**Exercise 1 – Preparing the Data**

1. For age, I looked at their birth year and calculated age = 2019 – birth year. I chose to do 2019 instead of 2022 as I am using their age at the time of data collection (2019). There may be issues with months but I didn’t think too much about them. For work experience, there were, up to 11 different experiences we had to consider. I added up all the non-empty data and divided by 52 to get the equivalent years of work experience. Please refer to my code for results (added to dat\_A4 data frame).
2. **I dropped the data that didn’t have information on all the education variables**. For total years of schooling, I looked at the education for biological and residential mom and dad as well as individual education. For all parent education, it was already in years (for the most part) so I just recoded ungraded (95) to 0 as I assume ungraded education is equivalent to 0 years of education. For the individual, I recoded the categorical variables into equivalent years of schooling (general case):
   1. None = 0
   2. GED or HS degree = 12
   3. Associate = 14
   4. Bachelors = 16
   5. Master = 18 (assumed 2 years after bachelor)
   6. PHD = 22 (assumed 6 years after bachelor)
   7. Professional Degree = 19 (assumed 3 years after bachelor).
3. Visualizations
   1. Plotting
      1. Plots for when income is positive by age, gender, and number of children:

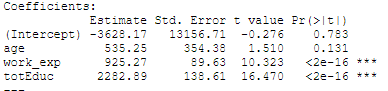
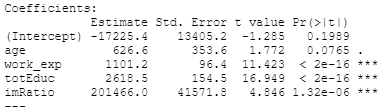
|  |  |  |
| --- | --- | --- |
|  |  |  |
| Age It seems as though there is rough equality in incomes between ages. Highest median at 37 but otherwise the IQR and general distribution seems similar. | Gender For gender it looked as though the distribution in income was higher and (slightly) more varied for males (1) versus females (2) as the IQR seemed slightly larger. | Number of Children For the number of children, there seemed to be a wave for the median income where from 0 to 2 it increased and then from 2 to 6 it decreased and then it increased up to 8. The IQR seemed roughly equal for 0 to 7 children but was tiny for 8 . Interestingly, some had missing data on children but that was included in the graphs (**as I didn’t drop those with missing children**). |

* 1. Table

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Age By age, a small portion of people had no income at age 35. That value decreased at 36 and went up a bit at 37. At 38, the value jumped up (slightly lower than at age 35) and then at 39, it dropped down to the lowest proportion of no income | Gender For gender, a slightly smaller proportion of males (1) had no income relative to female proportion (but very similar). | Number of Children  As the number of children increased (starting at 1 child), the proportion of people with no income decreased (from 1 to 2, it was roughly the same but 2 was slightly higher and 3 dropped a lot). Beginning at 4 kids, the proportion was 0 up to 10ish decimal places. Less people with no kids had 0 income compared to those with 1-3 kids. |

* 1. Did Above

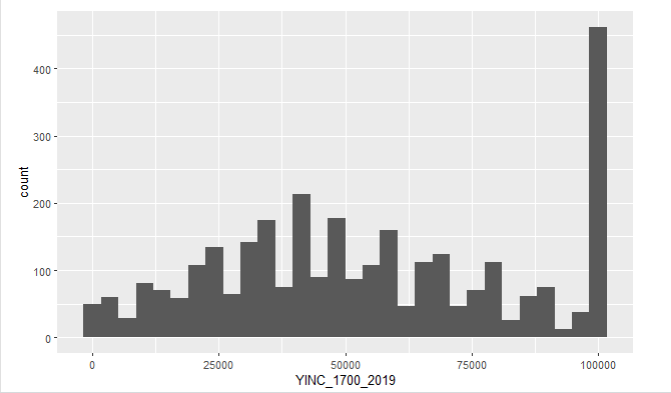
**Exercise 2 – Heckman Model**

1. . Used data for when income was positive and **filtered out data where age, work experience, total education, and any parent education was missing.** Got the following OLS results for positive income:  
     
   Says that increasing age by 1 year corresponds to a $535 increase in income (for those with positive income). And an increase in work experience and total education by 1 unit (year in both cases) corresponds to an increase in $925 and $2282 in income, respectively.  
     
   There might be issues of top-coded income data. Ie, individuals who make over 100k are coded as making 100k which could lead to issues of validity and inconsistent parameter estimates. Furthermore, there could also be issues due to truncation as we are only looking at positive incomes (and not including 0) so we are truncating at 0 (and not seeing those values) which could lead to issues as we are doing non-random selection.
2. Heckman offers a way to correct for non-random selected samples which is an issue we have since we are looking at positive income data. Truncated data where we do not see data for those with 0 income which can have a selection issue.
3. For selection, I am including the parental education but don’t include them in the second part. For the first step, I do a probit model for selection (positive income or not). Then, I calculate the inverse mills ratio and include it in the OLS regression.  
   

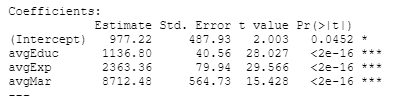
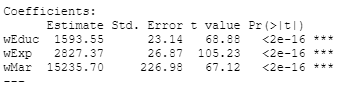
Now, the coefficients for age, work experience and education are much larger (but similar interpretations as before for normal OLS just with different values). The positive imRatio indicates that the observed income tends to be higher (which seems to make sense in this situation). Since the test statistic is large, we can reject the null and conclude there are some selection issues.

Why a difference?  
Firstly, the estimates are different now but the signs are the same. The Heckman estimates look to be larger too. This can be because we are now accounting selection issues (positive income) in the Heckman Model due to non-random missing data (only positive income). All three variables chosen (age, work\_exp, and total education) now seem to be statistically significant (namely age becomes more significant) and this can be because we now account for the non-random sampling which led to selection issues.

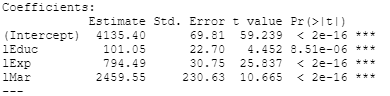
**Exercise 3 – Censoring**

1.   
   For the most part, it seemed relatively symmetric around 40k ish but then there is a huge spike at 100k which indicates that the data may be top-coded at 100k. This means that individuals making over 100k are seen as making exactly 100k in the data even though they make more. So the 100k frequency includes all people making at least 100k.
2. I will use the same regressors as I did for OLS. I won’t use the parental education data as I cannot rationalize why that would explain individuals making over 100k and with the tobit model, the regressors need to be the same (and adding variables for the sake of adding them seems unnecessary). Because of convergence issues, I used a packaged program and set my starting guess around that (added some perturbations). I followed the steps in class which include using both pdf and cdf in the likelihood as well as denoting, in the data, an indicator d for when income was below 100k (coded as 1 and 0 if higher).  
   y = y\* if y\* <100k and y = 100k if y\* 100k.  
   Likelihood Function: . I believe I should use 1-cdf as we are topcoding (example in notes uses cdf but is for bottomcoding.
3. Intercept, age, work experiences, and education: -5413.738 4054.260 1235.846 3013.883.
4. Again, increasing age, work experience, and education have similar interpretations but now the values seem much larger (for each estimate in terms of magnitude). This could indicate topcoding is an issue here because it dilutes effects.

**Exercise 4 – Panel Data**

1. This ability bias can be seen as the actual effect of additional schooling on earnings and the observed gap among workers with varying education levels. There can be issues such as those with higher ability tend to be in school longer but also tend to be more efficient (and get higher wages) which violates OLS assumptions. It is also difficult because ability is not easily measured. Panel methods, assuming ability is fixed, can get 'netted out'. For the data, I created little sub datas for income, marital status, education, and work experience. I then converted the experience into years of experience (by year – there were potential issues with the data regarding if it is cumulative or not but I went ahead and assumed the experience in job 01 was cumulative). I then converted all these sub datas into long format and merged them together to get relData. I then **dropped any observations where there were missing values**. I also **recoded the marital status as 0 if they were not married in that year and 1 if they were married**. Important to note that, since I removed all rows with incomplete data, the years between each observation may be inconsistent and when I checked which observations had 19 years worth of data, the result was 0 but TA said to proceed in this fashion. MAKES NO SENSE TO AVERAGE MARITAL STATUS THOUGH BUT I DID IT.
2. The model depends on the estimator chosen:
   1. Between Estimator – just take averages for each individual  
      I grouped the data by individual and simply calculated the means to get betData.  
        
      
   2. Within Estimator – subtract each value by the average  
      Important to note that within estimator does not include an intercept. I basically just subtracted the average from each value (in my long data) to get withinData.  
      
   3. First Difference Estimator - Current Value – Previous Value

Because of dropping NAs, sometimes the lag was more than a year ( ie if person had missing data in 2000 but had data in 99 and 01. Using the lag function to get these lagged differences after grouping by individual to get lagData.



1. The estimates all say that education, experience, and marital status are positively associated with income (but to varying degrees) and are all statistically significant. The first discusses, on average, the relationship between Y and X. The second discusses adjusted relationship between Y and X (adjusted in the sense that all values are the difference from the means). The third discusses the trend over time (as we are taking differences in consecutive observations). These parameters are different because the process for constructing them is inherently different and each come with their own drawbacks. Between estimators which just use the averages so it completely disregards time variation. Within-estimators disregard the variations in data. First difference can have problems depending on the difference used (ie first versus lagged versus final) and there needs to be enough variation for this to be good.