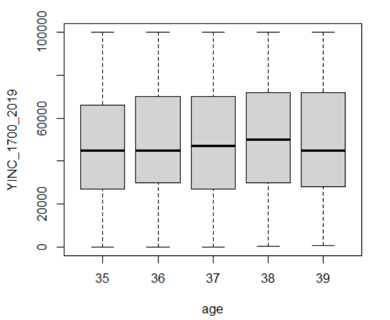
Exercise 1 Preparing the Data

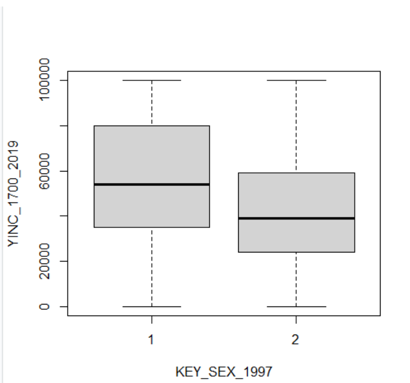
Q1 & Q2: see dataset “dat_A4”

Q3:

Plot the income data by age groups:



By gender groups:



By number of children:

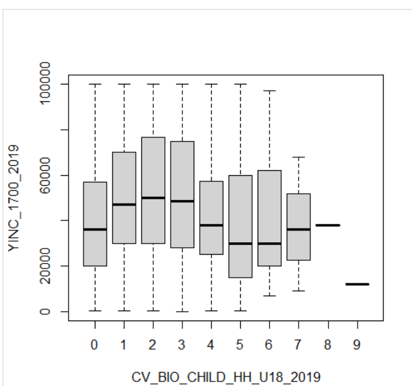


Table the share of 0 income by age groups:

	age	share
1	37	0.005420054
2	36	0.006300630
3	38	0.008960573
4	35	0.009293680
5	39	0.002994012

By gender groups:

	KEY_SEX_1997	share
1	1	0.007500000
2	2	0.005742726

By number of children:

	CV_BIO_CHILD_HH_U18_2019	share
1	3	0.008025682
2	1	0.007846556
3	2	0.005743001
4	0	0.014897579
5	4	0.000000000
6	5	0.000000000
7	6	0.000000000
8	9	0.000000000
9	7	0.000000000
10	8	0.000000000

By marital status:

	CV_MARSTAT_COLLAPSED_2019	share
1	0	0.005649718
2	1	0.007454342
3	3	0.001538462
4	4	0.000000000
5	2	0.043010753

Interpret the visualization above:

Plot 1: people with age 38 have highest average income

Plot 2: the male group has higher average income

Plot 3: households with two children have highest average income

Table 1: people with age 35 are more likely to have zero income

Table 2: the male group is more likely to have zero income

Table 3: households with more than 3 children do not have zero income

Table 4: the widowed group may not have zero income

Exercise 2 Heckman Selection Model

Q1: the OLS results are shown below

```
Residuals:
   Min     1Q Median     3Q    Max 
-64397 -20167  -2587  21007  78666 

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)    
(Intercept)   -1218.12   12801.83  -0.095               0.924    
age             485.50     346.31   1.402               0.161    
work_exp        904.13      87.58  10.323 <0.0000000000000002 ***
ind_schooling  2274.55     110.03  20.672 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpret the estimation results:

Keep other variables constant, every one week increase on work experience leads to 904.13 dollar increase on income in average. Keep other variables constant, every one-year increase on individual schooling leads to 2274.55 increase on income in average. The coefficient on age is insignificant.

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

There might be a selection problem because we use sub-sample with positive income. So, this sample is not selected randomly, which leads to a bias when estimating an OLS.

Q2: Explain why the Heckman Model can deal with the selection problem

Heckman selection model uses two stages method to solve the selection problem. In the first stage, Heckman model applies a probit regression to obtain the probability of working. In the second stage, Heckman model corrects for self-selection by transforming predicted probability of working into inverse mills ratio as a new independent variable in OLS estimation. This new OLS equation indicates Heckman's view on sample selection problem as omitted-variables bias.

Q3:

Likelihood function is “heckman_ll”

The optimized results are shown in “heck_optim”.

Below are coefficients from Heckman selection model.

```
                Estimate Std. Error t value      Pr(>|t|)
(Intercept)      26809.5    17626.3   1.521        0.1283
age                756.4      460.9   1.641        0.1008
work_exp          -526.7      291.2  -1.809        0.0705 .
ind_schooling     1323.8      224.1   5.907 0.00000000379 ***
```

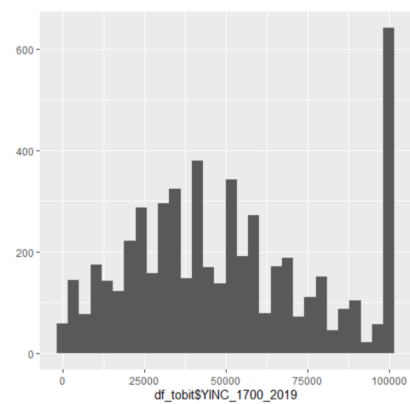
Interpret the results from Heckman selection model and compare the results to OLS results.

Why does there exist a difference?

In the Heckman selection model, keep other variables constant, every one-year increase on individual schooling leads to 1323.8 dollar increase on income in average. The coefficients on age and work experience are both insignificant at 5% level. Comparing to the OLS results, the effect of individual schools is less in Heckman model and more importantly, the effect of work experience is not significant in Heckman model. In the biased OLS, because the sample only contains people with positive income, which may indicate more work experience, the effect of work experience may be enlarged.

Exercise 3 Censoring

Q1: Plot a histogram to check whether the distribution of the income variable. What might be the censored value here?



The censored value is 100000

Q2: propose a model to deal with the censoring problem

Tobit model

Q3: Estimate the appropriate model with the censored data

The likelihood function is “tobit_ll”

The following are the results of tobit model

```
Coefficients:
                 Estimate  Std. Error z value            Pr(>|z|)
(Intercept)   -2828.69919 10656.21364  -0.265              0.7907
age             502.25965   287.15163   1.749              0.0803 .
work_exp       1069.59643    74.60646  14.337 <0.0000000000000002 ***
ind_schooling  2206.49463    82.19468  26.845 <0.0000000000000002 ***
Log(scale)       10.27375     0.01064 965.665 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Q4: Interpret the results above and compare to those when not correcting for the censored data

Keep other variables constant, every one week increase on work experience leads to 1069.60 dollar increase on income in average. Keep other variables constant, every one-year increase on individual schooling leads to 2206.49 dollar increase on income in average.

OLS

(Intercept)	double [1]	3142.581
age	double [1]	382.1197
work_exp	double [1]	1004.432
ind_schooling	double [1]	2000.877

Tobit

(Intercept)	double [1]	3143.567
age	double [1]	418.7482
work_exp	double [1]	1013.253
ind_schooling	double [1]	2016.951

Comparing to the results of OLS, the effects of all variables are larger in tobit model. The reason is that OLS ignores the effect of wages higher than $100000.

Exercise 4 Panel Data

Q1: Explain the potential ability bias when trying to explain to understand the determinants of wages

The potential ability bias in this case is the greater innate skills kept by people who have more education so that they may earn more even without additional year of schooling.

Q2: Estimate the model using the following strategy:

Within Estimator:

```
Coefficients:
                Estimate Std. Error t value            Pr(>|t|)
work_within      49.1330     0.5016  97.948 < 0.0000000000000002 ***
degree_within  1584.9639    22.0170  71.988 < 0.0000000000000002 ***
mar1_within   18753.6557   230.6412  81.311 < 0.0000000000000002 ***
mar2_within   15394.1364   838.7581  18.353 < 0.0000000000000002 ***
mar3_within   19370.3137   454.2332  42.644 < 0.0000000000000002 ***
mar4_within   10473.8183  2626.4722   3.988           0.0000667 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Between Estimator:

```
Coefficients:
                Estimate Std. Error t value           Pr(>|t|)
work_between      39.461      0.442  89.273 < 0.0000000000000002 ***
degree_between  1435.928      8.600 166.969 < 0.0000000000000002 ***
mar1_between    9876.891    179.366  55.066 < 0.0000000000000002 ***
mar2_between    5872.002   1170.031   5.019   0.000000521372500 ***
mar3_between    3467.483    415.937   8.337 < 0.0000000000000002 ***
mar4_between  -18008.259   2455.792  -7.333   0.000000000000227 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Difference Estimator:

```
Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
work_diff     25.3231     0.5618  45.076 <0.0000000000000002 ***
degree_diff  429.1127    20.8879  20.544 <0.0000000000000002 ***
mar1_diff   7042.0704   265.9069  26.483 <0.0000000000000002 ***
mar2_diff   6939.7485   656.1777  10.576 <0.0000000000000002 ***
mar3_diff   9913.6975   492.8792  20.114 <0.0000000000000002 ***
mar4_diff   4290.4264  2463.0709   1.742              0.0815 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mar1 = married

Mar2 = separated

Mar3 = divorced

Mar4 = widowed

Q3: Interpret the results from each model and explain why different models yield different parameter estimates

Within Estimator:

Holding other variables constant, every one week increase on work experience leads to 49.13 dollar increase on income in average. Holding other variables constant, every one-year increase on individual schooling leads to 1584.96 dollar increase on income in average. Married people earn 18753.66 dollar more than unmarried people on average. Separated people earn 15394.14 dollar more than unmarried people on average. Divorced people earn 19370.31 dollar more than unmarried people on average. Widowed people earn 10473.82 dollar more than unmarried people on average.

Within estimator measures the association between individual-specific deviations of regressors from their time averaged values and individual-specific deviations of the dependent variable from its time-averaged value. A special feature of this estimator is that it yields consistent estimates of beta in the fixed effects model.

Between Estimator:

Holding other variables constant, every one week increase on work experience leads to 39.46 dollar increase on income in average. Holding other variables constant, every one-year increase on individual schooling leads to 1435.93 dollar increase on income in average. Married people earn 9876.89 dollar more than unmarried people on average. Separated people earn 5872.00 dollar more than unmarried people on average. Divorced people earn 3467.48 dollar more than unmarried people on average. Widowed people earn 18008.26 dollar less than unmarried people on average.

Between estimator uses variation between different individuals. It is consistent if the regressors are independent of the composite error. This will be the case for the constant-coefficients model and the random effects model.

Difference Estimator:

Holding other variables constant, every one week increase on work experience leads to 25.32 dollar increase on income in average. Holding other variables constant, every one-year increase on individual schooling leads to 429.11 dollar increase on income in average. Married people earn 7042.07 dollar more than unmarried people on average. Separated people earn 6939.75 dollar more than unmarried people on average. Divorced people earn 9913.70 dollar more than unmarried people on average. The coefficient of widowed is not significant at 5% level.

First difference estimator measures the association between individual-specific one-period changes in regressors and individual-specific one-period changes in the dependent variable. It yields consistent estimates of beta in the fixed effects model, though the coefficients of time-invariant regressors are not identified.