Homework 4

Jiaxin Li

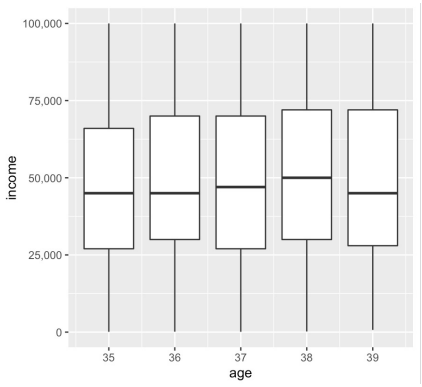
Exercise 1

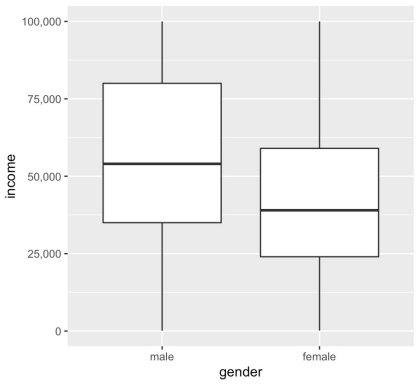
1.1 & 1.2

Created in programming

1.3

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





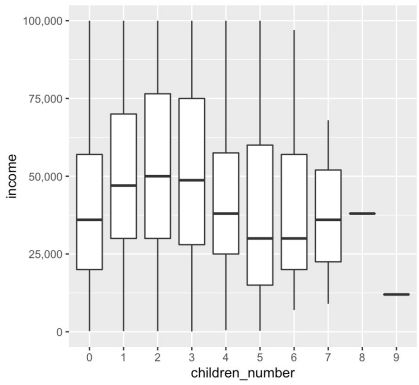


Table the share of "0" in the income data by i) age groups, ii) gender groups, iii) number of children and marital status

	age	income_zero	income_na
	<dbl>	<dbl>	<dbl>
1	35	0.00565	0.392
2	36	0.00387	0.385
3	37	0.00326	0.399
4	38	0.00534	0.404
5	39	0.00177	0.407

	gender	income_zero	income_na
	<fct>	<dbl>	<dbl>
1	male	0.00457	0.391
2	female	0.00342	0.404

```
# A tibble: 47 × 4
# Groups:   children_number [11]
   children_number marital_status income_zero income_na
   <fct>           <fct>                <dbl>     <dbl>
 1 0               never              0           0.424
 2 0               married            0.0265      0.219
 3 0               seperated          0.0833      0.389
 4 0               divorced           0.00483     0.290
 5 0               widowed            0           0
 6 0               NA                 0           0.5
 7 1               never              0.00832     0.249
 8 1               married            0.00710     0.129
 9 1               seperated          0           0.348
10 1               divorced           0           0.169
# … with 37 more rows
```

Interpret the visualizations from above:

In general, older people get relatively higher income, but the average income of 39 year-old people is lower.

The male get higher income than the female.

People have 2-4 children get relatively higher income than the rest.

The share of unobservable samples is large.

Exercise 2

2.1

All variables:

```
Residuals:
   Min     1Q Median     3Q    Max
-70559 -17319  -1711  18012  77224

Coefficients:
                 Estimate Std. Error t value    Pr(>|t|)
(Intercept)       1989.77   14323.24   0.139      0.8895
age                102.51     382.82   0.268      0.7889
genderfemale    -19733.51    1082.45 -18.230     < 2e-16 ***
children_number    533.73     509.79   1.047      0.2952
self_edu_year     2282.04     156.98  14.537     < 2e-16 ***
parent_edu_year    279.98      55.90   5.009 0.000000593 ***
work_exp          1030.26      98.93  10.415     < 2e-16 ***
married           5906.90    1485.91   3.975 0.000072613 ***
separated         4317.14    4815.10   0.897      0.3700
divorced          3723.55    1904.52   1.955      0.0507 .
widowed           6849.10   12349.95   0.555      0.5792
black            -1520.57    1574.60  -0.966      0.3343
hispanic           321.46    1521.29   0.211      0.8327
mixed            -4226.59    6187.53  -0.683      0.4946
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24530 on 2156 degrees of freedom
  (3206 observations deleted due to missingness)
Multiple R-squared:  0.2982,    Adjusted R-squared:  0.294
F-statistic: 70.48 on 13 and 2156 DF,  p-value: < 2.2e-16
```

Created variables (only including married dummy variable):

```
Residuals:
   Min     1Q Median     3Q    Max
-68530 -17332  -1698  18105  80321

Coefficients:
                 Estimate Std. Error t value    Pr(>|t|)
(Intercept)        750.98   14220.52   0.053    0.957889
age                185.24     380.79   0.486    0.626693
genderfemale    -19707.36    1078.51 -18.273     < 2e-16 ***
children_number    519.68     508.77   1.021    0.307155
self_edu_year     2298.82     156.74  14.667     < 2e-16 ***
parent_edu_year    267.62      49.85   5.368 0.000000088 ***
work_exp          1021.46      98.57  10.362     < 2e-16 ***
married           4450.85    1203.50   3.698    0.000222 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24530 on 2162 degrees of freedom
Multiple R-squared:  0.2961,    Adjusted R-squared:  0.2939
F-statistic: 129.9 on 7 and 2162 DF,  p-value: < 2.2e-16
```

Interpretation:

Holding other variables constant, (1) the female tend to get lower income than the male; (2) people with higher education tend to get higher income; (3) people whose parents have higher education tend to get higher income, but the effect is smaller than their education level; (4) people who have more working experience tend to get higher income; (5) married people tend to get higher income.

However, our sample only includes information of positive incomes, but no information of zero or missing incomes which is unobservable. The bias caused by people who have zero income or do not have a job/report their income is the sample selection bias. Therefore there might be a selection problem which results in biased estimation.

2.2

Selection equation: the individual's probability of being selected

Outcome equation: the conditional expectation of the outcome variable.

The bias of directly evaluating outcome equation can be viewed as the inverse mills ratio. Then, add the inverse mills ratio calculated from the predicted values of the selection equation as a control variable in outcome equation, we can get the unbiased estimation.

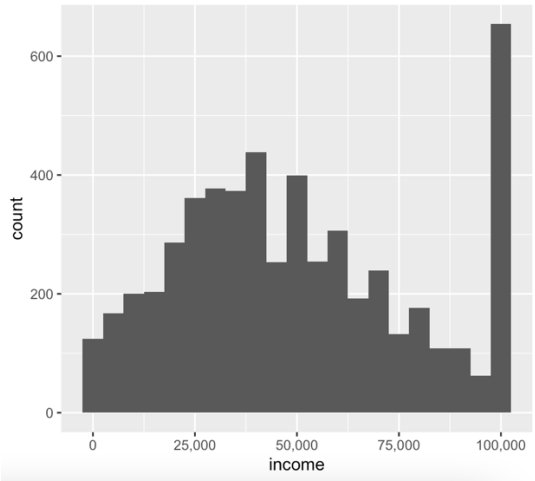
2.3

```
> first_stage
                heckit : est  heckit :se without package : est without package :se
(Intercept)     -0.184087187 0.819807553          -0.184087349         0.813025901
age             -0.001871219 0.021883655          -0.001871205         0.021737322
genderfemale    -0.365052789 0.064831569          -0.365054018         0.064291666
children_number -0.027025914 0.027254666          -0.027026376         0.027136354
self_edu_year    0.044678375 0.007874475           0.044678395         0.007725243
parent_edu_year  0.002062454 0.002794970           0.002062371         0.002797776
work_exp         0.112225119 0.007047268           0.112226796         0.007657978
married          0.176057144 0.067066283           0.176056531         0.066487607
> second_stage
                heckit : est  heckit :se without package : est without package :se
(Intercept)       28512.6281 19116.25260            28512.2964          19086.9707
age                 315.9654   488.62485              315.9661            487.2936
genderfemale     -14401.1326  1807.91826           -14401.1743           1812.4078
children_number     971.8441   643.81304              971.8453            642.7284
self_edu_year      1504.3479   258.93649             1504.3549            258.9863
parent_edu_year     239.3577    63.52285              239.3590             63.5577
work_exp           -254.0938   300.90055             -254.0934            301.3419
married            1818.4409  1623.57134             1818.4812           1622.0162
invMillsRatio    -43672.4954  8670.31562           -43672.1832           8751.3873
```

Compared with OLS model, gender, education and marriage have the same effect on income, but the influence degrees vary due to the selection bias caused by OLS.

Exercise 3

3.1



censored value is 100,000

3.2

Solve the censored problem with a Tobit model. First, consider the censored value is zero. Let income = 0, if income = 100,000

Results of Tobit model:

```
Observations: (3206 observations deleted due to missingness)
         Total  Left-censored     Uncensored Right-censored
          2170              0           1828            342

Coefficients:
                    Estimate   Std. Error z value    Pr(>|z|)
(Intercept)      -8927.68394  16354.88933  -0.546    0.585154
age                316.22930    440.96031   0.717    0.473289
genderfemale    -22426.05965   1247.24038 -17.981     < 2e-16 ***
children_number    737.69855    593.43492   1.243    0.213831
self_edu_year     2585.38255    193.03957  13.393     < 2e-16 ***
parent_edu_year    328.58268     58.80083   5.588 0.000000023 ***
work_exp          1111.31187    123.67787   8.986     < 2e-16 ***
married           5046.93023   1363.66090   3.701    0.000215 ***
Log(scale)          10.24897      0.01778 576.413     < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 28254

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -2.17e+04 on 9 Df
Wald-statistic: 896.3 on 7 Df, p-value: < 2.22e-16
```

The signs of coefficients are the same, but the magnitudes are different due to the change in censored value.

Exercise 4

4.1

There are unobservable determinants such as capacity, and the unobserved errors are not independent over periods. Both problems cause the bias.

4.2

Within:

```
Unbalanced Panel: n = 7775, T = 1-18, N = 70935

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-140559.35   -7912.22     167.78    7324.04  263774.56

Coefficients:
               Estimate Std. Error t-value  Pr(>|t|)
age            2058.166     21.794  94.435 < 2.2e-16 ***
marriedothers -5908.934    247.560 -23.869 < 2.2e-16 ***
work_exp        916.001     32.860  27.876 < 2.2e-16 ***
edu            3517.277     70.269  50.054 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    43230000000000
Residual Sum of Squares: 24858000000000
R-Squared:      0.42498
Adj. R-Squared: 0.35417
F-statistic: 11669.3 on 4 and 63156 DF, p-value: < 2.22e-16
```

Between:

```
Unbalanced Panel: n = 7775, T = 1-18, N = 70935
Observations used in estimation: 7775

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-54095.3  -9114.9  -2079.2   6063.6 265385.8

Coefficients:
               Estimate Std. Error t-value  Pr(>|t|)
(Intercept)   -57300.944   2390.110 -23.974 < 2.2e-16 ***
age             936.002     63.443  14.754 < 2.2e-16 ***
gendermale     9217.061    362.394  25.434 < 2.2e-16 ***
marriedothers -7976.417    574.192 -13.892 < 2.2e-16 ***
work_exp       1574.161     76.606  20.549 < 2.2e-16 ***
edu            4159.405    124.962  33.285 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    2738400000000
Residual Sum of Squares: 1942100000000
R-Squared:      0.29079
Adj. R-Squared: 0.29034
F-statistic: 637.093 on 5 and 7769 DF, p-value: < 2.22e-16
```

First difference:

```
Unbalanced Panel: n = 7775, T = 1-18, N = 70935
Observations used in estimation: 63160

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-212896.0   -6021.5   -1692.7    4594.0  321338.3

Coefficients:
               Estimate Std. Error t-value        Pr(>|t|)
(Intercept)    1214.593    113.653 10.6868       < 2.2e-16 ***
age            1829.993     54.058 33.8523       < 2.2e-16 ***
marriedothers -1567.378    246.397 -6.3612 0.0000000002015 ***
work_exp        677.709     33.029 20.5184       < 2.2e-16 ***
edu             872.541     87.337  9.9905       < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    20481000000000
Residual Sum of Squares: 19757000000000
R-Squared:      0.035353
Adj. R-Squared: 0.035292
F-statistic: 578.634 on 4 and 63155 DF, p-value: < 2.22e-16
```

Interpretation:

The signs of coefficients are the same, but the magnitudes are different which is due to the difference in groups. Within estimators indicate the effects on individual level. Between estimators indicate the effects between different individuals. First difference controls the individual heterogeneity.