HW4 Wenxuan Wang

```
setwd("/Users/wenxuan/Desktop/A4/Data")
library(data.table)
dat_A4 = fread("dat_A4.csv")
```

#==========

Exercise 1:

#==========

1.Create additional variable for the age of the agent ”age”, total work experience measured in years ”work exp”

```
#calculate age
dat_A4$age=2019-dat_A4$KEY_BDATE_Y_1997
#work experience
#first, replace NA==0
dat_A4[which(is.na(dat_A4[,18])==TRUE),'CV_WKSWK_JOB_DLI.01_2019'] = 0
dat_A4[which(is.na(dat_A4[,19])==TRUE),'CV_WKSWK_JOB_DLI.02_2019'] = 0
dat_A4[which(is.na(dat_A4[,20])==TRUE),'CV_WKSWK_JOB_DLI.03_2019'] = 0
dat_A4[which(is.na(dat_A4[,21])==TRUE),'CV_WKSWK_JOB_DLI.04_2019'] = 0
dat_A4[which(is.na(dat_A4[,22])==TRUE),'CV_WKSWK_JOB_DLI.05_2019'] = 0
dat_A4[which(is.na(dat_A4[,23])==TRUE),'CV_WKSWK_JOB_DLI.06_2019'] = 0
dat_A4[which(is.na(dat_A4[,24])==TRUE),'CV_WKSWK_JOB_DLI.07_2019'] = 0
dat_A4[which(is.na(dat_A4[,25])==TRUE),'CV_WKSWK_JOB_DLI.08_2019'] = 0
dat_A4[which(is.na(dat_A4[,26])==TRUE),'CV_WKSWK_JOB_DLI.09_2019'] = 0
dat_A4[which(is.na(dat_A4[,27])==TRUE),'CV_WKSWK_JOB_DLI.10_2019'] = 0
dat_A4[which(is.na(dat_A4[,28])==TRUE),'CV_WKSWK_JOB_DLI.11_2019'] = 0
```

```
#calculate years
as.numeric(dat_A4$CV_WKSWK_JOB_DLI.11_2019)
as.numeric(dat_A4$CV_WKSWK_JOB_DLI.10_2019)
as.numeric(dat_A4$CV_WKSWK_JOB_DLI.09_2019)
as.numeric(dat_A4$CV_WKSWK_JOB_DLI.08_2019)
as.numeric(dat_A4$CV_WKSWK_JOB_DLI.07_2019)
as.numeric(dat_A4$CV_WKSWK_JOB_DLI.06_2019)
as.numeric(dat_A4$CV_WKSWK_JOB_DLI.05_2019)
as.numeric(dat_A4$CV_WKSWK_JOB_DLI.04_2019)
as.numeric(dat_A4$CV_WKSWK_JOB_DLI.03_2019)
as.numeric(dat_A4$CV_WKSWK_JOB_DLI.02_2019)
as.numeric(dat_A4$CV_WKSWK_JOB_DLI.01_2019)

dat_A4$work_exp=(dat_A4$CV_WKSWK_JOB_DLI.01_2019+dat_A4$CV_WKSWK_JOB_DLI.02_2019+
                   dat_A4$CV_WKSWK_JOB_DLI.03_2019+dat_A4$CV_WKSWK_JOB_DLI.04_2019+
                   dat_A4$CV_WKSWK_JOB_DLI.05_2019+dat_A4$CV_WKSWK_JOB_DLI.06_2019+
                   dat_A4$CV_WKSWK_JOB_DLI.07_2019+dat_A4$CV_WKSWK_JOB_DLI.08_2019+
                   dat_A4$CV_WKSWK_JOB_DLI.09_2019+dat_A4$CV_WKSWK_JOB_DLI.10_2019+dat_A4$CV_WKSWK_JOB_DLI.11_2019)/52
```

The top 15 rows of the data are as follows:

age	work_exp
38	0.0000000
37	12.4230769
36	1.6923077
38	1.9230769
37	13.4615385
37	2.2500000
36	2.3653846
38	4.1923077
37	3.2307692
35	5.0769231
37	11.9423077
38	14.9230769
35	0.0000000
39	0.0000000
36	9.5961538

#2.Create additional education variables indicating total years of schooling from all variables related to education

```
#biological father
dat_A4$biodad=dat_A4$CV_HGC_BIO_DAD_1997
dat_A4[which(dat_A4$CV_HGC_BIO_DAD_1997==95),'biodad'] = 0

# biological mother
dat_A4$biomom=dat_A4$CV_HGC_BIO_MOM_1997
dat_A4[which(dat_A4$CV_HGC_BIO_MOM_1997==95),'biomom'] = 0

# residential father
dat_A4$resdad=dat_A4$CV_HGC_RES_DAD_1997
dat_A4[which(dat_A4$CV_HGC_RES_DAD_1997==95),'resdad'] = 0

# residential mother
dat_A4$resmom=dat_A4$CV_HGC_RES_MOM_1997
dat_A4[which(dat_A4$CV_HGC_RES_MOM_1997==95),'resmom'] = 0

#self
dat_A4[which(dat_A4$YSCH.3113_2019==1),'self'] = 0
dat_A4[which(dat_A4$YSCH.3113_2019==2),'self'] = 4
dat_A4[which(dat_A4$YSCH.3113_2019==3),'self'] = 12
dat_A4[which(dat_A4$YSCH.3113_2019==4),'self'] = 14
dat_A4[which(dat_A4$YSCH.3113_2019==5),'self'] = 16
dat_A4[which(dat_A4$YSCH.3113_2019==6),'self'] = 18
dat_A4[which(dat_A4$YSCH.3113_2019==7),'self'] = 23
dat_A4[which(dat_A4$YSCH.3113_2019==8),'self'] = 22
```

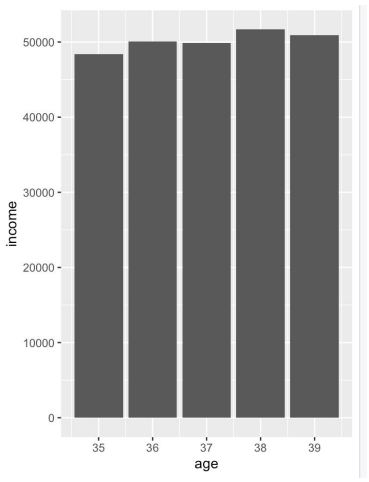
The top 15 rows of the data are as follows:

biodad	biomom	resdad	resmom	self
16	8	16	8	*NA*
17	15	14	15	12
NA	12	*NA*	12	16
12	12	*NA*	12	12
12	12	12	12	12
NA	12	*NA*	12	12
NA	12	*NA*	12	0
6	12	6	12	16
6	12	6	12	18
6	12	6	12	18
12	14	*NA*	14	16
NA	12	12	12	4
NA	6	*NA*	6	4
12	12	12	12	*NA*
13	12	13	12	14
10	11	*NA*	11	16
NA	14	10	14	4

#3.1Plot the income data (where income is positive) by i) age groups, ii) gender groups and iii)number of children

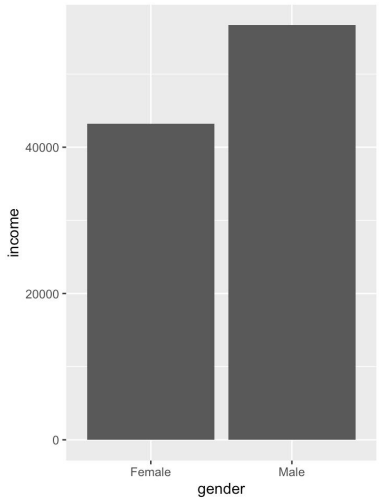
```
#if income = NA or 0, then delete them
dat_NA=dat_A4[-which(is.na(dat_A4$YINC_1700_2019))]
dat_NA=dat_NA[-which(dat_NA$YINC_1700_2019==0)]

#i) age groups
data2=aggregate(dat_NA$YINC_1700_2019, by=list(type=dat_NA$age),mean)
colnames(data2)=c("age","income")
ggplot(data2, aes(x = age, y = income)) +
  geom_bar(stat = "identity")
```



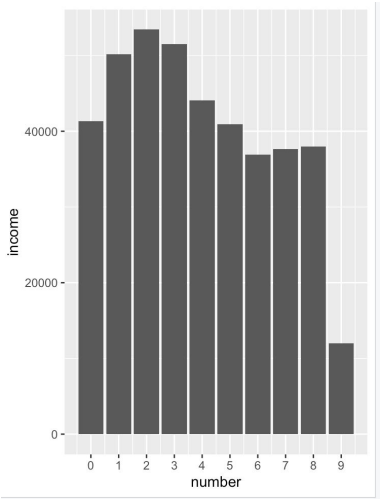
Interpret the visualizations from above：There is a weak positive correlation between age and income. 38 years old people earn the most.

```
#ii) gender groups
data1=aggregate(dat_NA$YINC_1700_2019, by=list(type=dat_NA$KEY_SEX_1997),mean)
colnames(data1)=c("gender","income")
data1[which(data1$gender==1),'gender'] = "Male"
data1[which(data1$gender==2),'gender'] = "Female"
ggplot(data1, aes(x = gender, y = income)) +
  geom_bar(stat = "identity")
```



Interpret the visualizations from above：The average income of male is higher than that of female.

```r
#iii)number of children
data3=aggregate(dat_NA$YINC_1700_2019, by=list(type=dat_NA$CV_BIO_CHILD_HH_U18_201),mean)
colnames(data3)=c("number","income")
ggplot(data3, aes(x = number, y = income)) +
  geom_bar(stat = "identity")+scale_x_continuous(breaks=seq(0,9,1))
```



Interpret the visualizations from above：Families with one to four children have more income. Families with two children had the highest incomes.

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

i) age group

```r
#delete NA
#i) age groups
data4=dat_A4[-which(is.na(dat_A4$YINC_1700_2019))]
data5=data4[which(data4$YINC_1700_2019==0)]
data6=aggregate(data5$YINC_1700_2019, by=list(type=data5$age),length)
data7=aggregate(data4$YINC_1700_2019, by=list(type=data4$age),length)
data6$x/data7$x
table1=data.frame(age=c(35,36,37,38,39),
                  share=c(data6$x/data7$x)
)
table1
```

```
> table1
  age       share
1  35 0.009293680
2  36 0.006300630
3  37 0.005420054
4  38 0.008960573
5  39 0.002994012
```

Interpret the visualizations from above：35 and 38 have a higher share of 0

ii) gender groups

```
#ii) gender groups
data8=aggregate(data5$YINC_1700_2019, by=list(type=data5$KEY_SEX_1997),length)
data9=aggregate(data4$YINC_1700_2019, by=list(type=data4$KEY_SEX_1997),length)
data8[which(data1$gender==1),'gender'] = "Male"
data8[which(data1$gender==2),'gender'] = "Female"
data9[which(data1$gender==1),'gender'] = "Male"
data9[which(data1$gender==2),'gender'] = "Female"
data8$x/data9$x
table2=data.frame(gender=c("Male","Female"),
                  share=c(data8$x/data9$x)
)
table2
```

```
> table2
  gender       share
1   Male 0.007500000
2 Female 0.005742726
```

Interpret the visualizations from above：Male have a higher share of 0

iii) number of children and marital status

The first is the combined of children and marital status.

```
#iii) number of children and marital status
#Create a new variable "children_marital"
data12=data4[-which(is.na(data4$CV_BIO_CHILD_HH_U18_2019))]
data12=data12[-which(is.na(data12$CV_MARSTAT_COLLAPSED_2019))]
data12$children_marital= paste(data12$CV_BIO_CHILD_HH_U18_2019,data12$CV_MARSTAT_COLLAPSED_2019)
data13=data12%>% group_by(children_marital) %>% mutate(share= length(which(YINC_1700_2019 == 0))/length(which(YINC_1700_2019 >= 0)))%>% ungroup
data14=cbind(data13[,38],data13[,39])
data15=unique(data14)
data15
```

children_marital	share
3 2	0.142857143
0 2	0.136363636
0 1	0.033898305
3 0	0.017699115
1 0	0.011080332
2 1	0.008456660
1 1	0.008156607
0 3	0.006802721
3 1	0.004597701
0 0	0.000000000
0 4	0.000000000
1 2	0.000000000

The second is number of children.

```
#iii)number of children
data10=aggregate(data5$YINC_1700_2019, by=list(type=data5$CV_BIO_CHILD_HH_U18_201),length)
data11=aggregate(data4$YINC_1700_2019, by=list(type=data4$CV_BIO_CHILD_HH_U18_201),length)
table3 = data.frame(
 children = c(0,1,2,3,4,5,6,7,8,9),
  share= c(8/537,9/1147,8/1393,5/623,0,0,0,0,0,0)
)
table3
```

```
> table3
   children       share
1         0 0.014897579
2         1 0.007846556
3         2 0.005743001
4         3 0.008025682
5         4 0.000000000
6         5 0.000000000
7         6 0.000000000
8         7 0.000000000
9         8 0.000000000
10        9 0.000000000
```

The third is the marital status.

```
#iiii)marital status
data12=aggregate(data5$YINC_1700_2019, by=list(type=data5$CV_MARSTAT_COLLAPSED_2019),length)
data13=aggregate(data4$YINC_1700_2019, by=list(type=data4$CV_MARSTAT_COLLAPSED_2019),length)
table4 = data.frame(
  marital_status = c(" Never-marrie","Married","Separated", "Divorced","Widowed"),
  share= c(11/1947,20/2683,4/93,1/650,0)
)
table4
```

```
> table4
  marital_status       share
1   Never-marrie 0.005649718
2        Married 0.007454342
3      Separated 0.043010753
4       Divorced 0.001538462
5        Widowed 0.000000000
```

Interpret the visualizations from above：Individuals with fewer children have a higher share of 0. Separated individuals have a higher share of 0

#==========

Exercise 2:

#==========

#2.1Specify and estimate an OLS model to explain the income variable (where income is positive).

```
#since income is positive, we need to drop 0
dat_OLS=dat_A4[-which(dat_A4$YINC_1700_2019==0)]
reg1=lm(dat_OLS$YINC_1700_2019~dat_OLS$age + dat_OLS$work_exp + dat_OLS$self)
summary(reg1)
```

```
Call:
lm(formula = dat_OLS$YINC_1700_2019 ~ dat_OLS$age + dat_OLS$work_exp +
    dat_OLS$self)

Residuals:
   Min     1Q Median     3Q    Max
-72758 -19234  -3535  17798  80860

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2051.13    9463.05   0.217    0.828
dat_OLS$age        407.64     254.99   1.599    0.110
dat_OLS$work_exp  1001.80      66.25  15.123   <2e-16 ***
dat_OLS$self      2012.28      73.03  27.556   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpret the estimation results

All else is equal, if age adds one year, the income goes up by 407. 64; if work experience increases 1 year, the income increases 1001.8; if year of education increases one year, the income increases 2012.28.

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

Explanation: In OLS, it will delete the missing value in the regression. Since the missing value is not a random missing value. For example, if you choose not to work, you don't get paid and this data will not be included. This situation may be related to the independent variables and thus causes endogeneity problem.

2.2 Explain why the Heckman model can deal with the selection problem.

First, estimate the probability of job participation, which can be derived from empirical data models. Then, delete the samples of people who do not work, and add the Inverse Mill’s Ratio, the

remaining sample points are shifted vertically downward according to their different working probability. Finally, get the new regression line. This regression line is consistent with the true regression line.

2.3 Estimate a Heckman selection model

```
#delete independent variables' NA

dat_OLS=dat_OLS[-which(is.na(dat_OLS$self))]
dat_OLS=dat_OLS[-which(is.na(dat_OLS$biodad))]
dat_OLS=dat_OLS[-which(is.na(dat_OLS$biomom))]
dat_OLS=dat_OLS[-which(is.na(dat_OLS$resdad))]
dat_OLS=dat_OLS[-which(is.na(dat_OLS$resmom))]
dat_OLS=dat_OLS[-which(is.na(dat_OLS$CV_BIO_CHILD_HH_U18_2019))]
```

```
#create a dummy variable observed_index, if income is NA, observed_index is 0; if income is not NA, observed_index is 1
dat_OLS[is.na(dat_OLS$YINC_1700_2019),'observed_index'] = 0
dat_OLS[which(dat_OLS$YINC_1700_2019 > 0),'observed_index'] = 1
#run the probit model, which includes many independent variables created above
probit = glm(observed_index ~ age +work_exp + self+biodad + biomom
             + resdad + resmom  +CV_BIO_CHILD_HH_U18_2019,
             family = binomial(link = 'probit'),data=dat_OLS)

summary(probit)
```

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.169688   0.814276  -0.208   0.8349
age                      -0.005231   0.021793  -0.240   0.8103
work_exp                  0.120507   0.007908  15.239  < 2e-16 ***
self                      0.039080   0.006425   6.082 1.18e-09 ***
biodad                    0.008149   0.029471   0.277   0.7822
biomom                   -0.111947   0.057425  -1.949   0.0512 .
resdad                    0.006343   0.029452   0.215   0.8295
resmom                    0.114983   0.056820   2.024   0.0430 *
CV_BIO_CHILD_HH_U18_2019 -0.046483   0.025307  -1.837   0.0662 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
#to calculate the inverse mills ratio
probit1=predict(probit)
millsratio=dnorm(probit1)/pnorm(probit1)
summary(millsratio)
```

```
> summary(millsratio)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
0.0000056 0.1123113 0.3044885 0.3307259 0.5031871 1.1941378
```

```
#linear regression
lm_select = lm(dat_OLS$YINC_1700_2019 ~  age +work_exp + self+biodad + biomom
               + resdad + resmom  +CV_BIO_CHILD_HH_U18_2019+millsratio,data=dat_OLS )
summary(lm_select)
```

```
> summary(lm_select)

Call:
lm(formula = dat_OLS$YINC_1700_2019 ~ age + work_exp + self +
    biodad + biomom + resdad + resmom + CV_BIO_CHILD_HH_U18_2019 +
    millsratio, data = dat_OLS)

Residuals:
   Min     1Q Median     3Q    Max
-67422 -19498  -2086  21016  83487

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               25329.99   15958.46   1.587  0.11260
age                         385.80     405.34   0.952  0.34131
work_exp                   -289.15     278.28  -1.039  0.29890
self                       1205.92     190.53   6.329 2.99e-10 ***
biodad                      439.18     544.21   0.807  0.41976
biomom                     2844.26     921.09   3.088  0.00204 **
resdad                      -92.75     541.52  -0.171  0.86402
resmom                    -2149.23     915.08  -2.349  0.01893 *
CV_BIO_CHILD_HH_U18_2019    348.02     534.28   0.651  0.51487
millsratio                -45002.90    8318.07  -5.410 6.99e-08 ***
---
```

```
function1 = function(par, X, Z, y, observed_index) {
  gamma     = par[1:9]
  probit1 = Z %*% gamma
  beta   = par[10:19]
  lp_lm = X %*% beta
  sigma = par[20]
  rho    = par[21]
  ll = sum(log(1-pnorm(probit1[!observed_index]))) - log(sigma) +
    sum(dnorm(y, mean = lp_lm, sd = sigma, log = TRUE)) +
    sum( pnorm((probit1[observed_index] + rho/sigma * (y-lp_lm)) / sqrt(1-rho^2),
           log.p = TRUE)
    )

  -ll
}
X = model.matrix(lm_select)
X
Z = model.matrix(probit)
Z
```

```r
# initial values
init = c(coef(probit), coef(lm_select),  1, 0)
fun = optim(
  init,
  function1,
  X = X,
  Z = Z,
  y = dat_OLS$YINC_1700_2019[which(dat_OLS$observed_index==1)],
  observed_index = dat_OLS$observed_index,
  method  = 'Nelder-Mead',
  control = list(maxit = 1000, reltol = 0),
  hessian = T
)
fun$par
```

```
> fun$par
          (Intercept)                   age              work_exp                  self                biodad
        -3.902718e-01         -1.084789e-01          4.148773e-01         -1.197953e-01         2.917845e-01
               biomom                resdad                resmom CV_BIO_CHILD_HH_U18_2019           (Intercept)
        -1.678280e-01         -7.920954e-02          6.881678e-02          4.135804e-01         2.532915e+04
                  age              work_exp                  self                biodad               biomom
         3.830657e+02         -2.888788e+02          1.205778e+03          4.420484e+02         2.844907e+03
               resdad                resmom CV_BIO_CHILD_HH_U18_2019            millsratio
        -9.159247e+01         -2.150466e+03          3.475849e+02         -4.500295e+04         4.502065e+03

        -6.115310e-03
```

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

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               25329.99   15958.46   1.587  0.11260
age                         385.80     405.34   0.952  0.34131
work_exp                   -289.15     278.28  -1.039  0.29890
self                       1205.92     190.53   6.329 2.99e-10 ***
biodad                      439.18     544.21   0.807  0.41976
biomom                     2844.26     921.09   3.088  0.00204 **
resdad                      -92.75     541.52  -0.171  0.86402
resmom                    -2149.23     915.08  -2.349  0.01893 *
CV_BIO_CHILD_HH_U18_2019    348.02     534.28   0.651  0.51487
millsratio                -45002.90    8318.07  -5.410 6.99e-08 ***
---
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2051.13    9463.05   0.217    0.828
dat_OLS$age        407.64     254.99   1.599    0.110
dat_OLS$work_exp  1001.80      66.25  15.123   <2e-16 ***
dat_OLS$self      2012.28      73.03  27.556   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The first result is the Heckman result, and the second is the OLS result. Interpret the results from the

Heckman selection model: all else is equal, if year of education increase 1 year, the income will increase 1205.92 on average.

Comparison: The coefficient of work experience in Heckman model is negative, but the coefficient of work experience in OLS is positive. They are both not significant. The reason may be that in OLS, the effect of work experience is magnified because it considers fewer variables and may have endogeneity problems. The effect of years of education is also magnified in OLS.

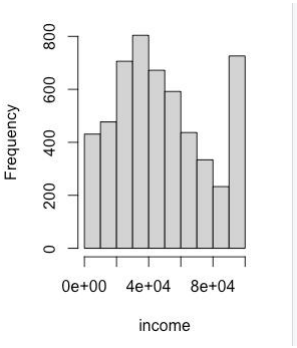
#==========

Exercise 3:

#==========

#3.1 Plot a histogram to check whether the distribution of the income variable.

```
hist(dat_A4$YINC_1700_2019,xlab="income")
#100,000 is the censored value here
```



100,000 is the censored value here

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

Tobit Model

3.3 Estimate the appropriate model with the censored data

```
#delete if income=0/NA
dat_OLS=dat_A4[-which(dat_A4$YINC_1700_2019==0)]
dat_OLS=dat_OLS[-which(is.na(dat_OLS$YINC_1700_2019))]
dat_OLS=dat_OLS[-which(is.na(dat_OLS$self))]


tobit = function(par, X, y, ul) {
  n=length(par)
  sigma = exp(par[n])
  beta  = par[-n]
    limit = ul
    indicator = y <=ul
    m = X %*% beta
  ll = sum(indicator * log((1/sigma)*dnorm((y-m)/sigma)) ) +
    sum((1-indicator) * log(pnorm((m-limit)/sigma)))
  -ll
}
```

```
init1 = lm(dat_OLS$YINC_1700_2019 ~  age +work_exp + self,data=dat_OLS)
X = model.matrix(init1)
X
init = c(coef(init1), log_sigma = log(summary(init1)$sigma))

res=optim(
     par = init,
     tobit1,
     y = dat_OLS$YINC_1700_2019,
     X = X,
     method  ="Nelder-Mead",
     ul = 100000,
     control = list(maxit = 16000, reltol = 1e-15)
   )
res$par
init1$coefficients
```

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

Interpret: All else is equal, when age increases 1 year, the income will increase about 529; when work experience increases 1 year, the income will increase about 1067; when the years of education increases 1 year, the income will increase about 2218.

Comparison: Compare to those when not correcting for the censored data, the effect of "age", "work_exp", and "years of education" all increase. Age had the largest effect, up to 30 percent, while years of education increased by about 10 percent.

```
> res$par
(Intercept)         age    work_exp        self   log_sigma
-3961.18161   528.67968  1066.88016  2218.44284    10.27339
> init1$coefficients
(Intercept)         age    work_exp        self
  2051.1273    407.6367   1001.7983   2012.2843
```

#==========

Exercise 4:

#==========

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

People with high ability are more successful and productive, and therefore earn more. However, even if we know that ability is part of the real model to understand the determinants of wages, we cannot include it because we do not have a reasonable and reliable measure of ability.

4.2Exploit 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

Organize the data as a panel.

```
#Organizing the Data as a Panel
dat_A4_panel = fread("dat_A4_panel.csv")
dat_A4_panel=dat_A4_panel[,-1]
#from wide data to long data
```

First, from wide data to long data.

```
data_panel1997=dat_A4_panel[,1:15]
data_panel1997=data_panel1997[,c(1,2,3,6:13)]
data_panel1997$KEY_SEX_1997=NA
data_panel1998=dat_A4_panel[,c(1,16:27)]
data_panel1999=dat_A4_panel[,c(1,28:39)]
data_panel2000=dat_A4_panel[,c(1,40:51)]
data_panel2001=dat_A4_panel[,c(1,52:62)]
data_panel2002=dat_A4_panel[,c(1,63:76)]
data_panel2003=dat_A4_panel[,c(1,89,77:88)]
data_panel2004=dat_A4_panel[,c(1,99,90:98)]
data_panel2005=dat_A4_panel[,c(1,111,100:110)]
data_panel2006=dat_A4_panel[,c(1,123,112:122)]
data_panel2007=dat_A4_panel[,c(1,134,124:133)]
data_panel2008=dat_A4_panel[,c(1,145,135:144)]
data_panel2009=dat_A4_panel[,c(1,157,146:156)]
data_panel2010=dat_A4_panel[,c(1,170,159:169)]
data_panel2011=dat_A4_panel[,c(1,187,172:186)]
data_panel2013=dat_A4_panel[,c(1,201,188:200)]
data_panel2013=data_panel2013[,-3]
data_panel2015=dat_A4_panel[,c(1,216,202:215)]
data_panel2017=dat_A4_panel[,c(1,234,217:233)]
data_panel2019=dat_A4_panel[,c(1,248,235:247)]
```

Second, create the year variables, that is representing which year this data is surveyed.

```
#create the year variable
data_panel1997$year=matrix(rep(1997,nrow(data_panel1997)))
data_panel1998$year=matrix(rep(1998,nrow(data_panel1998)))
data_panel1999$year=matrix(rep(1999,nrow(data_panel1999)))
data_panel2000$year=matrix(rep(2000,nrow(data_panel2000)))
data_panel2001$year=matrix(rep(2001,nrow(data_panel2001)))
data_panel2002$year=matrix(rep(2002,nrow(data_panel2002)))
data_panel2003$year=matrix(rep(2003,nrow(data_panel2003)))
data_panel2004$year=matrix(rep(2004,nrow(data_panel2004)))
data_panel2005$year=matrix(rep(2005,nrow(data_panel2005)))
data_panel2006$year=matrix(rep(2006,nrow(data_panel2006)))
data_panel2007$year=matrix(rep(2007,nrow(data_panel2007)))
data_panel2008$year=matrix(rep(2008,nrow(data_panel2008)))
data_panel2009$year=matrix(rep(2009,nrow(data_panel2009)))
data_panel2010$year=matrix(rep(2010,nrow(data_panel2010)))
data_panel2011$year=matrix(rep(2011,nrow(data_panel2011)))
data_panel2013$year=matrix(rep(2013,nrow(data_panel2013)))
data_panel2015$year=matrix(rep(2015,nrow(data_panel2015)))
data_panel2017$year=matrix(rep(2017,nrow(data_panel2017)))
data_panel2019$year=matrix(rep(2019,nrow(data_panel2019)))
```

Third, calculate the sum of work experience.

```
#calculate the sum of work experience
#first, replace working experience's NA==0
data_panel1997[which(is.na(data_panel1997[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_1997'] = 0
data_panel1997[which(is.na(data_panel1997[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_1997'] = 0
data_panel1997[which(is.na(data_panel1997[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_1997'] = 0
data_panel1997[which(is.na(data_panel1997[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_1997'] = 0
data_panel1997[which(is.na(data_panel1997[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_1997'] = 0
data_panel1997[which(is.na(data_panel1997[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_1997'] = 0
data_panel1997[which(is.na(data_panel1997[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_1997'] = 0

data_panel1998[which(is.na(data_panel1998[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_1998'] = 0
data_panel1998[which(is.na(data_panel1998[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_1998'] = 0
data_panel1998[which(is.na(data_panel1998[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_1998'] = 0
data_panel1998[which(is.na(data_panel1998[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_1998'] = 0
data_panel1998[which(is.na(data_panel1998[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_1998'] = 0
data_panel1998[which(is.na(data_panel1998[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_1998'] = 0
data_panel1998[which(is.na(data_panel1998[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_1998'] = 0
data_panel1998[which(is.na(data_panel1998[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_1998'] = 0
data_panel1998[which(is.na(data_panel1998[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_1998'] = 0

data_panel1999[which(is.na(data_panel1999[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_1999'] = 0
data_panel1999[which(is.na(data_panel1999[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_1999'] = 0
data_panel1999[which(is.na(data_panel1999[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_1999'] = 0
data_panel1999[which(is.na(data_panel1999[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_1999'] = 0
data_panel1999[which(is.na(data_panel1999[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_1999'] = 0
data_panel1999[which(is.na(data_panel1999[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_1999'] = 0
data_panel1999[which(is.na(data_panel1999[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_1999'] = 0
data_panel1999[which(is.na(data_panel1999[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_1999'] = 0
data_panel1999[which(is.na(data_panel1999[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_1999'] = 0


data_panel2000[which(is.na(data_panel2000[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2000'] = 0
data_panel2000[which(is.na(data_panel2000[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2000'] = 0
data_panel2000[which(is.na(data_panel2000[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2000'] = 0
data_panel2000[which(is.na(data_panel2000[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2000'] = 0
data_panel2000[which(is.na(data_panel2000[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2000'] = 0
data_panel2000[which(is.na(data_panel2000[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2000'] = 0
data_panel2000[which(is.na(data_panel2000[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2000'] = 0
data_panel2000[which(is.na(data_panel2000[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2000'] = 0
data_panel2000[which(is.na(data_panel2000[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2000'] = 0

data_panel2001[which(is.na(data_panel2001[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2001'] = 0
data_panel2001[which(is.na(data_panel2001[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2001'] = 0
data_panel2001[which(is.na(data_panel2001[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2001'] = 0
data_panel2001[which(is.na(data_panel2001[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2001'] = 0
data_panel2001[which(is.na(data_panel2001[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2001'] = 0
data_panel2001[which(is.na(data_panel2001[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2001'] = 0
data_panel2001[which(is.na(data_panel2001[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2001'] = 0
data_panel2001[which(is.na(data_panel2001[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2001'] = 0

data_panel2002[which(is.na(data_panel2002[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2002'] = 0
data_panel2002[which(is.na(data_panel2002[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2002'] = 0
data_panel2002[which(is.na(data_panel2002[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2002'] = 0
data_panel2002[which(is.na(data_panel2002[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2002'] = 0
data_panel2002[which(is.na(data_panel2002[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2002'] = 0
data_panel2002[which(is.na(data_panel2002[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2002'] = 0
data_panel2002[which(is.na(data_panel2002[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2002'] = 0
data_panel2002[which(is.na(data_panel2002[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2002'] = 0
data_panel2002[which(is.na(data_panel2002[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2002'] = 0
data_panel2002[which(is.na(data_panel2002[,14])==TRUE),'CV_WKSWK_JOB_DLI.10_2002'] = 0
data_panel2002[which(is.na(data_panel2002[,15])==TRUE),'CV_WKSWK_JOB_DLI.11_2002'] = 0

data_panel2003[which(is.na(data_panel2003[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2003'] = 0
data_panel2003[which(is.na(data_panel2003[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2003'] = 0
data_panel2003[which(is.na(data_panel2003[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2003'] = 0
data_panel2003[which(is.na(data_panel2003[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2003'] = 0
data_panel2003[which(is.na(data_panel2003[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2003'] = 0
data_panel2003[which(is.na(data_panel2003[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2003'] = 0
data_panel2003[which(is.na(data_panel2003[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2003'] = 0
data_panel2003[which(is.na(data_panel2003[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2003'] = 0
data_panel2003[which(is.na(data_panel2003[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2003'] = 0
data_panel2003[which(is.na(data_panel2003[,14])==TRUE),'CV_WKSWK_JOB_DLI.10_2003'] = 0


data_panel2004[which(is.na(data_panel2004[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2004'] = 0
data_panel2004[which(is.na(data_panel2004[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2004'] = 0
data_panel2004[which(is.na(data_panel2004[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2004'] = 0
data_panel2004[which(is.na(data_panel2004[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2004'] = 0
data_panel2004[which(is.na(data_panel2004[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2004'] = 0
data_panel2004[which(is.na(data_panel2004[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2004'] = 0
data_panel2004[which(is.na(data_panel2004[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2004'] = 0

data_panel2005[which(is.na(data_panel2005[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2005'] = 0
data_panel2005[which(is.na(data_panel2005[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2005'] = 0
data_panel2005[which(is.na(data_panel2005[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2005'] = 0
data_panel2005[which(is.na(data_panel2005[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2005'] = 0
data_panel2005[which(is.na(data_panel2005[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2005'] = 0
data_panel2005[which(is.na(data_panel2005[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2005'] = 0
data_panel2005[which(is.na(data_panel2005[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2005'] = 0
data_panel2005[which(is.na(data_panel2005[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2005'] = 0
data_panel2005[which(is.na(data_panel2005[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2005'] = 0

data_panel2006[which(is.na(data_panel2006[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2006'] = 0
data_panel2006[which(is.na(data_panel2006[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2006'] = 0
data_panel2006[which(is.na(data_panel2006[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2006'] = 0
data_panel2006[which(is.na(data_panel2006[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2006'] = 0
data_panel2006[which(is.na(data_panel2006[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2006'] = 0
data_panel2006[which(is.na(data_panel2006[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2006'] = 0
data_panel2006[which(is.na(data_panel2006[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2006'] = 0
data_panel2006[which(is.na(data_panel2006[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2006'] = 0
data_panel2006[which(is.na(data_panel2006[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2006'] = 0
```

```
data_panel2007[which(is.na(data_panel2007[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2007'] = 0
data_panel2007[which(is.na(data_panel2007[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2007'] = 0
data_panel2007[which(is.na(data_panel2007[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2007'] = 0
data_panel2007[which(is.na(data_panel2007[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2007'] = 0
data_panel2007[which(is.na(data_panel2007[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2007'] = 0
data_panel2007[which(is.na(data_panel2007[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2007'] = 0
data_panel2007[which(is.na(data_panel2007[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2007'] = 0
data_panel2007[which(is.na(data_panel2007[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2007'] = 0

data_panel2008[which(is.na(data_panel2008[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2008'] = 0
data_panel2008[which(is.na(data_panel2008[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2008'] = 0
data_panel2008[which(is.na(data_panel2008[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2008'] = 0
data_panel2008[which(is.na(data_panel2008[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2008'] = 0
data_panel2008[which(is.na(data_panel2008[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2008'] = 0
data_panel2008[which(is.na(data_panel2008[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2008'] = 0
data_panel2008[which(is.na(data_panel2008[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2008'] = 0
data_panel2008[which(is.na(data_panel2008[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2008'] = 0

data_panel2009[which(is.na(data_panel2009[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2009'] = 0
data_panel2009[which(is.na(data_panel2009[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2009'] = 0
data_panel2009[which(is.na(data_panel2009[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2009'] = 0
data_panel2009[which(is.na(data_panel2009[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2009'] = 0
data_panel2009[which(is.na(data_panel2009[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2009'] = 0
data_panel2009[which(is.na(data_panel2009[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2009'] = 0
data_panel2009[which(is.na(data_panel2009[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2009'] = 0
data_panel2009[which(is.na(data_panel2009[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2009'] = 0
data_panel2009[which(is.na(data_panel2009[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2009'] = 0


data_panel2010[which(is.na(data_panel2010[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2010'] = 0
data_panel2010[which(is.na(data_panel2010[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2010'] = 0
data_panel2010[which(is.na(data_panel2010[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2010'] = 0
data_panel2010[which(is.na(data_panel2010[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2010'] = 0
data_panel2010[which(is.na(data_panel2010[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2010'] = 0
data_panel2010[which(is.na(data_panel2010[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2010'] = 0
data_panel2010[which(is.na(data_panel2010[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2010'] = 0
data_panel2010[which(is.na(data_panel2010[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2010'] = 0
data_panel2010[which(is.na(data_panel2010[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2010'] = 0


data_panel2011[which(is.na(data_panel2011[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,14])==TRUE),'CV_WKSWK_JOB_DLI.10_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,15])==TRUE),'CV_WKSWK_JOB_DLI.11_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,16])==TRUE),'CV_WKSWK_JOB_DLI.12_2011'] = 0
data_panel2011[which(is.na(data_panel2011[,17])==TRUE),'CV_WKSWK_JOB_DLI.13_2011'] = 0

data_panel2013[which(is.na(data_panel2013[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2013'] = 0
data_panel2013[which(is.na(data_panel2013[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2013'] = 0
data_panel2013[which(is.na(data_panel2013[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2013'] = 0
data_panel2013[which(is.na(data_panel2013[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2013'] = 0
data_panel2013[which(is.na(data_panel2013[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2013'] = 0
data_panel2013[which(is.na(data_panel2013[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2013'] = 0
data_panel2013[which(is.na(data_panel2013[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2013'] = 0
data_panel2013[which(is.na(data_panel2013[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2013'] = 0
data_panel2013[which(is.na(data_panel2013[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2013'] = 0
data_panel2013=data_panel2013[,-14]


data_panel2015[which(is.na(data_panel2015[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2015'] = 0
data_panel2015[which(is.na(data_panel2015[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2015'] = 0
data_panel2015[which(is.na(data_panel2015[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2015'] = 0
data_panel2015[which(is.na(data_panel2015[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2015'] = 0
data_panel2015[which(is.na(data_panel2015[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2015'] = 0
data_panel2015[which(is.na(data_panel2015[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2015'] = 0
data_panel2015[which(is.na(data_panel2015[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2015'] = 0
data_panel2015[which(is.na(data_panel2015[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2015'] = 0
data_panel2015[which(is.na(data_panel2015[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2015'] = 0
data_panel2015[which(is.na(data_panel2015[,14])==TRUE),'CV_WKSWK_JOB_DLI.10_2015'] = 0
data_panel2015[which(is.na(data_panel2015[,15])==TRUE),'CV_WKSWK_JOB_DLI.11_2015'] = 0


data_panel2017[which(is.na(data_panel2017[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,14])==TRUE),'CV_WKSWK_JOB_DLI.10_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,15])==TRUE),'CV_WKSWK_JOB_DLI.11_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,16])==TRUE),'CV_WKSWK_JOB_DLI.12_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,17])==TRUE),'CV_WKSWK_JOB_DLI.13_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,18])==TRUE),'CV_WKSWK_JOB_DLI.14_2017'] = 0
data_panel2017[which(is.na(data_panel2017[,19])==TRUE),'CV_WKSWK_JOB_DLI.15_2017'] = 0


data_panel2019[which(is.na(data_panel2019[,5])==TRUE),'CV_WKSWK_JOB_DLI.01_2019'] = 0
data_panel2019[which(is.na(data_panel2019[,6])==TRUE),'CV_WKSWK_JOB_DLI.02_2019'] = 0
data_panel2019[which(is.na(data_panel2019[,7])==TRUE),'CV_WKSWK_JOB_DLI.03_2019'] = 0
data_panel2019[which(is.na(data_panel2019[,8])==TRUE),'CV_WKSWK_JOB_DLI.04_2019'] = 0
data_panel2019[which(is.na(data_panel2019[,9])==TRUE),'CV_WKSWK_JOB_DLI.05_2019'] = 0
data_panel2019[which(is.na(data_panel2019[,10])==TRUE),'CV_WKSWK_JOB_DLI.06_2019'] = 0
data_panel2019[which(is.na(data_panel2019[,11])==TRUE),'CV_WKSWK_JOB_DLI.07_2019'] = 0
data_panel2019[which(is.na(data_panel2019[,12])==TRUE),'CV_WKSWK_JOB_DLI.08_2019'] = 0
data_panel2019[which(is.na(data_panel2019[,13])==TRUE),'CV_WKSWK_JOB_DLI.09_2019'] = 0
data_panel2019[which(is.na(data_panel2019[,14])==TRUE),'CV_WKSWK_JOB_DLI.10_2019'] = 0
data_panel2019[which(is.na(data_panel2019[,15])==TRUE),'CV_WKSWK_JOB_DLI.11_2019'] = 0
```

```r
#sum the working experience
data_panel1997$experience=data_panel1997[,5]+data_panel1997[,6]+data_panel1997[,7]+
  data_panel1997[,8]+data_panel1997[,9]+data_panel1997[,10]+data_panel1997[,11]

data_panel1998$experience=data_panel1998[,5]+data_panel1998[,6]+data_panel1998[,7]+
  data_panel1998[,8]+data_panel1998[,9]+data_panel1998[,10]+data_panel1998[,11]+data_panel1998[,12]+
  data_panel1998[,13]

data_panel1999$experience=data_panel1999[,5]+data_panel1999[,6]+data_panel1999[,7]+
  data_panel1999[,8]+data_panel1999[,9]+data_panel1999[,10]+data_panel1999[,11]+data_panel1999[,12]+
  data_panel1999[,13]

data_panel2000$experience=data_panel2000[,5]+data_panel2000[,6]+data_panel2000[,7]+
  data_panel2000[,8]+data_panel2000[,9]+data_panel2000[,10]+data_panel2000[,11]+data_panel2000[,12]+
  data_panel2000[,13]

data_panel2001$experience=data_panel2001[,5]+data_panel2001[,6]+data_panel2001[,7]+
  data_panel2001[,8]+data_panel2001[,9]+data_panel2001[,10]+data_panel2001[,11]+data_panel2001[,12]

data_panel2002$experience=data_panel2002[,5]+data_panel2002[,6]+data_panel2002[,7]+
  data_panel2002[,8]+data_panel2002[,9]+data_panel2002[,10]+data_panel2002[,11]+data_panel2002[,12]+
  data_panel2002[,13]+data_panel2002[,14]+data_panel2002[,15]

data_panel2003$experience=data_panel2003[,5]+data_panel2003[,6]+data_panel2003[,7]+
  data_panel2003[,8]+data_panel2003[,9]+data_panel2003[,10]+data_panel2003[,11]+data_panel2003[,12]+
  data_panel2003[,13]+data_panel2003[,14]

data_panel2004$experience=data_panel2004[,5]+data_panel2004[,6]+data_panel2004[,7]+
  data_panel2004[,8]+data_panel2004[,9]+data_panel2004[,10]+data_panel2004[,11]

data_panel2005$experience=data_panel2005[,5]+data_panel2005[,6]+data_panel2005[,7]+
  data_panel2005[,8]+data_panel2005[,9]+data_panel2005[,10]+data_panel2005[,11]+data_panel2005[,12]+
  data_panel2005[,13]

data_panel2006$experience=data_panel2006[,5]+data_panel2006[,6]+data_panel2006[,7]+
  data_panel2006[,8]+data_panel2006[,9]+data_panel2006[,10]+data_panel2006[,11]+data_panel2006[,12]+
  data_panel2006[,13]

data_panel2007$experience=data_panel2007[,5]+data_panel2007[,6]+data_panel2007[,7]+
  data_panel2007[,8]+data_panel2007[,9]+data_panel2007[,10]+data_panel2007[,11]+data_panel2007[,12]

data_panel2008$experience=data_panel2008[,5]+data_panel2008[,6]+data_panel2008[,7]+
  data_panel2008[,8]+data_panel2008[,9]+data_panel2008[,10]+data_panel2008[,11]+data_panel2008[,12]

data_panel2009$experience=data_panel2009[,5]+data_panel2009[,6]+data_panel2009[,7]+
  data_panel2009[,8]+data_panel2009[,9]+data_panel2009[,10]+data_panel2009[,11]+data_panel2009[,12]+
  data_panel2009[,13]

data_panel2010$experience=data_panel2010[,5]+data_panel2010[,6]+data_panel2010[,7]+
  data_panel2010[,8]+data_panel2010[,9]+data_panel2010[,10]+data_panel2010[,11]+data_panel2010[,12]+
  data_panel2010[,13]

data_panel2011$experience=data_panel2011[,5]+data_panel2011[,6]+data_panel2011[,7]+
  data_panel2011[,8]+data_panel2011[,9]+data_panel2011[,10]+data_panel2011[,11]+data_panel2011[,12]+
  data_panel2011[,13]+data_panel2011[,14]+data_panel2011[,15]+data_panel2011[,16]+data_panel2011[,17]


data_panel2013$experience=data_panel2013[,5]+data_panel2013[,6]+data_panel2013[,7]+
  data_panel2013[,8]+data_panel2013[,9]+data_panel2013[,10]+data_panel2013[,11]+data_panel2013[,12]+
  data_panel2013[,13]

data_panel2015$experience=data_panel2015[,5]+data_panel2015[,6]+data_panel2015[,7]+
  data_panel2015[,8]+data_panel2015[,9]+data_panel2015[,10]+data_panel2015[,11]+data_panel2015[,12]+
  data_panel2015[,13]+data_panel2015[,14]+data_panel2015[,15]


data_panel2017$experience=data_panel2017$experience=data_panel2017[,5]+data_panel2017[,6]+data_panel2017[,7]+
  data_panel2017[,8]+data_panel2017[,9]+data_panel2017[,10]+data_panel2017[,11]+data_panel2017[,12]+
  data_panel2017[,13]+data_panel2017[,14]+data_panel2017[,15]+data_panel2017[,16]+data_panel2017[,17]+data_pa

data_panel2019$experience=data_panel2019[,5]+data_panel2019[,6]+data_panel2019[,7]+
  data_panel2019[,8]+data_panel2019[,9]+data_panel2019[,10]+data_panel2019[,11]+data_panel2019[,12]+
  data_panel2019[,13]+data_panel2019[,14]+data_panel2019[,15]
```

Fourthly, keep all the datset with year,income,education,married status,weeks of education.

```r
#keep all the datset with year,income,education,married status,weeks of education
data_panel1997=data_panel1997[,c(1,12,2,3,4,13)]
data_panel1998=data_panel1998[,c(1,14,2,3,4,15)]
data_panel1999=data_panel1999[,c(1,14,2,3,4,15)]
data_panel2000=data_panel2000[,c(1,14,2,3,4,15)]
data_panel2001=data_panel2001[,c(1,13,2,3,4,14)]
data_panel2002=data_panel2002[,c(1,16,2,3,4,17)]
data_panel2003=data_panel2003[,c(1,15,2,3,4,16)]
data_panel2004=data_panel2004[,c(1,12,2,3,4,13)]
data_panel2005=data_panel2005[,c(1,14,2,3,4,15)]
data_panel2006=data_panel2006[,c(1,14,2,3,4,15)]
data_panel2007=data_panel2007[,c(1,13,2,3,4,14)]
data_panel2008=data_panel2008[,c(1,13,2,3,4,14)]
data_panel2009=data_panel2009[,c(1,14,2,3,4,15)]
data_panel2010=data_panel2010[,c(1,14,2,3,4,15)]
data_panel2011=data_panel2011[,c(1,18,2,3,4,19)]
data_panel2013=data_panel2013[,c(1,14,2,3,4,15)]
data_panel2015=data_panel2015[,c(1,17,2,3,4,18)]
data_panel2017=data_panel2017[,c(1,20,2,3,4,21)]
data_panel2019=data_panel2019[,c(1,16,2,3,4,17)]
```

Fifthly, combine the 1997-2019 into one dataset

```r
#combine the 1997-2019 into one dataset
colnames(data_panel1997)=c("num","year","income","education","marital_status","experience")
colnames(data_panel1998)=c("num","year","income","education","marital_status","experience")
colnames(data_panel1999)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2000)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2001)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2002)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2003)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2004)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2005)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2006)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2007)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2008)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2009)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2010)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2011)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2013)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2015)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2017)=c("num","year","income","education","marital_status","experience")
colnames(data_panel2019)=c("num","year","income","education","marital_status","experience")
```

```r
panel=rbind.data.frame(data_panel1997,data_panel1998,data_panel1999,data_panel2000,data_panel2001,data_panel2
                       ,data_panel2003,data_panel2004,data_panel2005
                       ,data_panel2006,data_panel2007,data_panel2008
                       ,data_panel2009,data_panel2010,data_panel2011
                       ,data_panel2013,data_panel2015,data_panel2017,data_panel2019)
```

```r
#transfer the degree to year of education
#1997:NA
panel[which(panel$education==0),'education'] = 0
panel[which(panel$education==1),'education'] = 4
panel[which(panel$education==2),'education'] = 12
panel[which(panel$education==3),'education'] = 14
panel[which(panel$education==4),'education'] = 16
panel[which(panel$education==5),'education'] = 18
panel[which(panel$education==6),'education'] = 23
panel[which(panel$education==7),'education'] = 22
```

Since marital status is not a continuous variable, I change all the 5 status into 5 dummy variables, which represent whether the individual is in this kind of marital status.

```
#drop NA
panel_nonna=na.omit(panel)
#transfer the marital_status to five dummy variables,
#marital_status1=nevermarried
#marital_status2=married
#marital_status3=separated
#marital_status4=divorced
#marital_status5=widowed

panel_nonna[which(panel_nonna$marital_status==0),' marital_status1'] = 1
panel_nonna[which(panel_nonna$marital_status!=0),' marital_status1'] = 0
panel_nonna[which(panel_nonna$marital_status==1),'  marital_status2'] = 1
panel_nonna[which(panel_nonna$marital_status!=1),'  marital_status2'] = 0
panel_nonna[which(panel_nonna$marital_status==2),'  marital_status3'] = 1
panel_nonna[which(panel_nonna$marital_status!=2),'  marital_status3'] = 0
panel_nonna[which(panel_nonna$marital_status==3),'   marital_status4'] = 1
panel_nonna[which(panel_nonna$marital_status!=3),'   marital_status4'] = 0
panel_nonna[which(panel_nonna$marital_status==4),'   marital_status5'] = 1
panel_nonna[which(panel_nonna$marital_status!=4),'   marital_status5'] = 0

panel_nonna=arrange(panel_nonna,num)
```

Finally, the panel data is as follows.

	num	year	income	education	marital_status	experience	marital_status1	marital_status2	marital_stat
1	1	1998	475	0	0	72	1	0	
2	1	2000	8000	12	0	91	1	0	
3	1	2001	7000	12	0	221	1	0	
4	1	2002	8000	12	0	77	1	0	
5	1	2003	15000	16	0	65	1	0	
6	1	2005	10000	16	0	172	1	0	
7	1	2006	80471	16	0	221	1	0	
8	1	2007	112215	16	0	278	1	0	
9	1	2008	42500	16	0	100	1	0	
10	1	2009	43500	16	0	150	1	0	

The first is within estimator.


```
#within estimator
library(dplyr)
panel_nonna.centered = panel_nonna %>%
  group_by(num) %>%
  mutate_at(.vars = vars(income,education, ` marital_status1` ,`  marital_status2`,
                         `  marital_status3`,`   marital_status4`,`   marital_status5`, experience),
            .funs = funs('dm' = . - mean(.)))

within = lm(formula=income~education_dm+`  marital_status2_dm`+`  marital_status3_dm`+`   marital_status4_dm`
            +`   marital_status5_dm`+experience_dm,
             data = panel_nonna.centered)

summary(within)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.642e+04  9.014e+01 293.146  < 2e-16 ***
education_dm             1.300e+03  2.715e+01  47.887  < 2e-16 ***
`  marital_status2_dm`   1.941e+04  3.055e+02  63.526  < 2e-16 ***
`  marital_status3_dm`   1.547e+04  1.108e+03  13.956  < 2e-16 ***
`   marital_status4_dm`  1.992e+04  6.032e+02  33.031  < 2e-16 ***
`   marital_status5_dm`  9.802e+03  3.510e+03   2.793  0.00523 **
experience_dm            5.017e+01  6.646e-01  75.496  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#between estimator

```
#between estimator
panel_nonna.mean = panel_nonna %>%
  group_by(num) %>%
  mutate_at(.vars = vars(income,education, ` marital_status1` ,`  marital_status2`,
                         `  marital_status3`,`   marital_status4`,`   marital_status5`, experience),
            .funs = funs('dm' = mean(.)))

between = lm(formula=income~education_dm+`  marital_status2_dm`+`  marital_status3_dm`+`   marital_status4_dm`
             +`   marital_status5_dm`+experience_dm,
             data = panel_nonna.mean)
summary(between)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.403e+03  3.729e+02  11.807  < 2e-16 ***
education_dm             1.167e+03  2.785e+01  41.910  < 2e-16 ***
`  marital_status2_dm`   8.666e+03  3.435e+02  25.224  < 2e-16 ***
`  marital_status3_dm`  -4.954e+03  2.211e+03  -2.240   0.0251 *
`   marital_status4_dm` -9.641e+02  7.930e+02  -1.216   0.2241
`   marital_status5_dm` -2.114e+04  4.665e+03  -4.531 5.89e-06 ***
experience_dm            3.233e+01  9.088e-01  35.576  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#Difference (any) Estimator

```
panel_nonna.Difference <- panel_nonna %>%
  group_by(num) %>%
  mutate(across(c(income,education, ` marital_status1` ,`  marital_status2`,
                  `  marital_status3`,`   marital_status4`,`   marital_status5`, experience),
                ~.x-dplyr::lag(.x), .names = "{col}_dm")) %>%
  ungroup()


Difference = lm(formula=income_dm~education_dm+`  marital_status2_dm`+`  marital_status3_dm`+`   marital_status4_dm`
                +`   marital_status5_dm`+experience_dm,
                data = panel_nonna.Difference)
summary(Difference)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3942.8311    69.3273  56.873  < 2e-16 ***
education_dm               84.5845    20.5753   4.111 3.94e-05 ***
`  marital_status2_dm`   3817.5444   269.6951  14.155  < 2e-16 ***
`  marital_status3_dm`   2972.6733   650.8680   4.567 4.95e-06 ***
`   marital_status4_dm`  5111.8820   494.6723  10.334  < 2e-16 ***
`   marital_status5_dm`  -805.5087  2452.2549  -0.328    0.743
experience_dm              18.2306     0.5694  32.018  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpreting:

Within Estimator

All else is equal, an extra year of education adds $1,300 to your income. An extra week of experience adds $50.17 to your earnings. Getting married increases income by $19,400 compared to not getting married.

Between Estimator.

All else is equal, an additional year of education increases income by $1,167; An extra week of experience adds $32 to your income. Getting married increases income by $8,666 compared to not getting married.

Difference (any) Estimator

All else is equal, an extra year of schooling adds $84 to your income. An extra week of experience adds $18 to your income. Getting married increases your income by $3,817 compared to not getting married.

Comparison:

In Within Estimator model, it uses the fixed effect, so that it can get more consistent result. The Between Estimator is in general biased, since it is actually a cross section regression, thus may have the same problem with OLS.

The Difference Estimator model can get the consistent result because it wipes out the time invariant omitted variables by the difference.