A4 YS

Yonghan Shi

4/13/2022

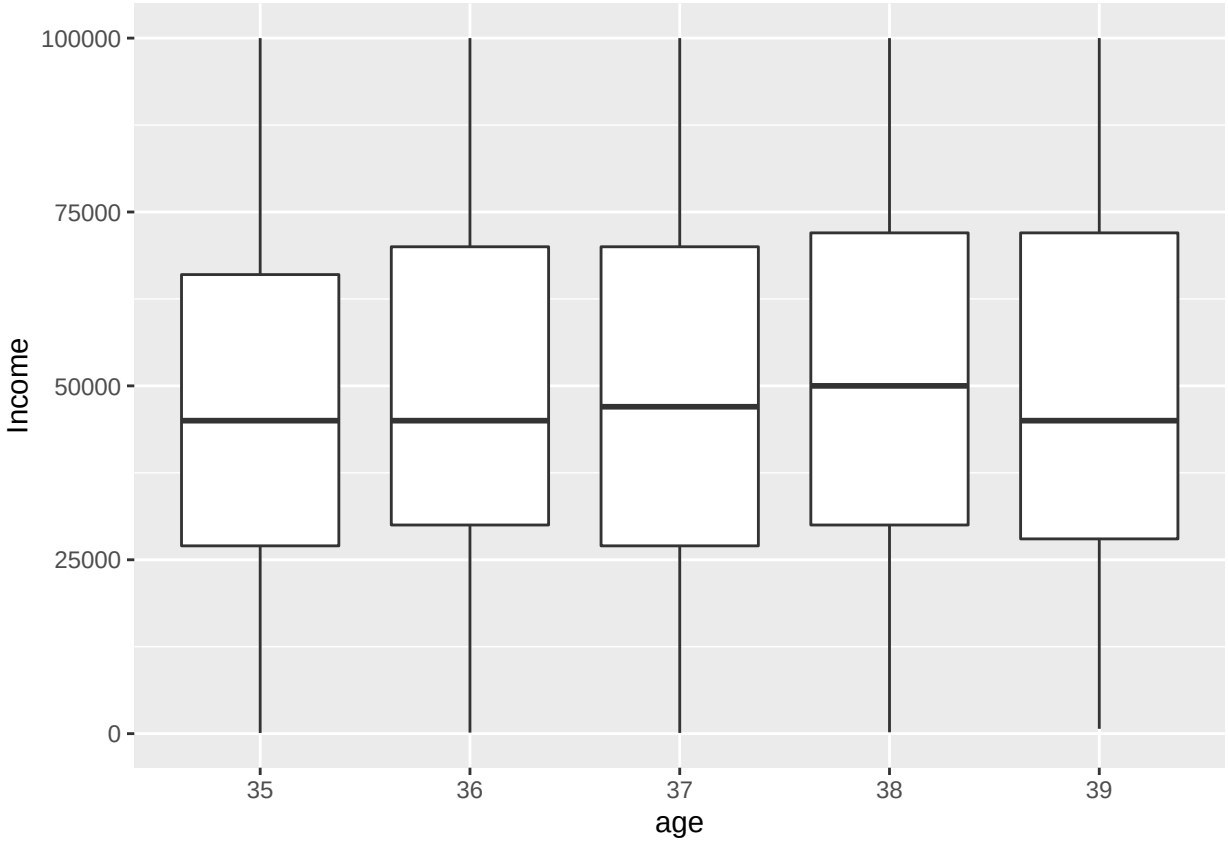
Exercise 1 Preparing the Data

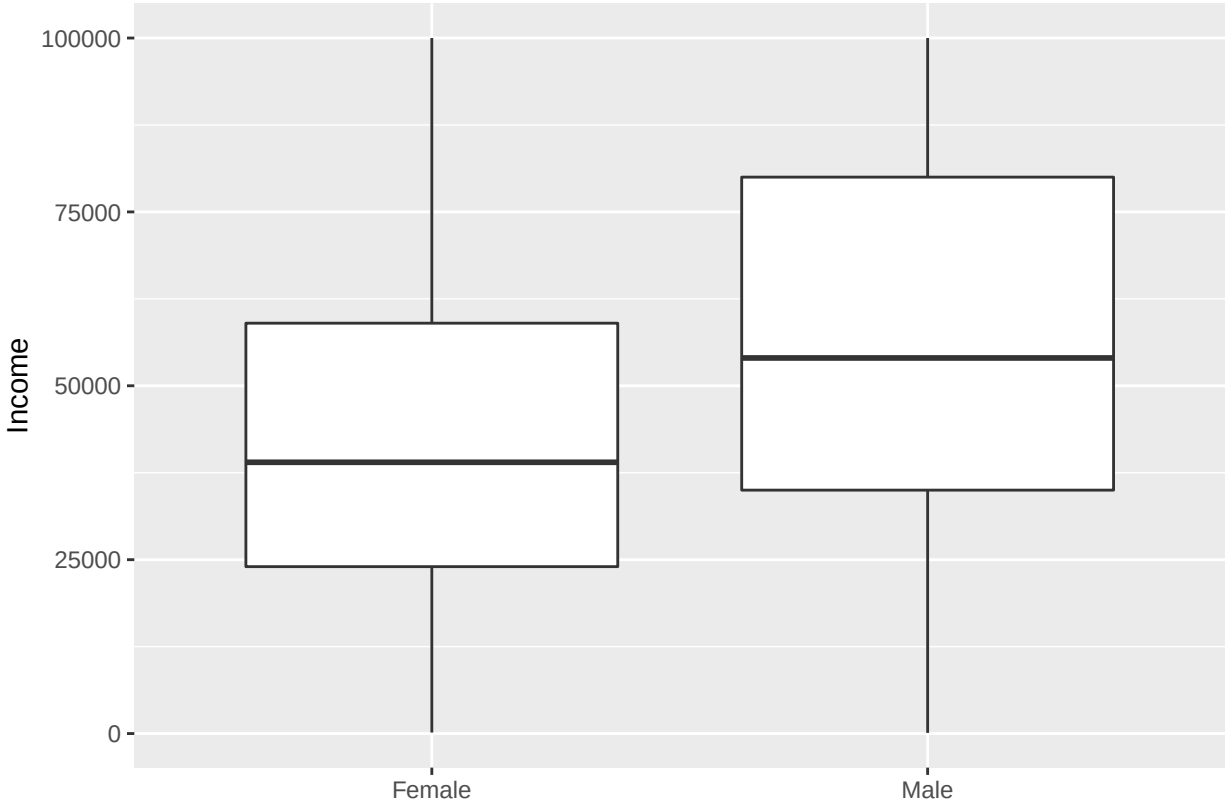
• Create additional variable for the age of the agent "age", total work experience measured in years "work exp". Hint: "CV WKSWK JOB DLI.01" denotes the number of weeks a person ever worked at JOB 01.

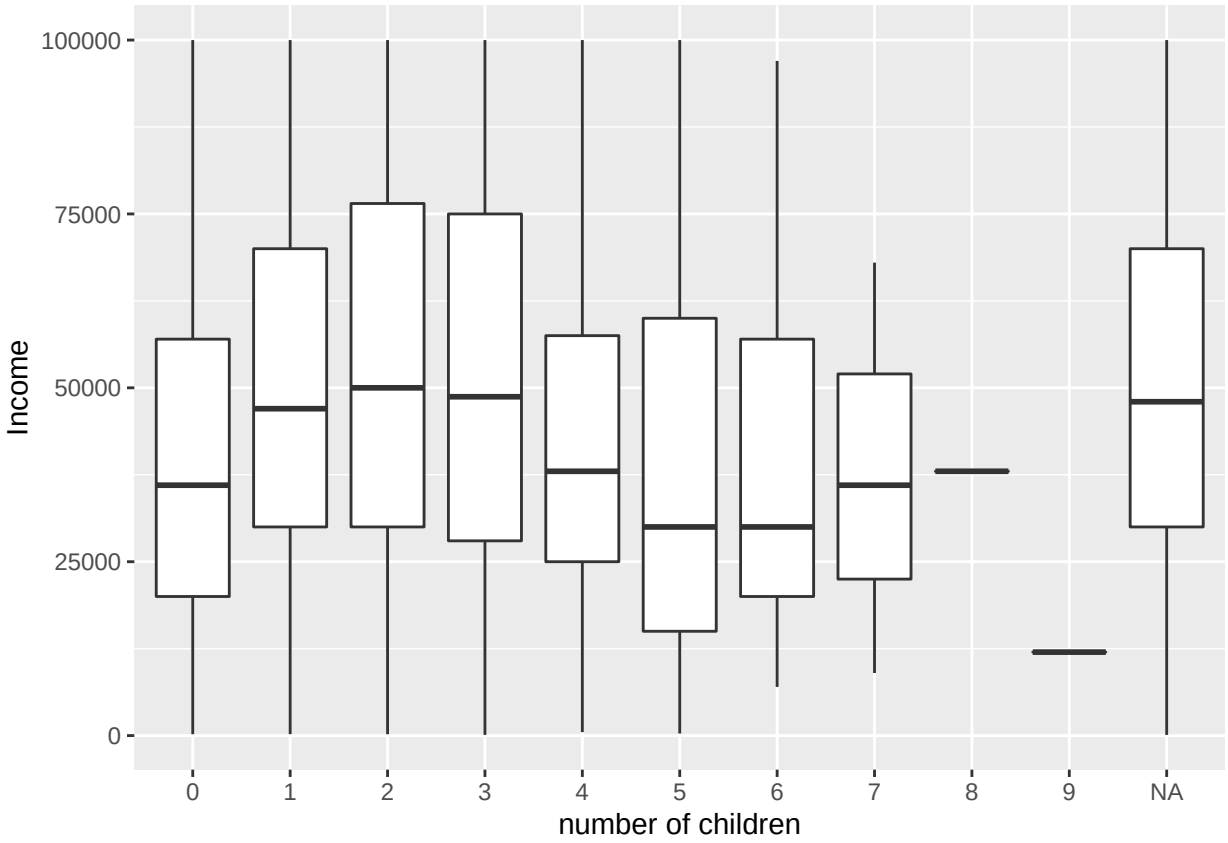
• Create additional education variables indicating total years of schooling from all variables related to education (eg, "BIOLOGICAL FATHERS HIGHEST GRADE COMPLETED") in our dataset.

• Provide the following visualizations.

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

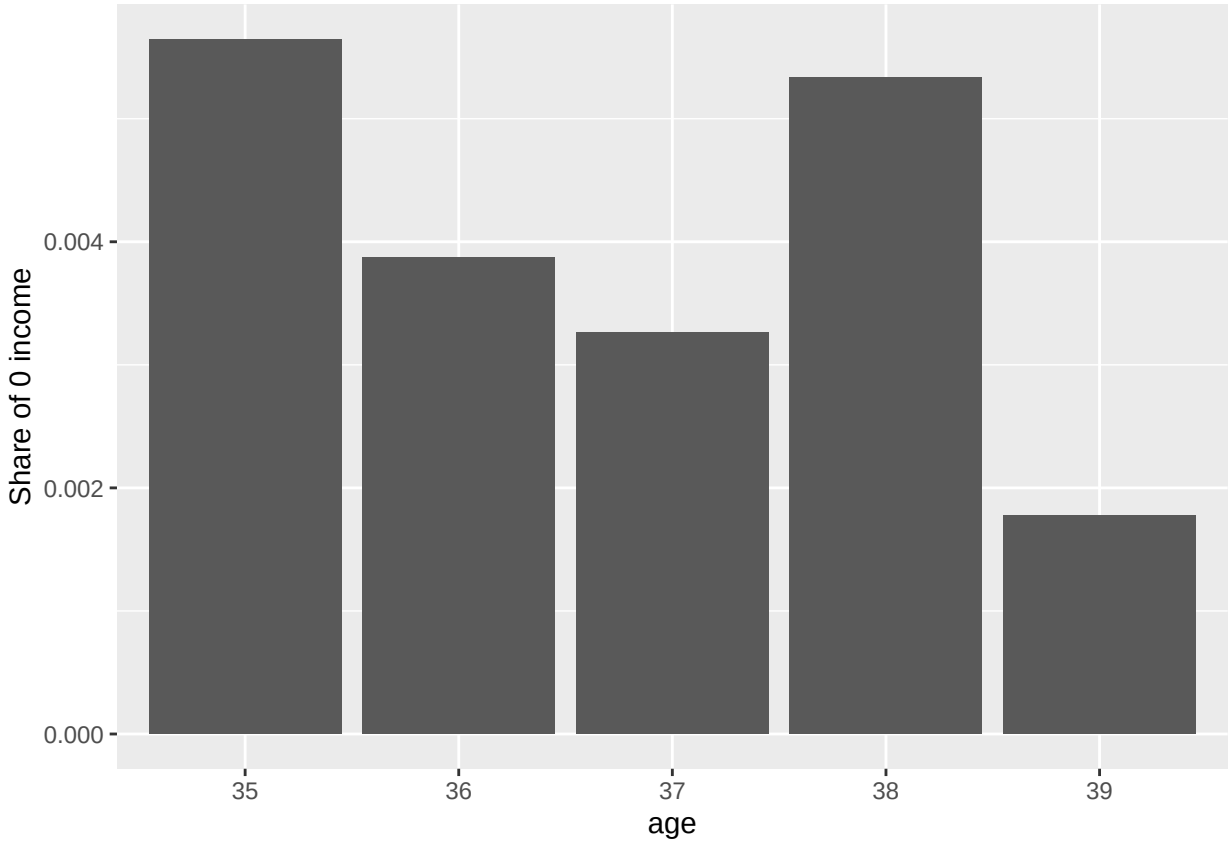


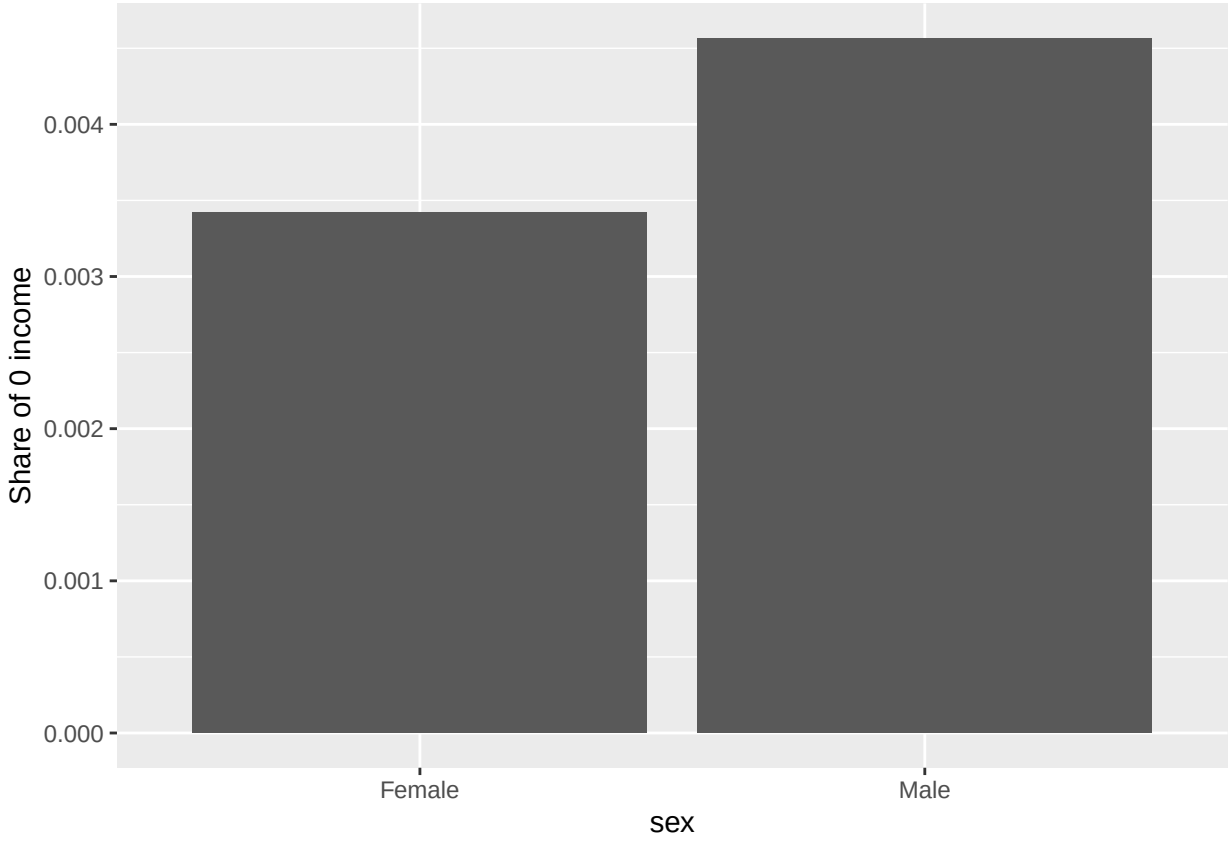


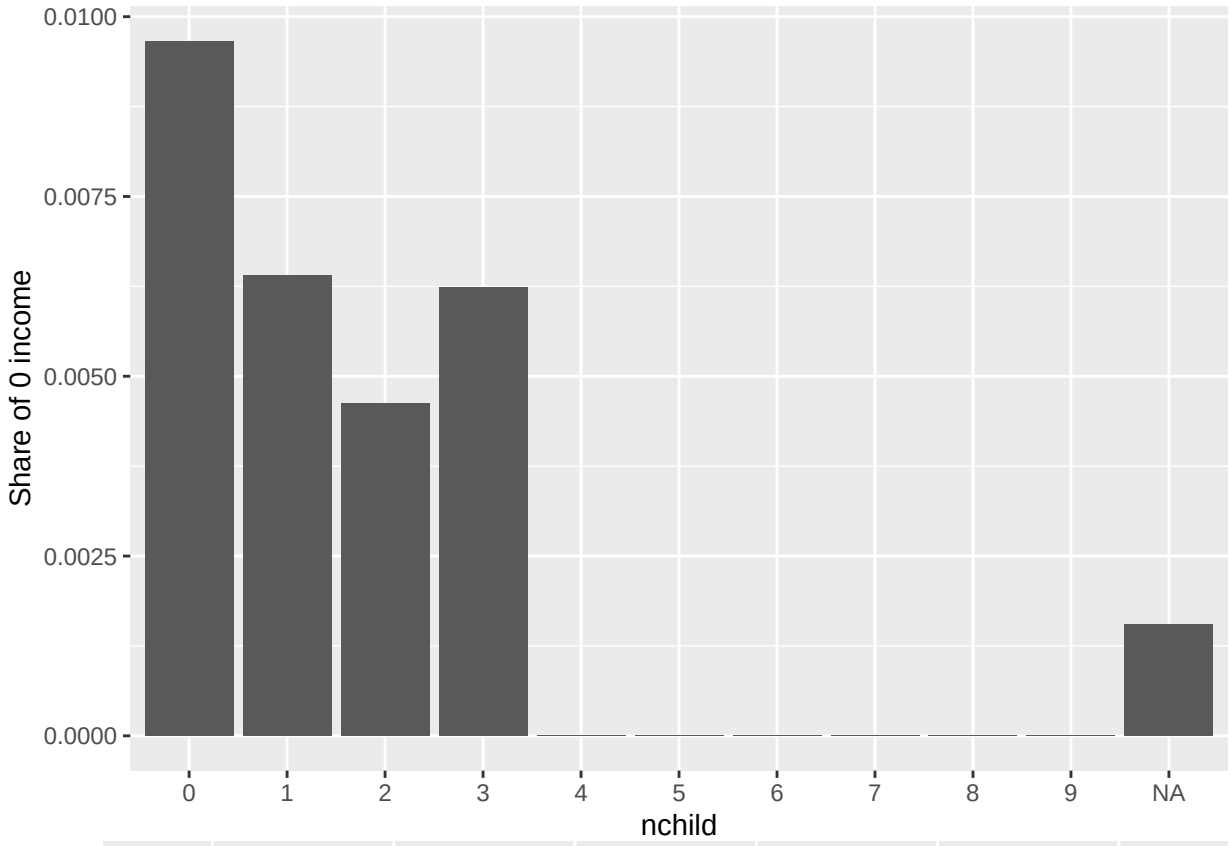


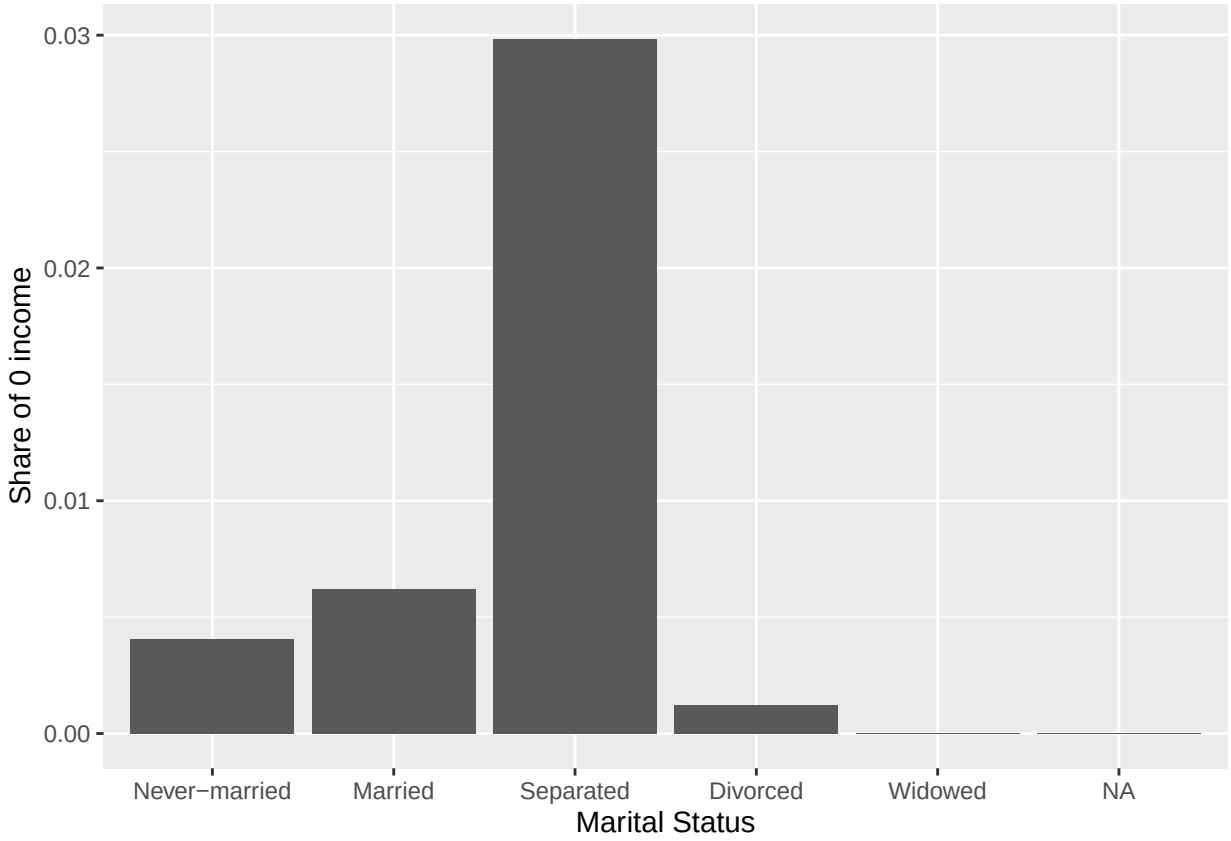
– Table the share of "0" in the income data by i) age groups, ii) gender groups, iii) number of children and

marital status









– Interpret the visualizations from above

Exercise 2 Heckman Selection Model

- Specify and estimate an OLS model to explain the income variable (where income is positive).

```
##
## Call:
## lm(formula = YINC_1700_2019 ~ work_exp + I(work_exp^2) + ysch +
##     gender + math + CV_URBAN.RURAL_2019 + mari, data = dat2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73910 -16407   -247  18832  67341
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -34373.899   5972.209  -5.756 1.16e-08 ***
## work_exp              1953.413    411.993   4.741 2.44e-06 ***
## I(work_exp^2)          -58.487     18.956  -3.085  0.00209 **
## ysch                  3613.613    347.641  10.395  < 2e-16 ***
## gender               12240.013   1618.395   7.563 9.20e-14 ***
## math                    36.779      7.593   4.844 1.48e-06 ***
## CV_URBAN.RURAL_2019   3548.950   2197.379   1.615  0.10662
## mari1                 4683.336   1796.742   2.607  0.00929 **
## mari2                 -854.990   6330.261  -0.135  0.89259
## mari3                 2674.995   3033.842   0.882  0.37815
## mari4                -8850.552  24623.239  -0.359  0.71935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24480 on 960 degrees of freedom
##   (4441 observations deleted due to missingness)
## Multiple R-squared:  0.2591, Adjusted R-squared:  0.2514
## F-statistic: 33.58 on 10 and 960 DF,  p-value: < 2.2e-16
```

– Interpret the estimation results

The results imply that education, gender, math score and working experience has a significant influence on income. Specifically, holding all other variables constant, a year more of education would result in 3613 dollars increase in income; being a male would result in a 12240 more income; a year more of working experience would result in approximately 1900 more dollars of income, this effect would dcrease with the increase of working experience.

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

The OLS is measuring with samples that have positive income, which means they have jobs and are relatively advantaged among society.

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

The two-stage model assumes a normal distribution in the first stage in order to depict the possibility of being selected (in this case, have job / income), and then include it in the second stage OLS to correct the bias.

- Estimate a Heckman selection model (Note:You cannot use a pre-programmed Heckman selection package. 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?

OLS:

```
##
```

```
## Call:
## lm(formula = YINC_1700_2019 ~ work_exp + I(work_exp^2) + ysch +
##     gender + math + CV_URBAN.RURAL_2019 + mari, data = dat2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73910 -16407   -247  18832  67341
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -34373.899   5972.209  -5.756 1.16e-08 ***
## work_exp              1953.413    411.993   4.741 2.44e-06 ***
## I(work_exp^2)          -58.487     18.956  -3.085  0.00209 **
## ysch                  3613.613    347.641  10.395  < 2e-16 ***
## gender               12240.013   1618.395   7.563 9.20e-14 ***
## math                    36.779      7.593   4.844 1.48e-06 ***
## CV_URBAN.RURAL_2019   3548.950   2197.379   1.615  0.10662
## mari1                 4683.336   1796.742   2.607  0.00929 **
## mari2                 -854.990   6330.261  -0.135  0.89259
## mari3                 2674.995   3033.842   0.882  0.37815
## mari4                -8850.552  24623.239  -0.359  0.71935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24480 on 960 degrees of freedom
##   (4441 observations deleted due to missingness)
## Multiple R-squared:  0.2591, Adjusted R-squared:  0.2514
## F-statistic: 33.58 on 10 and 960 DF,  p-value: < 2.2e-16
```

Heckman:

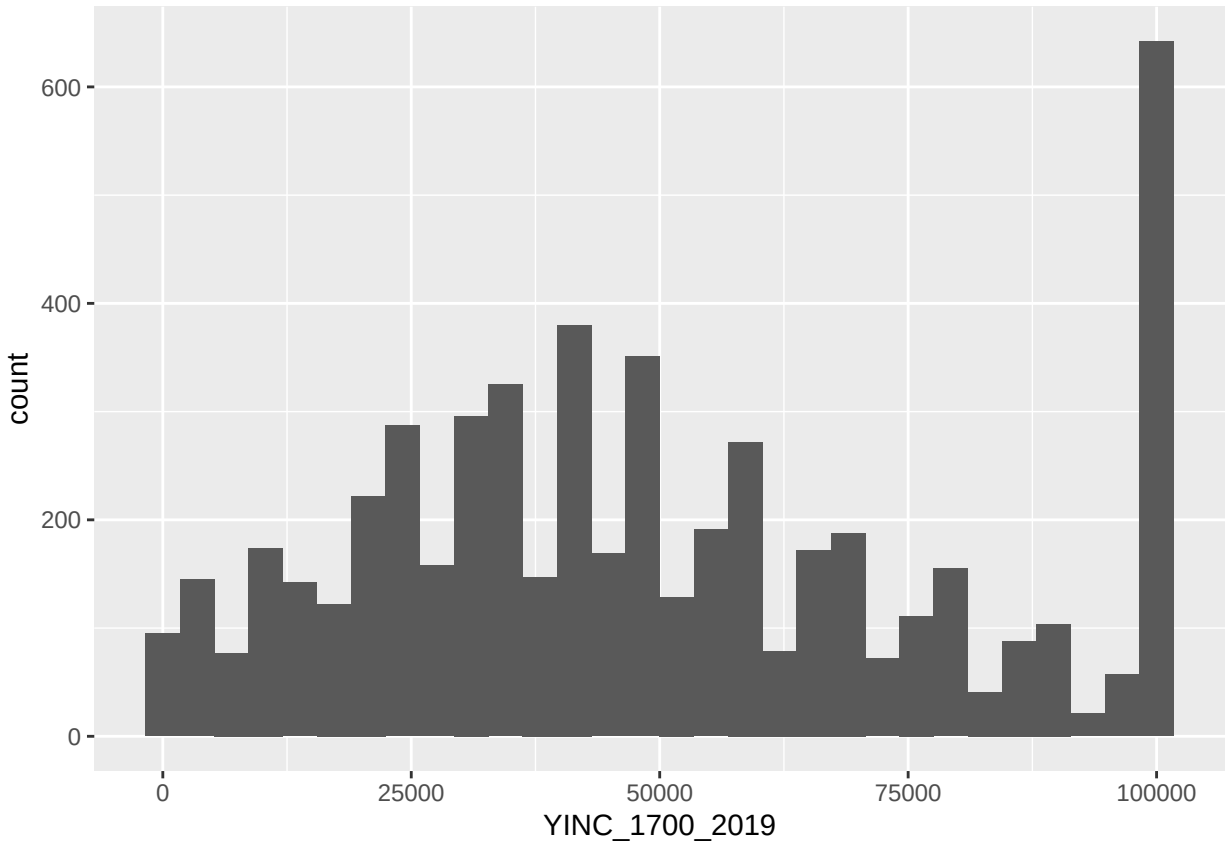
```
##
## Call:
## lm(formula = YINC_1700_2019 ~ -1 + work_exp + I(work_exp^2) +
##     ysch + gender + math + CV_URBAN.RURAL_2019 + mari + IMR,
##     data = dat2, subset = (inlf == 1))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67654 -16624   -334  18668  65455
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## work_exp              1841.888    408.580   4.508 7.36e-06 ***
## I(work_exp^2)          -51.714     18.760  -2.757 0.005953 **
## ysch                  3586.119    410.776   8.730  < 2e-16 ***
## gender               11654.723   1750.656   6.657 4.70e-11 ***
## math                    39.195      7.524   5.209 2.33e-07 ***
## CV_URBAN.RURAL_2019   3907.983   2169.171   1.802 0.071924 .
## mari0               -35192.907   9449.043  -3.724 0.000207 ***
## mari1               -30272.272   9332.119  -3.244 0.001220 **
## mari2               -36172.398  10806.892  -3.347 0.000848 ***
## mari3               -32648.489   9543.774  -3.421 0.000651 ***
## mari4               -44878.312  25581.701  -1.754 0.079698 .
## IMR                  28741.023 252852.842   0.114 0.909526
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24150 on 954 degrees of freedom
##   (4410 observations deleted due to missingness)
## Multiple R-squared:  0.8797, Adjusted R-squared:  0.8782
## F-statistic: 581.3 on 12 and 954 DF,  p-value: < 2.2e-16
```

The absolute values of the coefficients in heckman is higher, as after including people without any income, the effects would be even bigger than before.

Excercise 3 Censoring

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



The values over 100000 are censored.

• Propose a model to deal with the censoring problem.

We can use the Tobit Model to deal with the censoring problem.

• Estimate the appropriate model with the censored data (please write down the likelihood function and optimize yourself without using the pre-programmed package)

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

The absolute values of the coefficients are higher than that of OLS regression. This is basically because of Tobit model 'recreate' the censored data in a model sense.

Excercise 4 Panel Data

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

A person's upbringing, family characteristics, innate ability and demographics (except age) can influence wage and they are potential ability bias.

- Exploit the panel dimension of the data to propose a model to correct for the ability bias. Estimate the model using the following strategy.

– Within Estimator.

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = wage ~ educ + mari + work_exp, data = dat4, model = "within")
##
## Unbalanced Panel: n = 18, T = 1764-5609, N = 82008
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -78612.5 -11112.4  -2625.2   6721.7 291750.8
##
## Coefficients:
##            Estimate Std. Error t-value  Pr(>|t|)
## educ        605.176     17.697 34.1969 < 2.2e-16 ***
## mari1      8421.234    206.320 40.8165 < 2.2e-16 ***
## mari2     -1555.850    890.447 -1.7473  0.080594 .
## mari3      1204.306    425.766  2.8286  0.004677 **
## mari4    -15883.181   2511.369 -6.3245 2.553e-10 ***
## work_exp   1352.358     27.431 49.3002 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5.0781e+13
## Residual Sum of Squares: 4.7379e+13
## R-Squared:      0.067006
## Adj. R-Squared: 0.066744
## F-statistic: 981.324 on 6 and 81984 DF, p-value: < 2.22e-16
```

– Between Estimator

```
## Oneway (individual) effect Between Model
##
## Call:
## plm(formula = wage ~ educ + mari + work_exp, data = dat4, model = "between")
##
## Unbalanced Panel: n = 18, T = 1764-5609, N = 82008
## Observations used in estimation: 18
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -1371.12  -478.71   -12.51   708.07  1190.18
##
## Coefficients:
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept)    16725.12     4157.76  4.0226  0.002007 **
```

```
## educ            -967.75      234.56 -4.1258  0.001684 **
## mari1           4778.40     8349.43  0.5723  0.578626
## mari2         257228.43   137114.91  1.8760  0.087424 .
## mari3         394209.78    61551.59  6.4045 5.049e-05 ***
## mari4       -2274432.35  1128238.71 -2.0159  0.068898 .
## work_exp         481.62     1179.87  0.4082  0.690961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    4655400000
## Residual Sum of Squares: 10711000
## R-Squared:      0.9977
## Adj. R-Squared: 0.99644
## F-statistic: 795.019 on 6 and 11 DF, p-value: 7.5076e-14
```

– Difference (any) Estimator

```
## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = wage ~ educ + mari + work_exp, data = dat4, model = "fd")
##
## Unbalanced Panel: n = 18, T = 1764-5609, N = 82008
## Observations used in estimation: 81990
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -328374.561  -12050.899     -77.057   11886.673  331198.068
##
## Coefficients:
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept)     -6.1783    114.3657 -0.0540    0.9569
## educ           410.6255     17.4654 23.5107 < 2.2e-16 ***
## mari1         7435.0358    205.3421 36.2080 < 2.2e-16 ***
## mari2          228.0550    861.4948  0.2647    0.7912
## mari3         2335.0898    416.3777  5.6081 2.052e-08 ***
## mari4       -10580.2992   2416.6730 -4.3780 1.199e-05 ***
## work_exp      1272.0482     26.8011 47.4626 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    9.2644e+13
## Residual Sum of Squares: 8.7918e+13
## R-Squared:      0.051013
## Adj. R-Squared: 0.050943
## F-statistic: 734.5 on 6 and 81983 DF, p-value: < 2.22e-16
```

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

The between model has the best goodness of fit. In the within model, the results implies the direct relationship while including fixed effect. In the between model, it shows the relationship of the dependent variable in a period. In the difference model, it means that one unit increase of the independent variable in a period would result in a bigger difference in the dependent variable in that period. For example, having one more year in education in the period would result in 460 more dollars earning.

For the different estimators, the independent variables X and predicted Y values are calculated in different

ways, thus yield different parameter estimates. For First difference, it would be $Y_{diff} = (Y_{t_i} - Y_{t_{i-1}})$, for between model, it would be $Y_{t_i} - \bar{Y}_i$, for the between model, it would be just $\bar{Y}$.