ECON613: Applied Econometrics in Microeconomics

Problem Set # 4

Josh(Yueru) Li

Due Date: 11 PM, Wed April 20

Problem 1. Preparing the Data

1. Create additional variable for the age of agent, total work experience measured in years.

   Please refer to the code for implementation. For age, I calculated the age by months and divided by 12 to get years. I then round the decimal to nearest integer. For work experience in years, I summed up all the work experience at different jobs by weeks and divided by 52 to get years, abstracting away from the fact that a year is usually not exactly 52 weeks.

2. Create additional education variable indicating total years of schooling from all variables related to education.

   Please refer to the code for implementation. I dropped all the rows where there's an NA value among Biological parents' education, Residential parents' education and self education. For education of biological or residential parents, I re-code 95 to 0, assuming that ungraded equal to 0 years of education. For self-education, I make the following conversion. I treat none as 0 years of education, GED and high school as 12 years of education, Associate as 14 years of education, Bachelor as 16 years of education, Masters as 18 years of education, PhD as 22 years of education and professional degree as 20 years of education.

3. Provide the following visualization.

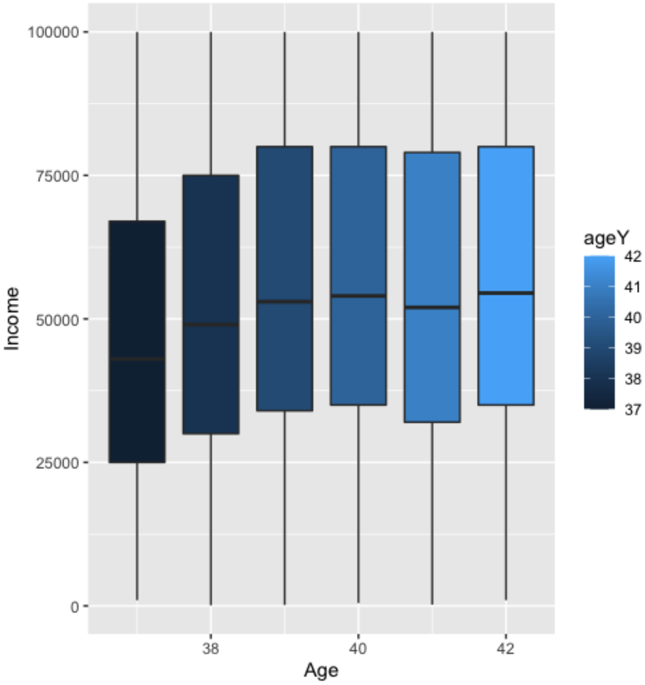

Figure 1: Income by Age Group

(a) Plot the data (where income is positive) by age group, gender groups and number of children.

Please refer to figure 1, 2 and 3. Notice that for the figure grouped by number of children, the number -1 correspond to the entries in the data where the number of children field is NA.

(b) Table the share of 0 in the inome data by age group, gender groups, number of children and marital status.

I first convert those with income as NA to 0, and then create the following tables.

Age Table

Income	Age 37	Age 38	Age 39	Age 40	Age 41	Age 42
Non-Zero	78.81%	80.03%	80.05%	81.05%	82.44%	80.60%
Zero	21.19%	19.97%	19.95%	18.95%	17.56%	19.40%

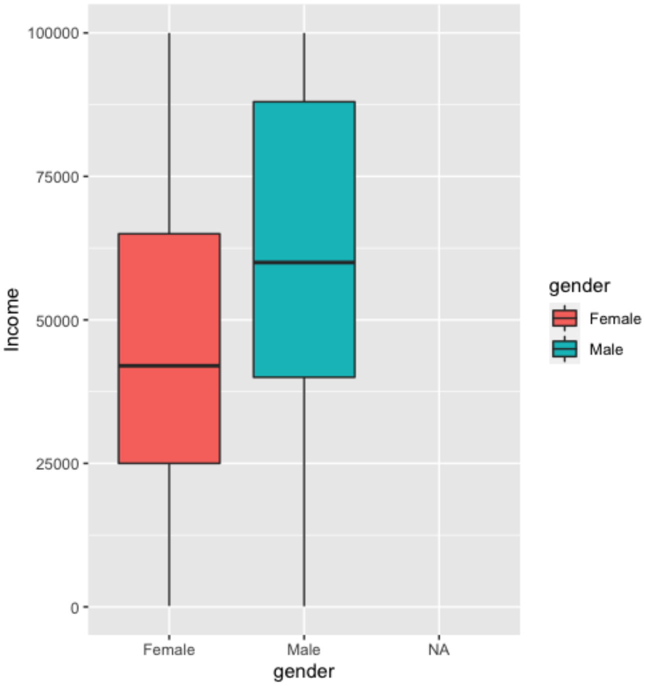

Figure 2: Income by Gender Group

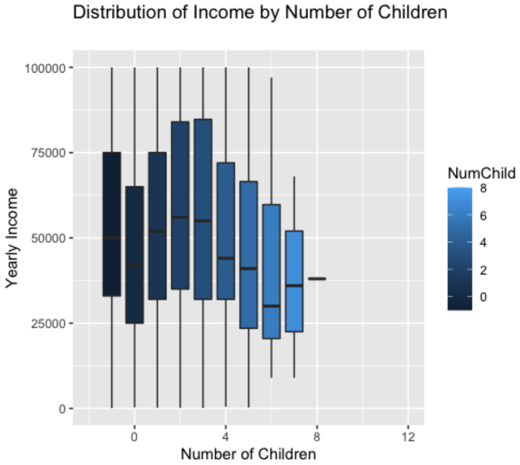

Figure 3: Income by Number of Children

Gender Table		
Income	Male	Female
Non-Zero	76.62%	84.83%
Zero	23.38%	14.17%

For the following number of children table, I keep less significant digit to fit the table in one line.

Number of Children Table									
Income	NA	0	1	2	3	4	5	6	7	8
Non-Zero	82%	68%	86%	83%	80%	65%	69%	63%	75%	50%
Zero	18%	32%	14%	17%	20%	35%	31%	36%	25%	50%

The last table is the zero income proportion by marital status table

Marital Status Table						
Income	NA	Never Married	Married	Separated	Divorced	Widowed
Non-Zero	41.18%	75.45%	84.62%	66.04%	81.17%	77.78%
Zero	58.82%	24.55%	15.38%	33.96%	18.83%	22.22%

Finally, I generate the following table for number of children and marital status. The number of each cell correspond to the percentage that make zero income conditioning on the variable values specified by that row and column. NA correspond lacking data to calculate percentage in that cell.

Number of Children and Marital Status Table										
Marital(NumChild)	NA	0	1	2	3	4	5	6	7	8
NA	60%	25%	50%	100%	0%	0%	NA	NA	NA	NA
NeverMarried	21%	43%	22%	24%	21%	29%	50%	67%	100%	NA
Married	10%	18%	11%	15%	20%	32%	32%	25%	0%	50%
Separated	13%	50%	50%	14%	25%	75%	0%	NA	NA	NA
Divorced	19%	23%	9%	21%	18%	50%	0%	NA	NA	NA
Widowed	25%	0%	33%	0%	NA	NA	0%	NA	NA	NA

(c) interpret the visualizations from above.

For those with jobs making positive income, older people and Male tend to make more money. The number of children is correlated with income in a quadratic way, with people having 2 to 3 children making the most money and those making less or more make less, with income decreasing with the deviation from the 2 to 3 center.

For the propensity to work and make positive income, I make the following observation.

Regardless of age, the percentage of people making positive income tend to be around 80%, with those at the age of 37 slightly lower at 78% and those at the age of 41 slightly higher at 82%.

Females are more likely to be making zero wage.

For the trend associated with number of children, it seems that more than 80% of people with small number of children 1 to 3 make positive income, with a peak at having 1 children. As people have more children, the probability of making positive income decreases to below 70%, until rising again at 7 children to 75%. The number for having 8 children is not representative because of the small sample size.

For the trend associated with marital status. Those with NA are mostly likely to be making 0 income, at 58.85%, much higher than any other category. The married people are the most likely to be making non-zero wage, followed by divorced. The separated category are least likely to be making positive wage among those with Non-NA value. It is hard to observe any interpretable trend here.

For the joint trend on marital status and number of children. It appears that conditioning on having the same number of children, the married people are less likely to be making zero income.

Problem 2. Heckman Selection Model

I recode the marital status to a binary varibal with 1 indicating married and 0 otherwise.

1. Specify and estimate an OLS model to explain the income variable(where income is positive) The model I estimate is the following:

$$\begin{aligned} Income &= \alpha_0 + \alpha_1 \text{marital} + \alpha_2 \text{Age} \\ &\quad + \alpha_3 \text{gender} + \alpha_4 \text{numChild} + \alpha_5 \text{work-exp} + \alpha_6 \text{school-yr} \end{aligned}$$

   The regression result of a OLS on the portion of data with positive income yields the following result in table 1. The independent variable I used are marital status, age in years, gender, number of children, work experience and education in years. The dependent variable is Income. The data is truncated to have only the positive income subset.

   (a) Interpret the estimation results

      I only interpret the statistically significant coefficients here. All else equal, compared to married individual, not married individuals make \$5664.72 less. All else equal, men make \$16,669.88 more than women. All else equal, another week of work experience increases income by \$16.642. All else equal, another year of schooling increases income by \$2758.65.

   (b) Explain why there might be a selection problem when estimating an OLS this way.

      People naturally select whether or not they want to work and make positive wages. For those observations of independent variables associated with 0 income, had they chosen to work, they would make some non-zero income. It might be that their market wage are too low compared to their reservation wage, it might be that at minimum wage the market does not have demand for their skills. Therefore we are regressing on a selected sample, which might bias our result.

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

   One can imagine the Heckman selection process as assuming that conditioning on the independent variables, wages are normally distributed. So we can utilize the observed portion of the wages to fill in the unobserved portion of the wages had they participated in the labor market.

3. Estimate a Heckman selection model. Interpret the results from the Heckman selection model and compare the results to OLS results. Why does there exists a differences?

For the Heckman two step model, the selection step result is summarized in table 2. For the selection step, which is a probit regression, the independent variables I used are marital status, age in years, gender, education in years, education of residential parents.

Using the result from selection step, I generate inverse mill ratios and run OLS as the second step. The results are summarized in table 3. I used the same set of independent random variable as the OLS model on the truncated OLS.

Again, I only interpret the statistically significant coefficients. All else equal, married individuals make \$5235.6 more compared to unmarried ones. All else equal, men make \$16,674.34 more than women. All else equal, individuals with an extra child make \$607.192 more. All else equal, another year of schooling increases income by \$2457.80.

Compared to the original OLS on truncated data, we see that the estimates of the statistically significant coefficients do not change much. However, the coefficient on number of children becomes statistically significant in the heckman two step model. This is intuitive as people might opt to quit working because of having more children, which creates serious selection issues. The heckman selection model to provide better estimate by taking this selection issue into consideration.

Problem 3. Censoring

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

   I subset the data to only look at the those with positive income so that for the purpose of this exercise, I only look at the top censoring. Please refer to figure 4 for the histogram of income. Notice that the censoring threshold is $100,000$.

2. Propose a model to deal with the censoring problem.

Table 1: OLS on Truncated Data

	Dependent variable:
	Income
marital	5,664.717***
	(1,021.342)
ageY	324.656
	(319.063)
gender	16,669.880***
	(932.424)
NumChild	460.544
	(319.572)
wk_exp_wk	16.642***
	(1.602)
School_yr	2,758.649***
	(127.870)
Constant	$-$15,955.100
	(12,741.440)
Observations	3,039
R^2	0.242
Adjusted R^2	0.241
Residual Std. Error	25,147.310 (df = 3032)
F Statistic	161.592*** (df = 6; 3032)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Heckman Two Step: Selection

	Dependent variable:
	NonZeroIncome
marital	0.235***
	(0.049)
ageY	0.009
	(0.017)
gender	0.395***
	(0.049)
School_yr	0.066***
	(0.006)
ResDadSch	−0.004
	(0.010)
ResMomSch	0.004
	(0.011)
Constant	−0.652
	(0.671)
Observations	3,763
Log Likelihood	−1,734.908
Akaike Inf. Crit.	3,483.817
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: Heckman Two Step: Second Step

	Dependent variable:
	Income
marital	5,235.600***
	(1,012.244)
ageY	202.993
	(316.143)
gender	16,674.340***
	(922.827)
NumChild	607.192*
	(316.811)
wk_exp_wk	16.505***
	(1.586)
School_yr	2,457.801***
	(131.991)
InverseMill	353.125***
	(44.006)
Constant	−9,104.935
	(12,639.160)
Observations	3,039
R^2	0.258
Adjusted R^2	0.256
Residual Std. Error	24,888.470 (df = 3031)
F Statistic	150.602*** (df = 7; 3031)
Note:	*p<0.1; **p<0.05; ***p<0.01

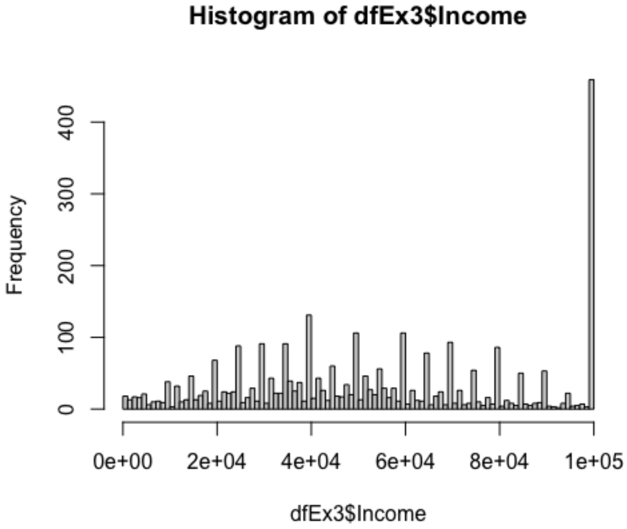

Figure 4: Histogram of Income

I propose to fit a tobit model with top-censoring.

$$\begin{aligned} y^* &= \beta X + \epsilon \\ y &= y^* \quad \text{if} y^* \leq 100,000 \\ y &= 100,000 \quad \text{otherwise} \end{aligned}$$

The x variable include marital status, age in years, gender, number of children, work experience and education in years. Please refer to the likelihood function in my r implementation.

3. Estimate the appropriate model with the censored data.
   The estimated coefficients are

   (a) Intercept: -18222.82

   (b) Coefficient on age 241.41

   (c) Coefficient on work experience in weeks 18.09

   (d) Coefficient on education by year 3132.60

   (e) Coefficient on marital status 7920.68

   (f) Coefficient on gender (1 if men and 0 if women) 19089.85

   (g) Coefficient on number of children 642.92

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

   The result suggests that all else equal, an individual that is one year older will make \$241.41 more. All else equal, an individual with one more week of work experience will make \$18.09 more. All else equal, an individual with one extra year of education would make \$3132.60 more. All else equal, an married individual would make \$7920.68 more. All else equal, individual with one extra child will make \$642.92 more. All else equal, men make \$19089.85 more.

   For comparison purposes, I also run OLS on the data set without consideration for censoring. The results are summarised in table 4. Notice that compared to the Tobit regression result, the OLS over estimates the constant, the effect of an extra child, an extra year in age and under estimates the effect of being married, being men, education and working experience.

Problem 4. Panel Data

1. Explain the potential ability bias when trying to understand the determinants of wages.

   Suppose we want to understand the effect of education on wages. Innate ability of a person is positively correlated with both educational attainment and income. Thus the estimated effect of education is likely to be over-estimated, picking up the effect of ability.

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

   (a) Between Estimator

   I estimate the following between estimator

$$\bar{Income}_i = \alpha_0 + \alpha_1 \bar{Education}_i + \alpha_2 \bar{Age}_i + \alpha_3 \bar{WorkExperience}_i + \epsilon_i$$

   Notice that it does not make sense to average over marital status, so I cannot include it in to this model. The result of this between estimator is summarzed in table 5. It appears that between estimator does a terrible job, yielding coefficients are are essentially 0 and statistically insignificant.

Table 4: Plain OLS for Exercise 3

	Dependent variable:
	Income
marital	5,664.717***
	(1,021.342)
ageY	324.656
	(319.063)
gender	16,669.880***
	(932.424)
NumChild	460.544
	(319.572)
wk_exp_wk	16.642***
	(1.602)
School_yr	2,758.649***
	(127.870)
Constant	−15,955.100
	(12,741.440)
Observations	3,039
R^2	0.242
Adjusted R^2	0.241
Residual Std. Error	25,147.310 (df = 3032)
F Statistic	161.592*** (df = 6; 3032)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5: Panel Data: Between Estimator

	Dependent variable:
	Avg_Income
avg_Edu	0.000
	(0.000)
avg_Age	0.000
	(0.000)
avg_wk_exp	0.000
	(0.000)
Constant	−16.118***
	(0.000)
Observations	8,984
R^2	0.500
Adjusted R^2	0.500
Residual Std. Error	0.000 (df = 8980)
F Statistic	2,993.320*** (df = 3; 8980)
Note:	*p<0.1; **p<0.05; ***p<0.01

(b) Within Estimator

I estimate the following within estimator.

$$(\text{Income}_i - \text{Avg-Income}_i) = \alpha_0 + \alpha_1(\text{Edu}_i - \text{Avg-Edu}_i) + \alpha_2(\text{Age}_i - \text{Avg-Age}_i) \\ + \alpha_3(\text{wk-exp}_i - \text{Avg-wk-exp}_i) + \epsilon_i$$

It does not make sense to take average of marital status, we cannot incorporate it into the model for estimation. The results of estimation are summarized in table 6. As an improvement over the between estimator, we now get statistically significant estimates. Since it does not make sense to average marital status, I can't include that in the model as well.

I drop the data in year 1997 for lack of education record. This is because it is unclear what is the value we should impute for education in 1997, a consequence of coding education in the original data set as highest degree attained instead of number of years of education.

Table 6: Panel Data: Within Estimator

	Dependent variable:
	(Income - Avg_Income)
Edu	817.512*** (17.066)
Age	1,973.246*** (17.683)
work_exp	24.967*** (0.514)
Constant	−38,657.100*** (381.572)
Observations	85,942
R^2	0.344
Adjusted R^2	0.344
Residual Std. Error	23,604.310 (df = 85938)
F Statistic	14,991.220*** (df = 3; 85938)
Note:	*p<0.1; **p<0.05; ***p<0.01

(c) Difference Estimator

I do the following difference estimator, subtracting the value at $t = 1098$. In particular, I estimate the following model.

$$\begin{aligned}\text{Income}_{i,t} - \text{Income}_{i,98} &= \alpha_0 + \alpha_1(\text{Edu}_{i,t} - \text{Edu}_{i,98}) \\ &+ \alpha_2(\text{wk-exp}_{i,t} - \text{wk-exp}_{i,98}) \\ &+ \alpha_3(\text{Marital}_{i,t} - \text{Marital}_{i,98}) \\ &+ \alpha_4(\text{Age}_{i,t} - \text{Age}_{i,98})\end{aligned}$$

This is the only model where I can introduce marital status. For simplicity, I code the marital status to be a binary variable with 1 meaning married and 0 otherwise. Notice that I cannot control for age in this difference regime as the age difference will always be the same for every unit.

There is no direct data on the education level in 1997, So I drop the observations in year 1997. This is because it is unclear what is the value we should impute for education in 1997, a consequence of coding education in the original data set as highest degree attained instead of number of years of education.

The result of the estimation can be found in table 7

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

The between estimator basically says that non of the variables we observe make any impact on income.

The Within estimator yields statistically significant estimates on all variables we regress on. Because the averages are taken for each specific unit, we should still interpret the coefficient linearly. The within estimate tells us that, all else equal, one year increase in education increases income by \$817.51. All else equal, a one year increase in age increases income by 1973.25. All else equal, a one week more work experience increases income by \$24.97. It is reasonable to doubt this estimate as it suggests that age is more important than work experience. As one year of work experience would increase income by $52 * 24.97 \approx 1300 < 1973$. I suspect that the age variable might be picking

Table 7: Panel Data: Difference Estimator

	Dependent variable:
	Income_diff
Wk_exp_diff	44.664^{***}
	(0.744)
MaritalChange No Change	$-17{,}143.560^{***}$
	(315.158)
MaritalChange Separation	$-12{,}949.650^{***}$
	(3,252.331)
Edu_diff	$1{,}074.135^{***}$
	(23.045)
Constant	$21{,}487.450^{***}$
	(381.307)
Observations	39,287
R^2	0.258
Adjusted R^2	0.258
Residual Std. Error	27,073.320 (df = 39282)
F Statistic	$3{,}419.798^{***}$ (df = 4; 39282)
Note:	*p<0.1; **p<0.05; ***p<0.01

up some other effects.

The difference estimator tells a very difference story. It suggests that on average, all else equal, one more week of work experience increases income by $44.66. All else equal, one extra year of education increases income by $1074.14. The estimates on changes in marital status should be interpreted with caution as they are merely capturing the change of marital status compared to based year 1998. I would lean toward only qualitatively interpret them as the following. It appears there is considerable premium associated with going from unmarried to married.

The three panel data estimators yield very different results because they each allow us to control for different things. The between estimator tries to average out individual fixed effect and only look at the mean. This is going to masks some non-linear effects of the variable. The within estimator still requires us to calculate the mean, but allows for a time dimension. Thus the within estimator allows for richer dynamics. The difference estimator is the only estimator that allow us to not to take the average of variables. Thus it allows us to look at dynamics of variables where it does not make any sense to take average, which, in our example, would be marital status. However, it prevents us from looking at the effect of variables that evolve exogeneously with time, like age.