Homework 4: Censoring and Panel Data

Yuzhe Wang

2022/4/20

Exercise 1 Preparing the Data

1.1.1 age

```
dat_nlsy97 = fread('./data/dat_A4.csv')
dat_nlsy97$age = 2019 - dat_nlsy97$KEY_BDATE_Y_1997
```

1.1.2work_exp

```
job_week = dat_nlsy97[,paste('CV_WKSWK_JOB_DLI.',
                        c(paste(0,1:9,sep=''),10,11),
                        '_2019',
                        sep='')]
job_week[is.na(job_week)] = 0
dat_nlsy97$work_exp = apply(job_week,1,sum)*7/365
```

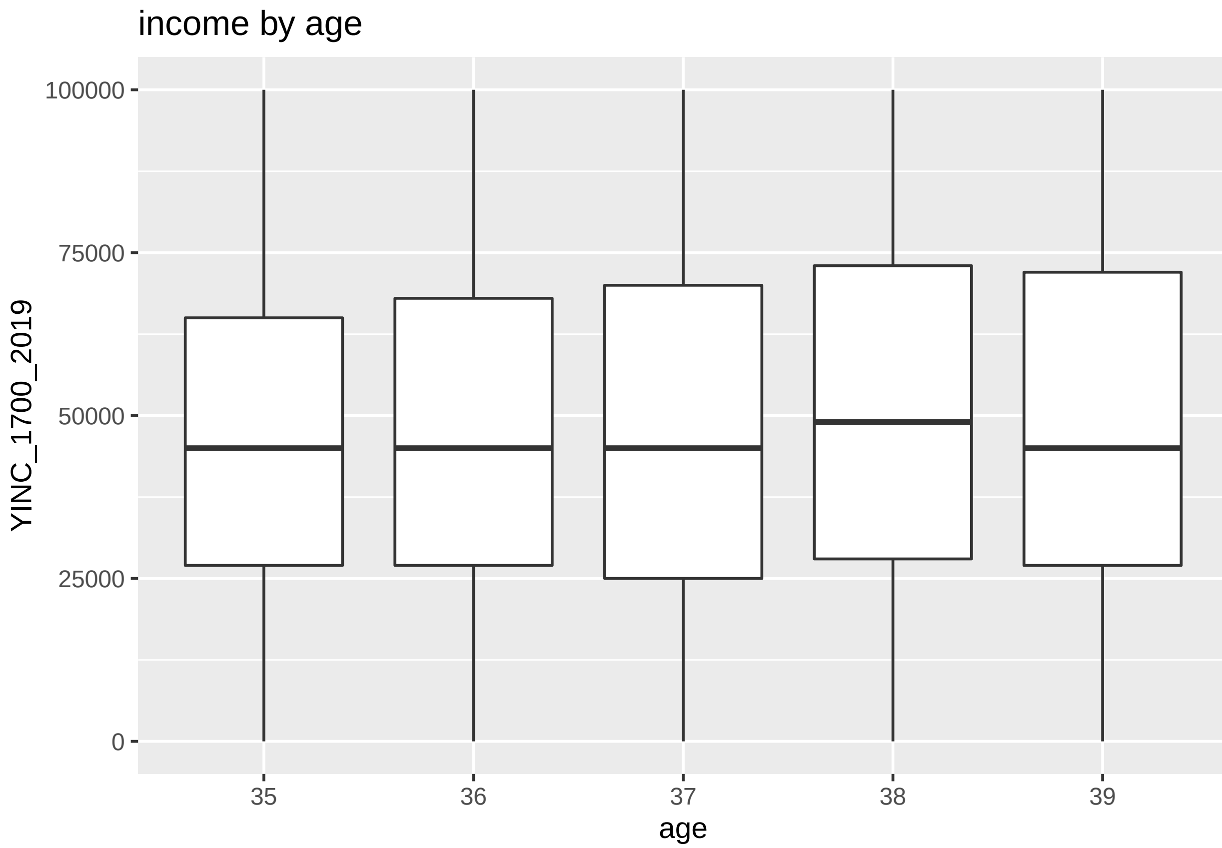
1.2 years of schooling

```
# drop na
dat_nlsy97 = dat_nlsy97[!is.na(dat_nlsy97$YSCH.3113_2019)]

# Assumptions:
# GED == 4 years of schooling
# phd == 23 years of schooling
# Professional degree == 21 years of schooling
fun_ex1_2 = function(x){
  if(x == 1){y = 0}
  if(x == 2){y = 4}
  if(x == 3){y = 12}
  if(x == 4){y = 15}
  if(x == 5){y = 16}
  if(x == 6){y = 18}
  if(x == 7){y = 23}
  if(x == 8){y = 21}
  return(y)
```

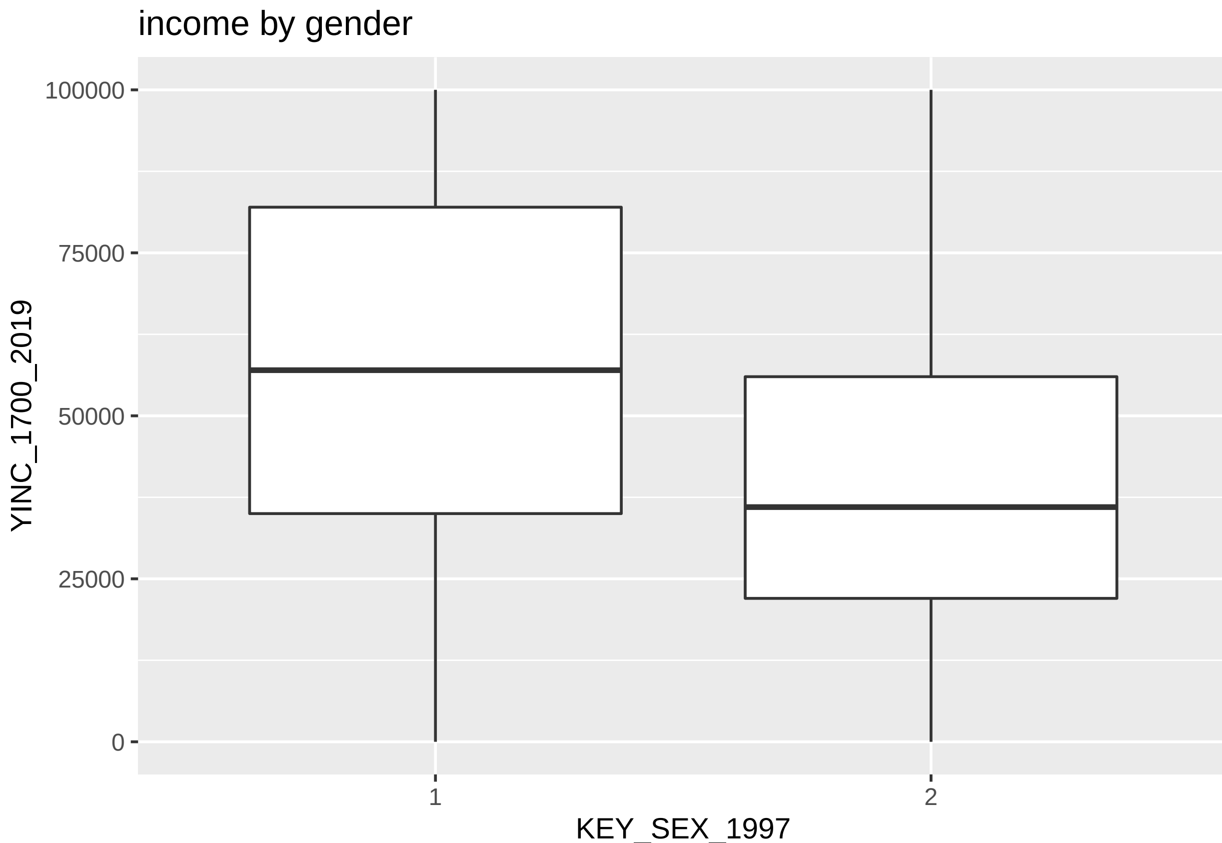
```
}
dat_nlsy97$edu = sapply(dat_nlsy97$YSCH.3113_2019,fun_ex1_2)
```

1.3 visualizations

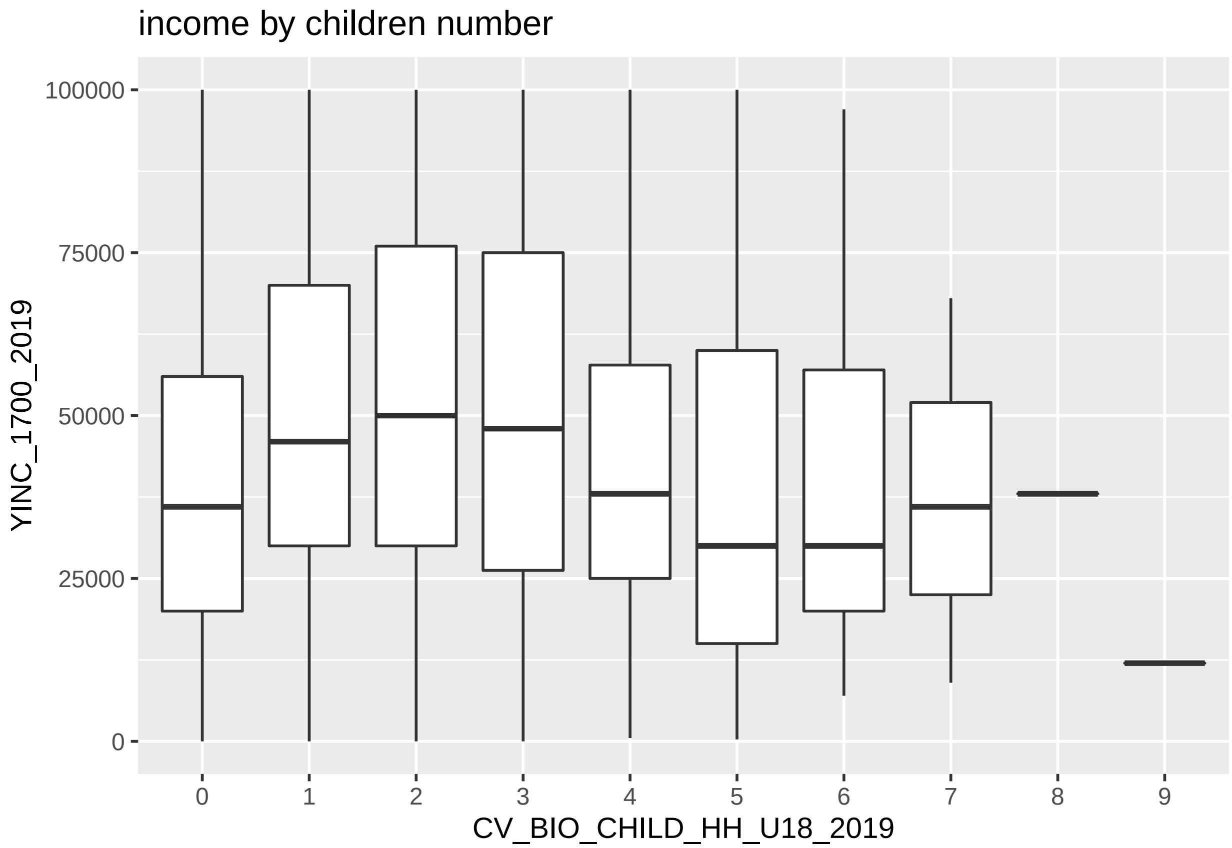
```
## 1.3.1 Plot the income data (where income is positive) by i) age groups, ii) gender groups and iii) n
# drop na
dat_nlsy97 = dat_nlsy97[!is.na(dat_nlsy97$YINC_1700_2019)]
dat_nlsy97 = dat_nlsy97[!is.na(dat_nlsy97$CV_BIO_CHILD_HH_U18_2019)]

dat_nlsy97$age = as.factor(dat_nlsy97$age)
dat_nlsy97$KEY_SEX_1997 = as.factor(dat_nlsy97$KEY_SEX_1997)
dat_nlsy97$CV_BIO_CHILD_HH_U18_2019 = as.factor(dat_nlsy97$CV_BIO_CHILD_HH_U18_2019)

p1 = ggplot(dat_nlsy97,aes(x=age,y=YINC_1700_2019))+
  geom_boxplot()+
  ggtitle(label='income by age')

p2 = ggplot(dat_nlsy97,aes(x=KEY_SEX_1997,y=YINC_1700_2019))+
  geom_boxplot()+
  ggtitle(label='income by gender')

p3 = ggplot(dat_nlsy97,aes(x=CV_BIO_CHILD_HH_U18_2019,y=YINC_1700_2019))+
  geom_boxplot()+
  ggtitle(label='income by children number')
p1
```


income by age
100000
75000
50000
25000
0
YINC_1700_2019
35
36
37
38
39
age

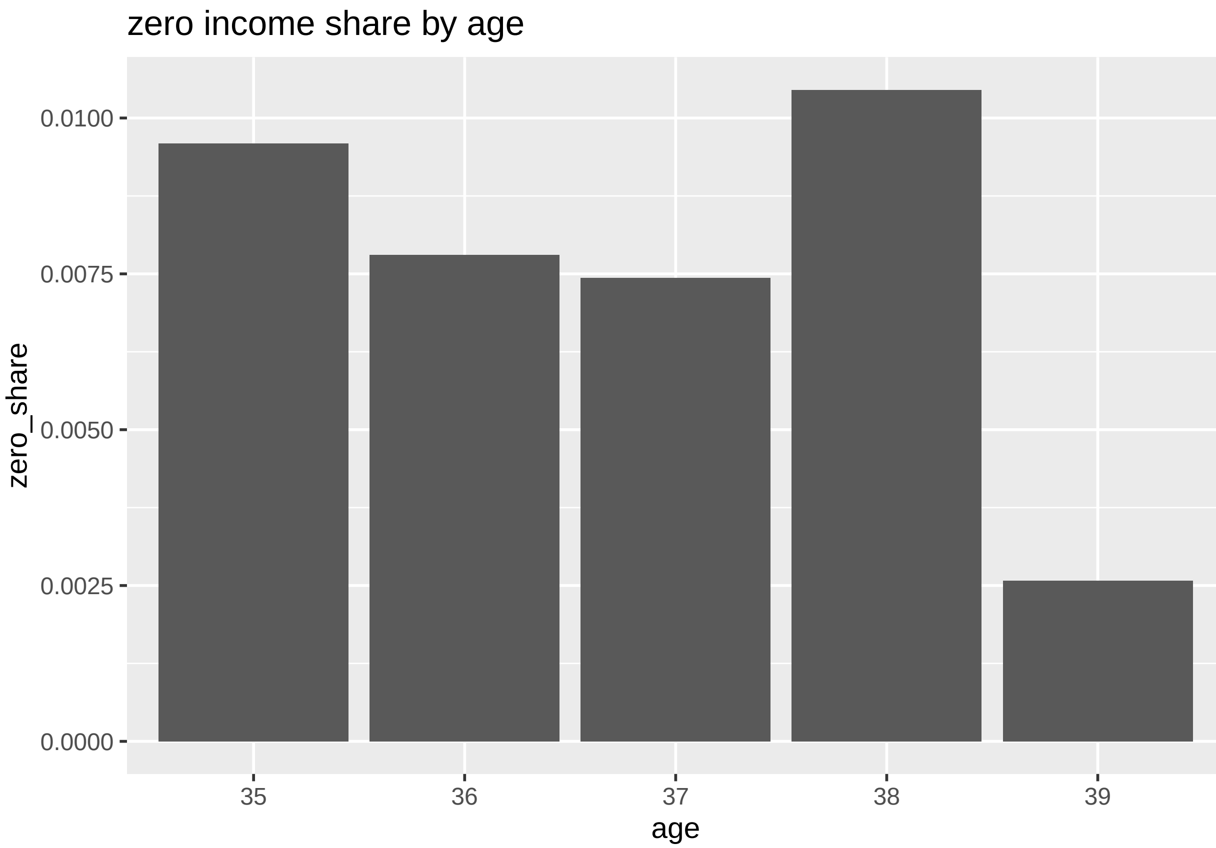
```
p2
```


income by gender
100000
75000
50000
25000
0
YINC_1700_2019
1
2
KEY_SEX_1997

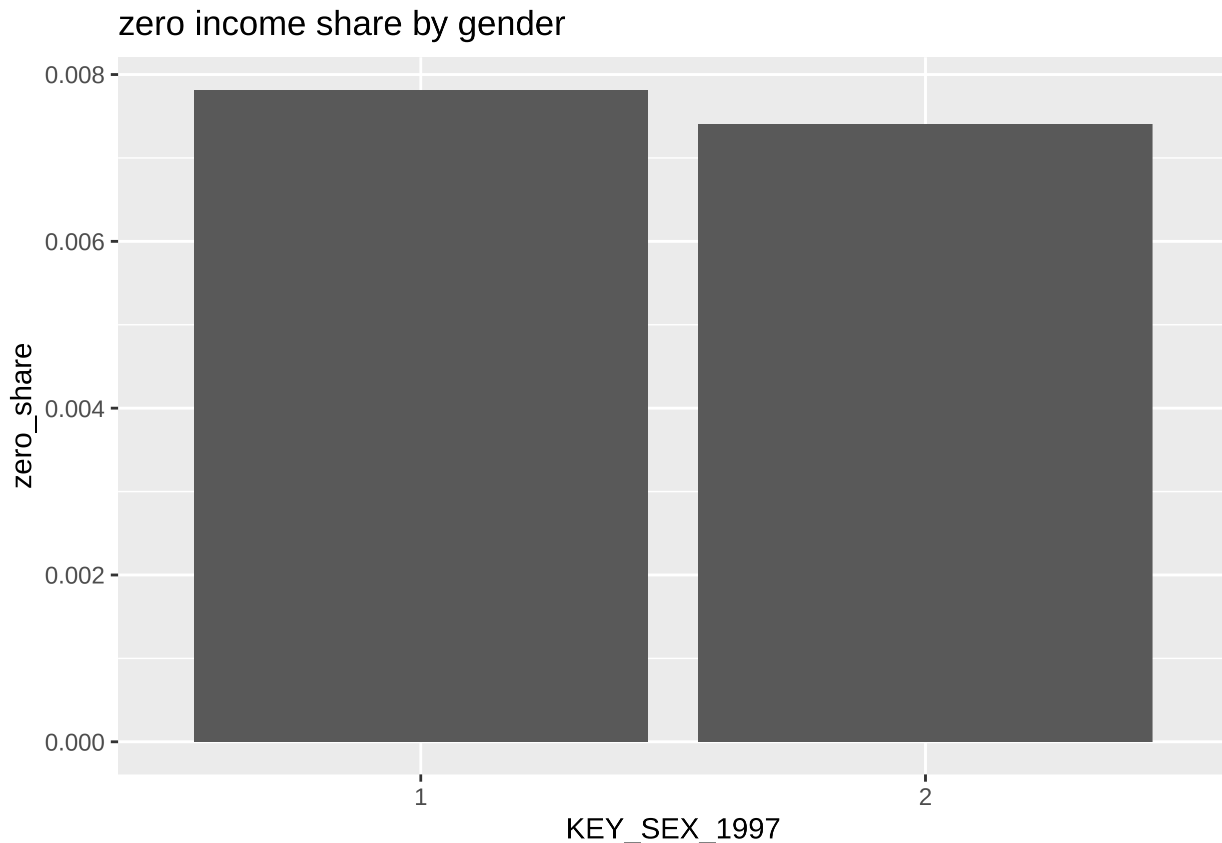
```
p3
```



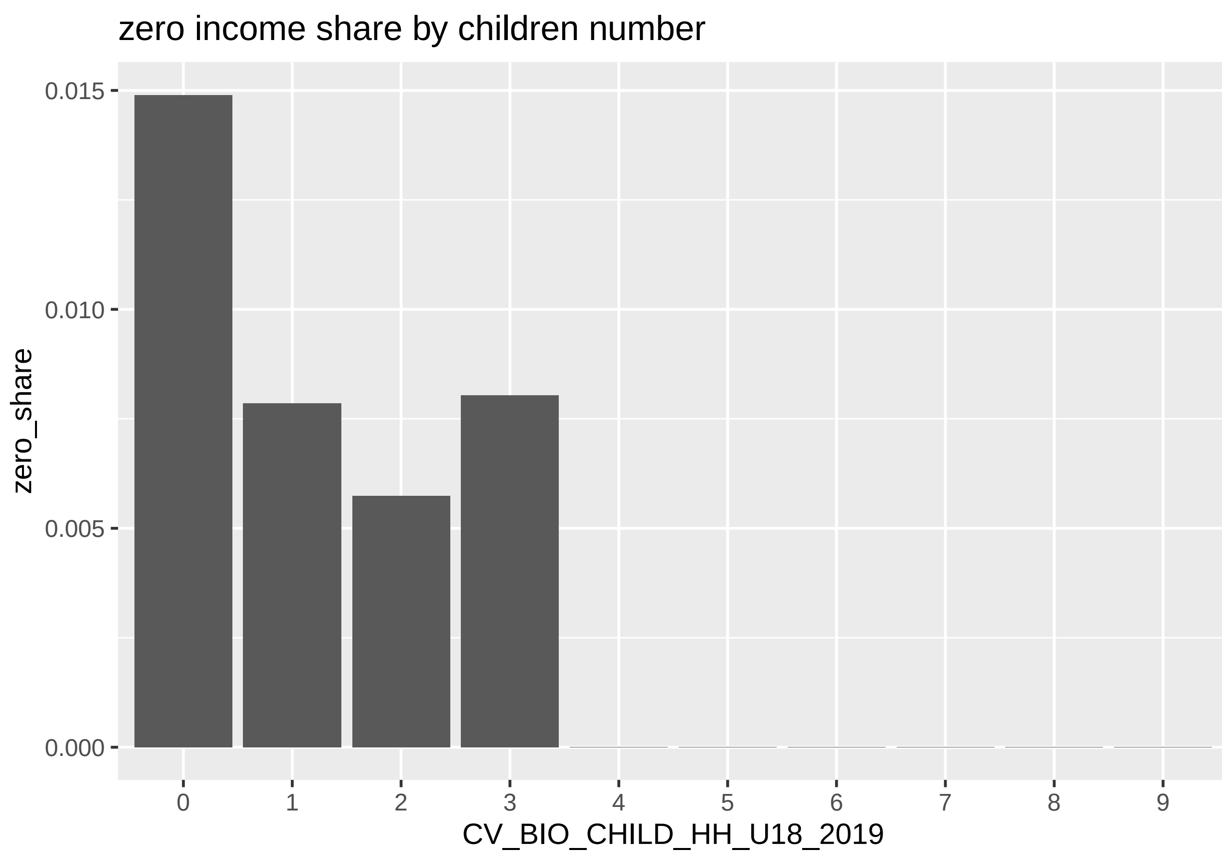

```
## 1.3.2 - Table the share of "0" in the income data by i) age groups, ii) gender groups, iii) number o
fun_ex1_3 = function(x){
  return(sum(x == 0))
}
p4 = summarise(group_by(dat_nlsy97, age),
          zero_share = fun_ex1_3(YINC_1700_2019)/length(YINC_1700_2019))%>%
  ggplot(aes(x=age,y=zero_share))+
  geom_bar(stat = 'identity')+
  ggtitle(label='zero income share by age')

p5 = summarise(group_by(dat_nlsy97, KEY_SEX_1997),
               zero_share = fun_ex1_3(YINC_1700_2019)/length(YINC_1700_2019))%>%
  ggplot(aes(x=KEY_SEX_1997,y=zero_share))+
  geom_bar(stat = 'identity')+
  ggtitle(label='zero income share by gender')

p6 = summarise(group_by(dat_nlsy97, CV_BIO_CHILD_HH_U18_2019),
               zero_share = fun_ex1_3(YINC_1700_2019)/length(YINC_1700_2019))%>%
  ggplot(aes(x=CV_BIO_CHILD_HH_U18_2019,y=zero_share))+
  geom_bar(stat = 'identity')+
  ggtitle(label='zero income share by children number')
p4
```


zero income share by age
0.0100
0.0075
0.0050
0.0025
0.0000
zero_share
35
36
37
38
39
age

```
p5
```


zero income share by gender
0.008
0.006
0.004
0.002
0.000
zero_share
1
2
KEY_SEX_1997

```
p6
```



Exercise 2 Heckman Selection Model

```
## 2.1 Specify and estimate an OLS model to explain the income variable (where income is positive).
dat_nlsy97$age = as.numeric(as.character(dat_nlsy97$age))
dat_nlsy97$CV_BIO_CHILD_HH_U18_2019 = as.numeric(as.character(dat_nlsy97$CV_BIO_CHILD_HH_U18_2019))
dat_nlsy97$KEY_SEX_1997 = as.numeric(as.character(dat_nlsy97$KEY_SEX_1997))
dat_nlsy97$KEY_SEX_1997 = dat_nlsy97$KEY_SEX_1997-1

dat_nlsy97_positive_income = dat_nlsy97[dat_nlsy97$YINC_1700_2019>0]
ols_reg = lm(YINC_1700_2019~age+work_exp+edu+KEY_SEX_1997+CV_BIO_CHILD_HH_U18_2019,data=dat_nlsy97_posi
summary(ols_reg)
```

```
##
## Call:
## lm(formula = YINC_1700_2019 ~ age + work_exp + edu + KEY_SEX_1997 +
##     CV_BIO_CHILD_HH_U18_2019, data = dat_nlsy97_positive_income)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -85300 -16944  -2355  16791  91700
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)               7096.81   10339.36   0.686  0.49251
## age                        453.59     277.45   1.635  0.10216
## work_exp                  1042.93      73.99  14.096  < 2e-16 ***
## edu                       2119.08      77.74  27.258  < 2e-16 ***
## KEY_SEX_1997            -19836.18     785.59 -25.250  < 2e-16 ***
## CV_BIO_CHILD_HH_U18_2019   1136.23     342.92   3.313  0.00093 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24070 on 3908 degrees of freedom
## Multiple R-squared:  0.2957, Adjusted R-squared:  0.2948
## F-statistic: 328.2 on 5 and 3908 DF,  p-value: < 2.2e-16
```

Interpret the estimation results:

When work experience increases 1 year, the income increases 1042 dollars.When years of schooling increases 1 year, the income increases 2119 dollars.If female, the income is 19836 dollars lower than male.When there is one more children, the income increases 1136 dollars.Age has insignificant effect to income.Explain why there might be a selection problem when estimating an OLS this way:There are some people in the sample have zero income. However, when estimate the effect of independent to the income.We actually presume that he/she has non-zero income.The people in the sample we selected mostly have non-zero income.Therefore, we overlook the effect to the zero income samples.

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

```
# Because Heckman model helps to estimate the potential income of those with zero income.
```

2.3 Estimate a Heckman selection model

```
## step1:probit
dat_nlsy97$income_positive = as.numeric(dat_nlsy97$YINC_1700_2019>0)
flike = function(beta,age,work_exp,edu,female,ch_num,y_real,y)
{
  x_beta = beta[1] + beta[2]*age +
    beta[3]*work_exp+beta[4]*edu +
    beta[5]*female+beta[6]*ch_num
  s = sd(y_real - x_beta)
  pr = pnorm(x_beta/s)
  pr[pr>0.999999] = 0.999999
  pr[pr<0.000001] = 0.000001
  likelihood = y*log(pr) + (1-y)*log(1-pr)
  return(-sum(likelihood))
}

set.seed(100)

ntry = 50
out = mat.or.vec(ntry,7)
for (i in 1:ntry){
```

```r
  start = c(runif(1,-15000,15000),runif(1,-1000,1000),runif(1,-10000,10000),
            runif(1,-10000,10000),runif(1,-30000,-10000),runif(1,-10000,10000))
  capture.output(res <- optim(start,
                              fn=flike,
                              method="BFGS",
                              control=list(trace=6,maxit=2000),
                              age=dat_nlsy97$age,
                              work_exp=dat_nlsy97$work_exp,
                              edu=dat_nlsy97$edu,
                              female=dat_nlsy97$KEY_SEX_1997,
                              ch_num=dat_nlsy97$CV_BIO_CHILD_HH_U18_2019,
                              y_real=dat_nlsy97$YINC_1700_2019,
                              y=dat_nlsy97$income_positive))
  out[i,c(1:6)] = res$par
  out[i,7] = res$value
}
out = data.frame(out)
par = out[which(out$X7==min(out$X7)),1:6]

step1_fits = apply(dat_nlsy97[,c('age','work_exp','edu','KEY_SEX_1997','CV_BIO_CHILD_HH_U18_2019')],
        1, function(x) return(sum(x * as.numeric(par[2:6]))+as.numeric(par[1])))
sig = sd(dat_nlsy97$YINC_1700_2019-step1_fits)
imr = dnorm(step1_fits/sig)/pnorm(step1_fits/sig)

## step2:OLS estimation
X = as.matrix(cbind(rep(1,length(imr)),dat_nlsy97[,c('age','work_exp','edu','KEY_SEX_1997','CV_BIO_CHIL
Y = dat_nlsy97$YINC_1700_2019
step2_beta = solve(t(X)%*%X)%*%t(X)%*%Y
step2_ols_check = lm(dat_nlsy97$YINC_1700_2019~
                   dat_nlsy97$age+
                   dat_nlsy97$work_exp+
                   dat_nlsy97$edu+
                   dat_nlsy97$KEY_SEX_1997+
                   dat_nlsy97$CV_BIO_CHILD_HH_U18_2019+
                   imr)
summary(step2_ols_check)
```

```
##
## Call:
## lm(formula = dat_nlsy97$YINC_1700_2019 ~ dat_nlsy97$age + dat_nlsy97$work_exp +
##     dat_nlsy97$edu + dat_nlsy97$KEY_SEX_1997 + dat_nlsy97$CV_BIO_CHILD_HH_U18_2019 +
##     imr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84039 -16917  -2156  16978  92529
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          14673.60   13995.49   1.048    0.294
## dat_nlsy97$age                         326.17     326.02   1.000    0.317
## dat_nlsy97$work_exp                    978.04     103.36   9.462   <2e-16 ***
## dat_nlsy97$edu                        2087.61      81.27  25.688   <2e-16 ***
```

```
## dat_nlsy97$KEY_SEX_1997             -18596.93    1302.61 -14.277   <2e-16 ***
## dat_nlsy97$CV_BIO_CHILD_HH_U18_2019    495.49     762.64   0.650    0.516
## imr                                 -89600.41   81360.06  -1.101    0.271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24340 on 3937 degrees of freedom
## Multiple R-squared:  0.2906, Adjusted R-squared:  0.2895
## F-statistic: 268.7 on 6 and 3937 DF,  p-value: < 2.2e-16
```

```
step2_beta # my results
```

```
##                                [,1]
## V1                       14673.6037
## age                        326.1686
## work_exp                   978.0393
## edu                       2087.6097
## KEY_SEX_1997            -18596.9349
## CV_BIO_CHILD_HH_U18_2019   495.4852
## imr                     -89600.4145
```

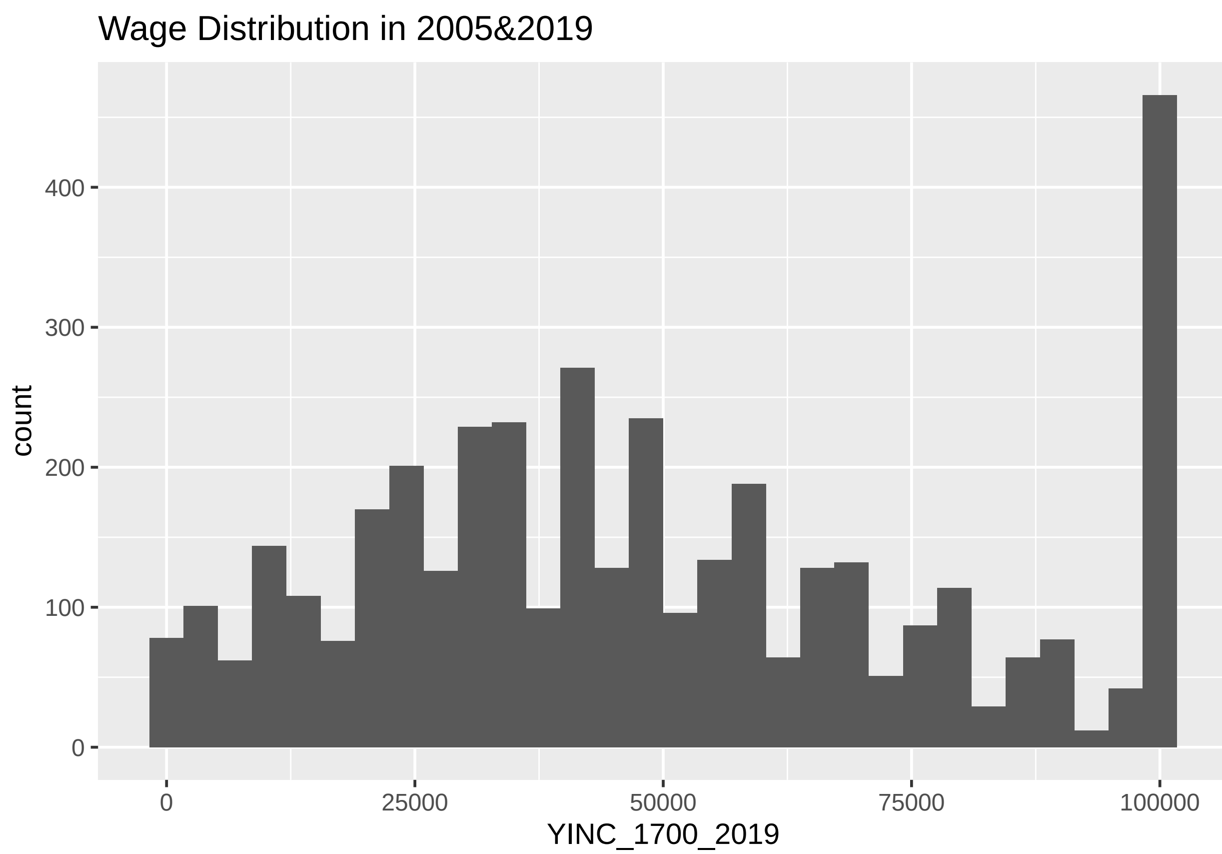
Interpret the results from the Heckman selection model and compare the results to OLS results.

Age now has smaller effect on income, which seems to be more reasonable.The effect of work experience is slightly smaller.The effects of education is slightly smaller.The effects of gender is slightly smaller.The effect of children number becomes insignificant while it is significantly positive in ols.The difference mainly comes from the problem of sample selection. When we include the samples with zero income, the slopes are overestimated.However, with Heckman method, we calculate the potential income for those with zero income, which lower the slope.

Exercise 3 Censoring

```
## 3.1 Plot a histogram to check whether the distribution of the income variable.
p7 = ggplot(data=dat_nlsy97,aes(YINC_1700_2019))+
  geom_histogram()+
  ggtitle(label='Wage Distribution in 2005&2019')
p7
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



3.2 Propose a model to deal with the censoring problem

```
# Censored regression model: Tobit model.
```

3.3 Estimate the appropriate model with the censored data

```
flike2 = function(beta,age,work_exp,edu,female,ch_num,y_real,y,i_a,i_b)
{
  a=0
  b=100000
  x_beta = beta[1] + beta[2]*age +
    beta[3]*work_exp+beta[4]*edu+
    beta[5]*female+beta[6]*ch_num

  s = sd(y_real - x_beta)
  pr_a = pnorm((a - x_beta)/s)
  pr_b = pnorm((x_beta - b)/s)

  pr_a[pr_a>0.999999] = 0.999999
  pr_a[pr_a<0.000001] = 0.000001
  pr_b[pr_b>0.999999] = 0.999999
  pr_b[pr_b<0.000001] = 0.000001
```

```
  likelihood = i_a*log(pr_a) + i_b*log(pr_b) +
    (1 - i_a - i_b) * (log(dnorm((y_real - x_beta)/s)) - log(s))
  return(-sum(likelihood))
}

dat_nlsy97$i_a = as.numeric(dat_nlsy97$YINC_1700_2019 == 0)
dat_nlsy97$i_b = as.numeric(dat_nlsy97$YINC_1700_2019 == 100000)

ntry = 50
out = mat.or.vec(ntry,7)
for (i in 1:ntry){
  start = c(runif(1,-15000,15000),runif(1,-1000,1000),runif(1,-10000,10000),
            runif(1,-10000,10000),runif(1,-30000,-10000),runif(1,-10000,10000))
  capture.output(res <- optim(start,
                              fn=flike2,
                              method="BFGS",
                              control=list(trace=6,maxit=2000),
                              age=dat_nlsy97$age,
                              work_exp=dat_nlsy97$work_exp,
                              edu=dat_nlsy97$edu,
                              female=dat_nlsy97$KEY_SEX_1997,
                              ch_num=dat_nlsy97$CV_BIO_CHILD_HH_U18_2019,
                              y_real=dat_nlsy97$YINC_1700_2019,
                              y=dat_nlsy97$income_positive,
                              i_a=dat_nlsy97$i_a,
                              i_b=dat_nlsy97$i_b))
  out[i,c(1:6)] = res$par
  out[i,7] = res$value
}
out = data.frame(out)
par_tobit = out[which(out$X7==min(out$X7)),1:6]
ols_coe = ols_reg$coefficients
heckman_coe = step2_beta[1:6]
tobit_coe = as.numeric(par_tobit[1:6])
reslts_compare = data.frame(ols= ols_coe, heckman=heckman_coe,tobit=tobit_coe)
reslts_compare
```

```
##                                 ols     heckman      tobit
## (Intercept)               7096.8144  14673.6037   2387.099
## age                        453.5915    326.1686    517.580
## work_exp                  1042.9302    978.0393   1163.599
## edu                       2119.0795   2087.6097   2293.738
## KEY_SEX_1997            -19836.1840 -18596.9349 -19216.583
## CV_BIO_CHILD_HH_U18_2019  1136.2335    495.4852   1049.561
```

As we can observe, heckman model gives smaller effect of independent variables than ols model, because we estimate the potential income of those with zero income.The tobit model gives larger effect than ols and heckman, because we estimate the potential income of those with income = 100000.People with 100000 actually have income larger than 100000. Therefore, if we still treat them as 100000, the efect will be underestimated.With Tobit model, we can see that the effect becomes larger becuase we consider the income > 100000.take the effect of working experience as an example:

```
working_experience = rep(dat_nlsy97$work_exp,3)
model = c(rep('OLS',length(dat_nlsy97$V1)),
          rep('Heckman',length(dat_nlsy97$V1)),
          rep('Tobit',length(dat_nlsy97$V1)))
x = cbind(rep(mean(dat_nlsy97$age),length(dat_nlsy97$V1)),
      dat_nlsy97$work_exp,
      rep(mean(dat_nlsy97$edu),length(dat_nlsy97$V1)),
      rep(mean(dat_nlsy97$KEY_SEX_1997),length(dat_nlsy97$V1)),
      rep(mean(dat_nlsy97$CV_BIO_CHILD_HH_U18_2019),length(dat_nlsy97$V1)))
income = c(apply(x,1, function(x) return(sum(x * ols_coe[2:6])+ols_coe[1])),
           apply(x,1, function(x) return(sum(x * heckman_coe[2:6])+heckman_coe[1])),
           apply(x,1, function(x) return(sum(x * tobit_coe[2:6])+tobit_coe[1])))
compare_dat = data.frame(working_experience=working_experience,
                         income=income,
                         model=model)
p8 = ggplot(data=compare_dat,aes(working_experience,income))+
  geom_line(aes(color=model))+
  ggtitle(label='fitted income with working experience, other independent variables take mean value')
p8
```

Exercise 4 Panel Data

```
dat_panel = fread('./data/dat_A4_panel.csv')
for(year in c(1997:2011,c(2013,2015,2017,2019))){
  dat_panel[,paste('age_',year,sep='')] = year - dat_panel$KEY_BDATE_Y_1997
}

fun = function(x){
  return(sum(x[!is.na(x)]))
}
dat_panel$we1997 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:7,sep='')),'_1997',sep='')],1,fun)*7/365
dat_panel$we1998 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:9,sep='')),'_1998',sep='')],1,fun)*7/365
dat_panel$we1999 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:9,sep='')),'_1999',sep='')],1,fun)*7/365
dat_panel$we2000 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:9,sep='')),'_2000',sep='')],1,fun)*7/365
dat_panel$we2001 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:8,sep='')),'_2001',sep='')],1,fun)*7/365
dat_panel$we2002 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:9,sep=''),10,11),'_2002',sep='')],1,fun)*7/365
dat_panel$we2003 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:9,sep=''),10),'_2003',sep='')],1,fun)*7/365
dat_panel$we2004 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:7,sep='')),'_2004',sep='')],1,fun)*7/365
dat_panel$we2005 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:9,sep='')),'_2005',sep='')],1,fun)*7/365
dat_panel$we2006 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:9,sep='')),'_2006',sep='')],1,fun)*7/365
dat_panel$we2007 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:8,sep='')),'_2007',sep='')],1,fun)*7/365
dat_panel$we2008 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:8,sep='')),'_2008',sep='')],1,fun)*7/365
dat_panel$we2009 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:9,sep='')),'_2009',sep='')],1,fun)*7/365
dat_panel$we2010 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                         c(paste(0,1:9,sep='')),'_2010',sep='')],1,fun)*7/365
dat_panel$we2011 =
```

```
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                       c(paste(0,1:9,sep=''),10,11,12,13),'_2011',sep='')],1,fun)*7/365
dat_panel$we2013 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                       c(paste(0,1:9,sep=''),10),'_2013',sep='')],1,fun)*7/365
dat_panel$we2015 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                       c(paste(0,1:9,sep=''),10,11,12),'_2015',sep='')],1,fun)*7/365
dat_panel$we2017 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                       c(paste(0,1:9,sep=''),10:15),'_2017',sep='')],1,fun)*7/365
dat_panel$we2019 =
  apply(dat_panel[,paste('CV_WKSWK_JOB_DLI.',
                       c(paste(0,1:9,sep=''),10,11),'_2019',sep='')],1,fun)*7/365

for(year in c(1997:2011,c(2013,2015,2017,2019))){
  for(marstat in 0:4){
    dat_panel[,paste('marstat_',marstat,'_',year,sep='')] =
      case_when(dat_panel[,paste('CV_MARSTAT_COLLAPSED_',year,sep=''),with=FALSE]==marstat~1)
  }
}

for(year in 1997:2009){
  temp = paste(substr(year+1,3,4),substr(year+2,3,4),sep='')
  dat_panel[,paste('edu',year,sep='')]=case_when(
    dat_panel[,paste('CV_HIGHEST_DEGREE_',temp,'_',year+1,sep=''),with=FALSE]==0~0,
    dat_panel[,paste('CV_HIGHEST_DEGREE_',temp,'_',year+1,sep=''),with=FALSE]==1~4,
    dat_panel[,paste('CV_HIGHEST_DEGREE_',temp,'_',year+1,sep=''),with=FALSE]==2~12,
    dat_panel[,paste('CV_HIGHEST_DEGREE_',temp,'_',year+1,sep=''),with=FALSE]==3~15,
    dat_panel[,paste('CV_HIGHEST_DEGREE_',temp,'_',year+1,sep=''),with=FALSE]==4~16,
    dat_panel[,paste('CV_HIGHEST_DEGREE_',temp,'_',year+1,sep=''),with=FALSE]==5~18,
    dat_panel[,paste('CV_HIGHEST_DEGREE_',temp,'_',year+1,sep=''),with=FALSE]==6~23,
    dat_panel[,paste('CV_HIGHEST_DEGREE_',temp,'_',year+1,sep=''),with=FALSE]==7~21,
  )
}

for(year in c(2010, 2011, 2013,2015,2017,2019)){
  dat_panel[,paste('edu',year,sep='')]=case_when(
    dat_panel[,paste('CV_HIGHEST_DEGREE_EVER_EDT_',year,sep=''),with=FALSE]==0~0,
    dat_panel[,paste('CV_HIGHEST_DEGREE_EVER_EDT_',year,sep=''),with=FALSE]==1~4,
    dat_panel[,paste('CV_HIGHEST_DEGREE_EVER_EDT_',year,sep=''),with=FALSE]==2~12,
    dat_panel[,paste('CV_HIGHEST_DEGREE_EVER_EDT_',year,sep=''),with=FALSE]==3~15,
    dat_panel[,paste('CV_HIGHEST_DEGREE_EVER_EDT_',year,sep=''),with=FALSE]==4~16,
    dat_panel[,paste('CV_HIGHEST_DEGREE_EVER_EDT_',year,sep=''),with=FALSE]==5~18,
    dat_panel[,paste('CV_HIGHEST_DEGREE_EVER_EDT_',year,sep=''),with=FALSE]==6~23,
    dat_panel[,paste('CV_HIGHEST_DEGREE_EVER_EDT_',year,sep=''),with=FALSE]==7~21,
  )
}

year = 1997
dat_panel2 = dat_panel[,c('PUBID_1997',paste(c('YINC-1700_','age_','edu','we',
                                        'marstat_0_','marstat_1_','marstat_2_',
                                        'marstat_3_','marstat_4_'),year,sep='')),with=FALSE]
```

```r
colnames(dat_panel2) = c('id','income','age','edu','work_exp','Nevermarried','Married',
                    'Separated','Divorced','Widowed')
dat_panel2 = dat_panel2[!is.na(dat_panel2$income)&!is.na(dat_panel2$age)&
                          !is.na(dat_panel2$edu)&!is.na(dat_panel2$work_exp)]
dat_panel2 = dat_panel2[!is.na(dat_panel2$Nevermarried)|!is.na(dat_panel2$Married)|
                           !is.na(dat_panel2$Separated)|!is.na(dat_panel2$Divorced)|
                           !is.na(dat_panel2$Widowed)]
dat_panel2[is.na(dat_panel2)] = 0
dat_panel2$year = year
for(year in c(1998:2011,c(2013,2015,2017,2019))){
  temp = dat_panel[,c('PUBID_1997',paste(c('YINC-1700_','age_','edu','we',
                            'marstat_0_','marstat_1_','marstat_2_',
                            'marstat_3_','marstat_4_'),year,sep='')),with=FALSE]
  colnames(temp) = c('id','income','age','edu','work_exp','Nevermarried','Married',
                     'Separated','Divorced','Widowed')
  temp = temp[!is.na(temp$income)&!is.na(temp$age)&
                            !is.na(temp$edu)&!is.na(temp$work_exp)]
  temp = temp[!is.na(temp$Nevermarried)|!is.na(temp$Married)|
                            !is.na(temp$Separated)|!is.na(temp$Divorced)|
                            !is.na(temp$Widowed)]
  temp[is.na(temp)] = 0
  temp$year = year
  dat_panel2 = rbind(dat_panel2,temp)
}
dat_panel2 = dat_panel2[order(dat_panel2$id),]
```

within

```r
mean_panel = mutate(group_by(dat_panel2, id),
                       income = mean(income),
                       age = mean(age),
                       edu = mean(edu),
                       work_exp = mean(work_exp),
                       Nevermarried = mean(Nevermarried),
                       Married = mean(Married),
                       Separated = mean(Separated),
                       Divorced = mean(Divorced),
                       Widowed = mean(Widowed),
                       year = mean(year))

within_panel = as.data.frame(dat_panel2) - as.data.frame(mean_panel)
within_results = lm(income~age+edu+work_exp+Married+Separated+Divorced+Widowed-1,data=within_panel)
summary(within_results)
```

```
##
## Call:
## lm(formula = income ~ age + edu + work_exp + Married + Separated +
##     Divorced + Widowed - 1, data = within_panel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -140453   -8034      34    7327  263141
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## age        2124.00      19.27 110.235  < 2e-16 ***
## edu         744.55      34.76  21.419  < 2e-16 ***
## work_exp    954.95      29.19  32.713  < 2e-16 ***
## Married    8078.46     244.35  33.060  < 2e-16 ***
## Separated  1457.00     816.03   1.785   0.0742 .
## Divorced   2933.49     462.85   6.338 2.34e-10 ***
## Widowed   -6003.19    2545.63  -2.358   0.0184 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18480 on 80139 degrees of freedom
## Multiple R-squared:  0.4235, Adjusted R-squared:  0.4234
## F-statistic:  8410 on 7 and 80139 DF,  p-value: < 2.2e-16
```

between

```
between = summarise(group_by(dat_panel2, id),
                 income = mean(income),
                 age = mean(age),
                 edu = mean(edu),
                 work_exp = mean(work_exp),
                 Nevermarried = mean(Nevermarried),
                 Married = mean(Married),
                 Separated = mean(Separated),
                 Divorced = mean(Divorced),
                 Widowed = mean(Widowed),
                 year = mean(year))
between = as.data.frame(between)
between_results = lm(income~age+edu+work_exp+Married+Separated+Divorced+Widowed,data=between)
summary(between_results)
```

```
##
## Call:
## lm(formula = income ~ age + edu + work_exp + Married + Separated +
##     Divorced + Widowed, data = between)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -51080  -8728  -2345   5685 275234
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -31773.34    1544.79 -20.568  < 2e-16 ***
## age           1479.23      61.79  23.940  < 2e-16 ***
## edu            924.83      35.47  26.071  < 2e-16 ***
## work_exp      1700.25      78.51  21.656  < 2e-16 ***
## Married       6949.29     579.32  11.996  < 2e-16 ***
## Separated     2490.21    2894.31   0.860    0.390
```

```
## Divorced      -895.31    1232.51  -0.726    0.468
## Widowed     -30613.62    7206.93  -4.248 2.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15690 on 8477 degrees of freedom
## Multiple R-squared:  0.2803, Adjusted R-squared:  0.2797
## F-statistic: 471.6 on 7 and 8477 DF,  p-value: < 2.2e-16
```

difference

```
diff_panel = mutate(group_by(dat_panel2, id),
                    income = lag(income,n=1,default = NA),
                    age = lag(age,n=1,default = NA),
                    edu = lag(edu,n=1,default = NA),
                    work_exp = lag(work_exp,n=1,default = NA),
                    Nevermarried = lag(Nevermarried,n=1,default = NA),
                    Married = lag(Married,n=1,default = NA),
                    Separated = lag(Separated,n=1,default = NA),
                    Divorced = lag(Divorced,n=1,default = NA),
                    Widowed = lag(Widowed,n=1,default = NA),
                    year = lag(year,n=1,default = NA))
diff_panel = na.omit(as.data.frame(dat_panel2) - as.data.frame(diff_panel))
diff_results = lm(income~age+edu+work_exp+Married+Separated+Divorced+Widowed-1,data=diff_panel)
summary(diff_results)
```

```
##
## Call:
## lm(formula = income ~ age + edu + work_exp + Married + Separated +
##     Divorced + Widowed - 1, data = diff_panel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -212126   -4999    -901    4890  322003
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## age        2088.18      30.74  67.935  < 2e-16 ***
## edu         -26.34      34.60  -0.761   0.4465
## work_exp    698.24      30.53  22.872  < 2e-16 ***
## Married    2737.67     274.17   9.985  < 2e-16 ***
## Separated  1170.28     662.89   1.765   0.0775 .
## Divorced   2898.87     501.83   5.777 7.65e-09 ***
## Widowed   -1176.51    2466.50  -0.477   0.6334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17000 on 71654 degrees of freedom
## Multiple R-squared:  0.1044, Adjusted R-squared:  0.1043
## F-statistic:  1193 on 7 and 71654 DF,  p-value: < 2.2e-16
```

compare

```r
dat_r = data.frame(within = c('',within_results$coefficients),
                   between = between_results$coefficients,
                   diff = c('',diff_results$coefficients))
dat_r
```

```
##                       within     between              diff
##                              -31773.3442
## age         2124.00406967266   1479.2269    2088.177814612
## edu         744.549966492504    924.8349 -26.3351057493448
## work_exp    954.950562101839   1700.2525  698.237511094955
## Married     8078.45710069201   6949.2887  2737.67147978256
## Separated    1457.0029439646   2490.2120  1170.28277750251
## Divorced     2933.4853166413   -895.3122  2898.86520213902
## Widowed    -6003.18624810924 -30613.6186 -1176.50922392795
```

Answer1: Potiential bias: for example, the ability. We cannot observe it, but it affects the wage, leading to bias.

Answer2: The results are different. Because, 1) Within and Diff model acctually normalize the observed variables, leading to similar results, 2) the between model only measure the effect on mean level, leading to a more different model.