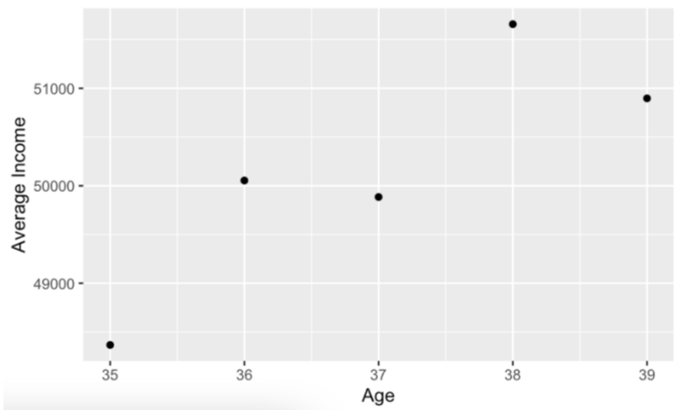
Econ 613 Homework 4
Hannah Marsho

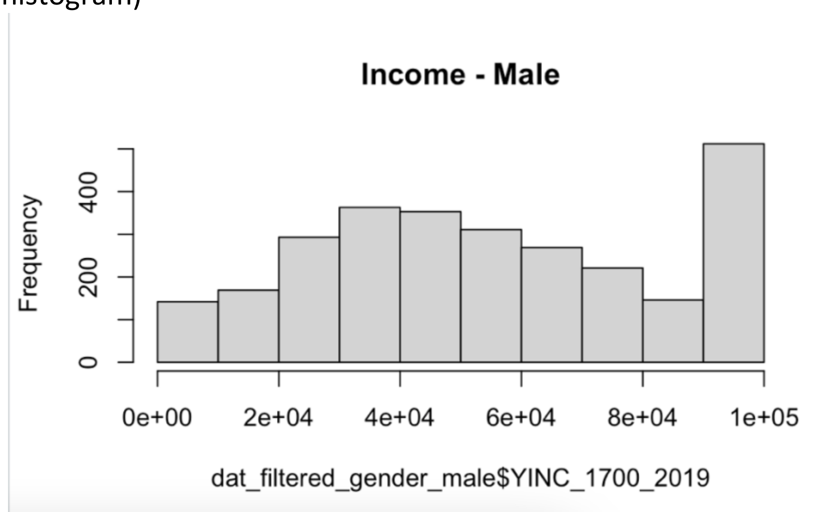
Exercise 1:

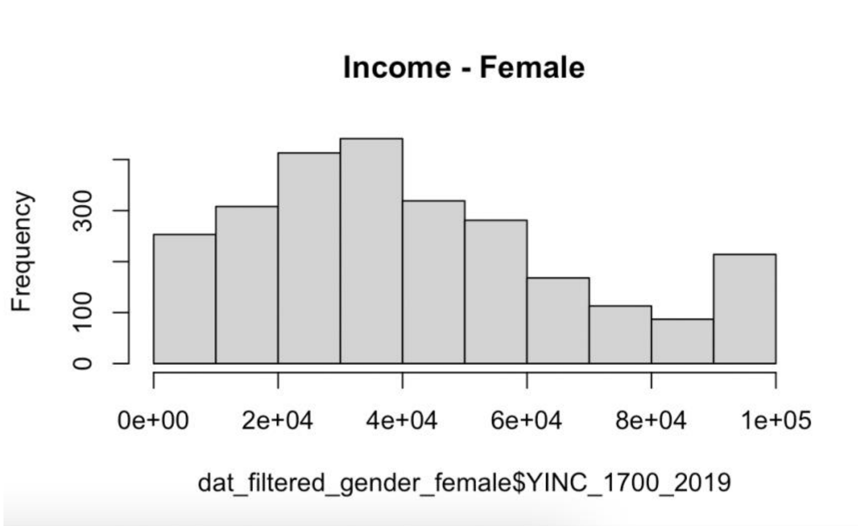
1a) The requested variables are called age_final work_exp and are found in data set dat.
1b) The requested variables are called average_grade_parent and years_education and are found in data set dat.
1ci) The visualizations are found below:

i) Positive income data plotted by age groups

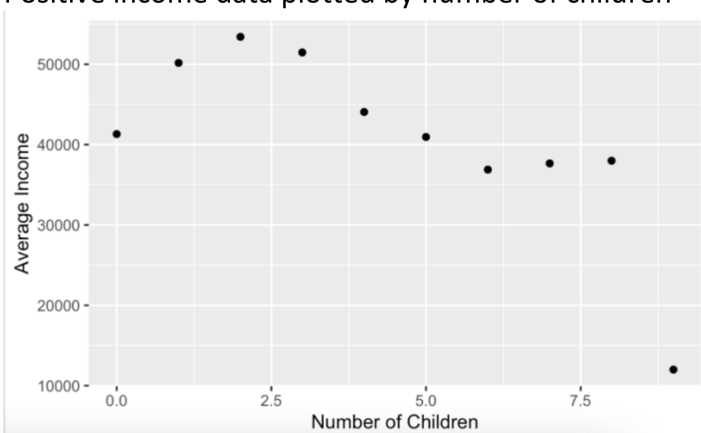


ii) Positive income data plotted by gender groups (each gender received its own histogram)





iii) Positive income data plotted by number of children



1cii) The visualizations are found below:

i) Tabled share of "0" in the income data by age groups

	age_final	N	num_zeros	share_zeros
1	35	1771	705	0.3980802
2	36	1807	703	0.3890426
3	37	1841	740	0.4019555
4	38	1874	768	0.4098186
5	39	1691	692	0.4092253

ii) Tabled share of "0" in the income data by gender groups

	KEY_SEX_1997	N	num_zeros	share_zeros
1	1	4599	1820	0.3957382
2	2	4385	1788	0.4077537

iii) Tabled share of "0" in the income data by number of children and marital status groups

	CV_BIO_CHILD_HH_U18_2019	CV_MARSTAT_COLLAPSED_2019	N	num_zeros	share_zeros
1	0	0	422	179	0.4241706
2	0	1	151	37	0.2450331
3	0	2	36	17	0.4722222
4	0	3	207	61	0.2946860
5	0	4	2	0	0.0000000
6	1	0	481	124	0.2577963
7	1	1	704	96	0.1363636
8	1	2	23	8	0.3478261
9	1	3	178	30	0.1685393
10	1	4	7	2	0.2857143
11	2	0	361	98	0.2714681
12	2	1	1131	193	0.1706454
13	2	2	32	8	0.2500000
14	2	3	188	38	0.2021277
15	2	4	8	2	0.2500000
16	3	0	164	53	0.3231707
17	3	1	542	109	0.2011070
18	3	2	9	3	0.3333333
19	3	3	82	15	0.1829268
20	4	0	64	28	0.4375000
21	4	1	167	43	0.2574850
22	4	2	8	3	0.3750000
23	4	3	25	7	0.2800000
24	5	0	19	7	0.3684211
25	5	1	45	16	0.3555556
26	5	2	2	1	0.5000000
27	5	3	3	1	0.3333333
28	5	4	1	0	0.0000000
29	6	0	10	6	0.6000000
30	6	1	12	3	0.2500000
31	6	2	1	0	0.0000000
32	7	0	1	1	1.0000000
33	7	1	4	1	0.2500000
34	8	1	2	1	0.5000000
35	9	0	1	0	0.0000000

1ciii) For our positive income data plots by group… As age increases, the mean income for that age also increases. Female mean income is lower than male mean income, but the distributions are also different. More of the male incomes are top-coded and the distribution is flatter. As number of children increases (while accounting for marital status), the mean income for that number of children decreases. For our tabled shares of “0” in the income data by group… For the age groups, the share of “0” is roughly the same across each age group. For the gender groups, the share of “0” is also roughly the same. For the number of children and marital status groups, there is much greater variety in the share of “0” across groups. Some groups have no zeros, whereas other groups can have upwards of over 40-50% zeros.

<u>Exercise 2:</u>

2a) The OLS estimates are found below:

```
Residuals:
   Min     1Q Median     3Q    Max
-93135 -15828  -1452  15458  77324

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                43681.01   10093.00   4.328 1.54e-05 ***
age_final                    257.83     268.74   0.959  0.33742
work_exp                    1058.48      71.28  14.849  < 2e-16 ***
average_grade_parent         616.75      98.92   6.235 5.01e-10 ***
years_education14           8530.12    1135.28   7.514 7.10e-14 ***
years_education16          19136.76    1011.05  18.928  < 2e-16 ***
years_education18          29082.35    1360.43  21.377  < 2e-16 ***
years_education21          36646.29    2316.55  15.819  < 2e-16 ***
KEY_SEX_1997              -19875.56     764.78 -25.988  < 2e-16 ***
CV_BIO_CHILD_HH_U18_2019    1250.90     331.03   3.779  0.00016 ***
CV_MARSTAT_COLLAPSED_2019   1520.06     403.97   3.763  0.00017 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23230 on 3888 degrees of freedom
  (1477 observations deleted due to missingness)
Multiple R-squared:  0.345,     Adjusted R-squared:  0.3433
F-statistic: 204.8 on 10 and 3888 DF,  p-value: < 2.2e-16
```

i) All variables except for the participant’s age in the final panel year have a highly significant impact on income. However, our R-squared value is low at only 0.345, so much of the variation is not captured in our model.

ii) By estimating OLS in this way, we can potentially run into a selection problem. This is because we have removed many of the income data points (due to them being NA’s or 0’s). There might have been some non-random reason for why those data points were NA’s or 0’s. Something about those participants could have been systematically different. If this were true, then we would have a selection problem and our standard OLS estimation would be biased.

2b) The Heckman selection model can help through correcting for any of those potential non-random reasons for why data points are NA’s or 0’s. The model achieves this through modelling the individual sampling probability for each participant and then creating the conditional expectation for our dependent variable (income).

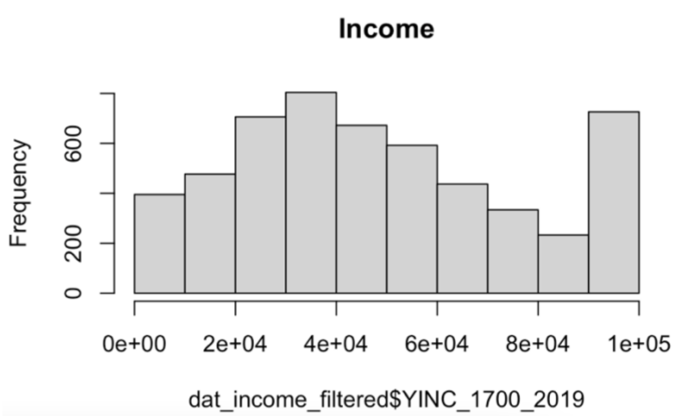
2c) The Heckman selection model estimates are found below:

```
Residuals:
   Min     1Q Median     3Q    Max 
-84349 -19852  -3510  17909  86032 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  21682.0    10509.9   2.063   0.0392 *  
x1            7239.2      753.9   9.602  < 2e-16 ***
x2           -6089.4      848.5  -7.176 8.19e-13 ***
x3           -3098.2      494.1  -6.271 3.89e-10 ***
x4          -32197.4     1789.2 -17.995  < 2e-16 ***
x5           18886.9     1758.7  10.739  < 2e-16 ***
x6          -11208.5     1469.9  -7.625 2.89e-14 ***
invM_ratio  -20777.5     2136.8  -9.724  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27610 on 5085 degrees of freedom
  (3891 observations deleted due to missingness)
Multiple R-squared:  0.2925,	Adjusted R-squared:  0.2915 
F-statistic: 300.4 on 7 and 5085 DF,  p-value: < 2.2e-16
```

The results are slightly different. Now, all the variables are highly significant, including the participant's age in the final panel year. Our R-squared has decreased to 0.2925 though. The signs on several of the coefficients have also switched.

<u>Exercise 3:</u>

3a) The requested histogram is found below:



From checking the data set, we can see that the highest possible value is $100,000 for income. So the top-coded/censored value/mass point is simply $100,000.

3b) To deal with the censoring problem, we can use a tobit model. With a tobit model, we modify the likelihood function to reflect the unequal sampling probability for each of the sample's participants depending on where the participant's dependent variable falls with respect to the mass point.

3c) The appropriate model is called result2 and can be found in the code and below:

```
> result2$par
[1]  0.2375039  7.0038668  5.7292566  7.4477772  3.3116237 -0.5911932  2.2177975
```

3d) The results are slightly different. The coefficient magnitudes are different, but the signs remained the same.

Exercise 4:

4a) For participants who have higher innate abilities, their wages will tend to be higher as well. This is because those participants are likely more productive, intelligent, charismatic, etc. However, we don't have a variable to account for ability in our data set. So, our estimates could potentially have an ability bias. This is an omitted variable bias where the beneficial effect of having higher innate abilities is falsely attributed to our other variables.
4b) The requested models created with each of the three estimation strategies are called within_regression, between_regression, and fd_regression and can be found in the code.
4b)

i) The within estimator regression results are found below:

```
Residuals:
    Min      1Q  Median      3Q     Max
-137919   -8679    -367    7282  281012

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1649.11      71.91  -22.93   <2e-16 ***
work_exp_diff        2337.48      24.64   94.85   <2e-16 ***
education_diff       6277.90      55.83  112.45   <2e-16 ***
marital_status_diff  7546.39     136.49   55.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19690 on 82004 degrees of freedom
  (88688 observations deleted due to missingness)
Multiple R-squared:  0.3452,    Adjusted R-squared:  0.3452
F-statistic: 1.441e+04 on 3 and 82004 DF,  p-value: < 2.2e-16
```

ii) The between estimator regression results are found below:

```
Residuals:
   Min     1Q Median     3Q    Max
-43661  -9017  -2670   5679 149206

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -43622.51    1686.73 -25.862  < 2e-16 ***
mean_work_exp        2675.58      91.76  29.158  < 2e-16 ***
mean_education       4593.97     133.81  34.333  < 2e-16 ***
mean_marital_status  2519.16     317.14   7.943 2.21e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14140 on 8693 degrees of freedom
  (287 observations deleted due to missingness)
Multiple R-squared:  0.2322,    Adjusted R-squared:  0.232
F-statistic: 876.5 on 3 and 8693 DF,  p-value: < 2.2e-16
```

iii) The first difference estimator regression results are found below:

```
Residuals:
    Min      1Q  Median      3Q     Max
-212047   -4998   -1847    3936  322684

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3688.15      70.84  52.062  < 2e-16 ***
work_exp_fd         590.35      34.56  17.080  < 2e-16 ***
education_fd       -215.79     106.08  -2.034  0.04194 *
marital_status_fd   625.68     191.60   3.266  0.00109 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16230 on 58592 degrees of freedom
  (112100 observations deleted due to missingness)
Multiple R-squared:  0.005218,  Adjusted R-squared:  0.005167
F-statistic: 102.4 on 3 and 58592 DF,  p-value: < 2.2e-16
```

4c) For the within estimator…

All else equal, an additional year of work experience increases income by an expected $2337.48.
All else equal, an additional year of education increases income by an expected $6277.90.
All else equal, getting married increases income by an expected $7546.39.

For the between estimator…

All else equal, an additional year of work experience increases income by an expected $2675.58.
All else equal, an additional year of education increases income by an expected $4593.97.
All else equal, getting married increases income by an expected $2519.16.

For the first difference estimator…

All else equal, an additional year of work experience increases income by an expected $590.35.
All else equal, an additional year of education decreases income by an expected $215.79.
All else equal, getting married increases income by an expected $625.68.

Each model produces significantly different results. For each model, each coefficient retains the same sign except for with our education variable. In the first difference model, our education coefficient becomes negative. The magnitudes of the coefficients are generally larger in the within estimator model compared to the other two models. The magnitudes of the coefficients in the first difference model are, however, much smaller than those in the other two models. These differences stem from the different ways that we calculated each model’s estimators. For the between estimator model, we removed all time variation. For the within estimator model, we removed all individual variation. For the first difference estimator model, we were able to preserve both forms of variation, possibly making this a more sensible model.