# Econ 613 - Applied Econometrics - 2022 Spring
# Homework 4

Yican Liu

April 20, 2022

Before I do analysis, I dropped the observations with NA or negative income.

## 1 Exercise 1

### 1.1

I use the current year (2019) minus birth year (KEY _BDATE_Y_1997) to measure the age. I aggregate all "CV_WKSWK_JOB_DLI" to create the total work experience.

### 1.2

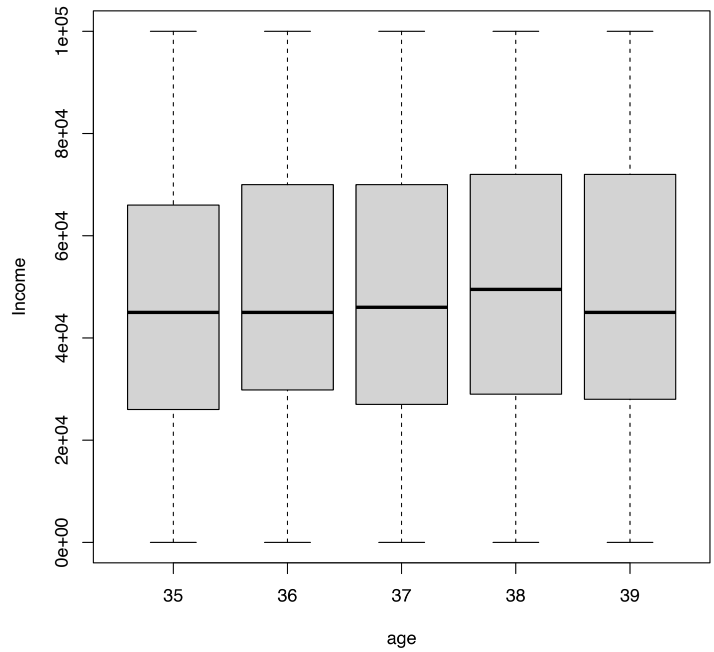
I transfer YSCH.3113_2019 to numerics. The schooling year is transfered as follows:

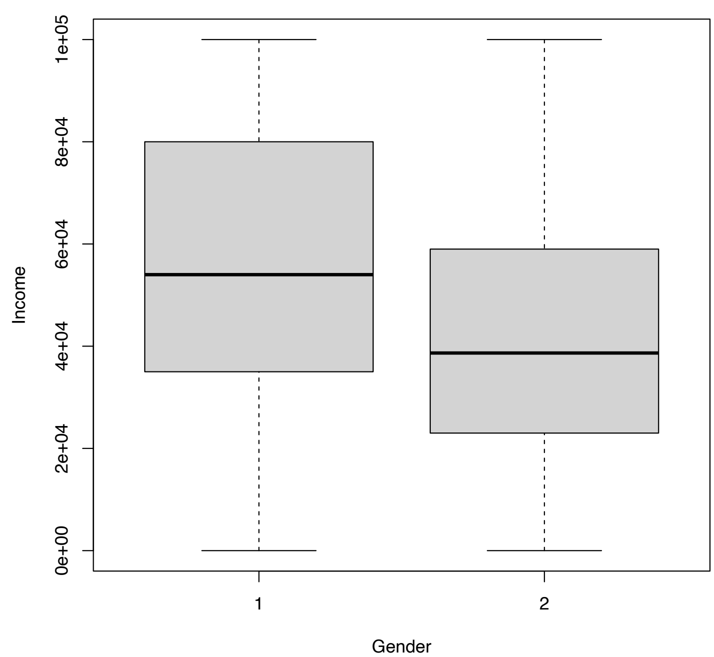
- GED: 4 years.
- High school: 12 years.
- Bachelor: 16 years.
- Master: 18 years.
- Ph.D.: 23 years.
- Professional: 22 years.
- AA: 14 years.

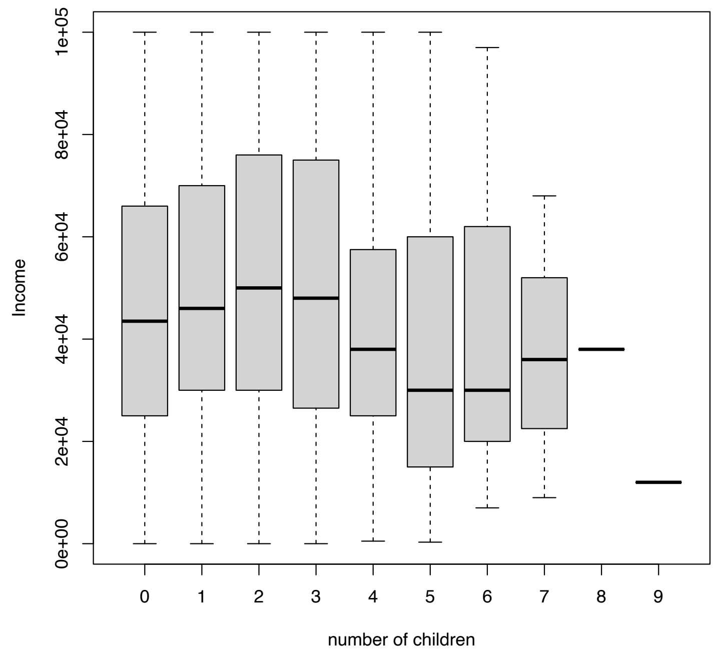
Other variables, such as those for parents education, are not transferred because these variables indicates the total education year. I generate a new variable, *total_edu*, as the transferred variable of YSCH.3113_2019. I use *total_edu* in the folloiwng analysis.

### 1.3

#### 1.3.1 Income by Groups






### 1.3.2 The Share of Zero in Income

The proportion of zero income by age, gender, marital status, and number of children is shown as follows.

| | age | % of zero income |
|---|---|---|
| 1 | 35 | 0.009293680 |
| 2 | 36 | 0.006300630 |
| 3 | 37 | 0.005420054 |
| 4 | 38 | 0.008960573 |
| 5 | 39 | 0.002994012 |

| | gender | % of zero income |
|---|---|---|
| 1 | 0 | 0.007500000 |
| 2 | 1 | 0.005742726 |

| | marital status | % of zero income |
|---|---|---|
| 1 | 0 | 0.005592272 |
| 2 | 1 | 0.007454342 |
| 3 | 2 | 0.043010753 |
| 4 | 3 | 0.001538462 |
| 5 | 4 | 0.000000000 |

| | number of children | % of zero income |
|---|---|---|
| 1 | 0 | 0.006993007 |
| 2 | 1 | 0.007846556 |
| 3 | 2 | 0.005743001 |
| 4 | 3 | 0.008025682 |
| 5 | 4 | 0.000000000 |
| 6 | 5 | 0.000000000 |
| 7 | 6 | 0.000000000 |
| 8 | 7 | 0.000000000 |
| 9 | 8 | 0.000000000 |
| 10 | 9 | 0.000000000 |

#### 1.3.3

To interpret these results, we know the following facts.

- In general, the average incomes of difference age groups (from 35 to 39) are almost the same. The age group of 36 has a slightly higher average income compared to the other age groups. The zero income rate also does not varies a lot across all these five groups.
- Male has a higher average income than female. In addition, female has higher zero income rate compared to male.
- People who have one, two, three, or four childrens tend to earn more than the other people. In addition, those who have no children has the highest zero income rate.

## 2 Exercise 2

### 2.1

#### 2.1.1

The OLS regression is shown as follows

$$\text{Income}_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{WorkExp}_i + \beta_3 \text{Education}_i + \epsilon_i \tag{1}$$

where $i$ indicates the observation (individual). I include age, work experience, and education variable in my model. The regression coefficients are as follows:

```
Call:
lm(formula = data.data4.posinc$YINC_1700_2019 ~ data.data4.posinc$age +
    data.data4.posinc$work_exp + data.data4.posinc$total_edu +
    data.data4.posinc$children + data.data4.posinc$female)

Residuals:
   Min     1Q Median     3Q    Max
-70981 -18433  -3724  16354  89175

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  12765.21    9462.54   1.349    0.177
data.data4.posinc$age          331.17     254.52   1.301    0.193
data.data4.posinc$work_exp     999.16      66.26  15.080  < 2e-16 ***
data.data4.posinc$total_edu   1554.28      76.14  20.413  < 2e-16 ***
data.data4.posinc$children    1323.13     287.47   4.603 4.27e-06 ***
data.data4.posinc$female    -15185.78     722.33 -21.023  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25920 on 5370 degrees of freedom
Multiple R-squared:  0.1728,    Adjusted R-squared:  0.172
F-statistic: 224.4 on 5 and 5370 DF,  p-value: < 2.2e-16
```

In this regression, the regression coefficients of work experience, number of children, work experience, and female are significant. However, the regression coefficient for age are not significant. This is not commonsensible. Thus, there might be biases when estimating the regression coefficient by OLS regression.

To interpret the results, for the people who have positive income, we know that: given everything else fixed, if work experience increase by 1 unit, the personal income will also increase by 999.16 dollar; if personal total education increased by 1 year, the personal income will also increase by 1554.28; if the number of children increase by 1, the personal income will also increase by 1323.13 dollar; and for female individuals, given anything else as controlled, they earn 15185.78 less than male individuals.

### 2.1.2

Selection problem exists in this model because we only focus on the people who has positive income. That is, we dropped the observations with zero income.

### 2.2

In our cases, endogeneity problem exists due to selection biases. Thus, by using Heckman methodology, we include a new variable $\text{IMR}_i$ in our regression analysis to control for the prabability of "having positive income." With the probability estimated by probit model as regressors, we do not need to drop any observations in Heckman model.

### 2.3

I estimate the Heckman Model as follows.

First, I run a Probit model to estimate the probability for whether $\text{Income}_i > 0$.

Second, for each observation $i$, I estimate the value of

$$\text{IMR}_i = \frac{\phi\left(\frac{x_i\beta}{\sigma}\right)}{\Phi(\frac{x_i\beta}{\sigma})}$$

and include this value in the OLS regression (1). The regression coefficient is

```
Call:
lm(formula = data.data4$YINC_1700_2019 ~ data.data4$age + data.data4$work_exp +
    data.data4$total_edu + data.data4$children + data.data4$female +
    heckman.glm.model.imr)

Residuals:
   Min     1Q Median     3Q    Max
-66870 -18491  -3619  16618  93363

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            137031.6    34462.9   3.976 7.09e-05 ***
data.data4$age          -2018.1      677.4  -2.979 0.002901 **
data.data4$work_exp       245.2      210.1   1.167 0.243205
data.data4$total_edu     1025.8      159.0   6.451 1.21e-10 ***
data.data4$children      -204.5      498.5  -0.410 0.681620
data.data4$female      -18705.6     1202.1 -15.561  < 2e-16 ***
heckman.glm.model.imr -1109853.7   290672.5  -3.818 0.000136 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26110 on 5405 degrees of freedom
Multiple R-squared:  0.1723,    Adjusted R-squared:  0.1714
F-statistic: 187.5 on 6 and 5405 DF,  p-value: < 2.2e-16
```

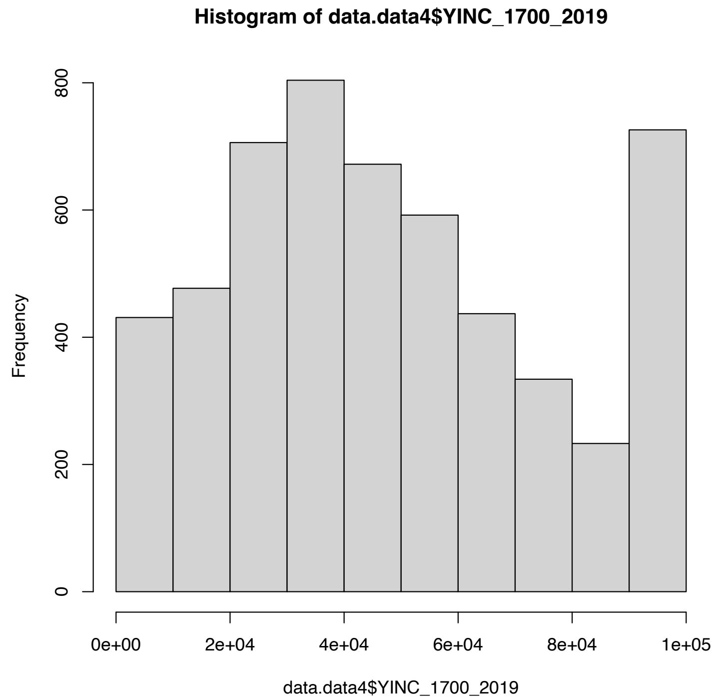
In this table, we can see that in the Heckman model, the regression coefficient for age, personal total education (generated by YSCH.3113_2019), female dummy, as well as a Heckman IMR has significant regression results. Comparing to our previous regression results, we can know that the Heckman IMR variable is an omitted variable in our previous empirical analysis.

In the OLS setting, the regression coefficients for work experience and number of children are significant. However, in the Heckman regression, these two variables are not significant. Age becomes significant in the Heckman regression, while it is not a significant regressor in the previous OLS regression. Given anything else fixed, if the age increase by 1, the personal income will decrease by 2018.1; if personal total education year increased by 1 year, the income will increase by 1025.8; and given everything else fixed, female individuals earn 18705.6 less than male individuals. After controlling for people's willingness to work and earn money (Heckman IMR), the effects from work experience become less significant.

# 3 Exercise 3

## 3.1

The histogram for income variable is as follows




As shown in this figure, the censored value might be 100000.00.

### 3.2

I use Tobit model to deal with such problem (the Tobit model for right-hand side censoring).

### 3.3

I use *optim* command in R to estimate the Tobit model regression coefficient. The results are shown as follows:

```
$par
                                    [,1]
(Intercept)                   12761.8020
data.data4.posinc$age           309.7582
data.data4.posinc$work_exp     1078.9516
data.data4.posinc$total_edu    1666.8389
data.data4.posinc$female     -15341.8900
data.data4.posinc$children     1396.4698
log_sigma                        10.2724

$value
[1] 56180.66

$counts
function gradient
     142      100

$convergence
[1] 1
```

### 3.4

To interpret the results above, we know that given everything else fixed, if age increased by 1, the personal income will increase by 309. On average, female individuals earn 15341.89 less than male individuals.

For comparison, I also use the OLS regression to estimate the regression coefficients. The results are shown as follows:

```
Call:
lm(formula = data.data4.posinc$YINC_1700_2019 ~ data.data4.posinc$age +
    data.data4.posinc$work_exp + data.data4.posinc$total_edu +
    data.data4.posinc$female + data.data4.posinc$children)

Residuals:
   Min     1Q Median     3Q    Max
-68563 -20516  -3870  18420  67533

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                36127.08    9763.34   3.700 0.000218 ***
data.data4.posinc$age        274.00     264.20   1.037 0.299743
data.data4.posinc$work_exp  1163.54      68.31  17.034  < 2e-16 ***
data.data4.posinc$total_edu   33.26      24.38   1.364 0.172479
data.data4.posinc$female  -13649.04     745.78 -18.302  < 2e-16 ***
data.data4.posinc$children  1175.44     298.61   3.936 8.38e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26900 on 5370 degrees of freedom
Multiple R-squared:  0.1089,    Adjusted R-squared:  0.1081
F-statistic: 131.3 on 5 and 5370 DF,  p-value: < 2.2e-16
```

The regression coefficient is slightly difference from the Tobit model.

## 4 Exersice 4

### 4.1

The potentially biases comes from something varies from person to person (for example, personality) and something varies from time to time (for example, macroeconomic conditions), which is unobservable but need to contrl. That is, including fixed effects might solve the potentially biases.

### 4.2

I reorganize the wide panel data to long panel data before I do these regressions.

#### 4.2.1 Within Estimator

I estimate the within estimator as

$$(\text{Income}_{it} - \overline{\text{Income}}_i) = \beta_0 + \beta_1(\text{Edu}_{it} - \overline{\text{Edu}}_i) + \beta_2(\text{WorkExp}_{it} - \overline{\text{WorkExp}}_i) + (\epsilon_{it} - \bar{\epsilon_i})$$

and the regression coefficient is

```
Call:
lm(formula = data.panel$income.demeaned ~ data.panel$edu.demeaned +
    data.panel$work.demeaned)

Residuals:
    Min      1Q  Median      3Q     Max
-110449   -7356       0    5690  304071

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -7.207e-10  4.401e+01     0.0        1
data.panel$edu.demeaned    1.091e+03  1.051e+01   103.8   <2e-16 ***
data.panel$work.demeaned  6.131e+01  3.552e-01   172.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17700 on 161709 degrees of freedom
Multiple R-squared:  0.2929,    Adjusted R-squared:  0.2929
F-statistic: 3.349e+04 on 2 and 161709 DF,  p-value: < 2.2e-16
```

### 4.2.2 Between Estimator

I estimate the between estimator as

$$\overline{\text{Income}}_i = \beta_0 + \beta_1 \overline{\text{Edu}}_i + \beta_2 \overline{\text{WorkExp}}_i + \bar{\epsilon}_i$$

and the regression coefficient is

```
Call:
lm(formula = data.panel.between$mean.income ~ data.panel.between$mean.total.edu +
    data.panel.between$mean.total.work)

Residuals:
   Min     1Q Median     3Q    Max
-38010  -5725  -1168   3195  87129

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                          12.089    216.500   0.056    0.955
data.panel.between$mean.total.edu  1027.447     26.264  39.120   <2e-16 ***
data.panel.between$mean.total.work   50.620      1.293  39.151   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10250 on 8981 degrees of freedom
Multiple R-squared:  0.3772,    Adjusted R-squared:  0.3771
F-statistic:  2720 on 2 and 8981 DF,  p-value: < 2.2e-16
```

### 4.2.3 Difference Estimator

I estiamte the difference estimator as

$$(\text{Income}_{it} - \text{Income}_{it-1}) = \beta_0 + \beta_1(\text{Edu}_{it} - \text{Edu}_{it-1}) + \beta_2(\text{WorkExp}_{it} - \text{WorkExp}_{it-1}) + (\epsilon_{it} - \epsilon_{it-1})$$

and the regression coefficient is

```
Call:
lm(formula = data.panel$income.diff ~ data.panel$total.edu.diff +
    data.panel$total.work.diff)

Residuals:
    Min      1Q  Median      3Q     Max
-236253   -3574   -1143    2641  345176

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1142.8981    43.0296   26.56   <2e-16 ***
data.panel$total.edu.diff   761.5424    11.5574   65.89   <2e-16 ***
data.panel$total.work.diff   36.8401     0.4115   89.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16650 on 152725 degrees of freedom
  (8984 observations deleted due to missingness)
Multiple R-squared:  0.1153,    Adjusted R-squared:  0.1153
F-statistic:  9950 on 2 and 152725 DF,  p-value: < 2.2e-16
```