HW4 Yida Xu

Exercise1

1、2、

```
setwd("C:/Users/yidax/OneDrive/Desktop/613/HW4/Data/Data")
library(gmodels)
library(dplyr)
library(data.table)
library(ggplot2)
library(tidyverse)
library(lubridate)
library(tidyr)
library(magrittr)
library(xlsx)
library(plm)
library(data.table)
#Exericse 1
#1 Create new variables(age and work experience)
dat_A4 <- fread("dat_A4.csv")
dat_A4$age <- 2019 - dat_A4$KEY_BDATE_Y_1997
dat_A4[,18:28][is.na(dat_A4[,18:28])] <- 0
dat_A4$work_exp <- rowSums(dat_A4[,18:28])/52
```

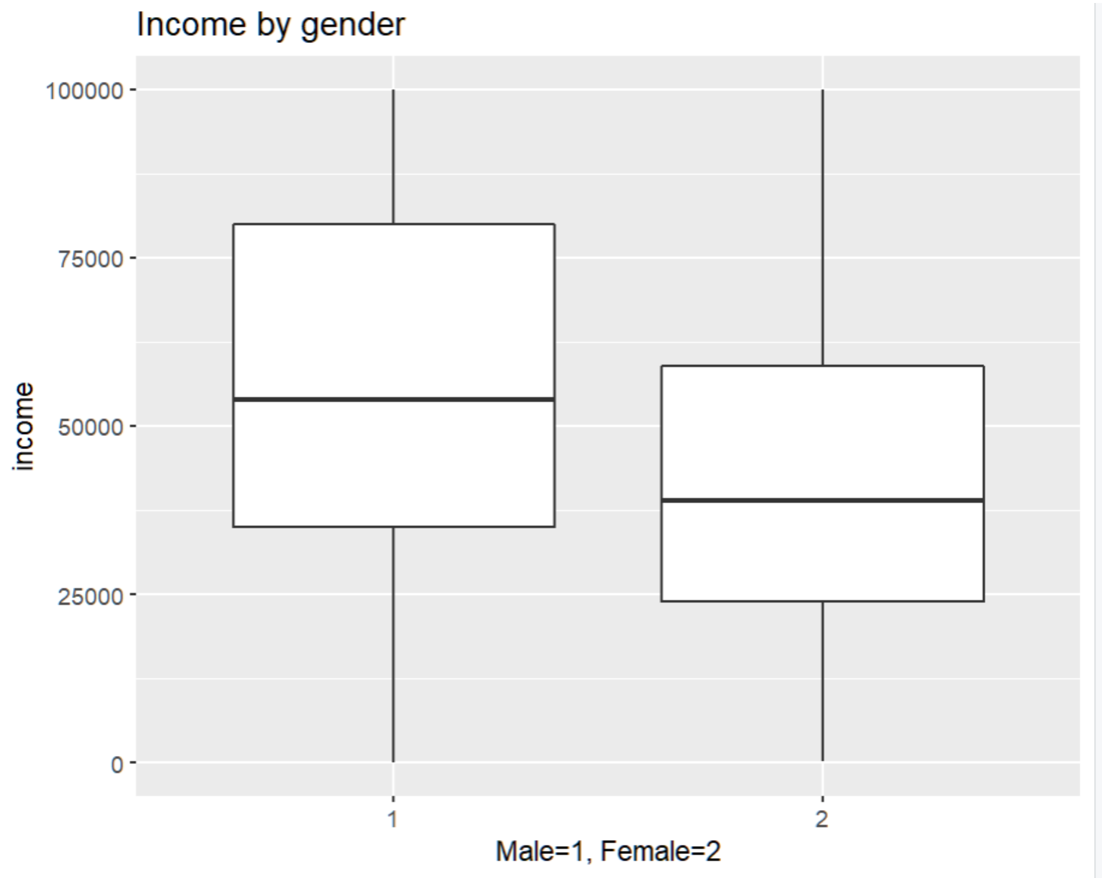
```
#2 Create additional education variable relating to "BIOLOGICAL FATHERS HIGHEST GRADE COMPLETED"
#We first deal with the "ungraded"
dat_A4$CV_HGC_BIO_DAD_1997[which(dat_A4$CV_HGC_BIO_DAD_1997==95)]<- NA
dat_A4$CV_HGC_BIO_MOM_1997[which(dat_A4$CV_HGC_BIO_MOM_1997==95)]<- NA
dat_A4$CV_HGC_RES_DAD_1997[which(dat_A4$CV_HGC_RES_DAD_1997==95)]<- NA
dat_A4$CV_HGC_RES_MOM_1997[which(dat_A4$CV_HGC_RES_MOM_1997==95)]<- NA
dat_A4[,8:11][is.na(dat_A4[,8:11])]<- 0
dat_A4$BioDad<-dat_A4$CV_HGC_BIO_DAD_1997
dat_A4$BioMom<-dat_A4$CV_HGC_BIO_MOM_1997
dat_A4$ResDad<-dat_A4$CV_HGC_RES_DAD_1997
dat_A4$ResMom<-dat_A4$CV_HGC_RES_MOM_1997
```

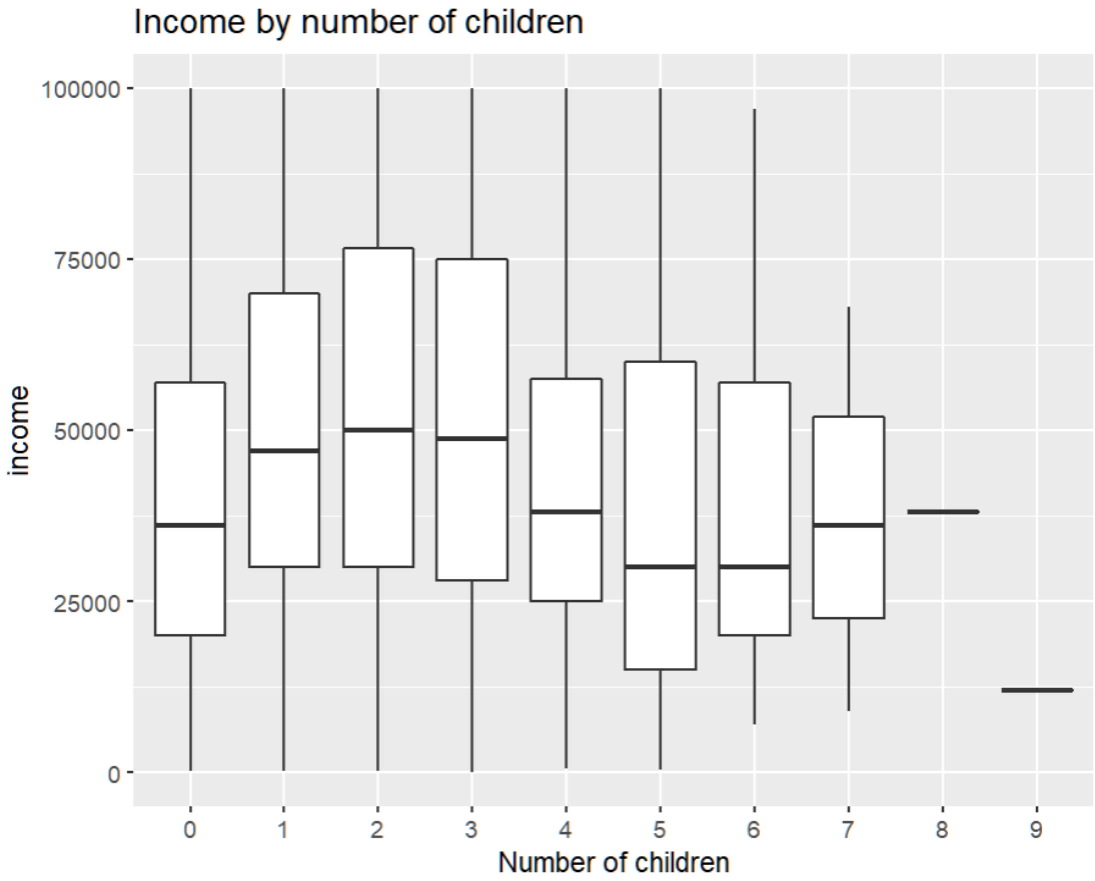
age	work_exp	BioDad	BioMom	ResDad	ResMom
37	12.4230769	17	15	14	15
38	1.9230769	12	12	0	12
38	14.9230769	0	12	12	12
35	0.0000000	0	6	0	6
36	9.5961538	13	12	13	12
37	2.9038462	10	11	0	11
38	12.3653846	0	14	10	14
37	0.0000000	10	15	0	15
35	2.5192308	10	15	0	15
37	3.1153846	11	13	0	13
39	0.0000000	0	12	0	12
35	0.0000000	9	16	9	16
36	12.9038462	0	0	0	0
38	10.3269231	11	0	11	0
38	0.0000000	0	0	0	0
35	6.1153846	0	8	0	8
35	0.0000000	0	8	0	8
37	1.2307692	0	11	0	0
36	1.9615385	0	11	0	0

3、

```
#3 Visualization1
dat_A4[,30][is.na(dat_A4[,30])]<- 0
dat_A4 %>% filter(YINC_1700_2019 > 0) %>% ggplot(aes(x = as.factor(age), y = YINC_1700_2019, )) +
  geom_boxplot() + labs(x = "age", y = "income", title = "Income by age")
dat_A4 %>% filter(YINC_1700_2019 > 0) %>% ggplot(aes(x = as.factor(KEY_SEX_1997), y = YINC_1700_2019)) +
  geom_boxplot() + labs(x="Male=1, Female=2", y="income", title = "Income by gender")
#We need to filter one more step to eliminate the NA
dat_A4 %>% filter(YINC_1700_2019 > 0) %>% filter(CV_BIO_CHILD_HH_U18_2019 >= 0) %>%
  ggplot(aes(x = as.factor(CV_BIO_CHILD_HH_U18_2019), y = YINC_1700_2019)) + geom_boxplot() +
  labs(x="Number of children", y="income", title = "Income by number of children")
```


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


Income by gender
100000
75000
50000
25000
0
income
1
2
Male=1, Female=2

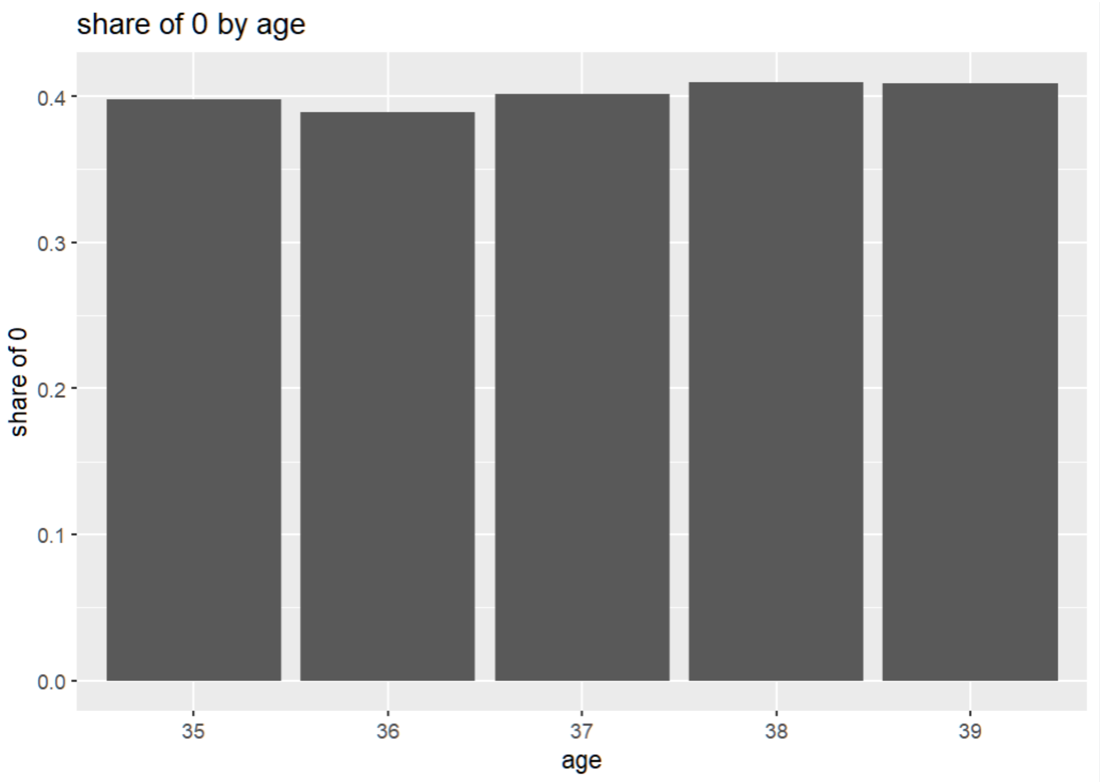



```r
# Visualization2
#by age
share_age<- group_by(dat_A4,age)%>%
summarise(shareage=length(which((YINC_1700_2019==0)=='TRUE'))/length(YINC_1700_2019))
share_age
ggplot(share_age, aes(x = as.factor(age), y = shareage))+geom_bar(stat = 'identity')+
  labs(x = "age", y = "share of 0", title = "share of 0 by age")
#by gender
share_gender<-group_by(dat_A4,KEY_SEX_1997)%>%
summarise(sharegender=length(which((YINC_1700_2019==0)=='TRUE'))/length(YINC_1700_2019))
share_gender
ggplot(share_gender, aes(x = as.factor(KEY_SEX_1997), y = sharegender))+
  geom_bar(stat = 'identity')+labs(x = "Male=1, Female=2", y = "share of 0", title = "share of 0 by gender")
#by number of children
share_children<-group_by(dat_A4,CV_BIO_CHILD_HH_U18_2019)%>%
filter(CV_BIO_CHILD_HH_U18_2019 >= 0)%>%
summarise(sharechildren=length(which((YINC_1700_2019==0)=='TRUE'))/length(YINC_1700_2019))
share_children
ggplot(share_children, aes(x = as.factor(CV_BIO_CHILD_HH_U18_2019), y = sharechildren))+
  geom_bar(stat = 'identity')+labs(x = "Number of children", y = "share of 0", title = "share of 0 by number of children")
#by martial status
share_martial<-group_by(dat_A4,CV_MARSTAT_COLLAPSED_2019)%>%
  filter(CV_MARSTAT_COLLAPSED_2019 >= 0)%>%
  summarise(sharemartial=length(which((YINC_1700_2019==0)=='TRUE'))/length(YINC_1700_2019))
share_martial
ggplot(share_martial, aes(x = as.factor(CV_MARSTAT_COLLAPSED_2019), y = sharemartial))+
  geom_bar(stat = 'identity')+labs(x = "martial status", y = "share of 0", title = "share of 0 by martial status")
```

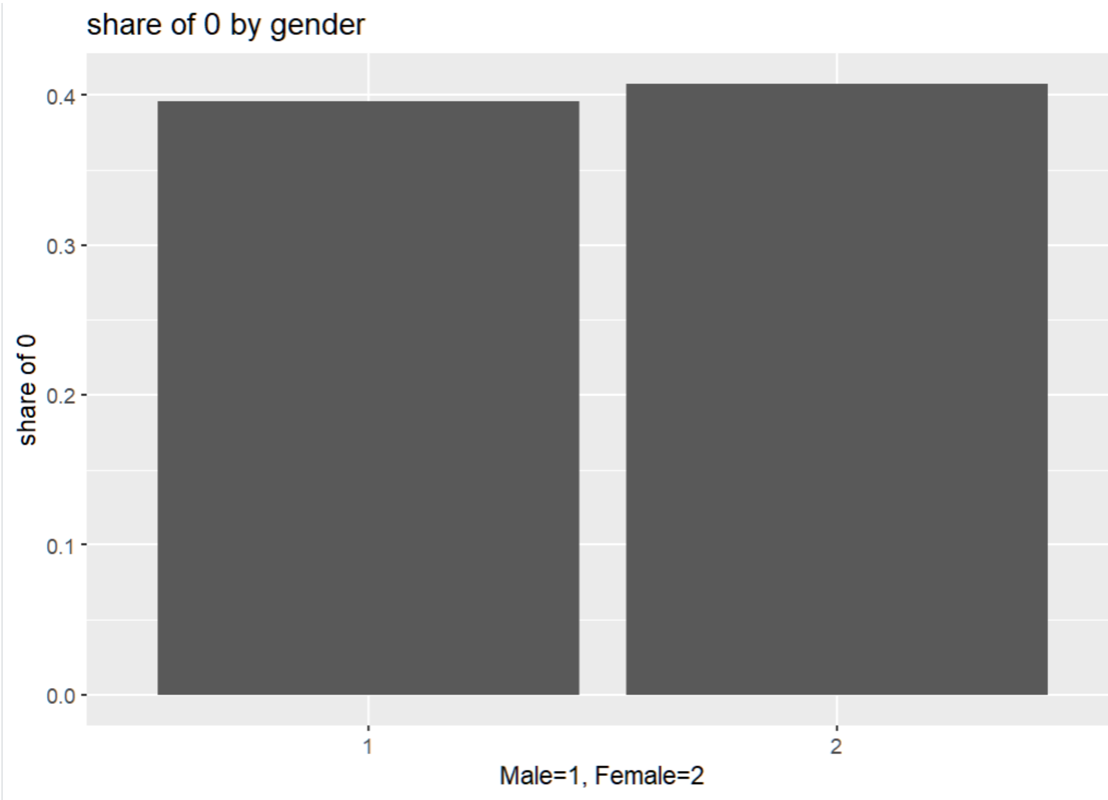
```r
#by number of children and martial status
share_children_martial<-group_by(dat_A4,CV_BIO_CHILD_HH_U18_2019,CV_MARSTAT_COLLAPSED_2019)
share_martial_children_edited<-share_children_martial%>%filter(CV_MARSTAT_COLLAPSED_2019 >= 0)%>%
  filter(CV_BIO_CHILD_HH_U18_2019 >= 0)%>%
summarise(sharechildrenmartial=length(which((YINC_1700_2019==0)=='TRUE'))/length(YINC_1700_2019))

ggplot(share_martial_children_edited, aes(x = as.factor(CV_MARSTAT_COLLAPSED_2019), fill = CV_BIO_CHILD_HH_U18_2019,y = sharechildrenmartial))+
  geom_bar(stat = 'identity')+labs(x = "Martial", fill = "Number of children", y = "share of 0", title = "share of 0 by number of children and martial stat
```

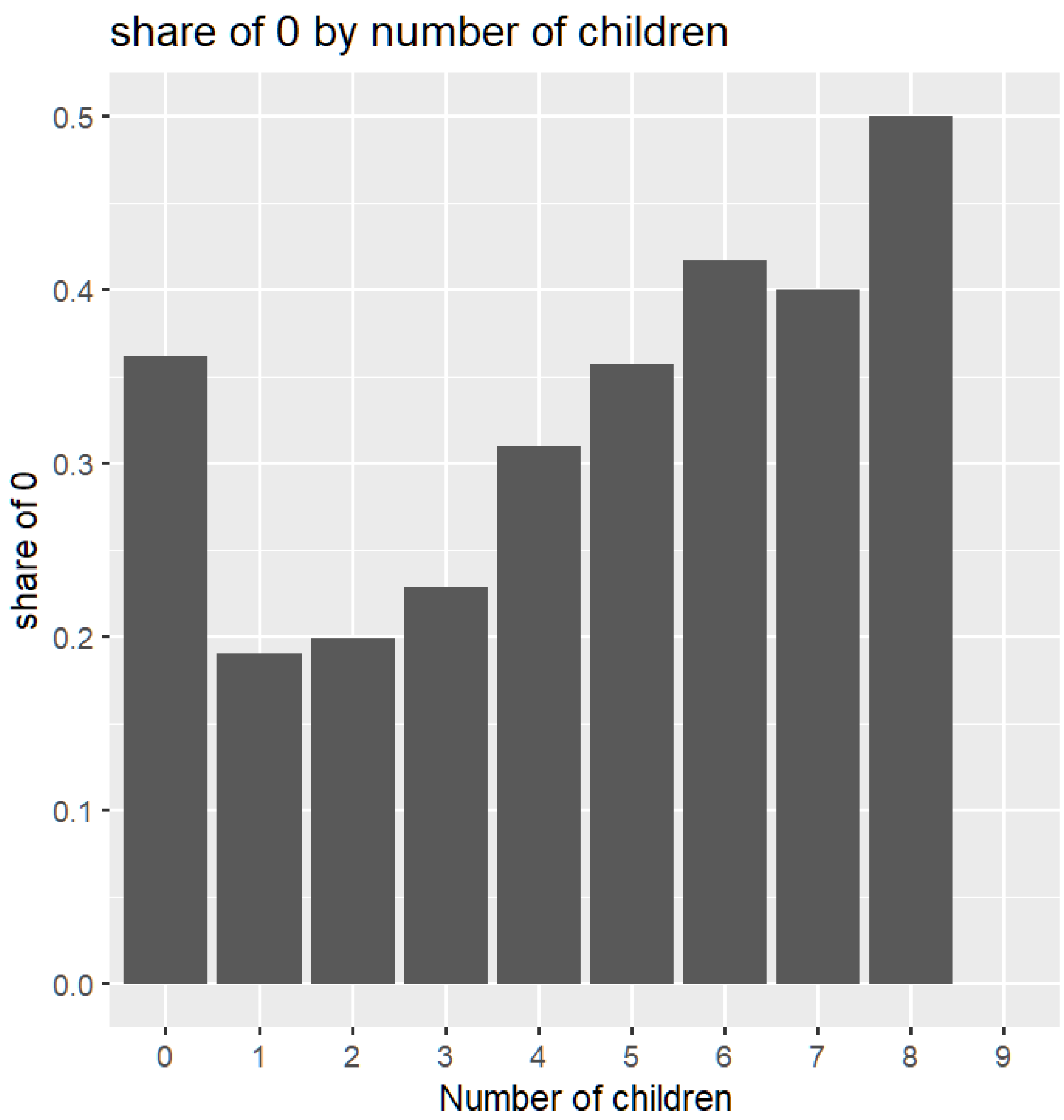
	age	shareage
1	35	0.3980802
2	36	0.3890426
3	37	0.4019555
4	38	0.4098186
5	39	0.4092253



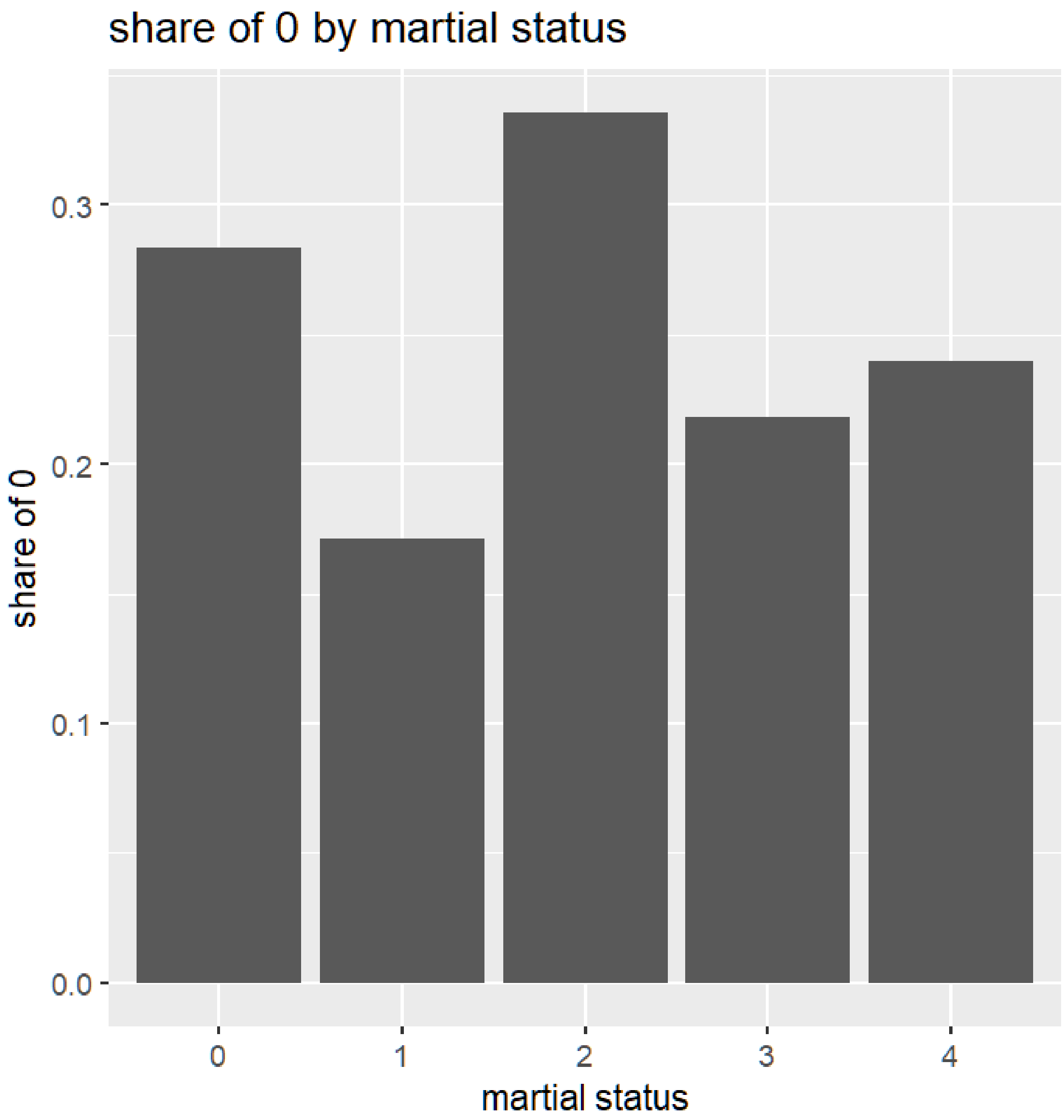
	KEY_SEX_1997	sharegender
1	1	0.3957382
2	2	0.4077537



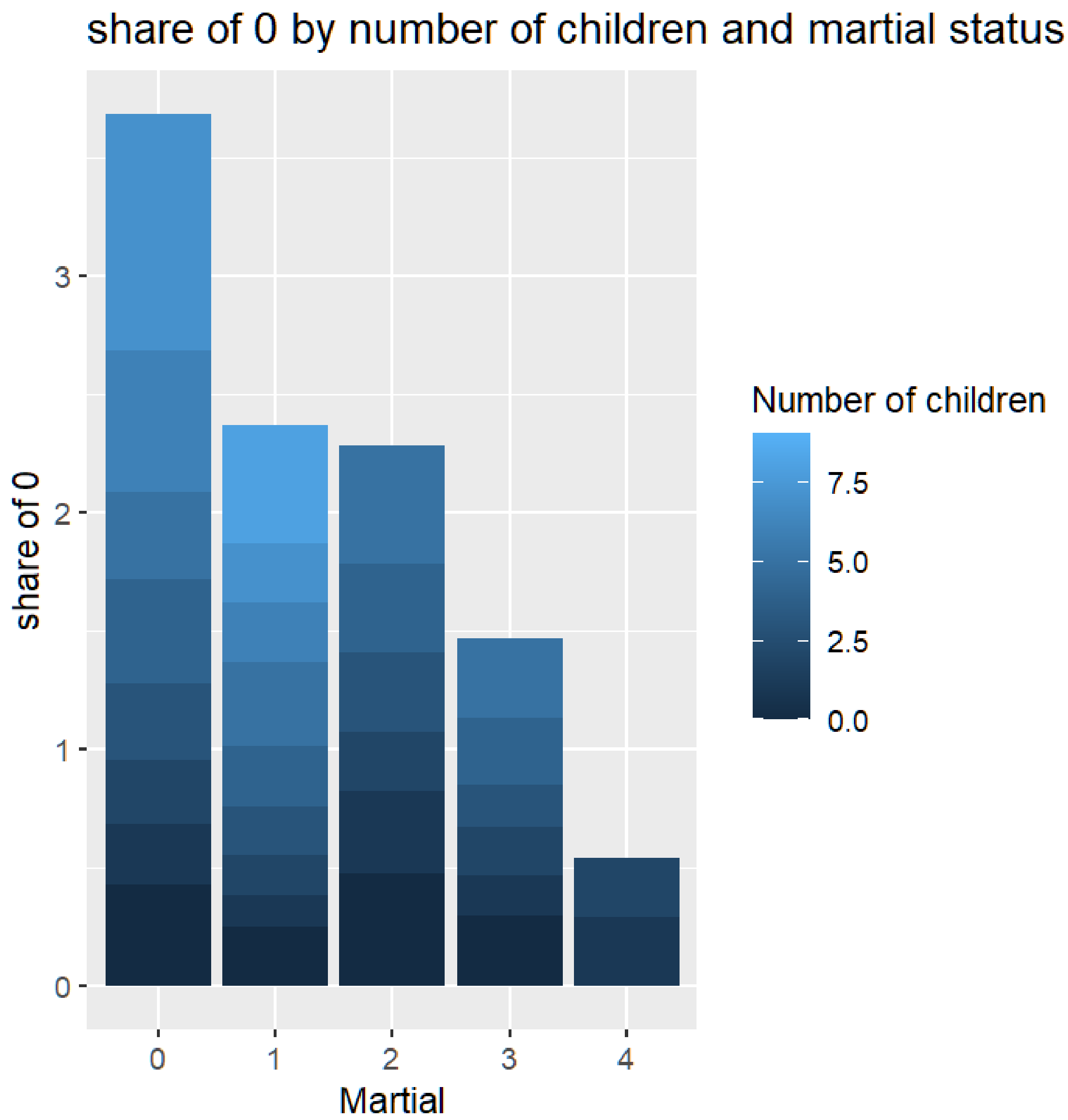
	CV_BIO_CHILD_HH_U18_2019	sharechildren
1	0	0.3611111
2	1	0.1894587
3	2	0.1984954
4	3	0.2284644
5	4	0.3094340
6	5	0.3571429
7	6	0.4166667
8	7	0.4000000
9	8	0.5000000
10	9	0.0000000



	CV_MARSTAT_COLLAPSED_2019	sharemartial
1	0	0.2832284
2	1	0.1711796
3	2	0.3358209
4	3	0.2180723
5	4	0.2400000


share of 0 by martial status
0.3
0.2
0.1
0.0
share of 0
0
1
2
3
4
martial status

	CV_BIO_CHILD_HH_U18_2019	CV_MARSTAT_COLLAPSED_2019	sharechildrenmartial
1	0	0	0.4241706
2	0	1	0.2450331
3	0	2	0.4722222
4	0	3	0.2946860
5	0	4	0.0000000
6	1	0	0.2577963
7	1	1	0.1363636
8	1	2	0.3478261
9	1	3	0.1685393
10	1	4	0.2857143
11	2	0	0.2714681
12	2	1	0.1706454
13	2	2	0.2500000
14	2	3	0.2021277
15	2	4	0.2500000
16	3	0	0.3231707
17	3	1	0.2011070
18	3	2	0.3333333
19	3	3	0.1829268
20	4	0	0.4375000



#interpret the visualizations from above
#1、Male earns more than female.
#2、There is no apparent relationship between age and income.
#3、Families with 1 to 3 children have higher incomes than those with no children and more. Among them, families with two children have the highest income
#4、The proportion of people with zero income increases slightly with age.
#5、The proportion of men earning 0 is greater than that of women.
#6、Unmarried and those with multiple children have the highest percentage of 0 income,Divorced and those with fewer children have the lowest percentage of 0 income.

#Exercise2

```
#Exercise2
#1
data1<-dat_A4%>%filter(YINC_1700_2019 >0)%>%filter(CV_MARSTAT_COLLAPSED_2019 >= 0)
y<-data1$YINC_1700_2019
x1<-data1$age
x2<-as.numeric(data1$KEY_SEX_1997)
x3<-data1$work_exp
x4<-as.numeric(data1$Bioeducation) # The sum of years of education that bio father and bio mother have in total

# Plan to run an OLS income = x0+X1age+X2gender+x3work+x4martial+x5edu.
OLSmodel1<-lm(y~x1+x2+x3+x4,data = data1)
summary(OLSmodel1)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = data1)

Residuals:
   Min     1Q Median     3Q    Max 
-70106 -19633  -3437  18686  76842 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  32594.19    9657.08   3.375 0.000743 ***
x1             367.91     258.83   1.421 0.155246    
x2          -12930.13     719.70 -17.966  < 2e-16 ***
x3            1117.88      66.88  16.714  < 2e-16 ***
x4             683.59      41.91  16.312  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26310 on 5351 degrees of freedom
Multiple R-squared:  0.1475,	Adjusted R-squared:  0.1469 
F-statistic: 231.5 on 4 and 5351 DF,  p-value: < 2.2e-16
```

#Interpret the estimation results

#All independent variables except "age" are significant.

#The p-value of "age" variable is 1.4 which is smaller than 1.96, thus insignificant.

#X2: if all other fixed, male earns 12930.13 than female.

#X3: if all other fixed, people with one more year of working experience earns 1117.88 more.

#X4: if all other fixed, those whose bio parents have one more year in total earns 683.59 more.

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

#The selection problem occurs when the selection of participants or the data is not random.

#In this situation, the selection problem coming from those who report income as 0 or NA (unwilling

#to report).

#SO the non-random sample population causes the selection problem.

2、

#2

Heckman model offers a two-step statistical approach to correct the non-randomly selected sample.

In the first stage, we perform a binary probit analysis on a seleciton equation. (Whether income>0 as dependent variable)

In the second stage, we perform outcome equation based on the first-stage binary probit model. We use the binary variable in the frist stage as the independent variable in the second

stage.

Thus, we can rule out the selection problem (income = 0).

3、

```
#3
#Step 1
#We transfer all NA into "0"
data2<-dat_A4
data2$dummy<-0
data2$dummy[which(data2$YINC_1700_2019>0)] <- 1
data2$dummy[is.na( data2$dummy ) == T] = 0
Dummy<-data2$dummy
data2$age[is.na(data2$age) == T] = 0
data2$YINC_1700_2019[is.na(data2$YINC_1700_2019) == T] = 0
data2$KEY_SEX_1997[is.na(data2$KEY_SEX_1997) == T] = 0
data2$work_exp[is.na(data2$work_exp) == T] = 0
data2$Bioeducation[is.na(data2$Bioeducation)== T] = 0
yy<-data2$YINC_1700_2019
xx1<-data2$age
xx2<-as.numeric(data2$KEY_SEX_1997)
xx3<-data2$work_exp
xx4<-as.numeric(data2$Bioeducation)
data2$Inter<-1
Inter<-data2$Inter
#run the first model
model_1<-glm(Dummy~xx1+xx2+xx3+xx4,family = binomial(link = "probit"), data = data2)
summary(model_1)
predict1<- -predict(model_1)
IMR <- dnorm(predict1)/pnorm(predict1)

likelihood1 <- runif(5,-1,1)
probitlikelihood1 = function(par,Inter,xx1,xx2,xx3,xx4,Dummy){
  yhat = par[1]*Inter + par[2]*xx1 + par[3]*xx2 + par[4]*xx3 + par[5]*xx4
  Prob = pnorm(yhat)
  Prob[Prob>0.999999] = 0.999999
  Prob[Prob<0.000001] = 0.000001
  like = Dummy*log(Prob) + (1-Dummy)*log(1-Prob)
  return(-sum(like))
}
result1<- optim(likelihood1,fn = probitlikelihood1,method="BFGS",control=list(trace=6,maxit=1000),Inter=Inter,xx1=xx1,xx2=xx2,xx3=xx3,xx4=xx4,Dummy=Dumm
result1$par
```

```
> result1$par
[1]  0.81665000 -0.85813835  0.01979600  0.08766433  0.54414536
```

```
#Step2
model_2 <- lm(yy~xx1+xx2+xx3+xx4+IMR)
summary(model_2)
# The results change a lot(both coefficience and significance).
#X2: if all other fixed, male earns 8422.67 than female.
#X3: if all other fixed, people with one more year of working experience earns 25955.71 more.
#X4: if all other fixed, those whose bio parents have one more year in total earns 1512.21 more.
# The difference exit might due to the missing data"NA", those whose income is too small or do not have a job will show "0" or "NA" in the original dataset.
# Thus the previous OLS model might be biased.
```

```
lm(formula = yy ~ xx1 + xx2 + xx3 + xx4 + IMR)

Residuals:
   Min     1Q Median     3Q    Max 
-64171 -13771  -4116  11257 101300 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  156146.05    7688.66   20.31   <2e-16 ***
xx1           -2615.38     196.78  -13.29   <2e-16 ***
xx2           -8422.67     528.63  -15.93   <2e-16 ***
xx3           25955.71     528.52   49.11   <2e-16 ***
xx4            1512.21      37.35   40.48   <2e-16 ***
IMR         -143138.23    3287.25  -43.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25030 on 8978 degrees of freedom
Multiple R-squared:  0.4261,	Adjusted R-squared:  0.4258 
F-statistic:  1333 on 5 and 8978 DF,  p-value: < 2.2e-16
```

The results change a lot(both coefficience and significance).

#X2: if all other fixed, male earns 8422.67 than female.

#X3: if all other fixed, people with one more year of working experience earns 25955.71 more.

#X4: if all other fixed, those whose bio parents have one more year in total education earns 1512.21 more.

The difference exit might due to the missing data"NA", those whose income is too small or do not have a job will show "0" or "NA" in the original dataset.

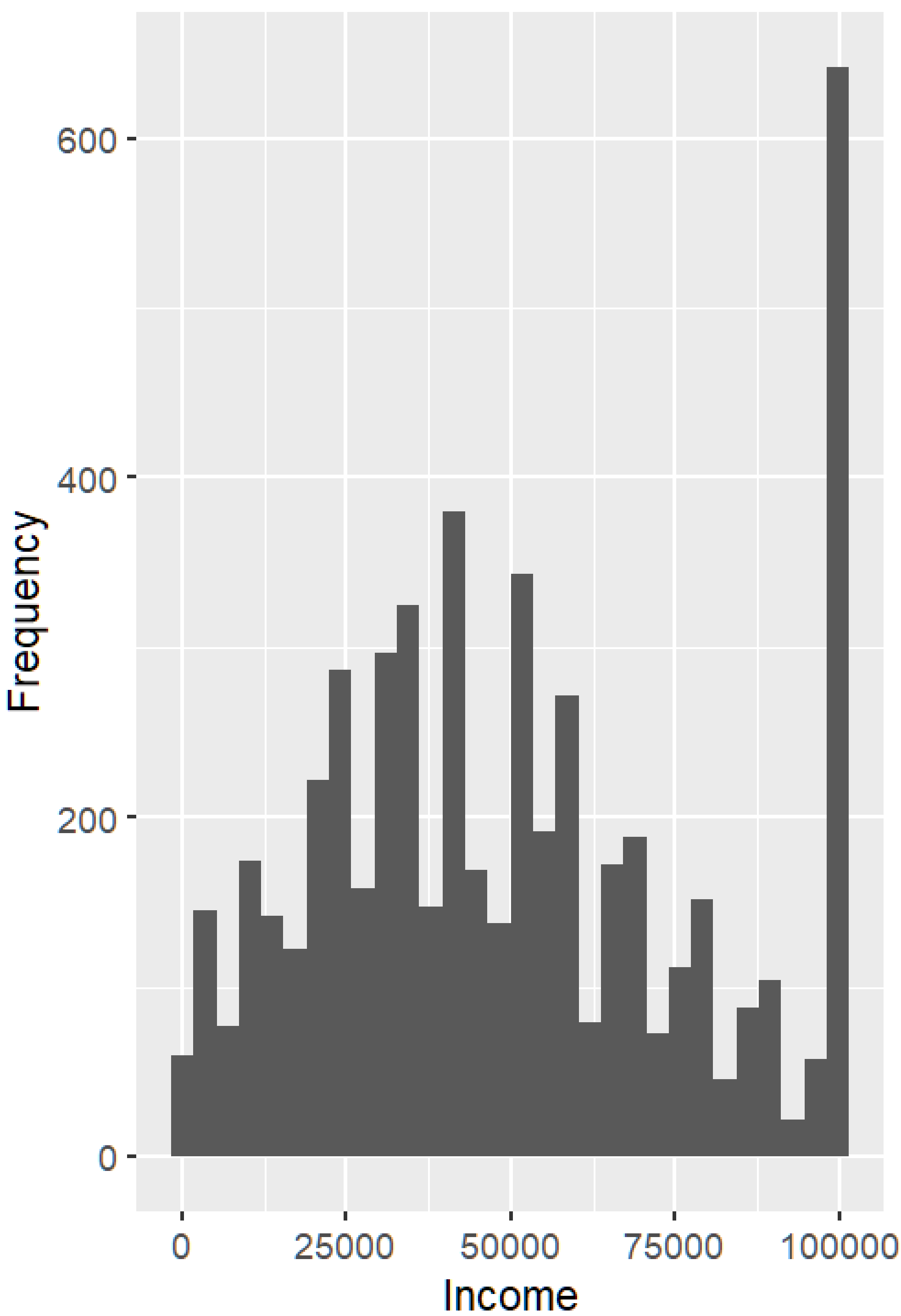
Thus the previous OLS model might be biased.

#Exercise3

1、

```
#Exercise3
#1 Plot a histogram to check whether the distribution of the income variable. What might be the censored value here?

dat_A4%>%filter(YINC_1700_2019 > 0) %>%ggplot(aes(x = YINC_1700_2019)) +
  geom_histogram(bins = 30) +labs(x = "Income",y = "Frequency")
# The censored value could be top-coded at 100000.
```



2、3、

```r
data3<-subset(dat_A4,dat_A4$YINC_1700_2019!='NA')
data3$dummy3 <- ifelse(data3$YINC_1700_2019 >= 100000, 0,1)
dummy3<-data3$dummy3
yyy<-data3$YINC_1700_2019
xxx1<-data3$age
xxx2<-as.numeric(data3$KEY_SEX_1997)
xxx3<-data3$work_exp
xxx4<-as.numeric(data3$Bioeducation)
model_3 <- tobit(yyy ~ xxx1 + xxx2 + xxx3 + xxx4,left=-Inf,right = 100000)
summary(model_3)
# The coefficient of log(scale) is 1.029e+01.
parm <- as.vector(c(model_3$coefficients,10.29))
data3$inter<-1
inter<-data3$inter
likelihood2<-parm + runif(6,-10,10)
tobitlikelihood2 <- function(parm,inter,xxx1,xxx2,xxx3,xxx4,dummy3,yyy){
  XB = parm[1]*inter + parm[2]*xxx1 + parm[3]*xxx2 + parm[4]*xxx3 + parm[5]*xxx4
  resid = yyy - XB
  stand = (100000-XB)/exp(parm[6])
  like = dummy3*log(dnorm(resid/exp(parm[6]))/exp(parm[6])) + (1-dummy3)*log(1 - pnorm(stand))
  return(-sum(like))
}
result2 <- optim(likelihood2,fn=tobitlikelihood2,method="BFGS",control=list(trace=6,REPORT=1,maxit=1000),inter=inter,xxx1=xxx1,xxx2=xxx2,xxx3=xxx3,xxx4=xxx4,dum
```

```
    right = 1e+05)

Observations:
         Total  Left-censored     Uncensored Right-censored 
          5412              0           4775            637 

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  2.676e+04  1.090e+04   2.455   0.0141 *  
xxx1         5.479e+02  2.922e+02   1.875   0.0608 .  
xxx2        -1.426e+04  8.127e+02 -17.542   <2e-16 ***
xxx3         1.209e+03  7.559e+01  15.991   <2e-16 ***
xxx4         7.628e+02  4.737e+01  16.105   <2e-16 ***
Log(scale)   1.029e+01  1.060e-02 971.646   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 29574 

Gaussian distribution
Number of Newton-Raphson Iterations: 3 
Log-likelihood: -5.67e+04 on 6 Df
Wald-statistic:   875 on 4 Df, p-value: < 2.22e-16 
```

4、

Tobit:

```
> result2$par
[1]  26764.14742    545.67276 -14257.08448   1205.54005    767.55052     10.29466
```

OLS:

```
Call:
lm(formula = yyy ~ xxx1 + xxx2 + xxx3 + xxx4, data = data3)

Residuals:
   Min     1Q Median     3Q    Max 
-77557 -19629  -3352  18819  77047 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  30276.74    9691.67   3.124  0.00179 ** 
xxx1           420.16     259.74   1.618  0.10580    
xxx2        -12803.69     722.32 -17.726  < 2e-16 ***
xxx3          1127.73      67.16  16.791  < 2e-16 ***
xxx4           673.80      42.05  16.023  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26540 on 5407 degrees of freedom
Multiple R-squared:  0.1445,	Adjusted R-squared:  0.1438 
F-statistic: 228.3 on 4 and 5407 DF,  p-value: < 2.2e-16
```

```
#4 Interpret the results above and compare to those when not correcting for the censored data
#xxx2: if all other fixed, male earns -14257.08 than female.
#xxx3: if all other fixed, people with one more year of working experience earns 1205.54 more.
#xxx4: if all other fixed, those whose bio parents have one more year in total earns 767.55 more.
#Since the previous model I made in question 2 assumes that NA be treated as 0, we neeed to re-do the model
compareols<-lm(yyy~xxx1+xxx2+xxx3+xxx4,data = data3)
summary(compareols)

#OLS: coefficients for age、gender、work_exp、and edu year of bio parents are 420.16、-12803.69、1127.73、673.80.
#Tobit: coefficients for age、gender、work_exp、and edu year of bio parents are 545.67、-14257.08、1205.54、767.55
#(keep the same data process of OLS tobit)
#From the above comparison, we can find that after dealing with the censoring problem, the coefficients of tobit is larger
#than that of OLS. That is, the effect of independent variables become bigger.
```

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

#xxx2: if all other fixed, male earns -14257.08 than female.

#xxx3: if all other fixed, people with one more year of working experience earns 1205.54 more.

#xxx4: if all other fixed, those whose bio parents have one more year in total earns 767.55 more.

#Since the previous model I made in question 2 assumes that NA be treated as 0, we neeed

#OLS: coefficients for age、gender、work_exp、and edu year of bio parents are 420.16、-12803.69、1127.73、673.80.

#Tobit：coefficients for age、gender、work_exp、and edu year of bio parents are 545.67、-14257.08、1205.54、767.55

#(keep the same data process of OLS tobit)

#From the above comparison, we can find that after dealing with the censoring problem, the coefficients of tobit is larger than that of OLS. That is, the effect of independent variables become bigger.

#Exercise4

```
#Exercise4
#We are interested in the effect of education,marital status, experience and education on wages.
#1
# The potential ability bias could be the difference of people's ability to work. It's unobservable, but it truely affect
#one' wage. For example, people who have disabilities must strive to behave like a normal or get the same salary as the normal.
```

```r
#2
paneldata <- fread("dat_A4_panel.csv")
#We need to use pacakge "panelr"
install.packages('panelr')
library(panelr)

# To start we first rename the highest degree in order to better class and eliminate years.
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_1998=CV_HIGHEST_DEGREE_9899_1998)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_1999=CV_HIGHEST_DEGREE_9900_1999)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_2000=CV_HIGHEST_DEGREE_0001_2000)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_2001=CV_HIGHEST_DEGREE_0102_2001)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_2002=CV_HIGHEST_DEGREE_0203_2002)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_2003=CV_HIGHEST_DEGREE_0304_2003)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_2004=CV_HIGHEST_DEGREE_0405_2004)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_2005=CV_HIGHEST_DEGREE_0506_2005)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_2006=CV_HIGHEST_DEGREE_0607_2006)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_2007=CV_HIGHEST_DEGREE_0708_2007)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_2008=CV_HIGHEST_DEGREE_0809_2008)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_2009=CV_HIGHEST_DEGREE_0910_2009)
paneldata <-  rename(paneldata,CV_HIGHEST_DEGREE_2010=CV_HIGHEST_DEGREE_1011_2010)
#This function takes wide format panels as input and converts them to long format.
# We will deal with NA later
```

```r
paneldataedited <- long_panel(paneldata, prefix='_', begin  = 1997, end = 2019,id = "id", label_location = "end")
#Then we eliminate the year of 2012,2014,2016
panelnew <-paneldataedited %>% subset(wave!='2012'&wave!='2014'&wave!='2016'&wave!='2018')
panelnew <-  rename(panelnew,edu=CV_HIGHEST_DEGREE)
panelnew <-  rename(panelnew,income= 'YINC-1700')
# Separate the age
panelnew$age <- panelnew$wave - panelnew$KEY_BDATE_Y
# I want to separate out the martial status.
panelnew$Nevermarried<-ifelse(panelnew$CV_MARSTAT_COLLAPSED==0,1,0)
panelnew$Married<-ifelse(panelnew$CV_MARSTAT_COLLAPSED==1,1,0)
panelnew$Separated<-ifelse(panelnew$CV_MARSTAT_COLLAPSED==2,1,0)
panelnew$Divorced<-ifelse(panelnew$CV_MARSTAT_COLLAPSED==3,1,0)
panelnew$Widowed<-ifelse(panelnew$CV_MARSTAT_COLLAPSED==4,1,0)
#separate the work_expr
work_expr<-as.matrix(panelnew[,c(10:16,23:30)])
#We treat NA as 0 and add them up.
work_expr[is.na(work_expr)]<-0
work_expr<-as.data.frame(work_expr)
for (i in 3:17) {
  work_expr[,i]<-as.numeric(work_expr[,i])
}
panelnew$work_exp <- rowSums(work_expr[,3:17])/52
panelmodel<-panelnew[,c('PUBID','income','age','edu','work_exp','Nevermarried',
                        'Married','Separated','Divorced','Widowed')]
Panelmodel<-na.omit(panelmodel)
```

```r
#Within
panel_mean<-mutate(group_by(Panelmodel,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))
Mincome<-Panelmodel$income-panel_mean$income
Mage<-Panelmodel$age-panel_mean$age
Medu<-Panelmodel$edu-panel_mean$edu
Mwork<-Panelmodel$work_exp-panel_mean$work_exp
MNevermarried<-Panelmodel$Nevermarried-panel_mean$Nevermarried
MMarried<-Panelmodel$Married-panel_mean$Married
MSeparated<-Panelmodel$Separated-panel_mean$Separated
MDivorced<-Panelmodel$Divorced-panel_mean$Divorced
MWidowed<-Panelmodel$Widowed-panel_mean$Widowed
Within_model<- lm(Mincome~Mage+Medu+Mwork+MMarried+MSeparated+MDivorced+MWidowed)#Here we let nevermarried as reference
summary(Within_model)
```

```
Call:
lm(formula = Mincome ~ Mage + Medu + Mwork + MMarried + MSeparated +
    MDivorced + MWidowed)

Residuals:
   Min     1Q Median     3Q    Max
-68953  -5440     -9   5011  93531

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.788e-12  4.540e+01   0.000 1.000000
Mage         2.274e+03  2.250e+01 101.062  < 2e-16 ***
Medu         2.735e+03  7.639e+01  35.807  < 2e-16 ***
Mwork        8.753e+02  3.140e+01  27.878  < 2e-16 ***
MMarried     3.824e+03  1.925e+02  19.866  < 2e-16 ***
MSeparated   2.516e+03  6.976e+02   3.606 0.000311 ***
MDivorced    1.611e+03  4.377e+02   3.680 0.000233 ***
MWidowed    -5.199e+03  2.845e+03  -1.827 0.067674 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10590 on 54411 degrees of freedom
Multiple R-squared:  0.4526,    Adjusted R-squared:  0.4526
F-statistic:  6428 on 7 and 54411 DF,  p-value: < 2.2e-16
```

```
#Between
panel_between<-summarise(group_by(Panelmodel,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))
Between<-as.data.frame(panel_between)
Between_model<-lm(income~age+edu+work_exp+Married+Separated+Divorced+Widowed,data = Between)
summary(Between_model)
```

```
Call:
lm(formula = income ~ age + edu + work_exp + Married + Separated +
    Divorced + Widowed, data = Between)

Residuals:
   Min     1Q Median     3Q    Max
-48085  -6119  -1293   4566  81724

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23257.42    1285.50 -18.092  < 2e-16 ***
age           1322.74      58.17  22.741  < 2e-16 ***
edu           2538.97     110.34  23.010  < 2e-16 ***
work_exp      1639.64      67.07  24.445  < 2e-16 ***
Married       4457.15     374.90  11.889  < 2e-16 ***
Separated      764.68    2034.37   0.376  0.70702
Divorced      2235.10    1007.10   2.219  0.02649 *
Widowed     -17599.49    6818.19  -2.581  0.00986 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9766 on 8371 degrees of freedom
Multiple R-squared:  0.2573,    Adjusted R-squared:  0.2567
F-statistic: 414.3 on 7 and 8371 DF,  p-value: < 2.2e-16
```

```
#First difference
Panel3<-as.matrix(Panelmodel)
Panel3<-as.data.frame(Panel3)
Panel3$id<-as.numeric(Panel3$id)
Panel3$income<-as.numeric(Panel3$income)
Panel3$age<-as.numeric(Panel3$age)
Panel3$edu<-as.numeric(Panel3$edu)
Panel3$work_exp<-as.numeric(Panel3$work_exp)
Panel3$Nevermarried<-as.numeric(Panel3$Nevermarried)
Panel3$Married<-as.numeric(Panel3$Married)
Panel3$Separated<-as.numeric(Panel3$Separated)
Panel3$Divorced<-as.numeric(Panel3$Divorced)
Panel3$Widowed<-as.numeric(Panel3$Widowed)

Panel3$Fi <- ave(Panel3$income, Panel3$id, FUN=function(x)dplyr::lag(x))
Panel3$Fa <- ave(Panel3$age, Panel3$id, FUN=function(x)dplyr::lag(x))
Panel3$Fe <- ave(Panel3$edu, Panel3$id, FUN=function(x)dplyr::lag(x))
Panel3$Fw <- ave(Panel3$work_exp, Panel3$id, FUN=function(x)dplyr::lag(x))
Panel3$FN <- ave(Panel3$Nevermarried, Panel3$id, FUN=function(x)dplyr::lag(x))
Panel3$FM <- ave(Panel3$Married, Panel3$id, FUN=function(x)dplyr::lag(x))
Panel3$FS <- ave(Panel3$Separated, Panel3$id, FUN=function(x)dplyr::lag(x))
Panel3$FD <- ave(Panel3$Divorced, Panel3$id, FUN=function(x)dplyr::lag(x))
Panel3$FWi <- ave(Panel3$Widowed, Panel3$id, FUN=function(x)dplyr::lag(x))
Panel3$Dincome<-Panel3$income-Panel3$Fi
Panel3$Dage<-Panel3$age-Panel3$Fa
Panel3$Dedu<-Panel3$edu-Panel3$Fe
Panel3$Dwork<-Panel3$work_exp-Panel3$Fw
Panel3$DNevermarried<-Panel3$Nevermarried-Panel3$FN
Panel3$DMarried<-Panel3$Married-Panel3$FM
Panel3$DSeparated<-Panel3$Separated-Panel3$FS
Panel3$DDivorced<-Panel3$Divorced-Panel3$FD
Panel3$DWidowed<-Panel3$Widowed-Panel3$FWi
Differencemodel<-lm(Dincome~Dage+Dedu+Dwork+DMarried+DSeparated+DDivorced+DWidowed,data = Panel3)
summary(Differencemodel)
```

```
Call:
lm(formula = Dincome ~ Dage + Dedu + Dwork + DMarried + DSeparated + 
    DDivorced + DWidowed, data = Panel3)

Residuals:
    Min      1Q  Median      3Q     Max 
-125137   -4535   -1349    3765  118722 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   917.70      89.28  10.279  < 2e-16 ***
Dage         1776.03      54.52  32.579  < 2e-16 ***
Dedu          439.71      82.85   5.307 1.12e-07 ***
Dwork         588.94      34.36  17.138  < 2e-16 ***
DMarried     1860.44     228.44   8.144 3.91e-16 ***
DSeparated   2303.14     635.14   3.626 0.000288 ***
DDivorced    1742.50     522.47   3.335 0.000853 ***
DWidowed    -3635.44    2874.56  -1.265 0.205988    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11460 on 47744 degrees of freedom
  (因为不存在，8431个观察量被删除了)
Multiple R-squared:  0.03865,	Adjusted R-squared:  0.0385 
F-statistic: 274.2 on 7 and 47744 DF,  p-value: < 2.2e-16
```

#3

#Within

#As all else equal, one more year of age increases one's wage by 2274.

#As all else equal, one more year of education increases one's wage by 2735.

#As all else equal, one more year of work experience increases one's wage by 875.3.

#As all else equal, those who married earns 3824 more than those who never married.

#As all else equal, those who separated earns 2516 more than those who never married.

#As all else equal, those who divorced earns 1611 more than those who never married.

#Between
#As all else equal, one more year of age increases one's wage by 1322.74.
#As all else equal, one more year of education increases one's wage by 2538.97.
#As all else equal, one more year of work experience increases one's wage by 1639.64.
#As all else equal, those who married earns 4457.15 more than those who never married.
#As all else equal, those who separated earns 764.68 more than those who never married.
#As all else equal, those who divorced earns 2235.1p0 more than those who never married.

#First difference:
#As all else equal, one more year of age increases one's wage by 1776.03.
#As all else equal, one more year of education increases one's wage by 439.71.
#As all else equal, one more year of work experience increases one's wage by 588.94.
#As all else equal, those who married earns 1860.44 more than those who never married.
#As all else equal, those who separated earns 2303.14 more than those who never married.
#As all else equal, those who divorced earns 1742.50 more than those who never married.

different models yield different parameter estimates because they process data in a different way.
The within model eliminate the individual effect and normalized the data.
The Between model eliminate the time effect and basically provide the effect of mean level.
The Difference model keeps the variation of both individual and time effect.