Exercise 1

(a)

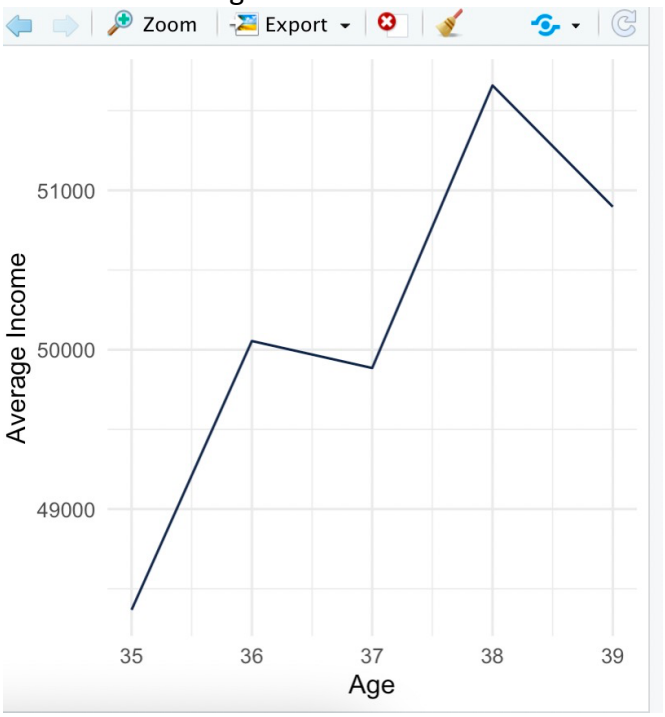
The required variables are presented in R.

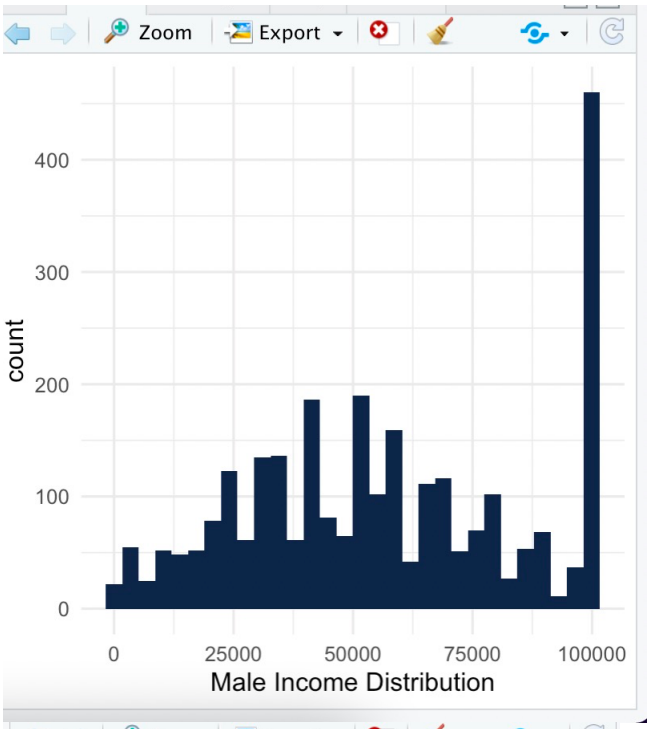
(b)

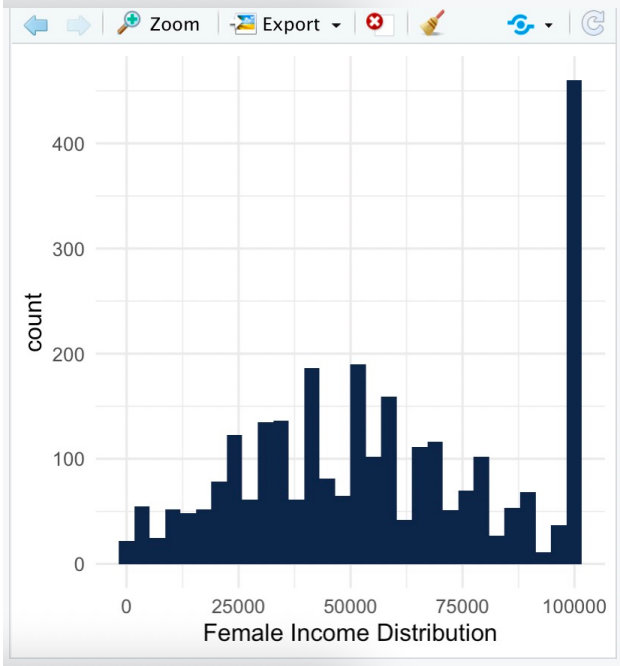
I created two variables, parent_edu represents the average years of education of their parents received and highest_degree represents the highest degree the worker received.

(c)

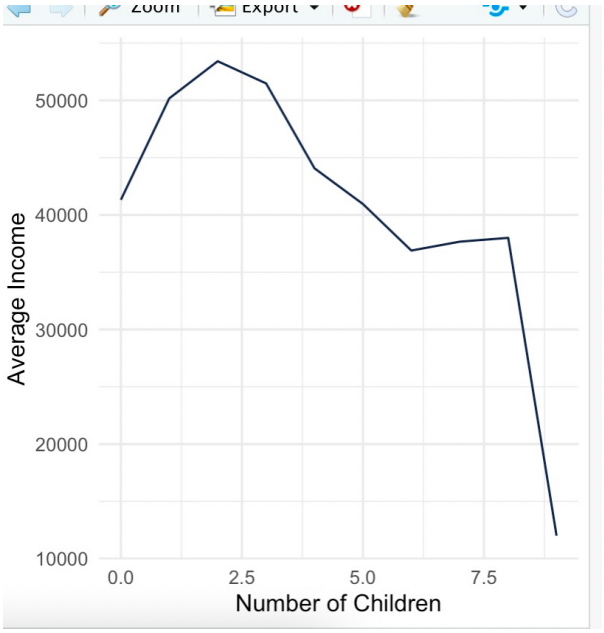
The average income is not that different between age groups. I think it's because they are about the same age.







I graphed the histogram for both male and female. Males have higher average income than female. Also, there are more males whose income are higher than 10,0000.



When having more children, the average income increases initially (from 0 to 2) and then decreases.

```
table
```

A tibble: 5 × 4

age <dbl>	Number_of_People <int>	No_income_People <int>	Share <dbl>
35	1771	705	0.398
36	1807	703	0.389
37	1841	740	0.402
38	1874	768	0.410
39	1691	692	0.409

Exercise 2

(a)

```
Call:
lm(formula = YINC_1700_2019 ~ age + work_exp + parent_edu + KEY_SEX_1997 +
    CV_MARSTAT_COLLAPSED_2019, data = income_data)

Residuals:
   Min     1Q Median     3Q    Max
-72360 -19357  -3415  18593  75103

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 29297.1     9598.7   3.052  0.00228 **
age                           371.5      257.3   1.444  0.14882
work_exp                     1019.5       61.5  16.577  < 2e-16 ***
parent_edu                   1605.1       86.7  18.513  < 2e-16 ***
KEY_SEX_1997               -12900.0      714.9 -18.045  < 2e-16 ***
CV_MARSTAT_COLLAPSED_2019    1653.9      380.1   4.351 1.38e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26070 on 5350 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.1628,    Adjusted R-squared:  0.162
F-statistic:   208 on 5 and 5350 DF,  p-value: < 2.2e-16

> |
```

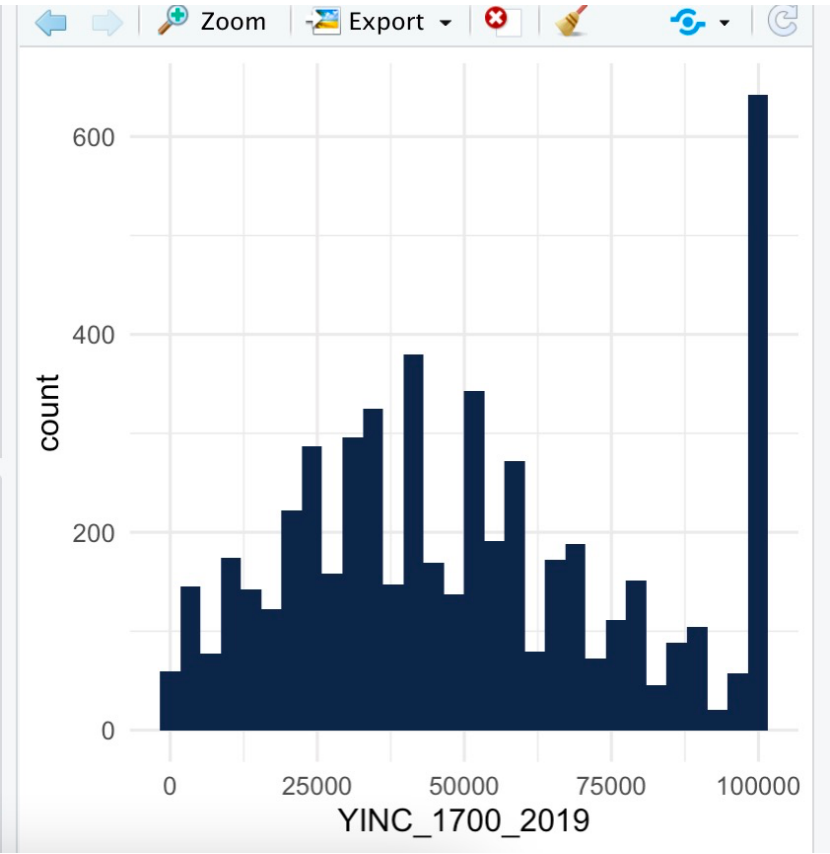
Holding other variables constant, age is not significant.
Holding other variables constant, work experience increases wage by 1019.5 dollars.
Holding other variables constant, an additional year of average parents' education increases wage by roughly 1605.1 dollars.
Holding other variables constant, being a female deceases wage by -12900 dollars.
Holding other variables constant, getting married increases wage by 1653.9 dollars

The regression above omits all NAs and 0 income people, causing a sample selection bias. Heckman selection model considers the probability of getting a positive income and incorporates Inverse M ratio variable to correct the bias.
(b)
After incorporating the Inverse M ratio variable, the results are different.

```
Call:
lm(formula = dat_A4$YINC_1700_2019 ~ x2 + x3 + x4 + x5 + x6 +
    Inv_M_ratio)

Residuals:
   Min     1Q Median     3Q    Max
-92974 -21462  -3502  18976  87469

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.136e+04  9.315e+03   2.293   0.0219 *
x2           1.366e+02  2.494e+02   0.548   0.5838
x3           1.870e+03  5.826e+01  32.096  < 2e-16 ***
x4           1.634e+03  8.321e+01  19.632  < 2e-16 ***
x5          -1.242e+04  6.959e+02 -17.854  < 2e-16 ***
x6           2.438e+03  3.643e+02   6.693 2.36e-11 ***
Inv_M_ratio -5.743e+27  1.026e+28  -0.560   0.5757
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28730 on 6896 degrees of freedom
  (2081 observations deleted due to missingness)
Multiple R-squared:  0.2278,    Adjusted R-squared:  0.2271
F-statistic:   339 on 6 and 6896 DF,  p-value: < 2.2e-16
```

Holding other variables constant, age is not significant.
Holding other variables constant, work experience increases wage by 1870 dollars.
Holding other variables constant, an additional year of average parents’ education increases wage by roughly 1634 dollars.
Holding other variables constant, being a female deceases wage by -12420 dollars.
Holding other variables constant, getting married increases wage by 2438 dollars.

Exercise 3
(a)



I exclude all the NAs and 0 income people for this problem.

(b)

We can use Tobit model to fix the censored problem.

```
Call:
tobit(formula = dat_A4$YINC_1700_2019 ~ x2 + x3 + x4 + x5 + x6,
    left = -Inf, right = 10000)

Observations: (20 observations deleted due to missingness)
         Total  Left-censored     Uncensored Right-censored
          5356              0            302           5054

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) 23191.2658  8527.1655   2.720  0.00653 **
x2             10.7842   227.1903   0.047  0.96214
x3            980.6220    92.9076  10.555  < 2e-16 ***
x4            337.8647    77.8798   4.338 1.44e-05 ***
x5          -3794.1352   672.3989  -5.643 1.67e-08 ***
x6            373.1627   326.8468   1.142  0.25358
Log(scale)      9.2928     0.0515 180.440  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 10859

Gaussian distribution
Number of Newton-Raphson Iterations: 6
Log-likelihood: -3918 on 7 Df
Wald-statistic: 137.9 on 5 Df, p-value: < 2.22e-16
```

(c)

```
> result2$par
[1] 23191.26577    10.78422   980.62202   337.86472 -3794.13519   373.16274    10.00000
```

The coefficients are different, but the signs are the same.

Exercise 4

Potential ability bias means that individuals have more abilities tend to study for a longer period.

Here is the between estimator.

```
Call:
lm(formula = m_income ~ m_work_exp + m_education + m_marital_status,
    data = estimate)

Residuals:
   Min     1Q Median     3Q    Max
-41063  -8953  -2621   5712 159328

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       7597.62      79.96   95.02   <2e-16 ***
m_work_exp        2629.49      22.78  115.44   <2e-16 ***
m_education       3612.40      24.10  149.90   <2e-16 ***
m_marital_status  2030.53      69.54   29.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14480 on 185583 degrees of freedom
  (21045 observations deleted due to missingness)
Multiple R-squared:  0.1966,    Adjusted R-squared:  0.1966
F-statistic: 1.514e+04 on 3 and 185583 DF,  p-value: < 2.2e-16
```

Holding other variables constant, work experience increases wage by 2629.49 dollars.
Holding other variables constant, an additional year of education increases wage by roughly 3612.4 dollars.
Holding other variables constant, getting married increases wage by 2430.53 dollars.

Here is the within estimator.

```
Call:
lm(formula = diff_income ~ diff_education + diff_work_exp + diff_marital_status,
    data = estimate)

Residuals:
    Min      1Q  Median      3Q     Max
-135030  -13185   -4914    6455  271460

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         10124.54     219.87   46.05   <2e-16 ***
diff_education      10597.32     526.71   20.12   <2e-16 ***
diff_work_exp        1397.74      38.97   35.86   <2e-16 ***
diff_marital_status  4436.72     266.73   16.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26320 on 31031 degrees of freedom
  (175597 observations deleted due to missingness)
Multiple R-squared:  0.06752,   Adjusted R-squared:  0.06743
F-statistic:   749 on 3 and 31031 DF,  p-value: < 2.2e-16
```

Holding other variables constant, work experience increases wage by 1397.74 dollars.
Holding other variables constant, an additional year of education increases wage by roughly 10124.54 dollars.
Holding other variables constant, getting married increases wage by 4436.72 dollars.

Here is the difference estimator.

```
Call:
lm(formula = first_diff_income ~ first_diff_education + first_diff_work_exp +
    first_diff_marital_status, data = estimate)

Residuals:
   Min     1Q Median     3Q    Max
-95798  -4229  -1159   4037 144505

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2401.8      217.4  11.048  < 2e-16 ***
first_diff_education         -68.3     1001.0  -0.068    0.946
first_diff_work_exp          528.1      111.0   4.756 2.03e-06 ***
first_diff_marital_status   1209.7      660.1   1.832    0.067 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13980 on 4358 degrees of freedom
  (202270 observations deleted due to missingness)
Multiple R-squared:  0.006007,  Adjusted R-squared:  0.005322
F-statistic: 8.778 on 3 and 4358 DF,  p-value: 8.418e-06

> |
```

Holding other variables constant, work experience increases wage by 528.1 dollars.
Holding other variables constant, an additional year of education increases wage by roughly -68.3 dollars.
Holding other variables constant, getting married increases wage by 1209.7 dollars.
The outcome are different because the estimators are different.