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:

1. Positive income data plotted by age groupsChart, scatter chart

   Description automatically generated
2. Positive income data plotted by gender groups (each gender received its own histogram)

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

1. Positive income data plotted by number of children Chart, scatter chart

   Description automatically generated

1cii) The visualizations are found below:

1. Tabled share of “0” in the income data by age groups

Table

Description automatically generated

1. Tabled share of “0” in the income data by gender groupsTable

   Description automatically generated
2. Tabled share of “0” in the income data by number of children and marital status groupsTable

   Description automatically generatedTable

   Description automatically generated

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.

Exercise 2:

2a) The OLS estimates are found below:Table

Description automatically generated

1. 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.
2. 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:

Exercise 3:

3a) The requested histogram is found below:

Chart, histogram

Description automatically generated

From checking the data set, we can see that the highest possible value is $100,000

# for income. So the top-coded/censored value is simply $100,000.

3b)