A4_Tian

Tian Chen

2022/04/29

Set up environment and import the Data

```
rm(list= ls())
require(tidyverse)
```

```
## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
require(hrbrthemes)
```

```
## Loading required package: hrbrthemes

## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.

##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and

##       if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow
```

```
dat_A4_panel <- read.csv("~/Desktop/Duke study/Econ613/A4/Data/dat_A4_panel.csv")
dat_A4 <- read.csv("~/Desktop/Duke study/Econ613/A4/Data/dat_A4.csv")
```

Exercise 1 Preparing the Data

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

```
dat_A4_tbl <- dat_A4 %>% as_tibble()
#creating age
dat_A4_tbl <- dat_A4_tbl %>% mutate(age = 2019 - KEY_BDATE_Y_1997)
#creating total work experience
work_name <- names(select(dat_A4_tbl,contains("CV_WKSWK_JOB_DLI")))
dat_A4_tbl <- dat_A4_tbl %>% rowwise(X) %>% mutate(work_exp = sum(c_across(work_name[1]:work_name[11]),
select(dat_A4_tbl, work_exp,age)
```

```
## # A tibble: 8,984 x 2
##    work_exp   age
##       <dbl> <dbl>
```

```
## 1      0       38
## 2     12.4     37
## 3      1.69    36
## 4      1.92    38
## 5     13.5     37
## 6      2.25    37
## 7      2.37    36
## 8      4.19    38
## 9      3.23    37
## 10     5.08    35
## # ... with 8,974 more rows
```

2.Create additional education variables indicating total years of schooling from all variables related to education (eg, "BIOLOGICAL FATHERS HIGHEST GRADE COMPLETED") in our dataset.

```
educ_names <- names(select(dat_A4_tbl, contains("CV_HGC")))
dat_A4_tbl$CV_HGC_BIO_DAD_1997[dat_A4_tbl$CV_HGC_BIO_DAD_1997 == 95] <- 0
dat_A4_tbl$CV_HGC_BIO_MOM_1997[dat_A4_tbl$CV_HGC_BIO_MOM_1997 == 95] <- 0
dat_A4_tbl$CV_HGC_RES_DAD_1997[dat_A4_tbl$CV_HGC_RES_DAD_1997 == 95] <- 0
dat_A4_tbl$CV_HGC_RES_MOM_1997[dat_A4_tbl$CV_HGC_RES_MOM_1997 == 95] <- 0
#change grade to numeric
dat_A4_tbl$YSCH.3113_2019[dat_A4_tbl$YSCH.3113_2019 == 1] <- 0 #None
dat_A4_tbl$YSCH.3113_2019[dat_A4_tbl$YSCH.3113_2019 == 2] <- 4 #GED
dat_A4_tbl$YSCH.3113_2019[dat_A4_tbl$YSCH.3113_2019 == 3] <- 12 #High
dat_A4_tbl$YSCH.3113_2019[dat_A4_tbl$YSCH.3113_2019 == 4] <- 14 #AA
dat_A4_tbl$YSCH.3113_2019[dat_A4_tbl$YSCH.3113_2019 == 5] <- 16 #BA
dat_A4_tbl$YSCH.3113_2019[dat_A4_tbl$YSCH.3113_2019 == 6] <- 18 #MA
dat_A4_tbl$YSCH.3113_2019[dat_A4_tbl$YSCH.3113_2019 == 7] <- 23#PhD
dat_A4_tbl$YSCH.3113_2019[dat_A4_tbl$YSCH.3113_2019 == 8] <- 22#JD,MD
dat_A4_tbl <- dat_A4_tbl %>% rowwise(X) %>% mutate(education = sum(c_across(educ_names[1]:educ_names[4]
select(dat_A4_tbl, education)
```

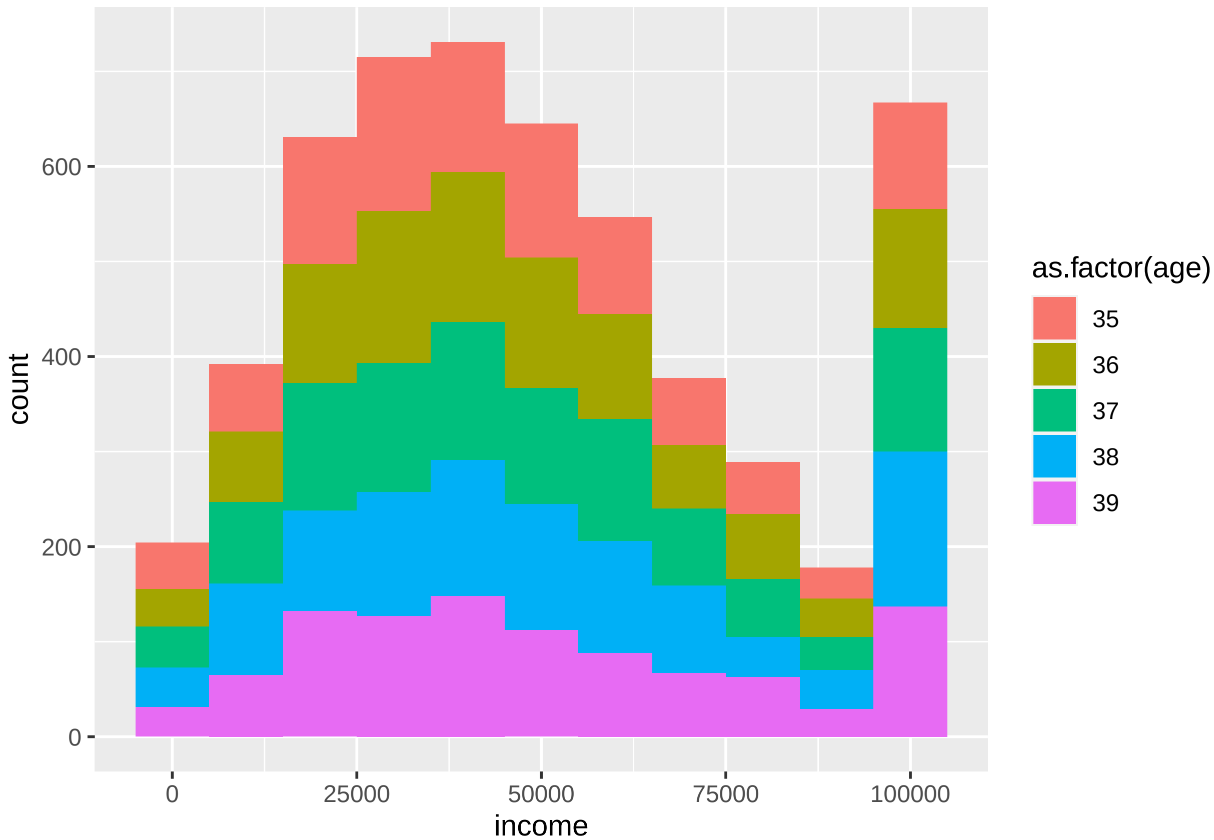
```
## Adding missing grouping variables: `X`
```

```
## # A tibble: 8,984 x 2
## # Rowwise:  X
##         X education
##     <int>     <dbl>
## 1      1        NA
## 2      2        73
## 3      3        40
## 4      4        48
## 5      5        60
## 6      6        36
## 7      7        24
## 8      8        52
## 9      9        54
## 10    10        54
## # ... with 8,974 more rows
```

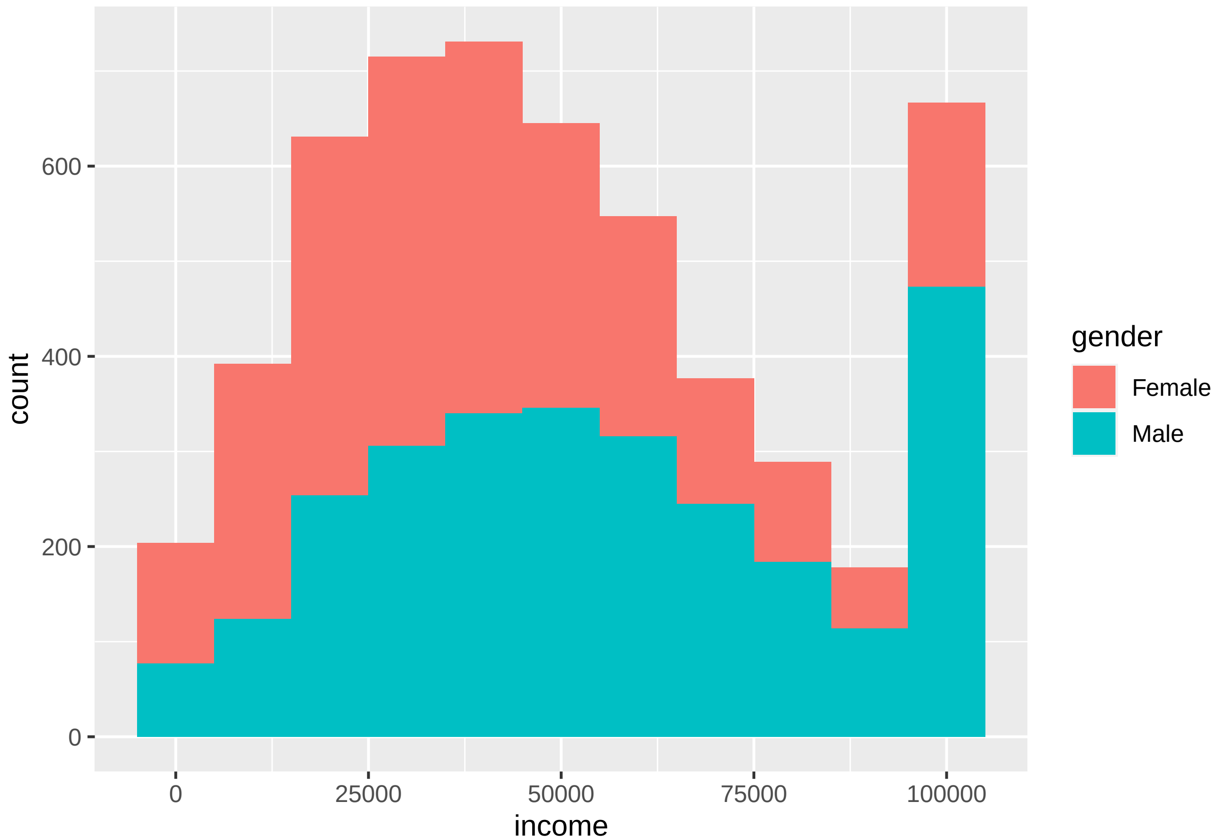
3.Provide the following visualizations 3.1. Plot the income data (where income is positive) by i) age groups, ii) gender groups and iii) number of children.

```
dat_A4_tbl <- dat_A4_tbl %>% mutate(gender = ifelse(KEY_SEX_1997 == 1, "Male", ifelse(KEY_SEX_1997 == 2
#age group
graph_age <- dat_A4_tbl %>% filter(YINC_1700_2019>0) %>% ggplot(aes(x = YINC_1700_2019, fill = as.facto
graph_gender <- dat_A4_tbl %>% filter(YINC_1700_2019>0) %>% ggplot(aes(x = YINC_1700_2019, fill = gende
```

```
graph_child <- dat_A4_tbl %>% filter(YINC_1700_2019>0) %>% ggplot(aes(x = YINC_1700_2019, fill = as.fac
graph_age
```

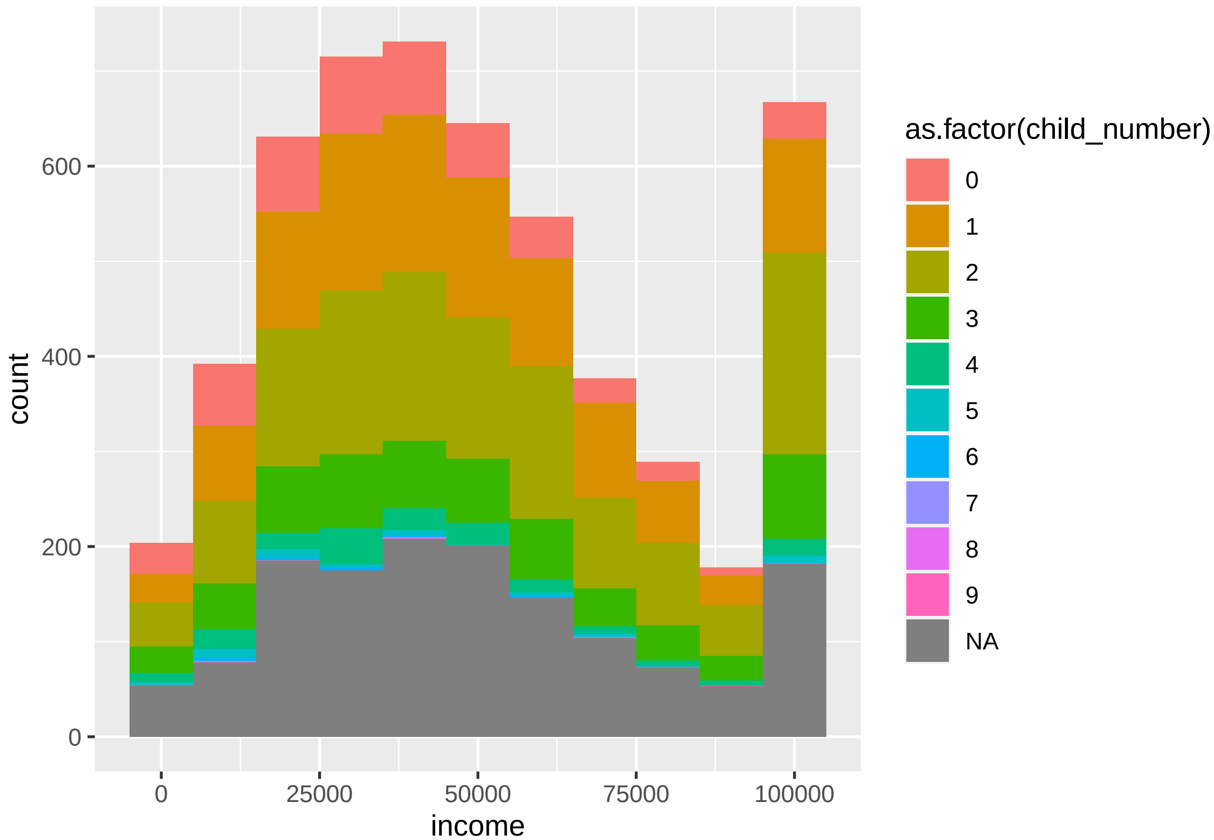



```
graph_gender
```



600
400
200
0
count
0
25000
50000
75000
100000
income
gender
Female
Male


```
graph_child
```



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

```
dat_A4_tbl <- dat_A4_tbl %>% mutate(zero_income = ifelse(YINC_1700_2019 == 0, 1, 0))

zero_income_age <- dat_A4_tbl %>% group_by(age) %>% summarize(zero_income_share = sum(zero_income, na.r
zero_income_age
```

```
## # A tibble: 5 x 2
##     age zero_income_share
##   <dbl>             <dbl>
## 1    35           0.00565
## 2    36           0.00387
## 3    37           0.00326
## 4    38           0.00534
## 5    39           0.00177
```

```
zero_income_gender <- dat_A4_tbl %>% group_by(gender) %>% summarize(zero_income_share = sum(zero_income
zero_income_gender
```

```
## # A tibble: 2 x 2
##   gender zero_income_share
##   <chr>              <dbl>
## 1 Female           0.00342
## 2 Male             0.00457
```

```
zero_income_age <- dat_A4_tbl  %>% group_by(child_number) %>% summarize(zero_income_share = sum(zero_in
zero_income_age
```

```
## # A tibble: 11 x 2
```

```
##    child_number zero_income_share
##           <int>             <dbl>
##  1            0           0.00966
##  2            1           0.00641
##  3            2           0.00463
##  4            3           0.00624
##  5            4           0
##  6            5           0
##  7            6           0
##  8            7           0
##  9            8           0
## 10            9           0
## 11           NA           0.00156
```

From those graphs, we can find that people with older age, people who are male and people who with higher number of children are more likely to get higher income. Also, the table shows that younger people, male and people with children are more likely to have zero income. ## Exercise 2 Heckman Selection Model

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

```
dat_regression <- dat_A4_tbl %>% ungroup() %>% select(YINC_1700_2019 ,age, work_exp, education, child_nu
dat_OLS <- dat_regression %>% drop_na() %>% filter(income > 0)
X <- as.matrix(select(dat_OLS, !income))
one <- rep(1, length(dat_OLS$income))
dim(one) <- c(length(dat_OLS$income), 1)
X <- cbind(one, X)
Y <- as.matrix(dat_OLS$income)
beta <- solve(t(X)%*%X)%*%t(X)%*%Y
rownames(beta) <- c("Intercept", "age", "work_exp", "education", "child_number", "gender_dummy")
colnames(beta) <- c("Coefficient")
beta
```

```
##              Coefficient
## Intercept    -14229.0553
## age             536.5608
## work_exp       1168.8565
## education       451.5384
## child_number   1659.7344
## gender_dummy  17160.0377
```

2.1.1 Interpret the estimation results The income is positively correlated with age, working experience and education level, child's number and gender. Whenever a person has higher age, more working experience, more children and higher education level of his or her parents and if the gender of such a person is male, the person will on average get a higher income. Among all of those dependent variables, working experience plays the most significant role.

2.1.2 Explain why there might be a selection bias when estimating an OLS this way In the data cleaning process, we removed many those respondents with NAs, making the sample non-random. In other words, the sample we use for the OLS regression includes only those reporting their income, who do not represent the total population. As a result, the OLS estimates will be biased by unobserved omitted variables we do not include into the the OLS estimation. Also, the incomparable control groups here makes the interpretation of estimates impossible.

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

In the Heckman model, we firstly use the probit model to calculate the probability that an individual is observed (inverse mills ratio or IMR) in the first-stage equation. Then we use the IMR as control variable in the second-stage equation. As a result, the IMR can correct the sample selection problem.

2.3 Estimate a Heckman selection model

```r
#we firstly calculate the dummy for sample observation
dat_heckman <- dat_regression
dat_heckman$income[is.na(dat_heckman$income)] <- -1
dat_heckman <- dat_heckman %>% mutate(y_dummy = ifelse(income>0, 1, 0)) %>% drop_na()

#First stage equation
#we choose age, working experience, education, the number of children, education, and gender as indepen
#we use the probit model to estimate the IMR
first_stage <- glm(y_dummy ~ age + work_exp + education + child_number + gender_dummy, family = binomia
    data = dat_heckman)
summary(first_stage)
```

```
##
## Call:
## glm(formula = y_dummy ~ age + work_exp + education + child_number +
##     gender_dummy, family = binomial(link = "probit"), data = dat_heckman)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8718   0.0792   0.4654   0.7760   1.5908
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.163897   0.565550  -2.058   0.0396 *
## age           0.016317   0.015082   1.082   0.2793
## work_exp      0.122433   0.005591  21.897  < 2e-16 ***
## education     0.010145   0.001132   8.963  < 2e-16 ***
## child_number  0.041657   0.017507   2.379   0.0173 *
## gender_dummy  0.261190   0.044180   5.912 3.38e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5592.8  on 5120  degrees of freedom
## Residual deviance: 4617.0  on 5115  degrees of freedom
## AIC: 4629
##
## Number of Fisher Scoring iterations: 7
```

```r
dat_heckman <- dat_heckman %>% mutate(predict_y = predict(first_stage, dat_heckman))
#calculate the inverse mills ratio
dat_heckman <- dat_heckman %>% mutate(IMR = dnorm(predict_y)/pnorm(predict_y))

#second stage equation (OLS)
X <- as.matrix(select(dat_heckman, !c(income,predict_y, y_dummy)))
one <- rep(1, length(dat_heckman$income))
dim(one) <- c(length(dat_heckman$income), 1)
X <- cbind(one, X)
Y <- as.matrix(dat_heckman$income)
beta <- solve(t(X)%*%X)%*%t(X)%*%Y
#rownames(beta) <- c("Intercept", "age", "work_exp", "education", "child_number", "gender_dummy")
#colnames(beta) <- c("Coefficient")
```

```r
beta
```

```
##                    [,1]
##              54578.8038
## age            233.7859
## work_exp      -859.0873
## education      109.2808
## child_number   660.9471
## gender_dummy  8358.3946
## IMR         -76643.0681
```

```r
summary(lm(income~age+work_exp+education+child_number+gender_dummy+IMR, data = dat_heckman))
```

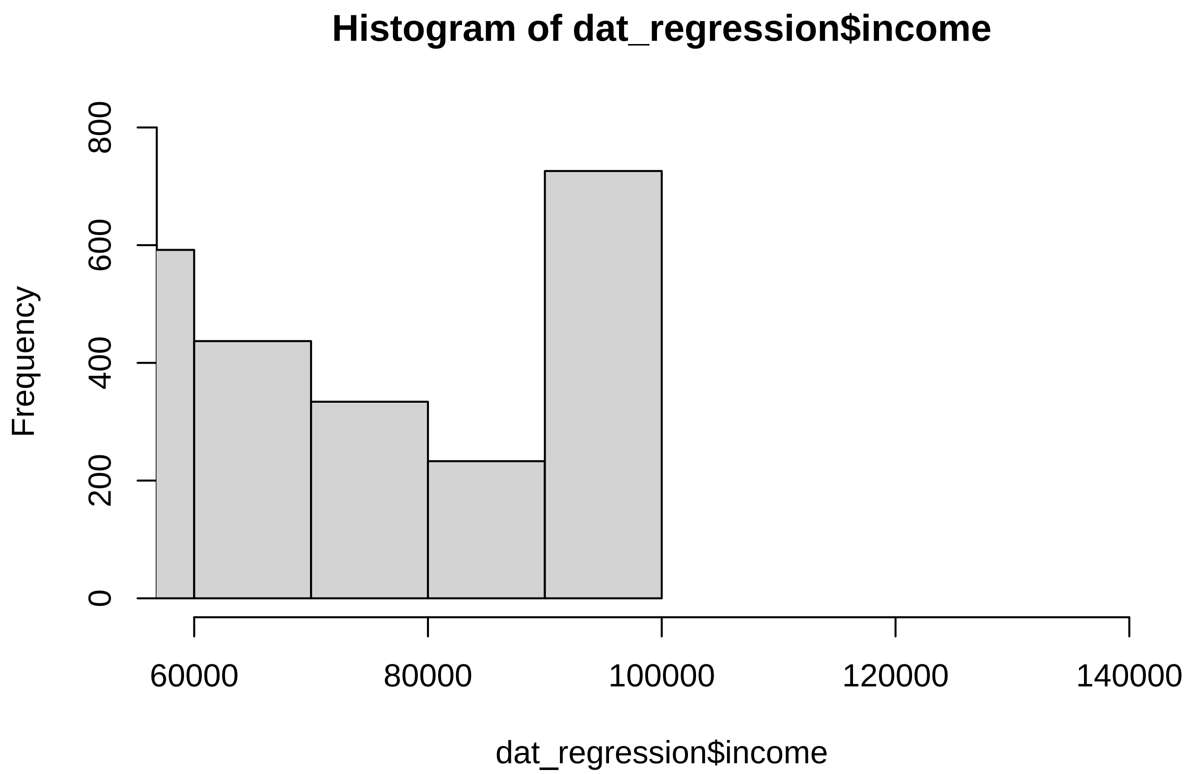
```
##
## Call:
## lm(formula = income ~ age + work_exp + education + child_number +
##     gender_dummy + IMR, data = dat_heckman)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -69804 -18572  -2439  17222  97200
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54578.80   11289.48   4.834 1.37e-06 ***
## age            233.79     270.69   0.864 0.387805
## work_exp      -859.09     187.15  -4.590 4.53e-06 ***
## education      109.28      29.05   3.762 0.000171 ***
## child_number   660.95     327.07   2.021 0.043351 *
## gender_dummy  8358.39     917.10   9.114  < 2e-16 ***
## IMR         -76643.07    4561.59 -16.802  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26770 on 5114 degrees of freedom
## Multiple R-squared:  0.3348, Adjusted R-squared:  0.334
## F-statistic:   429 on 6 and 5114 DF,  p-value: < 2.2e-16
```

Interpret: As we can see, the coefficient of IMR is negative and significant, indicating that there exists the sample selection bias. After heckman selection, the working experience is negatively associated with the income. For other variables, though the coefficients of them are still positive, but their impacts become smaller.

Exercise 3 Censoring

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

```r
hist(dat_regression$income,xlim = range(100000))
```



As we can see, the censopred value here is 100,000,

3.2 Propose a model to deal with the censoring problem I would like to use the heckman model as well. As Heckman argues, the censored data can be considered as a case of selcetion problem.

3.3 Estimate the appropriate model with the censored data

```
dat_heckman2 <- dat_heckman %>% filter(income > 0)
#we create censored dummy (if income)
dat_heckman2 <- dat_heckman2 %>% mutate(y_dummy2 = ifelse(income == 100000, 0, 1))
#calculate the heckman IMR for censored data
first_stage <- glm(y_dummy2 ~ age + work_exp + education + child_number + gender_dummy, family = binomia
    data = dat_heckman2)
summary(first_stage)
```

```
##
## Call:
## glm(formula = y_dummy2 ~ age + work_exp + education + child_number +
##     gender_dummy, family = binomial(link = "probit"), data = dat_heckman2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1991   0.2041   0.3563   0.5361   1.4418
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   6.047157   0.779263   7.760 8.49e-15 ***
## age          -0.073881   0.020602  -3.586 0.000336 ***
## work_exp     -0.025727   0.005064  -5.080 3.77e-07 ***
## education    -0.022653   0.001780 -12.725  < 2e-16 ***
## child_number -0.098691   0.024909  -3.962 7.43e-05 ***
## gender_dummy -0.706336   0.059789 -11.814  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2853.6  on 3913  degrees of freedom
## Residual deviance: 2445.2  on 3908  degrees of freedom
## AIC: 2457.2
##
## Number of Fisher Scoring iterations: 6
```

```r
dat_heckman2 <- dat_heckman2 %>% mutate(predict_y2 = predict(first_stage, dat_heckman2))
#calculate the inverse mills ratio
dat_heckman2 <- dat_heckman2 %>% mutate(IMR2 = dnorm(predict_y2)/pnorm(predict_y2))

#second stage equation (OLS)
X <- as.matrix(select(dat_heckman2, !c(income,predict_y, y_dummy, y_dummy2, predict_y2, IMR)))
one <- rep(1, length(dat_heckman2$income))
dim(one) <- c(length(dat_heckman2$income), 1)
X <- cbind(one, X)
Y <- as.matrix(dat_heckman2$income)
beta <- solve(t(X)%*%X)%*%t(X)%*%Y
#rownames(beta) <- c("Intercept", "age", "work_exp", "education", "child_number", "gender_dummy")
#colnames(beta) <- c("Coefficient")
beta
```

```
##                    [,1]
##              20780.5398
## age           -146.8791
## work_exp       925.6842
## education      269.9077
## child_number   665.8743
## gender_dummy 10995.6898
## IMR2         31040.5227
```

3.4 compare the results

```r
#ols
summary(lm(income ~ age + work_exp + education + child_number +
    gender_dummy, data = dat_heckman2))
```

```
##
## Call:
## lm(formula = income ~ age + work_exp + education + child_number +
##     gender_dummy, data = dat_heckman2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75631 -17717  -2965  17543  75212
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -14229.06   10774.51  -1.321   0.1867
## age             536.56     287.76   1.865   0.0623 .
## work_exp       1168.86      76.14  15.352  < 2e-16 ***
## education       451.54      22.12  20.415  < 2e-16 ***
```

```
## child_number   1659.73     354.43   4.683 2.93e-06 ***
## gender_dummy  17160.04     810.56  21.171  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24960 on 3908 degrees of freedom
## Multiple R-squared:  0.2426, Adjusted R-squared:  0.2416
## F-statistic: 250.3 on 5 and 3908 DF,  p-value: < 2.2e-16
```

```
summary(lm(income ~ age + work_exp + education + child_number +
    gender_dummy + IMR2, data = dat_heckman2))
```

```
##
## Call:
## lm(formula = income ~ age + work_exp + education + child_number +
##     gender_dummy + IMR2, data = dat_heckman2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83784 -17662  -3069  17047  72201
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20780.54   13391.89   1.552    0.121
## age           -146.88     326.70  -0.450    0.653
## work_exp       925.68      94.06   9.841  < 2e-16 ***
## education      269.91      46.95   5.749 9.65e-09 ***
## child_number   665.87     420.06   1.585    0.113
## gender_dummy 10995.69    1622.24   6.778 1.40e-11 ***
## IMR2         31040.52    7081.46   4.383 1.20e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24900 on 3907 degrees of freedom
## Multiple R-squared:  0.2463, Adjusted R-squared:  0.2451
## F-statistic: 212.8 on 6 and 3907 DF,  p-value: < 2.2e-16
```

As we can see, after the Heckman correction, the coefficient of age becomes negative. All other variables' coefficient are still positive but their impacts on income become smaller. Also, the significance of IMR indicats the censored data cause selection bias in the OLS regression.

Exercise 4 Panel Data

4.1 Explain the potential ability bias.

Ability bias indicates that the income returns to education may be caused by people's ability. Those people with the higher ability (or IQ) are more likely to have both higher educational levels and income because of their innate ability. In other words, even with less education, they can also earn more money than other people. In the OLS regression, without controlling for the ability, the estimate will be overestimated because the ability is positively correlated with both education and income.

4.2 Exploit the panel dimension of the data to propose a model to correct for the ability bias. Estimate the model using the following strategy.

Import data

```r
#education, marital status, experience, wages
#education: CV_HGC
#experience: CV_WKSWK_JOB_DLI
#wages: YINC_1700
#marital status: CV_MARSTAT
#id:X
dat_tbl <- dat_A4_panel %>% as_tibble()
#education
education_long <-dat_tbl %>% select(X,contains("CV_HIGHEST_DEGREE")) %>% pivot_longer(!X, names_to = "ye
year_name <- unique(education_long$year)
for (name in year_name) {
    x <- name
    year <- str_sub(x, -4, -1)
    education_long$year[education_long$year == x] <- year
}
education_long
```

```
## # A tibble: 188,664 x 3
##        X year  education
##    <int> <chr>     <int>
##  1     1 1998          0
##  2     1 1999          2
##  3     1 2000          2
##  4     1 2001          2
##  5     1 2002          2
##  6     1 2003          4
##  7     1 2004          4
##  8     1 2005          4
##  9     1 2006          4
## 10     1 2007          4
## # ... with 188,654 more rows
```

```r
#wages
wage_long<- dat_tbl %>% select(X, contains("YINC.1700")) %>% pivot_longer(!X, names_to = "year", values_
year_name <- unique(wage_long$year)
for (name in year_name) {
    x <- name
    year <- str_sub(x, -4, -1)
    wage_long$year[wage_long$year == x] <- year
}
wage_long
```

```
## # A tibble: 170,696 x 3
##        X year   wage
##    <int> <chr> <int>
##  1     1 1997     NA
##  2     1 1998    475
##  3     1 1999     NA
##  4     1 2000   8000
##  5     1 2001   7000
##  6     1 2002   8000
##  7     1 2003  15000
##  8     1 2004     NA
##  9     1 2005  10000
## 10     1 2006  80471
```

```
## # ... with 170,686 more rows
```

```r
#Marital status
marital_long<- dat_tbl %>% select(X, contains("CV_MARSTAT")) %>% pivot_longer(!X, names_to = "year", val
year_name <- unique(marital_long$year)
for (name in year_name) {
    x <- name
    year <- str_sub(x, -4, -1)
    marital_long$year[marital_long$year == x] <- year
}
marital_long
```

```
## # A tibble: 170,696 x 3
##        X year  marital_status
##    <int> <chr>          <int>
##  1     1 1997              NA
##  2     1 1998               0
##  3     1 1999               0
##  4     1 2000               0
##  5     1 2001               0
##  6     1 2002               0
##  7     1 2003               0
##  8     1 2004               0
##  9     1 2005               0
## 10     1 2006               0
## # ... with 170,686 more rows
```

```r
#Experience
experience_long <- dat_tbl %>% select(X, contains("CV_WKSWK_JOB_DLI"))
experience_long <- experience_long %>% rowwise(X) %>% mutate(work_exp_1997 = sum(c_across(names(select(c
experience_long <- experience_long %>% select(contains("work_exp")) %>% ungroup() %>% pivot_longer(!X,na
```

```
## Adding missing grouping variables: `X`
```

```r
year_name <- unique(experience_long$year)
for (name in year_name) {
    x <- name
    year <- str_sub(x, -4, -1)
    experience_long$year[experience_long$year == x] <- year
}
experience_long
```

```
## # A tibble: 206,632 x 3
##        X year  experience
##    <int> <chr>      <int>
##  1     1 1997           3
##  2     1 1998          72
##  3     1 1999         128
##  4     1 2000          91
##  5     1 2001         221
##  6     1 2002          77
##  7     1 2003          65
##  8     1 2004         121
##  9     1 2005         172
## 10     1 2006         221
## # ... with 206,622 more rows
```

```
dat_panel <- left_join(education_long, wage_long, by = c("X", "year")) %>% left_join(marital_long, by =

#replace the sperated, divorced, widowed to 0
dat_panel$marital_status[dat_panel$marital_status == 2] <- 0
dat_panel$marital_status[dat_panel$marital_status == 3] <- 0
dat_panel$marital_status[dat_panel$marital_status == 4] <- 0

dat_panel
```

```
## # A tibble: 161,867 x 6
##        X year  education   wage marital_status experience
##    <int> <chr>     <int>  <int>          <dbl>      <int>
##  1     1 1998          0    475              0         72
##  2     1 1999          2     NA              0        128
##  3     1 2000          2   8000              0         91
##  4     1 2001          2   7000              0        221
##  5     1 2002          2   8000              0         77
##  6     1 2003          4  15000              0         65
##  7     1 2004          4     NA              0        121
##  8     1 2005          4  10000              0        172
##  9     1 2006          4  80471              0        221
## 10     1 2007          4 112215              0        278
## # ... with 161,857 more rows
```

4.2 Exploit the panel dimension of the data to propose a model to correct for the ability bias. Estimate the model using the following strategy

I propose to use the fixed effect model. Because ability is time-invariant omitted variable, controlling for the individual fixed effect can solve the omitted variable bias.

```
#firstly, we create the year dummy and individual dummy
#there are some duplicated ids in the data, we remove all of them
require(plm)
```

```
## Loading required package: plm

##
## Attaching package: 'plm'

## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

```
unique_id <- dat_panel %>% select(X,year)
dat_panel_unique <- dat_panel %>% filter(!duplicated(unique_id))
PanelData <- pdata.frame(dat_panel_unique, index = c("X", "year"))
model<-wage ~ education + marital_status + experience
#within estimator
within_fe<- plm(model,data = PanelData, model='within', effect='twoways')

#between estimator
between_fe <- plm(model, data = PanelData, model = "between")

#difference estimator
diff_fe <- plm(model, data = PanelData, model = "fd")

require(stargazer)
```

```
## Loading required package: stargazer
##
## Please cite as:
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
stargazer(within_fe, between_fe, diff_fe,type='text',
          column.labels = c("Within","Between","Difference"))
```

```
##
## ==========================================================================================
##                                              Dependent variable:
##               ----------------------------------------------------------------------------
##                                                     wage
##                          Within                    Between                  Difference
##                           (1)                        (2)                       (3)
## ------------------------------------------------------------------------------------------
## education            5,640.627***              5,561.725***              1,307.476***
##                        (107.343)                 (151.713)                 (109.203)
##
## marital_status       6,732.962***              8,190.806***              2,285.990***
##                        (224.187)                 (553.333)                 (224.902)
##
## experience             19.313***                 38.207***                 18.249***
##                         (0.579)                   (1.422)                   (0.569)
##
## Constant                                       3,940.110***              3,896.617***
##                                                  (382.533)                  (68.788)
##
## ------------------------------------------------------------------------------------------
## Observations             82,008                     8,600                     73,408
## R2                       0.064                      0.272                     0.018
## Adjusted R2             -0.046                      0.272                     0.018
## F Statistic    1,668.945*** (df = 3; 73388) 1,071.677*** (df = 3; 8596) 437.532*** (df = 3; 73404)
## ==========================================================================================
## Note:                                                        *p<0.1; **p<0.05; ***p<0.01
```

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

Each model shows that education, marital status and experience are positively correlated with wage, indicating that on average people with higher education, partners and more working experience can get higher income. Why the results from difference estimators is so distinct from others? That's because in the difference estimators, we cannot include the year fixed effects, leading to the estimation bias caused by the time-variant omitted variables.