A4 -Zihao Zhang

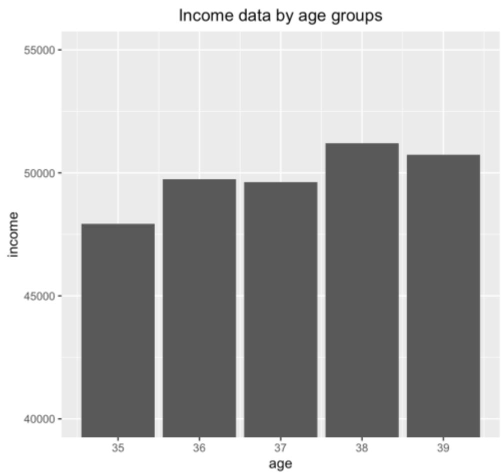
```r
setwd("/Users/zhangzihao/Desktop/duke/613/HW/a4/Data")
library(data.table)
library(tidyverse)
dat_A4_panel <- fread("dat_A4_panel.csv")
dat_A4 <- fread("dat_A4.csv")

# 1
# 1.1
dat_A4$age <- 2019 - dat_A4$KEY_BDATE_Y_1997
dat_A4$work_exp <- rowSums(dat_A4[,18:28],na.rm = "TRUE")
dat_A4$work_exp <- dat_A4$work_exp/52
dat_A4$work_exp <- round(dat_A4$work_exp,2)
```

CV_WKSWK_JOB_DLI.11_2019	YSCH.3113_2019	YINC_1700_2019	age	work_exp	ed
NA	NA	NA	38	0.00	
NA	3	100000	37	12.42	
NA	5	59000	36	1.69	
NA	3	27000	38	1.92	
NA	3	100000	37	13.46	
NA	3	17000	37	2.25	
NA	1	NA	36	2.37	
NA	5	60000	38	4.19	
NA	6	90000	37	3.23	

```r

# 1.2
# I am a little bit confused about the question.
# Should the variable represent the total years of schooling for both residential parents and biological parents?
# I Create three variables. total years of schooling of biological parents; total years of residential parents; total years of these four type persons.
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$edu_bio_parents <- rowSums(dat_A4[,8:9],na.rm = "TRUE")
dat_A4$edu_res_parents <- rowSums(dat_A4[,10:11],na.rm = "TRUE")
dat_A4$edu_parents <- rowSums(dat_A4[,8:11],na.rm = "TRUE")
```

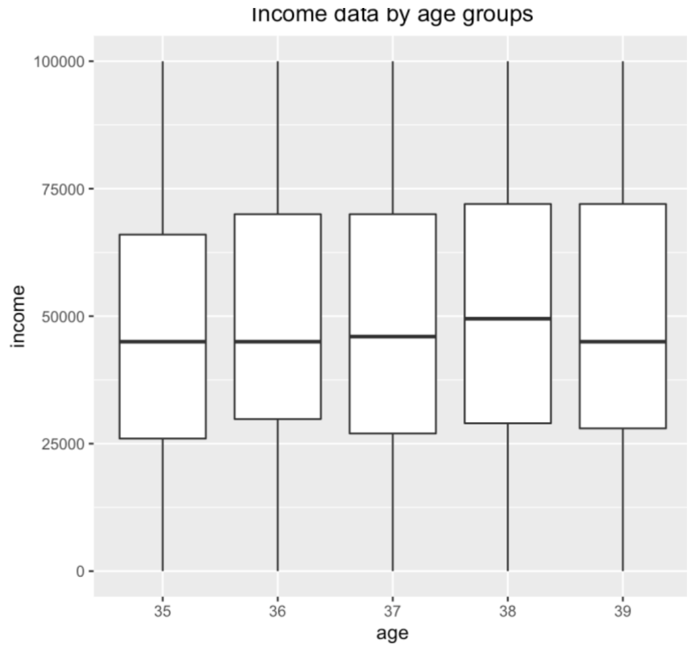
SWK_JOB_DLI.11_2019	YSCH.3113_2019	YINC_1700_2019	age	work_exp	edu_bio_parents	edu_res_parents	edu_parents
	NA	NA	38	0.00	24	24	48
	3	100000	37	12.42	32	29	61
	5	59000	36	1.69	12	12	24
	3	27000	38	1.92	24	12	36
	3	100000	37	13.46	24	24	48
	3	17000	37	2.25	12	12	24
	1	NA	36	2.37	12	12	24
	5	60000	38	4.19	18	18	36
	6	90000	37	3.23	18	18	36
	6	100000	35	5.08	18	18	36

```
# 1.3
library("ggplot2")
# 1.3.1
# firstly, i think i need to calculate the mean
# but then, i think box plot can be more intuitive
# income data by age groups
dat_plot <- dat_A4
income_age <- dat_A4 %>% group_by(age) %>% summarise(income = mean(YINC_1700_2019,na.rm = T))
p <- ggplot(data = income_age,
            mapping = aes(
              x = age,
              y = income,
              ))
p + geom_col() +
  coord_cartesian(ylim=c(40000,55000)) +
  ggtitle("Income data by age groups") +
  theme(plot.title = element_text(hjust = 0.5))
```

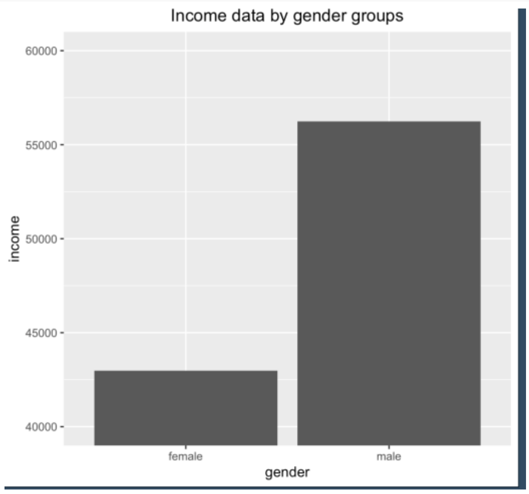



```
dat_A4 %>% dplyr::filter(!is.na(dat_A4$YINC_1700_2019)) %>%
  ggplot(aes(x = as.factor(age), y = YINC_1700_2019, )) + geom_boxplot() +
  ggtitle("Income data by age groups") +
  theme(plot.title = element_text(hjust = 0.5)) +
```

```r
  labs(x = "age", y = "income")
```

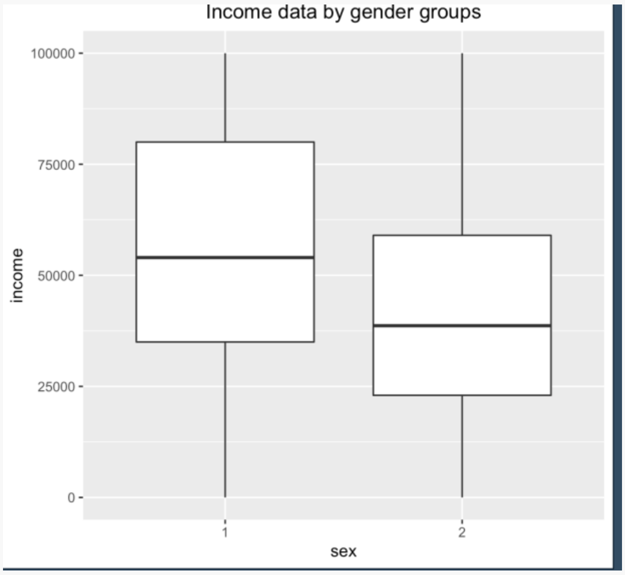



```r
# income data by gender groups
income_gender <- dat_A4 %>% group_by(KEY_SEX_1997) %>% summarise(income = mean(YINC_1700_2019,na.rm = T))
income_gender$KEY_SEX_1997 = c("male","female")
p <- ggplot(data = income_gender,
            mapping = aes(
              x = KEY_SEX_1997,
              y = income
            )) +
  labs(x = "gender")
p + geom_col() +
  coord_cartesian(ylim=c(40000,60000)) +
  ggtitle("Income data by gender groups") +
  theme(plot.title = element_text(hjust = 0.5))
```

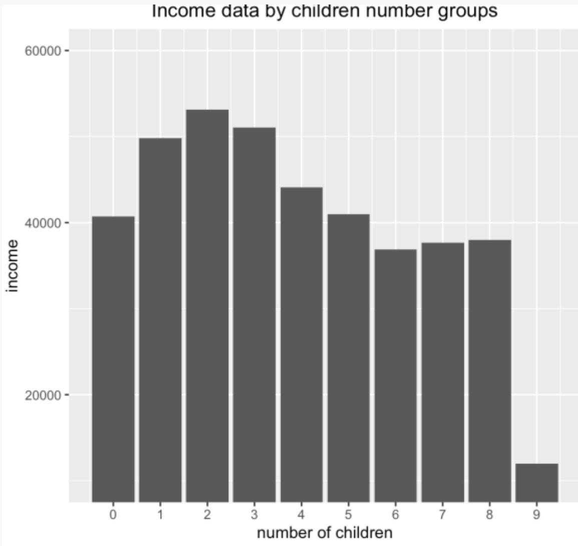



```r
dat_A4 %>% dplyr::filter(!is.na(dat_A4$YINC_1700_2019)) %>%
```

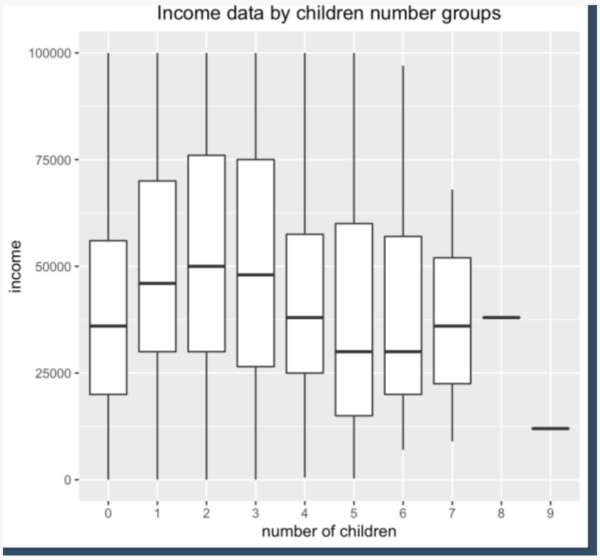
```
  ggplot(aes(x = as.factor(KEY_SEX_1997), y = YINC_1700_2019 )) +
  geom_boxplot() +
  labs(x = "sex", y = "income") +
  ggtitle("Income data by gender groups") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# income data by number of children
income_child <- dat_A4 %>% group_by(CV_BIO_CHILD_HH_U18_2019) %>% summarise(income
= mean(YINC_1700_2019,na.rm = T)) %>% na.omit()
p <- ggplot(data = income_child,
            mapping = aes(
              x = CV_BIO_CHILD_HH_U18_2019,
              y = income,
            ))+
  labs(x = "number of children")
p + geom_col() +
  coord_cartesian(ylim=c(10000,60000)) +
  scale_x_continuous(breaks = seq(0,9,1)) +
  ggtitle("Income data by children number groups") +
  theme(plot.title = element_text(hjust = 0.5))
```

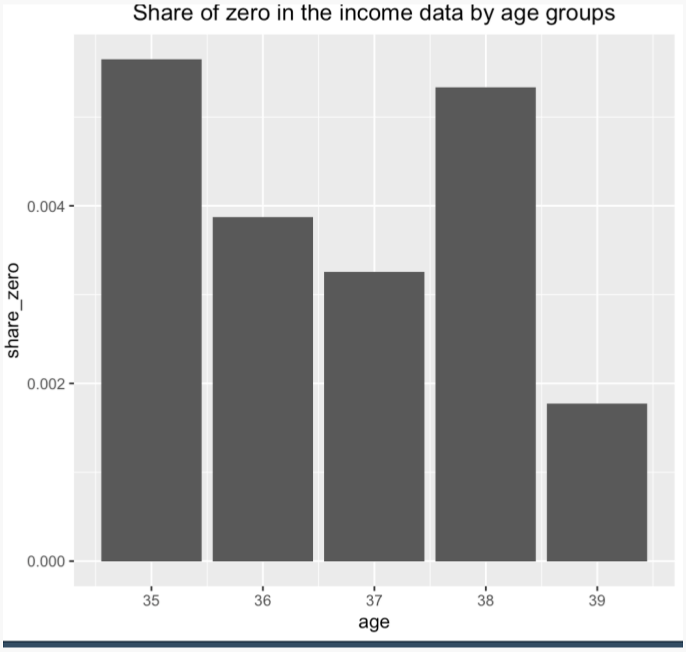



```
dat_A4 %>% dplyr::filter(!is.na(dat_A4$YINC_1700_2019)) %>%
  filter(CV_BIO_CHILD_HH_U18_2019 >= 0) %>%
  ggplot(aes(x = as.factor(CV_BIO_CHILD_HH_U18_2019), y = YINC_1700_2019 )) +
  geom_boxplot() +
  labs(x = "number of children", y = "income") +
  ggtitle("Income data by children number groups") +
  theme(plot.title = element_text(hjust = 0.5))
```

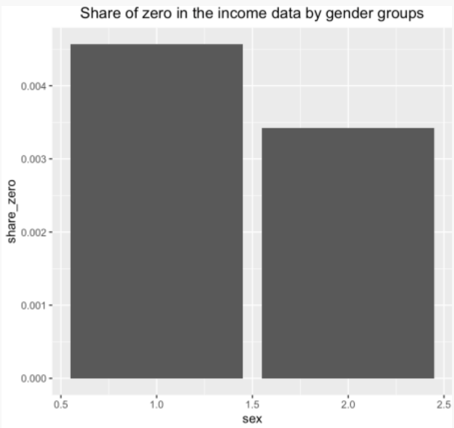



```
# 1.3.2
# share of "0" in the income data by age groups
share_inc_age <- dat_A4 %>% group_by(age) %>% summarize(share_zero = length(which((
YINC_1700_2019==0)=='TRUE'))/length(YINC_1700_2019))
p <- ggplot(data = share_inc_age,
            mapping = aes(
              x = age,
              y = share_zero,
            ))
```

```
p + geom_col() +
  ggtitle("Share of zero in the income data by age groups") +
  theme(plot.title = element_text(hjust = 0.5))
```

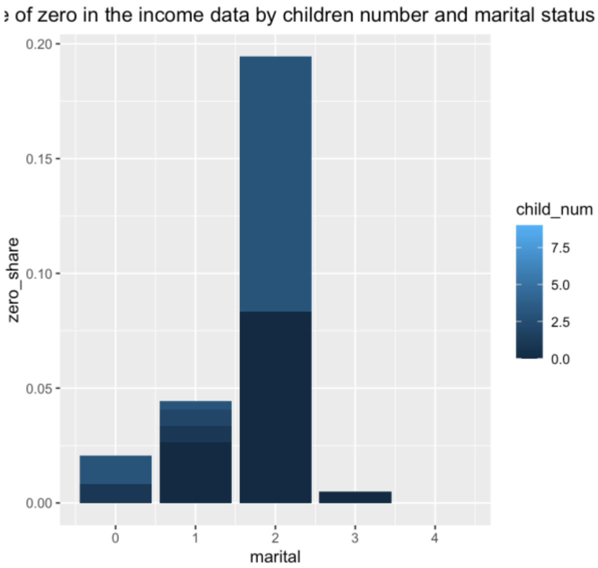



```
# share of "0" in the income data by gender groups
share_inc_gender <- dat_A4 %>% group_by(KEY_SEX_1997) %>% summarize(share_zero = length(which((YINC_1700_2019==0)=='TRUE'))/length(YINC_1700_2019))
p <- ggplot(data = share_inc_gender,
            mapping = aes(
              x = KEY_SEX_1997,
              y = share_zero,
            ))
p + geom_col() +
  ggtitle("Share of zero in the income data by gender groups") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "sex")
```

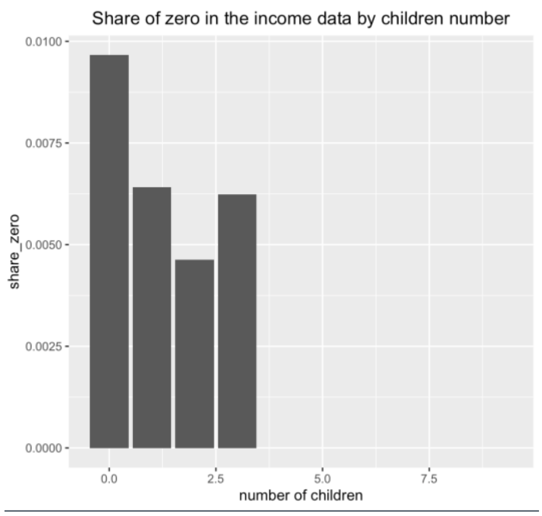



```
# share of "0" in the income data by number of children and marital status
```

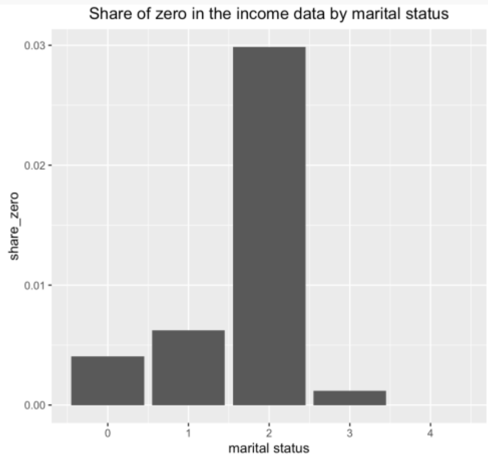
```
share_inc_child_marital <- dat_A4 %>% group_by(CV_MARSTAT_COLLAPSED_2019,CV_BIO_CHILD_HH_U18_2019) %>%
  summarize(share_zero = length(which((YINC_1700_2019==0)=='TRUE'))/length(YINC_1700_2019))
colnames(share_inc_child_marital) <- c("marital", "child_num", "zero_share")
share_inc_child_marital <- na.omit(share_inc_child_marital)
p <- ggplot(data = share_inc_child_marital,
            mapping = aes(
              x = marital,
              y = zero_share,
              fill = child_num
            ))
p + geom_col() +
  ggtitle("Share of zero in the income data by children number and marital status") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# share of "0" in the income data by number of children
share_inc_child <- dat_A4 %>% group_by(CV_BIO_CHILD_HH_U18_2019) %>%
  summarize(share_zero = length(which((YINC_1700_2019==0)=='TRUE'))/length(YINC_1700_2019))
share_inc_child <- na.omit(share_inc_child)
p <- ggplot(data = share_inc_child,
            mapping = aes(
              x = CV_BIO_CHILD_HH_U18_2019,
              y = share_zero,
            ))
p + geom_col() +
  ggtitle("Share of zero in the income data by children number") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "number of children")
```




```
# share of "0" in the income data by number of marital status
share_inc_marital <- dat_A4 %>% group_by(CV_MARSTAT_COLLAPSED_2019) %>%
  summarize(share_zero = length(which((YINC_1700_2019==0)=='TRUE'))/length(YINC_1700_2019))
share_inc_marital <- na.omit(share_inc_marital)
p <- ggplot(data = share_inc_marital,
            mapping = aes(
              x = CV_MARSTAT_COLLAPSED_2019,
              y = share_zero,
            ))
p + geom_col() +
  ggtitle("Share of zero in the income data by marital status") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "marital status")
```




```
# 1.3.3
# Age and income show a positive correlation, but the correlation is not very significant
# Men earn significantly more than women
# The number of children is positively correlated with income at the beginning (with 1-3 children), and then becomes negative
```

```r
# Families with 1-3 children seem to have the most income
# there is no obvious trend, it seems like that 35 and 38 years group has more zero proportion in income
# More men than women have no income
# Separated households have the largest share of no income
# More than half of these separated families have 2-3 children

# 2
# 2.1
data_2.1 <- dat_A4 %>% filter(YINC_1700_2019 >0)

# I set age/gender/work_exp/edu_parents as the independent variables
# using lm function to check
data_2.1 %>%
  lm(YINC_1700_2019 ~ work_exp + age + KEY_SEX_1997 + edu_parents, data =.) %>%
  summary()
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   27229.95    9589.08   2.840  0.00453 **
work_exp       1092.08      66.38  16.453  < 2e-16 ***
age             462.10     256.58   1.801  0.07176 .
KEY_SEX_1997 -12747.51     713.41 -17.868  < 2e-16 ***
edu_parents     402.48      21.66  18.581  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26120 on 5371 degrees of freedom
Multiple R-squared:   0.16,     Adjusted R-squared:  0.1593
F-statistic: 255.7 on 4 and 5371 DF,  p-value: < 2.2e-16
```

```r
# interpret the estimate results
# all independent variables are significant, The significance of "age" is relatively low, only significant at 95% significance
# work_exp: if other factors are fixed, workers earn more $1,092 for every additional year of work experience
# age: if other factors are fixed, workers earn $462 more if one year older
# sex: if other factors are fixed, female earns $12747 less than male
# parents' education: if other factors are fixed, if parents' education years increase one year, the income increses $402

# why there might be selection problem
# there are some interviewee reporting their incomes are zero or unwilling to report, which is may be not random.
# this phenomenon may influence the bias and the estimate of independent variables.

# 2.2
# why Heckman can solve the problem
```

```r
# To solve the selection problem, the Heckman model assumes that there are some other variables influencing dependent variable but not included in the independent variable sets.
# and the model estimates this part at the first stage
# then, using this part as a regressor at the second stage to avoid the problem

# 2.3
data_2.3 <- dat_A4

# create a dummy variables
for (i in 1:nrow(data_2.3)){
  if (isTRUE(data_2.3$YINC_1700_2019[i]>0)){
    data_2.3$dummy[i] = 1
  }else{
    data_2.3$dummy[i] = 0
  }
}
# missing value = 0

# create inter
data_2.3$inter <- 1

# create other variables
i <- data_2.3$inter
work_exp <- data_2.3$work_exp
edu <- data_2.3$edu_parents
age <- as.numeric(data_2.3$age)
sex <- as.numeric(data_2.3$KEY_SEX_1997)
d <- data_2.3$dummy

# first stage
heck1 <- glm(formula =  d ~ work_exp + age + edu + sex, family = binomial(link = "probit"), data = data_2.3)
summary(heck1)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.3910549  0.4096167   0.955  0.33974
work_exp     0.2044746  0.0048510  42.151  < 2e-16 ***
age         -0.0297541  0.0109349  -2.721  0.00651 **
edu          0.0050751  0.0009089   5.584 2.35e-08 ***
sex          0.0001196  0.0305432   0.004  0.99687
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12104.3  on 8983  degrees of freedom
Residual deviance:  8533.1  on 8979  degrees of freedom
AIC: 8543.1
```

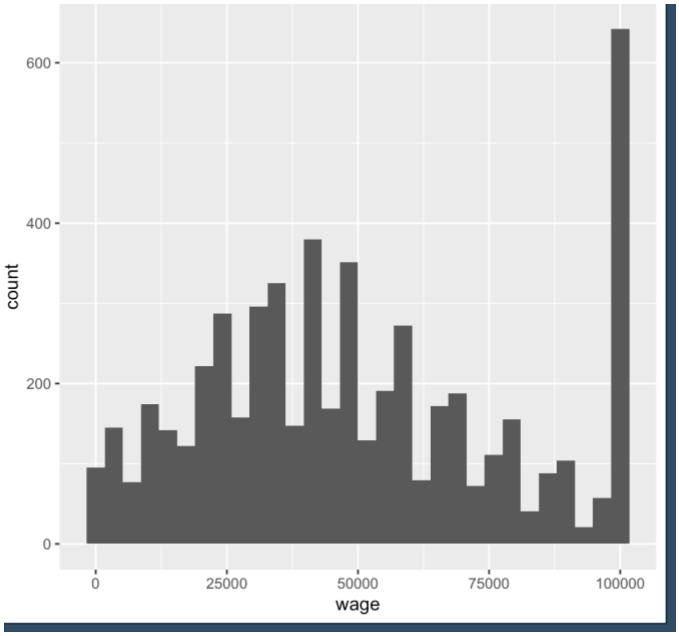
```r
predict1 <- -predict(heck1, newdata = NULL,
                     type = c("link", "response", "terms"),
                     se.fit = FALSE, dispersion = NULL, terms = NULL,
                     na.action = na.pass)

imr <- (1/(1-pnorm(predict1))) * dnorm(predict1)

heckfunc <- function (par, work_exp, edu, age, sex, imr)  {
  yhat  = par[1] + par[2]* work_exp + par[3]* edu + par[4] * age + par[5] * sex + par[6] * imr
  prob = pnorm(yhat)
  prob[prob>0.999999] = 0.999999
  prob[prob<0.000001] = 0.000001
  like = imr*log(prob) + (1-imr)*log(1-prob)
  return( - sum(like) )
}

datah <-  cbind(data_2.3,imr)
imr_reg = datah$imr

predictor <- lm(YINC_1700_2019 ~ work_exp + edu + age + sex + imr, data = datah)
summary(predictor)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  35910.81    9545.92   3.762  0.00017 ***
work_exp      -619.28     150.52  -4.114 3.94e-05 ***
edu            321.96      22.27  14.458  < 2e-16 ***
age           1006.09     257.44   3.908 9.42e-05 ***
sex         -12796.75     707.62 -18.084  < 2e-16 ***
imr         -30100.47    2369.76 -12.702  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25990 on 5406 degrees of freedom
  (3572 observations deleted due to missingness)
Multiple R-squared:    0.18,     Adjusted R-squared:  0.1792
F-statistic: 237.3 on 5 and 5406 DF,  p-value: < 2.2e-16
```

```r
start <- predictor$coefficients
results <- optim(start, fn = heckfunc, method = "BFGS",
                        control = list(trace = 6, maxit = 3000),
                        work_exp = work_exp, edu = edu, age = age, sex = sex, imr = imr_reg)
results$par
```

```
initial  value 39410.827908
final  value 39410.827908
converged
> results$par
(Intercept)    work_exp         edu         age         sex         imr
 35910.8089   -619.2836    321.9631   1006.0943 -12796.7481 -30100.4700
```

```r
# The difference is obvious, so we can assume that the absence of the data of low earning or zero may be not random.
 # It influences the result of OLS
 # the coefficients of work experience and sex are negative showing a negative correlation between these variables and income. And the edu and age have a positive correlation with income. Specifically, for example, when all other conditions fixed, male earns $12796 more than female.

# 3
# 3.1
p <- ggplot(data = dat_A4,
            mapping = aes(
              x = YINC_1700_2019,
            ))
p + geom_histogram() +
  labs(x = "wage")
```




```r
# what might be the censored value
# the income over 100000 should be top-coded

# 3.2
# we can use a two-step model to solve the censoring problem
# Similar with question 2.3, we can firstly explain top-coded incidents and then we can use the inverse mills ratio to estimate at the second stage.
 # 3.3
data_3.2 <- dat_A4

for (i in 1:nrow(data_3.2)){
```

```r
  if (isTRUE(data_3.2$YINC_1700_2019[i]<100000)){
    data_3.2$d[i] = 1
  }else{
    data_3.2$d[i] = 0
  }
}

data_3.2$inter <- 1
inter <- data_3.2$inter
d2 <- data_3.2$d
work_exp <- data_3.2$work_exp
edu <- data_3.2$edu_parents
age <- as.numeric(data_3.2$age)
sex <- as.numeric(data_3.2$KEY_SEX_1997)
inc <- data_3.2$YINC_1700_2019

# we may use Tobit model to solve the censoring problem
library(AER)
tobit <- tobit(inc ~ work_exp + edu + age + sex, right = 100000, data = data_3.2)
summary(tobit)

```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.053e+04  1.093e+04   1.878   0.0604 .
work_exp     1.184e+03  7.572e+01  15.633   <2e-16 ***
edu          4.481e+02  2.477e+01  18.092   <2e-16 ***
age          6.571e+02  2.926e+02   2.246   0.0247 *
sex         -1.401e+04  8.134e+02 -17.230   <2e-16 ***
Log(scale)   1.029e+01  1.066e-02 966.074   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 29576

Gaussian distribution
Number of Newton-Raphson Iterations: 3
Log-likelihood: -5.633e+04 on 6 Df
Wald-statistic: 943.5 on 4 Df, p-value: < 2.22e-16
```

```r
par <- tobit$coefficients
par[6] <- 10.295
par <- as.vector(c(par))

tlikefunc <- function(par, work_exp, edu, age, sex, d2, inc){
  yhat = par[1] + par[2]* work_exp + par[3]* edu + par[4] * age + par[5] * sex
  like = (1-d2) * log(1 - pnorm((100000 - yhat)/exp(par[7]))) + d2*log(dnorm((inc-yhat)/exp(par[7]))/exp(par[7]))
  return(-sum(like,na.rm=T))
}

start =  par + runif(6,-1,1)
```

```
results <- optim(start, fn = tlikefunc, method = "BFGS",
                 control = list(trace = 6, maxit = 3000),
                 work_exp = work_exp, edu = edu, age = age, sex = sex, d2 = d2, inc
 = inc)
results$par

```

```
initial  value -0.000000
final  value -0.000000
converged
> results$par
[1]  20524.76250   1183.41274    449.09457    657.10549 -14015.17576     10.83978
```

```
# not correcting for the censored data
ols <- lm(inc ~ work_exp + edu + age +sex, data = data_3.2)
summary(ols)
ols$coefficients
results$par

```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25150.3     9648.2   2.607  0.00917 **
work_exp      1100.0       66.8  16.468  < 2e-16 ***
edu            396.4       21.8  18.186  < 2e-16 ***
age            508.3      258.2   1.969  0.04904 *
sex         -12590.9      717.9 -17.540  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26370 on 5407 degrees of freedom
  (3572 observations deleted due to missingness)
Multiple R-squared:  0.1555,    Adjusted R-squared:  0.1549
F-statistic: 248.9 on 4 and 5407 DF,  p-value: < 2.2e-16

> ols$coefficients
(Intercept)    work_exp         edu         age         sex
 25150.3154   1100.0225    396.4440    508.2759 -12590.8694
```

```
# 3.4
# we can find the absolute value of the coefficient of each variable is smaller if
 we ignore the censoring problem
# it means that we underestimate the effect.

# 4
# 4.1
# Correlations between variables can lead to selection problems.
# For example, people with better educational background may have better family bac
kground and talents, which makes them more popular in the marriage market and easie
r to obtain better income.
# In addition, good marital status may allow people to focus more on work, leading
 to higher earnings.
# These situations can lead to the selection problem.

# 4.2
# data preparing
```

```
data4 <- dat_A4_panel

# it is hard to conduct based on initial data
# so i need to convert the wide data to long
# firstly i need to rename the variables
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_1998"] <-
  "CV_HIGHEST_DEGREE_9899_1998"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_1999"] <-
  "CV_HIGHEST_DEGREE_9900_1999"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_2000"] <-
  "CV_HIGHEST_DEGREE_0001_2000"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_2001"] <-
  "CV_HIGHEST_DEGREE_0102_2001"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_2002"] <-
  "CV_HIGHEST_DEGREE_0203_2002"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_2003"] <-
  "CV_HIGHEST_DEGREE_0304_2003"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_2004"] <-
  "CV_HIGHEST_DEGREE_0405_2004"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_2005"] <-
  "CV_HIGHEST_DEGREE_0506_2005"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_2006"] <-
  "CV_HIGHEST_DEGREE_0607_2006"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_2007"] <-
  "CV_HIGHEST_DEGREE_0708_2007"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_2008"] <-
  "CV_HIGHEST_DEGREE_0809_2008"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_2009"] <-
  "CV_HIGHEST_DEGREE_0910_2009"
names(data4)[names(data4) =="CV_HIGHEST_DEGREE_EVER_EDT_2010"] <-
  "CV_HIGHEST_DEGREE_1011_2010"

library(panelr)
data4 <- long_panel(
  data4,
  prefix = '_',
  suffix = NULL,
  begin = 1997,
  end = 2019,
  id = "id",
  wave = "wave",
  periods = NULL,
  label_location = "end",
  as_panel_data = TRUE,
```

```
  match = ".*",
  use.regex = FALSE,
  check.varying = TRUE
)

data4 <- data4 %>%
  filter(wave!='2012') %>%
  filter(wave!='2014') %>%
  filter(wave!='2016') %>%
  filter(wave!='2018')

# new variables
# work_experience
we <- data4[,c(10:16,20:27)]
we <- as.matrix(we)
we[is.na(we)] <- 0
we <- we[,3:17]
we <- as.data.frame(we)
for (i in 1:14) {
  we[,i]<-as.numeric(we[,i])
}

data4$work_exp <- rowSums(we[,1:14])/52
data4$work_exp <- round(data4$work_exp,2)

# age
data4$age <- data4$wave - data4$KEY_BDATE_Y

# between estimator: gender, work_exp, edu, marital status
m_gender <- data4 %>% group_by(id) %>% summarize(m_gender=mean(KEY_SEX,na.rm = TRUE))
m_work_exp  <- data4 %>% group_by(id) %>% summarize(m_work_exp=mean(work_exp,na.rm = TRUE))
m_ms  <- data4 %>% group_by(id) %>% summarize(m_ms=mean(CV_MARSTAT_COLLAPSED,na.rm = TRUE))

data4$'YINC-1700' <- as.numeric(data4$'YINC-1700')
m_inc <- data4 %>% group_by(id) %>% summarize(m_inc=mean(`YINC-1700`,na.rm = TRUE))
data4$CV_HIGHEST_DEGREE_EVER_EDT <- as.numeric(data4$CV_HIGHEST_DEGREE_EVER_EDT)
m_edu <- data4 %>% group_by(id) %>% summarize(m_edu=mean(CV_HIGHEST_DEGREE_EVER_EDT,na.rm = TRUE))

between <- lm(m_inc$m_inc~ m_edu$m_edu + m_work_exp$m_work_exp + m_ms$m_ms)
```

```r
summary(between)

```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           7647.95     387.56  19.734  < 2e-16 ***
m_edu$m_edu           3446.72     115.91  29.735  < 2e-16 ***
m_work_exp$m_work_exp 2302.74      96.63  23.830  < 2e-16 ***
m_ms$m_ms             1958.12     335.18   5.842 5.36e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14510 on 7998 degrees of freedom
  (982 observations deleted due to missingness)
Multiple R-squared:  0.1943,	Adjusted R-squared:  0.194
F-statistic: 643.1 on 3 and 7998 DF,  p-value: < 2.2e-16
```

```r
# Within Estimator: exper, edu, marital status
data4 <- left_join(data4,m_inc,by = "id")
data4 <- left_join(data4,m_edu,by='id')
data4 <- left_join(data4,m_work_exp,by='id')
data4 <- left_join(data4,m_ms,by='id')

data4$d_inc <- data4$'YINC-1700'-data4$m_inc
data4$d_edu <- data4$CV_HIGHEST_DEGREE_EVER_EDT - data4$m_edu
data4$d_work_exp <- data4$work_exp - data4$m_work_exp
data4$d_marital_status <- data4$CV_MARSTAT_COLLAPSED - data4$m_ms

within <- lm(d_inc ~ d_edu + d_work_exp + d_marital_status, data = data4)
summary(within)

```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      12749.84     256.98   49.61   <2e-16 ***
d_edu             9815.15     674.39   14.55   <2e-16 ***
d_work_exp        1405.78      49.28   28.53   <2e-16 ***
d_marital_status  4108.70     302.45   13.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27990 on 25813 degrees of freedom
  (144879 observations deleted due to missingness)
Multiple R-squared:  0.04942,	Adjusted R-squared:  0.04931
F-statistic: 447.3 on 3 and 25813 DF,  p-value: < 2.2e-16
```

```r
# first difference estimator
dataf <- data4
dataf$first_inc <- ave(dataf$'YINC-1700', dataf$id, FUN=function(x)dplyr::lag(x))
dataf$first_work_exp <- ave(dataf$work_exp, dataf$id, FUN=function(x)dplyr::lag(x))

dataf$first_edu <- ave(dataf$CV_HIGHEST_DEGREE_EVER_EDT, dataf$id, FUN=function(x)dplyr::lag(x))
dataf$first_ms <- ave(dataf$CV_MARSTAT_COLLAPSED, dataf$id, FUN=function(x)dplyr::lag(x))

```

```r
dataf$fd_inc <- dataf$'YINC-1700' - dataf$first_inc
dataf$fd_edu  <- dataf$CV_HIGHEST_DEGREE_EVER_EDT - dataf$first_edu
dataf$fd_work_exp <- dataf$work_exp - dataf$first_work_exp
dataf$fd_ms <- dataf$CV_MARSTAT_COLLAPSED - dataf$first_ms

fd <- lm(fd_inc ~ fd_work_exp + fd_edu + fd_ms, data = dataf)
summary(fd)

```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6190.28     207.64  29.813  < 2e-16 ***
fd_work_exp   459.25      69.77   6.582 4.77e-11 ***
fd_edu       -486.66     698.79  -0.696    0.486
fd_ms         232.11     419.11   0.554    0.580
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24660 on 16387 degrees of freedom
  (154305 observations deleted due to missingness)
Multiple R-squared:  0.002681,	Adjusted R-squared:  0.002498
F-statistic: 14.68 on 3 and 16387 DF,  p-value: 1.517e-09
```

```r
# 4.3
# all variables in between and within estimator are significant and positive.
# the results show that when other condition fixed, the increase of the increase leading to increase of income
# for example, the coefficient of work experience in between estimator is 1958, which means, when all conditions fixed, if the mean of work experience increase one year, the mean of income increase $1958
# two of variables in first estimator are not significant, it confused me.
# the difference may be caused by different method of dealing with NA
# This difference is more pronounced in the process of the first difference estimator, which may be the reason why this model is not significant
```