Econ 613 HW 4
4/20/2022

Exercise 1 Preparing the Data

1. Create additional variable for the age of the agent “age”, total work experience measured in years “work exp”

```
library(tidyverse)
library(readr)
library(dplyr)
library(data.table)
library(ggplot2)
library(VGAM)
library(AER)
library(panelr)
library(plm)

#Assume the final year of survey is 2019
dat <- dat %>% mutate(age = 2019 - KEY_BDATE_Y_1997)
#Assume there are 52 weeks in a year
dat$work_exp <- rowSums(dat[,18:28], na.rm = TRUE) / 52
```

2. Create additional education variables indicating total years of schooling from all variables related to education

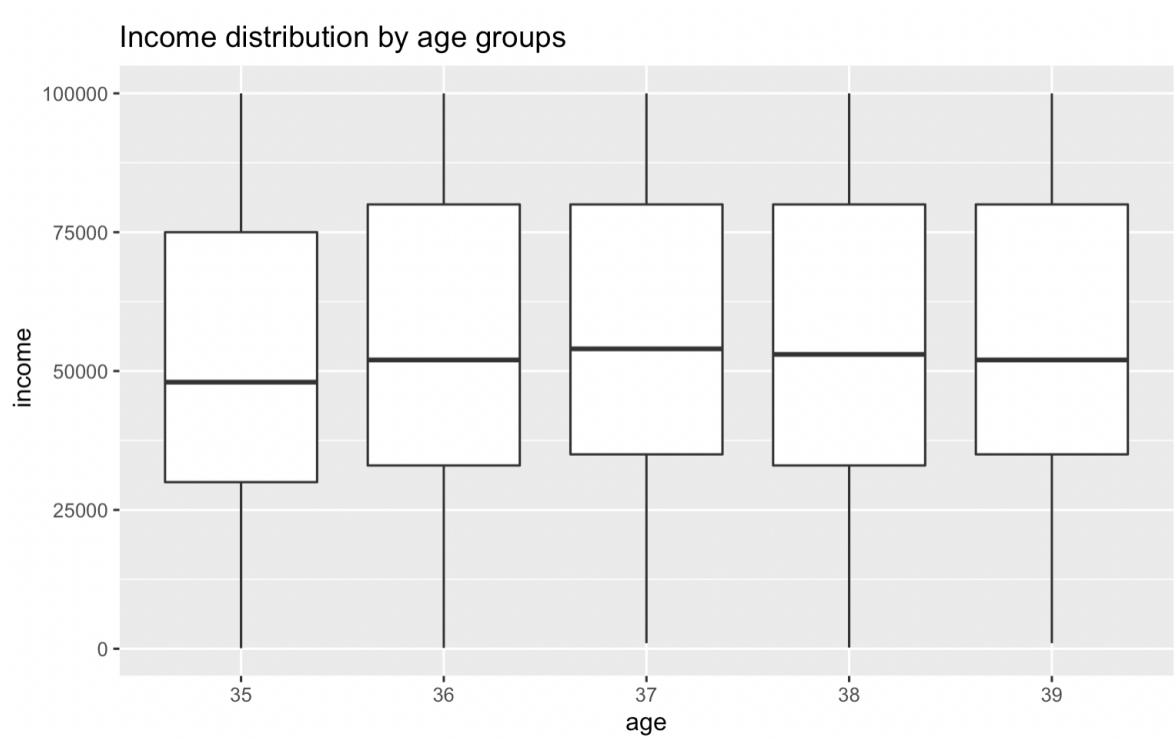
```
# Assume None: 0 years; GED: 12 years; High school: 12 years; Associate College: 14 years; Bachelor's degree:16 years; Master's degree: 18 years; PhD:23 years; Professional degree:20 years
BIO_DAD <- which(dat$CV_HGC_BIO_DAD_1997 == 95)
BIO_MOM <- which(dat$CV_HGC_BIO_MOM_1997 == 95)
RED_DAD <- which(dat$CV_HGC_RES_DAD_1997 == 95)
RED_MOM <- which(dat$CV_HGC_RES_MOM_1997 == 95)

dat$CV_HGC_BIO_DAD_1997[BIO_DAD] = 0
dat$CV_HGC_BIO_MOM_1997[BIO_MOM] = 0
dat$CV_HGC_RES_DAD_1997[RED_DAD] = 0
dat$CV_HGC_RES_MOM_1997[RED_MOM] = 0

dat1 <- dat %>% mutate(individual_edu = case_when(dat$YSCH.3113_2019 == 1 ~ "0",
                                                  dat$YSCH.3113_2019 == 2 ~ "12",
                                                  dat$YSCH.3113_2019 == 3 ~ "12",
                                                  dat$YSCH.3113_2019 == 4 ~ "14",
                                                  dat$YSCH.3113_2019 == 5 ~ "16",
                                                  dat$YSCH.3113_2019 == 6 ~ "18",
                                                  dat$YSCH.3113_2019 == 7 ~ "23",
                                                  dat$YSCH.3113_2019 == 8 ~ "20"))

dat1 <- dat1 %>% drop_na(CV_HGC_BIO_DAD_1997)
dat1 <- dat1 %>% drop_na(CV_HGC_BIO_MOM_1997)
dat1 <- dat1 %>% drop_na(CV_HGC_RES_DAD_1997)
dat1 <- dat1 %>% drop_na(CV_HGC_RES_MOM_1997)
dat1 <- dat1 %>% drop_na(individual_edu)
```

3. Provide the following visualizations.

(1) Plot the income data (where income is positive) by i) age groups, ii) gender groups and iii) number of children

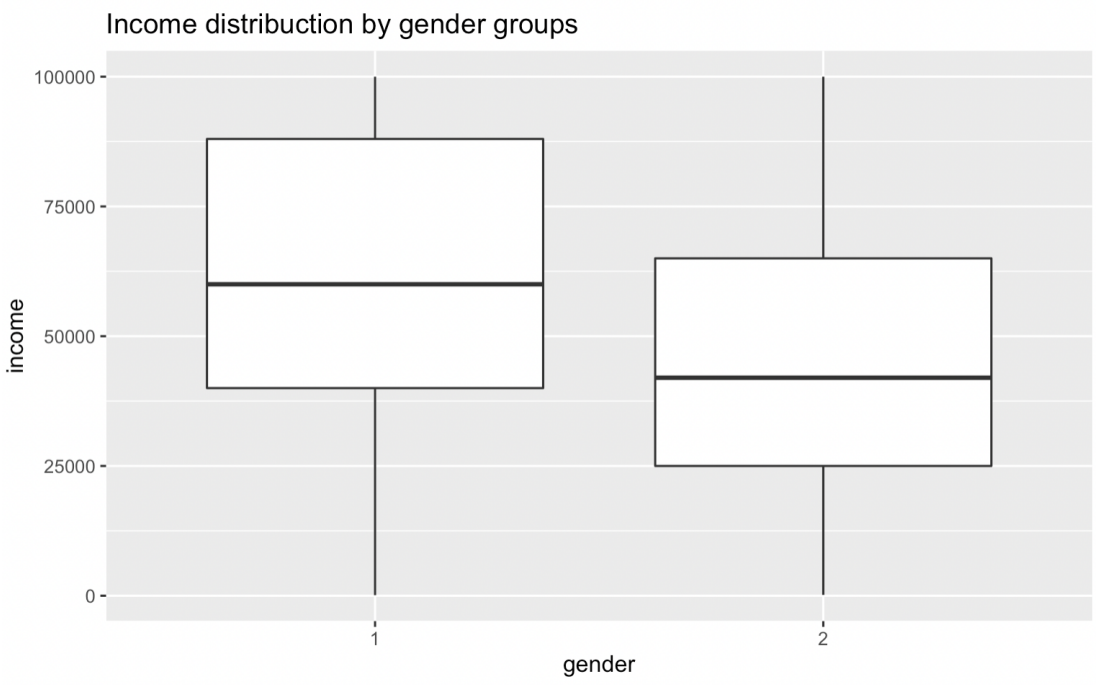
(i) Income data by age groups

```
dat1$YINC_1700_2019 <- as.factor(dat$YINC_1700_2019)
dat_posincome <- subset(dat1, dat1$YINC_1700_2019 > 0 & dat1$YINC_1700_2019 != "NA" )
ggplot(dat_posincome, aes(as.factor(age),
                          YINC_1700_2019,
                          group = age)) +
  geom_boxplot() +
  labs(title = "Income distribution by age groups",
       x = "age", y = " income")
```



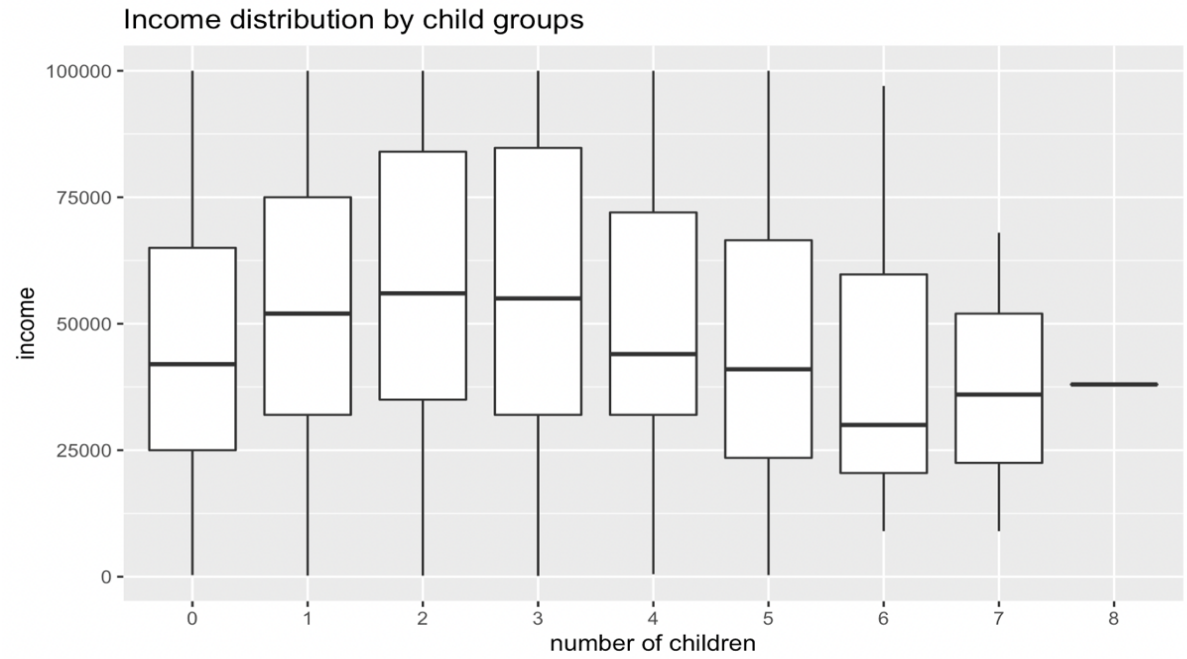
(ii) Income data by gender groups

```
dat_gender <- dat_posincome[!(is.na(dat_posincome$KEY_SEX_1997)),]
ggplot(dat_gender, aes(as.factor(KEY_SEX_1997),
                       YINC_1700_2019,
                       group = KEY_SEX_1997)) +
  geom_boxplot() +
  labs(title = "Income distribuction by gender groups", x = "gender", y = "income")
```



(iii) Income data by number of children

```
dat_child <- dat_posincome[!(is.na(dat_posincome$CV_BIO_CHILD_HH_U18_2019)),]
ggplot(dat_child, aes(as.factor(CV_BIO_CHILD_HH_U18_2019),
                      YINC_1700_2019,
                      group = CV_BIO_CHILD_HH_U18_2019)) +
        geom_boxplot() +
        labs(title = "Income distribution by child groups",
             x = "number of children", y = "income")
```



(2) Table the share of ”0” in the income data by i) age groups, ii) gender groups, iii) number of children and marital status

(i) age groups

```
age_share <- dat %>% group_by(age) %>% summarize(zero_income = length(which(YINC_1700_2019 == 0)), total = n())
%>% mutate(share = zero_income / total)
```

	age	zero_income	total	share
1	35	10	1771	0.005646527
2	36	7	1807	0.003873824
3	37	6	1841	0.003259098
4	38	10	1874	0.005336179
5	39	3	1691	0.001774098

(ii) gender groups

```
# gender groups
gender_share <- dat %>% group_by(KEY_SEX_1997) %>% summarize(zero_income = length(which(YINC_1700_2019 == 0)),
total = n()) %>% mutate(share = zero_income / total)
```

	KEY_SEX_1997	zero_income	total	share
1	1	21	4599	0.004566210
2	2	15	4385	0.003420753

(iii) number of children and marital status

```
dat$CV_MARSTAT_COLLAPSED_2019 <- as.factor(dat$CV_MARSTAT_COLLAPSED_2019)
childandmarital_share <- dat %>% group_by(CV_BIO_CHILD_HH_U18_2019, CV_MARSTAT_COLLAPSED_2019) %>% summarize(zero_income =
length(which(YINC_1700_2019 == 0)), total = n()) %>% mutate(share = zero_income / total)
```

	CV_BIO_CHILD_HH_U18_2019	CV_MARSTAT_COLLAPSED_2019	zero_income	total	share
1	0	0	0	422	0.000000000
2	0	1	4	151	0.026490066
3	0	2	3	36	0.083333333
4	0	3	1	207	0.004830918
5	0	4	0	2	0.000000000
6	0	*NA*	0	10	0.000000000
7	1	0	4	481	0.008316008
8	1	1	5	704	0.007102273
9	1	2	0	23	0.000000000
10	1	3	0	178	0.000000000
11	1	4	0	7	0.000000000
12	1	*NA*	0	11	0.000000000
13	2	0	0	361	0.000000000
14	2	1	8	1131	0.007073386
15	2	2	0	32	0.000000000

Showing 1 to 15 of 47 entries, 5 total columns

（3）interpret the visualizations from above

- From the boxplot of Income distribution by age groups, we can see that the distribution of income of different age groups is similar.
- From the boxplot of Income distribution by gender groups, we can see that males have higher income than females.
- From the boxplot of Income distribution by child groups, we can see that income is higher when the number of children is between 1 and 3. Besides, when the number of children is more than 3, the income is decreasing in general.
- From the table of share of ”0” in the income data by age groups, we can see that people aged at 35 have a large proportion of no income.
- From the table of share of ”0” in the income data by gender groups, we can see that males have a large proportion of no income than females.
- From the table of share of ”0” in the income data by children group and marital status, we can see that people are more likely to have zero income with the increase of children and married people are more likely to have zero income.

Exercise 2 Heckman Selection Model

1. Using the variables created above, estimate the following models.

(a) Specify and estimate an OLS model to explain the income variable

```
Residuals:
   Min     1Q Median     3Q    Max
-72626 -18405  -2153  19191  71824

Coefficients:
                           Estimate Std. Error t value             Pr(>|t|)
(Intercept)                18075.43   14686.53   1.231              0.21855
age                          287.70     392.66   0.733              0.46383
work_exp                    1123.46     101.78  11.038 < 0.0000000000000002 ***
total_edu                    643.96      39.92  16.132 < 0.0000000000000002 ***
KEY_SEX_1997               -18851.08    1099.47 -17.146 < 0.0000000000000002 ***
CV_BIO_CHILD_HH_U18_2019     1640.97     505.65   3.245              0.00119 **
CV_MARSTAT_COLLAPSED_2019    1091.78     633.73   1.723              0.08507 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25330 on 2164 degrees of freedom
  (861 observations deleted due to missingness)
Multiple R-squared:  0.2485,    Adjusted R-squared:  0.2464
F-statistic: 119.2 on 6 and 2164 DF,  p-value: < 0.00000000000000022
```

(b) Interpret the estimation results

Results show that income will increase 1123.46 when increasing one year of work experience. The income will increase 643.96 when increasing one year of education and the income of females is 18851.08 less than the males. The income will increase 1640.97 when having one more child.

(c) Explain why there might be a selection problem when estimating an OLS this way

Because the selection of the participants is not random. The income of employed people is observed while the income of unemployed people is unobserved and might not be included. The result of OLS regression will be biased.

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

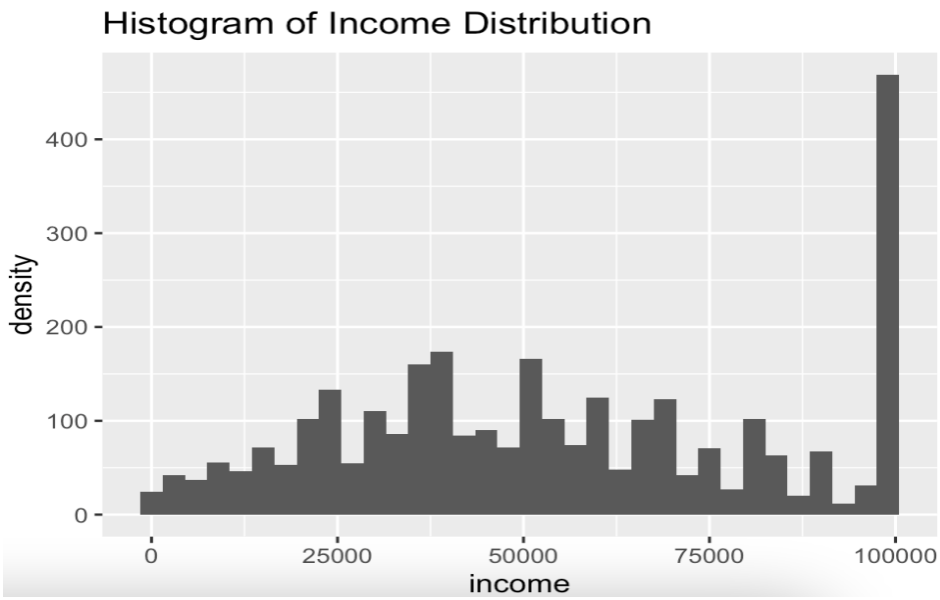
Since the Heckman model is a two-stage estimation model. It can help us solve the possible selection problem when we only include the positive income.

(3) Estimate a Heckman selection model (Please write down the likelihood and optimize the two-stage Heckman model). Interpret the results from the Heckman selection model and compare the results to OLS results. Why does there exist a difference?

Exercise 3 Censoring

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

```
ggplot(dat_posincome, aes(x = YINC_1700_2019)) +
  geom_histogram(binwidth = 3000) +
  labs(title = "Histogram of Income Distribution", x = "income", y = "density")
```



The censored value is $100,000.

2. Propose a model to deal with the censoring problem.

The Tobie model which is designed to estimate the linear relationships between variables when the censoring exists in the dependent variables can be used to deal with the censoring problem.

3. Estimate the appropriate model with the censored data

```
iter 985 value 21914.156510
iter 986 value 21914.155970
iter 987 value 21914.154488
iter 988 value 21914.150769
iter 989 value 21914.140891
iter 990 value 21914.115209
iter 991 value 21914.047958
iter 992 value 21913.970004
iter 993 value 21913.969268
iter 994 value 21913.968931
iter 995 value 21913.966473
iter 996 value 21913.962720
iter 997 value 21913.956322
iter 998 value 21913.951290
iter 999 value 21913.949313
iter1000 value 21913.948778
final  value 21913.948778
stopped after 1000 iterations
> tobit_fun$par
[1]    33.11946    19.78528  1324.82153   746.93665 -1321.65546   180.57424    10.33997
```

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

The income will increase when increasing additional one year of work experience, additional one year of education, increase one year old and so on. However, the income of females will still less than the males.

Exercise 4 Panel Data

1. Explain the potential ability bias when trying to explain to understand the determinants of wages.

The ability bias might happen in studying the effect of education on incomes. People who are more capable are interested in pursuing education and therefore have a higher education and income than other people.

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.

(1) Within Estimator.

```
Call:
lm(formula = income_diff ~ marital_diff + edu_diff + work_diff,
    data = mean_data)

Residuals:
    Min      1Q  Median      3Q     Max
-150461  -10237    -933    8476  283046

Coefficients: (1 not defined because of singularities)
                         Estimate           Std. Error t value            Pr(>|t|)
(Intercept)      0.00000000001582    74.31141829291471    0.00                   1
marital_diff 20774.27707397577615   228.89595161682189   90.76 <0.0000000000000002 ***
edu_diff      2166.61304953479794    23.01604408492161   94.14 <0.0000000000000002 ***
work_diff                      NA                    NA      NA                  NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21280 on 82005 degrees of freedom
Multiple R-squared:  0.2261,    Adjusted R-squared:  0.2261
F-statistic: 1.198e+04 on 2 and 82005 DF,  p-value: < 0.00000000000000022
```

```
> within_model <- plm(income ~ mar + edu + work_exp,mean_data, model = "within")
> summary(within_model)
Oneway (individual) effect Within Model

Call:
plm(formula = income ~ mar + edu + work_exp, data = mean_data,
    model = "within")

Unbalanced Panel: n = 8600, T = 1-18, N = 82008

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-150461.23  -10236.96    -933.28    8476.49  283046.36

Coefficients:
     Estimate Std. Error t-value              Pr(>|t|)
mar 20774.277    241.932  85.868 < 0.00000000000000022 ***
edu  2166.613     24.327  89.063 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    47989000000000
Residual Sum of Squares: 37137000000000
R-Squared:      0.22614
Adj. R-Squared: 0.13547
F-statistic: 10725.5 on 2 and 73406 DF, p-value: < 0.000000000000000222
```

(2) Between Estimator

```
> summary(between_model)
Oneway (individual) effect Between Model

Call:
plm(formula = income ~ edu + mar + work_exp, data = longp_dat,
    model = "between")

Unbalanced Panel: n = 8600, T = 1-18, N = 82008
Observations used in estimation: 8600

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-40009.7  -9462.3  -2653.1   5556.8 293896.6

Coefficients:
             Estimate Std. Error t-value              Pr(>|t|)
(Intercept) 3153.9693   481.0565  6.5563      0.00000000005832 ***
edu         1231.6685    41.4902 29.6857 < 0.00000000000000022 ***
mar         9057.5972   576.2751 15.7175 < 0.00000000000000022 ***
work_exp     119.3033     5.5146 21.6343 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    2703500000000
Residual Sum of Squares: 2133700000000
R-Squared:      0.21078
Adj. R-Squared: 0.21051
F-statistic: 765.264 on 3 and 8596 DF, p-value: < 0.000000000000000222
```

(3) Difference Estimator

```
> summary(difference_model)
Oneway (individual) effect First-Difference Model

Call:
plm(formula = income ~ edu + mar + work_exp, data = longp_dat,
    model = "fd")

Unbalanced Panel: n = 8600, T = 1-18, N = 82008
Observations used in estimation: 73408

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-215192.3   -5551.3   -2551.3    4448.7  322899.7

Coefficients:
            Estimate Std. Error t-value              Pr(>|t|)
(Intercept) 4551.320     67.073  67.856 < 0.00000000000000022 ***
edu          105.494     22.102   4.773           0.000001819 ***
mar         2509.688    226.545  11.078 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    21799000000000
Residual Sum of Squares: 21756000000000
R-Squared:      0.0019998
Adj. R-Squared: 0.0019726
F-statistic: 73.5445 on 2 and 73405 DF, p-value: < 0.000000000000000222
```