636	0	2
0.033898305	0	1
0.017699115	3	0
0.011080332	1	0
0.008456660	2	1
0.008156607	1	1
0.006802721	0	3
0.004597701	3	1

#Exercise2==========

#2.1==========

```
#Using the variables created above, estimate the following models.
#Specify and estimate an OLS model to explain the income variable (where income is positive).
#Drop the data where income is 0
dat_OLS = dat[-which(dat$YINC_1700_2019 == 0),]
#Use OLS regression
reg1 = lm(dat_OLS$YINC_1700_2019 ~ dat_OLS$age + dat_OLS$work_exp + dat_OLS$YSCH.3113_2019_year,dat_OLS)
summary(reg1)
```

```
> summary(reg1)

Call:
lm(formula = dat_OLS$YINC_1700_2019 ~ dat_OLS$age + dat_OLS$work_exp +
    dat_OLS$YSCH.3113_2019_year, data = dat_OLS)

Residuals:
   Min     1Q Median     3Q    Max
-72236 -19333  -3485  17909  80453

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1960.08    9463.42   0.207    0.836
dat_OLS$age                  410.27     254.99   1.609    0.108
dat_OLS$work_exp             998.12      66.11  15.097   <2e-16 ***
dat_OLS$YSCH.3113_2019_year 2014.01      73.02  27.582   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25960 on 5368 degrees of freedom
  (3576 observations deleted due to missingness)
Multiple R-squared:  0.1703,    Adjusted R-squared:  0.1699
F-statistic: 367.3 on 3 and 5368 DF,  p-value: < 2.2e-16
```

```
# - Interpret the estimation results
```

In this model, the parameter of age is not significant. When everything else is equal, the income will increase about an average of 410 units as age increases by 1 year. The income will increase about an average of 998 units as work experience increases 1 year. The income will increase about an average of 2014 units as the year of education increases by 1 year.

The linear regression model only run the data which has positive income. There exists selection bias since the data is not random selected.

```
# – Explain why there might be a selection problem when estimating an OLS this way
```

In a linear regression model, sample selection bias occurs when data on the dependent variable are missing nonrandomly, conditional on the independent variables. In this case, 5921 observations deleted due to missingness in OLS regression model. Therefore, the dependent variable "income" may be conditional on the independent variables.

#2.2==========

```
#Explain why the Heckman model can deal with the selection problem.
```

#The missing variable in the regression is called Inverse Mill's Ratio. Heckman model first calculates the IMR of all samples; Then, the missing variable IMR is substituted into the original regression equation, so that selection problem is resolved.

#2.3==========

```
#2.3================================================
#Estimate a Heckman selection model(Note:You can not use a pre-programmed Heckman selection package.
#Please write down the likelihood and optimize the two-stage Heckman model).
#Interpret the results from the Heckman selection model and compare the results to OLS results.
#Why does there exist a difference?

#Process the data, construct 0-1 variables to represent whether the income data can be observed or not
dat[is.na(dat$YINC_1700_2019),'YINC_1700_2019'] = 0
dat[which(dat$YINC_1700_2019 > 0),'selection'] = 1
dat[which(dat$YINC_1700_2019 == 0),'selection'] = 0
#Delete NA data in indenpendent variables
delete2= which(is.na(dat$CV_HGC_BIO_DAD_1997) == TRUE)
dat = dat[-delete2,]
delete3= which(is.na(dat$CV_HGC_BIO_MOM_1997) == TRUE)
dat = dat[-delete3,]
delete4= which(is.na(dat$CV_HGC_RES_DAD_1997) == TRUE)
dat = dat[-delete4,]
delete5= which(is.na(dat$CV_HGC_RES_MOM_1997) == TRUE)
dat = dat[-delete5,]
delete6= which(is.na(dat$YSCH.3113_2019_year) == TRUE)
dat = dat[-delete6,]
delete7= which(is.na(dat$CV_BIO_CHILD_HH_U18_2019) == TRUE)
dat = dat[-delete7,]

#The first step is to conduct a probit model regarding whether the income is observed or not
probit = glm(selection ~ age + work_exp + CV_HGC_BIO_DAD_1997 + CV_HGC_BIO_MOM_1997
             + CV_HGC_RES_DAD_1997 + CV_HGC_RES_MOM_1997 + YSCH.3113_2019_year +CV_BIO_CHILD_HH_U18_2019,
             family=binomial(link="probit"), data=dat)
summary(probit)
```

```
> summary(probit)

Call:
glm(formula = selection ~ age + work_exp + CV_HGC_BIO_DAD_1997 +
    CV_HGC_BIO_MOM_1997 + CV_HGC_RES_DAD_1997 + CV_HGC_RES_MOM_1997 +
    YSCH.3113_2019_year + CV_BIO_CHILD_HH_U18_2019, family = binomial(link = "probit"),
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4385   0.1540   0.4375   0.7106   1.4417

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.291137   0.802976  -0.363   0.7169
age                      -0.001146   0.021482  -0.053   0.9575
work_exp                  0.115829   0.007594  15.254  < 2e-16 ***
CV_HGC_BIO_DAD_1997      -0.016464   0.029560  -0.557   0.5775
CV_HGC_BIO_MOM_1997      -0.112155   0.057270  -1.958   0.0502 .
CV_HGC_RES_DAD_1997       0.028329   0.029561   0.958   0.3379
CV_HGC_RES_MOM_1997       0.114387   0.056673   2.018   0.0436 *
YSCH.3113_2019_year       0.039185   0.006365   6.156 7.44e-10 ***
CV_BIO_CHILD_HH_U18_2019 -0.044247   0.025048  -1.766   0.0773 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2694.0  on 2712  degrees of freedom
Residual deviance: 2229.4  on 2704  degrees of freedom
AIC: 2247.4

Number of Fisher Scoring iterations: 7
```

```
#Calculate Inverse Mills Ratio
dat$IMR <- dnorm(probit$linear.predictors)/pnorm(probit$linear.predictors)
#The step two is run the standard linear regression, income ~ age + work_exp + highest grade + IMR
outcome = lm(YINC_1700_2019 ~ age + work_exp + YSCH.3113_2019_year + IMR, data=dat,subset=(selection==1))
summary(outcome)
```

```
> summary(outcome)

Call:
lm(formula = YINC_1700_2019 ~ age + work_exp + YSCH.3113_2019_year +
    IMR, data = dat, subset = (selection == 1))

Residuals:
   Min     1Q Median     3Q    Max
-64239 -19789  -2220  21388  78957

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          38409.4    15951.2   2.408   0.0161 *
age                    314.9      405.8   0.776   0.4378
work_exp              -296.3      261.1  -1.134   0.2567
YSCH.3113_2019_year   1481.7      181.5   8.163 5.49e-16 ***
IMR                 -44897.2     7756.5  -5.788 8.14e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26310 on 2173 degrees of freedom
Multiple R-squared:  0.1889,    Adjusted R-squared:  0.1874
F-statistic: 126.5 on 4 and 2173 DF,  p-value: < 2.2e-16
```

```r
#Write the likelihood function
like = function(par, X, Z, y, observed_index) {
  gamma     = par[1:9]
  lp_probit = Z %*% gamma
  beta  = par[10:13]
  lp_lm = X %*% beta
  sigma = par[14]
  rho   = par[15]
  ll = sum(log(1-pnorm(lp_probit[!observed_index]))) +
    - log(sigma) + sum(dnorm(y, mean = lp_lm, sd = sigma, log = TRUE)) +
    sum(pnorm((lp_probit[observed_index] + rho/sigma * (y-lp_lm)) / sqrt(1-rho^2), log.p = TRUE))
  -ll
}
X = model.matrix(outcome)
Z = model.matrix(probit)
init = c(coef(probit), coef(outcome)[-5],1,0)
#Optimize the model
Heckman = optim(init,like,X = X[,-5],Z = Z,y = dat$YINC_1700_2019[which(dat$selection==1)],
                      observed_index = dat$selection, method  = 'Nelder-Mead',
                      control = list(maxit = 1000, reltol = 1e-15),hessian = T)
Heckman$par
```

```
> Heckman$par
            (Intercept)                     age                work_exp      CV_HGC_BIO_DAD_1997      CV_HGC_BIO_MOM_1997
           1.174856e+00           -4.830702e-01            5.453915e-01           -2.829975e-02            2.980521e-01
    CV_HGC_RES_DAD_1997      CV_HGC_RES_MOM_1997      YSCH.3113_2019_year CV_BIO_CHILD_HH_U18_2019              (Intercept)
          -7.783653e-02           -3.978629e-01            9.272542e-01            9.117933e-01            3.840882e+04
                    age                work_exp      YSCH.3113_2019_year
           2.893692e+02           -2.876479e+02            1.477482e+03            3.854205e+03           -1.186117e-02
```

```r
#Interpret the results from the Heckman selection model and compare the results to OLS results.
#Why does there exist a difference?
```

According to the results of Heckman selection model, when everything else is equal, the income will increase an average of 1481 units as the education year increases by 1 year.

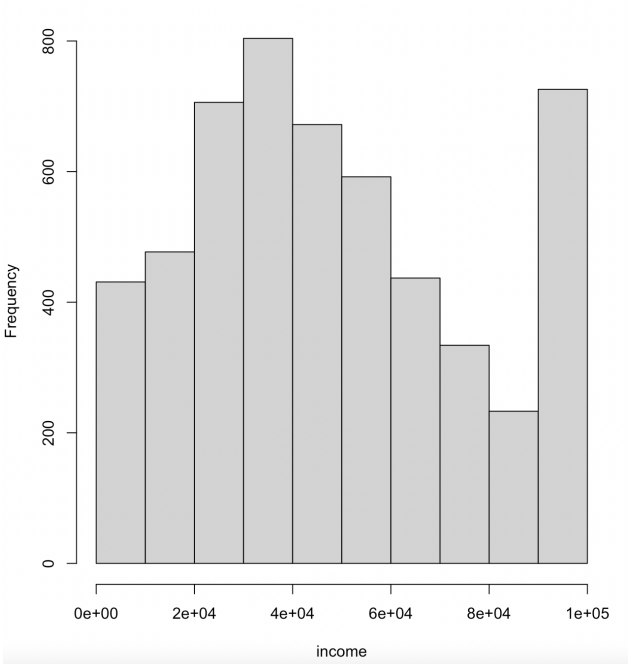
Compare to OLS results, the parameters of age are not significant in both models. And the work experience is significant in OLS model but not significant in Heckman selection model.

There exists a difference because people who have more work experience are more willing to work and have positive income. Thus, in OLS model, we don't take this selection bias into consideration.

#Exercise3==========

#3.1==========

```r
#Plot a histogram to check whether the distribution of the income variable.
#What might be the censored value here?

dat2 = fread("dat_A4.csv")
hist(dat2$YINC_1700_2019,xlab="income")
```



100,000 is the censored value

In some data sets we do not observe values above or below a certain magnitude, due to a censoring or truncation mechanism. In this data set, we cannot observe income values above 100,000, which is due to the censoring problem.

#3.2==========

```
#3.2==============================================
#Propose a model to deal with the censoring problem.
```

We can use the Tobit model to deal with the censoring problem.

#3.3==========

```
#Estimate the appropriate model with the censored data (please write down the likelihood function
#and optimize yourself without using the pre-programmed package)

#Run the original linear regression
reg2 = lm(dat2$YINC_1700_2019 ~ dat2$age + dat2$work_exp + dat2$YSCH.3113_2019_year,dat)
summary(reg2)
```

```
> summary(reg2)

Call:
lm(formula = dat$YINC_1700_2019 ~ dat$age + dat$work_exp + dat$YSCH.3113_2019_year,
    data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-77982 -24401  -2149  22170  89601

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             -3708.2    15121.4  -0.245    0.806
dat$age                   154.1      409.2   0.377    0.706
dat$work_exp             2024.2      103.3  19.593   <2e-16 ***
dat$YSCH.3113_2019_year  2209.4      120.3  18.365   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29700 on 2709 degrees of freedom
Multiple R-squared:  0.2424,    Adjusted R-squared:  0.2416
F-statistic:   289 on 3 and 2709 DF,  p-value: < 2.2e-16
```

```
#Write the likelihood of Tobit model
loglik = function(par,X,y,z)
{
  beta = par[1:4]
  sigma = par[5]
  result =  ifelse(y < z, dnorm(y, beta*X, sigma, log = T),
                   pnorm((z - (beta*X))/sigma, log = T))
  return(sum(result))
}
X = model.matrix(reg2)
init = c(coef(reg2),log_sigma = log(summary(reg2)$sigma))

#Optimize the Tobit model
tobit = optim(init,loglik, method  = 'CG', X = X, y = dat$YINC_1700_2019, z = 100000,
              control = list(fnscale = -1),hessian = T,)
tobit$par
```

```
> tobit$par
          (Intercept)               dat$age          dat$work_exp dat$YSCH.3113_2019_year
            -3664.2296              196.1304             2065.0003              2246.8511
            log_sigma
            12474.3865
```

#3.4==========

```
#Interpret the results above and compare to those when not correcting for the censored data
```

When everything else is equal, the income will increase about 196 units as age increases by 1 year. The income will increase about 2065 units as work experience increases 1 year. The income will increase about 2247 units as the year of education increases by 1 year.

#Exercise4==========

```
#Some variables used in previous exercises such as marital status, how many weeks of experience in each job,
#highest degree ever received, total income are selected in the new “dat A4 panel.csv” dataset.
#We now have the panel structure that includes 19 rounds of survey with these variables.
#Variable descriptions can be found in the “dat A4 panel/variables doc.pdf”.
#We are interested in the effect of education, marital status, experience and education on wages.

panel= fread("dat_A4_panel.csv")
```

#4.1==========

```
#4.1==============================================
#Explain the potential ability bias when trying to explain to understand the determinants of wages
```

People with traits the labor market values (intelligence, work ethic, conformity, etc.) tend to get more education. Since employers have some ability to detect these valued traits, people with more education would have earned above-average incomes even if their education were only average. Therefore, standard estimates overstate the effect of education on worker productivity and income.

#4.2==========

```
#Exploit the panel dimension of the data to propose a model to correct for the ability bias.
#Estimate the model using the following strategy.
# - Within Estimator.
# - Between Estimator
# - Difference (any) Estimator

panel= fread("dat_A4_panel.csv")

#First, turn it into panel data for each year and rowbind them
work_1997 = panel[,c(8:14)]
work_1997 = rowSums(work_1997,na.rm=TRUE)
dat_1997 = cbind(panel[,c(2,3)],NA,panel[,7],work_1997,1997)
dat_1997 = dat_1997 %>% drop_na(CV_MARSTAT_COLLAPSED_1997)
colnames(dat_1997) = c('ID','income','edu','marital','work','year')

work_1998 = panel[,c(20:28)]
work_1998 = rowSums(work_1998,na.rm=TRUE)
dat_1998 = cbind(panel[,c(2,17:19)],work_1998,1998)
dat_1998 = dat_1998 %>% drop_na(CV_MARSTAT_COLLAPSED_1998)
dat_1998 = dat_1998 %>% drop_na(CV_HIGHEST_DEGREE_9899_1998)
colnames(dat_1998) = c('ID','income','edu','marital','work','year')

work_1999 = panel[,c(32:40)]
work_1999 = rowSums(work_1999,na.rm=TRUE)
dat_1999 = cbind(panel[,c(2,29:31)],work_1999,1999)
dat_1999 = dat_1999 %>% drop_na(CV_HIGHEST_DEGREE_9900_1999)
colnames(dat_1999) = c('ID','income','edu','marital','work','year')

work_2000 = panel[,c(44:52)]
work_2000 = rowSums(work_2000,na.rm=TRUE)
dat_2000 = cbind(panel[,c(2,41:43)],work_2000,2000)
dat_2000 = dat_2000 %>% drop_na(CV_MARSTAT_COLLAPSED_2000)
dat_2000 = dat_2000 %>% drop_na(CV_HIGHEST_DEGREE_0001_2000)
colnames(dat_2000) = c('ID','income','edu','marital','work','year')
```

```r
work_2001 = panel[,c(56:63)]
work_2001 = rowSums(work_2001,na.rm=TRUE)
dat_2001 = cbind(panel[,c(2,53:55)],work_2001,2001)
dat_2001 = dat_2001 %>% drop_na(CV_MARSTAT_COLLAPSED_2001)
dat_2001 = dat_2001 %>% drop_na(CV_HIGHEST_DEGREE_0102_2001)
colnames(dat_2001) = c('ID','income','edu','marital','work','year')

work_2002 = panel[,c(67:77)]
work_2002 = rowSums(work_2002,na.rm=TRUE)
dat_2002 = cbind(panel[,c(2,64:66)],work_2002,2002)
dat_2002 = dat_2002 %>% drop_na(CV_MARSTAT_COLLAPSED_2002)
dat_2002 = dat_2002 %>% drop_na(CV_HIGHEST_DEGREE_0203_2002)
colnames(dat_2002) = c('ID','income','edu','marital','work','year')

work_2003 = panel[,c(80:89)]
work_2003 = rowSums(work_2003,na.rm=TRUE)
dat_2003 = cbind(panel[,c(2,78:79,90)],work_2003,2003)
dat_2003 = dat_2003 %>% drop_na(CV_MARSTAT_COLLAPSED_2003)
dat_2003 = dat_2003 %>% drop_na(CV_HIGHEST_DEGREE_0304_2003)
colnames(dat_2003) = c('ID','edu','marital','income','work','year')

work_2004 = panel[,c(93:99)]
work_2004 = rowSums(work_2004,na.rm=TRUE)
dat_2004 = cbind(panel[,c(2,91:92,100)],work_2004,2004)
dat_2004 = dat_2004 %>% drop_na(CV_MARSTAT_COLLAPSED_2004)
dat_2004 = dat_2004 %>% drop_na(CV_HIGHEST_DEGREE_0405_2004)
colnames(dat_2004) = c('ID','edu','marital','income','work','year')

work_2005 = panel[,c(103:111)]
work_2005 = rowSums(work_2005,na.rm=TRUE)
dat_2005 = cbind(panel[,c(2,101:102,112)],work_2005,2005)
dat_2005 = dat_2005 %>% drop_na(CV_MARSTAT_COLLAPSED_2005)
dat_2005 = dat_2005 %>% drop_na(CV_HIGHEST_DEGREE_0506_2005)
colnames(dat_2005) = c('ID','edu','marital','income','work','year')
```

```
work_2006 = panel[,c(115:123)]
work_2006 = rowSums(work_2006,na.rm=TRUE)
dat_2006 = cbind(panel[,c(2,113:114,124)],work_2006,2006)
dat_2006 = dat_2006 %>% drop_na(CV_MARSTAT_COLLAPSED_2006)
dat_2006 = dat_2006 %>% drop_na(CV_HIGHEST_DEGREE_0607_2006)
colnames(dat_2006) = c('ID','edu','marital','income','work','year')

work_2007 = panel[,c(127:134)]
work_2007 = rowSums(work_2007,na.rm=TRUE)
dat_2007 = cbind(panel[,c(2,125:126,135)],work_2007,2007)
dat_2007 = dat_2007 %>% drop_na(CV_MARSTAT_COLLAPSED_2007)
dat_2007 = dat_2007 %>% drop_na(CV_HIGHEST_DEGREE_0708_2007)
colnames(dat_2007) = c('ID','edu','marital','income','work','year')

work_2008 = panel[,c(138:145)]
work_2008 = rowSums(work_2008,na.rm=TRUE)
dat_2008 = cbind(panel[,c(2,136:137,146)],work_2008,2008)
dat_2008 = dat_2008 %>% drop_na(CV_MARSTAT_COLLAPSED_2008)
dat_2008 = dat_2008 %>% drop_na(CV_HIGHEST_DEGREE_0809_2008)
colnames(dat_2008) = c('ID','edu','marital','income','work','year')

work_2009 = panel[,c(149:157)]
work_2009 = rowSums(work_2009,na.rm=TRUE)
dat_2009 = cbind(panel[,c(2,147:148,158)],work_2009,2009)
dat_2009 = dat_2009 %>% drop_na(CV_MARSTAT_COLLAPSED_2009)
dat_2009 = dat_2009 %>% drop_na(CV_HIGHEST_DEGREE_0910_2009)
colnames(dat_2009) = c('ID','edu','marital','income','work','year')

work_2010 = panel[,c(162:170)]
work_2010 = rowSums(work_2010,na.rm=TRUE)
dat_2010 = cbind(panel[,c(2,160:161,171)],work_2010,2010)
dat_2010 = dat_2010 %>% drop_na(CV_MARSTAT_COLLAPSED_2010)
dat_2010 = dat_2010 %>% drop_na(CV_HIGHEST_DEGREE_1011_2010)
colnames(dat_2010) = c('ID','edu','marital','income','work','year')
```

```
work_2011 = panel[,c(175:187)]
work_2011 = rowSums(work_2011,na.rm=TRUE)
dat_2011 = cbind(panel[,c(2,173:174,188)],work_2011)
dat_2011 = dat_2011 %>% drop_na(CV_MARSTAT_COLLAPSED_2011)
dat_2011 = dat_2011 %>% drop_na(CV_HIGHEST_DEGREE_1112_2011)
colnames(dat_2011) = c('ID','edu','marital','income','work','year')

work_2013 = panel[,c(192:201)]
work_2013 = rowSums(work_2013,na.rm=TRUE)
dat_2013 = cbind(panel[,c(2,190:191,202)],work_2013,2013)
dat_2013 = dat_2013 %>% drop_na(CV_MARSTAT_COLLAPSED_2013)
dat_2013 = dat_2013 %>% drop_na(CV_HIGHEST_DEGREE_1314_2013)
colnames(dat_2013) = c('ID','edu','marital','income','work','year')

work_2015 = panel[,c(205:216)]
work_2015 = rowSums(work_2015,na.rm=TRUE)
dat_2015 = cbind(panel[,c(2,203:204,217)],work_2015,2015)
dat_2015 = dat_2015 %>% drop_na(CV_MARSTAT_COLLAPSED_2015)
dat_2015 = dat_2015 %>% drop_na(CV_HIGHEST_DEGREE_EVER_EDT_2015)
colnames(dat_2015) = c('ID','edu','marital','income','work','year')

work_2017 = panel[,c(220:234)]
work_2017 = rowSums(work_2017,na.rm=TRUE)
dat_2017 = cbind(panel[,c(2,218:219,235)],work_2017,2017)
dat_2017 = dat_2017 %>% drop_na(CV_MARSTAT_COLLAPSED_2017)
dat_2017 = dat_2017 %>% drop_na(CV_HIGHEST_DEGREE_EVER_EDT_2017)
colnames(dat_2017) = c('ID','edu','marital','income','work','year')

work_2019 = panel[,c(238:248)]
work_2019 = rowSums(work_2019,na.rm=TRUE)
dat_2019 = cbind(panel[,c(2,236:237,249)],work_2019,2019)
dat_2019 = dat_2019 %>% drop_na(CV_MARSTAT_COLLAPSED_2019)
dat_2019 = dat_2019 %>% drop_na(CV_HIGHEST_DEGREE_EVER_EDT_2019)
colnames(dat_2019) = c('ID','edu','marital','income','work','year')
```

```
datind = rbind(dat_1997,dat_1998,dat_1999,dat_2000,dat_2001,dat_2002,dat_2003,
               dat_2004,dat_2005,dat_2006,dat_2007,dat_2008,dat_2009,dat_2010,
               dat_2013,dat_2015,dat_2017,dat_2019)
view(datind)
```

	ID	income	edu	marital	work	year
1	14	NA	NA	0	0	1997
2	20	NA	NA	0	0	1997
3	26	NA	NA	0	0	1997
4	33	25	NA	0	2	1997
5	36	300	NA	0	50	1997
6	37	NA	NA	0	0	1997
7	39	NA	NA	0	0	1997
8	51	NA	NA	0	0	1997
9	62	NA	NA	0	0	1997
10	64	NA	NA	0	0	1997
11	66	NA	NA	0	0	1997
12	67	NA	NA	0	0	1997
13	71	NA	NA	0	14	1997
14	86	2500	NA	0	0	1997
15	94	464	NA	0	10	1997
16	97	NA	NA	0	0	1997
17	104	400	NA	0	124	1997
18	118	NA	NA	0	12	1997
19	123	2000	NA	0	84	1997
20	129	1000	NA	0	50	1997
21	131	NA	NA	0	0	1997
22	137	500	NA	0	24	1997
23	141	2000	NA	0	547	1997
24	153	NA	NA	0	14	1997
25	154	1500	NA	0	20	1997
26	155	NA	NA	0	122	1997

Showing 1 to 27 of 122,403 entries, 6 total columns

```
# - Within Estimator.
#Fix time
library(plm)
within = plm(datind$income~datind$edu + datind$marital + datind$work,datind,index=c("year"),model="within")
summary(within)
```

```
> summary(within)
Oneway (individual) effect Within Model

Call:
plm(formula = datind$income ~ datind$edu + datind$marital + datind$work,
    data = datind, model = "within", index = c("year"))

Unbalanced Panel: n = 18, T = 901-5609, N = 77608

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-81595.7 -10314.7  -2148.3   6247.1 293283.0

Coefficients:
                Estimate Std. Error t-value  Pr(>|t|)
datind$edu     5563.94074   69.74411  79.776 < 2.2e-16 ***
datind$marital 2387.25233  125.53047  19.017 < 2.2e-16 ***
datind$work      24.23905    0.53198  45.564 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    4.7274e+13
Residual Sum of Squares: 4.2198e+13
R-Squared:      0.10738
Adj. R-Squared: 0.10715
F-statistic: 3111.03 on 3 and 77587 DF, p-value: < 2.22e-16
```

```r
# - Between Estimator
#Fix individuals
between = plm(datind$income~datind$edu + datind$marital + datind$work,datind,index=c("ID"),model="between")
summary(between)
```

```
> summary(between)
Oneway (individual) effect Between Model

Call:
plm(formula = datind$income ~ datind$edu + datind$marital + datind$work,
    data = datind, model = "between", index = c("ID"))

Unbalanced Panel: n = 8680, T = 1-18, N = 76707
Observations used in estimation: 8680

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-51492.70  -5272.49   -844.18   4020.38 116949.00

Coefficients:
                 Estimate Std. Error t-value  Pr(>|t|)
(Intercept)    -8716.0345   277.2847 -31.433 < 2.2e-16 ***
datind$edu      7851.9893   156.1385  50.289 < 2.2e-16 ***
datind$marital  9287.8220   359.2368  25.854 < 2.2e-16 ***
datind$work       74.3343     1.3431  55.345 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    2.9322e+12
Residual Sum of Squares: 8.1841e+11
R-Squared:      0.72089
Adj. R-Squared: 0.72079
F-statistic: 7469.46 on 3 and 8676 DF, p-value: < 2.22e-16
```

```r
# - Difference (any) Estimator
difference = plm(datind$income~datind$edu + datind$marital + datind$work,datind,index=c("work"),model="fd")
summary(difference)
```

```
> summary(difference)
Oneway (individual) effect First-Difference Model

Call:
plm(formula = datind$income ~ datind$edu + datind$marital + datind$work,
    data = datind, model = "fd", index = c("work"))

Unbalanced Panel: n = 1191, T = 1-7376, N = 76707
Observations used in estimation: 75516

Residuals:
       Min.     1st Qu.      Median     3rd Qu.        Max.
-316246.233  -11543.703     -20.805   11418.726  333951.756

Coefficients:
                 Estimate Std. Error t-value Pr(>|t|)
(Intercept)      20.80528  115.67275  0.1799   0.8573
datind$edu     4665.81940   74.13207 62.9393   <2e-16 ***
datind$marital 2307.24946  125.47216 18.3885   <2e-16 ***
datind$work      23.08343    0.52979 43.5709   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    8.2836e+13
Residual Sum of Squares: 7.6299e+13
R-Squared:      0.078925
Adj. R-Squared: 0.078888
F-statistic: 2156.81 on 3 and 75512 DF, p-value: < 2.22e-16
```

#4.3==========

```
#4.3================================================
#Interpret the results from each model and explain why different models yield different parameter estimates
```

1. Within Estimator: When everything else is equal, income will increase by an average of 5564 units when education year increases by 1 year; people who get married will earn 2387 units more than people who don’t get married; income will increase by an average of 24 units when work experience increases 1 year.
2. Between Estimators: When everything else is equal, income will increase by an average of 7852 units when education year increases by 1 year; people who get married will earn 9287 units more than people who don’t get married; income will increase by an average of 74 units when work experience increases 1 year.
3. Difference Estimators: When everything else is equal, income will increase by an average of 4666 units when education year increases by 1 year; people who get married will earn 2307 units more than people who don’t get married; income will increase by an average of 23 units when work experience increases 1 year.
4. Why different models yield different parameter estimates? It is because in within estimator model, the time effect is fixed, while in between estimator model, we focus on the variation between different individuals. In FD model, we compute the difference of the variables between each periods and calculate the relationship between the differences of independent and dependent variables.