

Data Science

Problem Set 12

salimeh Birang

April 2020

Question 6

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
logwage	1,545	1.652	0.688	-0.956	1.201	2.120	4.166
hgc	2,229	12.455	2.444	5	11	14	18
exper	2,229	6.435	4.867	0.000	2.452	9.778	25.000
kids	2,229	0.429	0.495	0	0	1	1

Looking at table one we can conclude: hgc variable which is how many years of schooling each women has completed, on average each women completed 12 years of education at minimum it is 5 years of schooling. exper which is how long each women has worked on average it is 6.5 years and maximum years of work experience is 25 years. kids is an indicator shows if a woman has at least one kid or not. logwage is missing for 684, at almost 44%, I think it might MNAR or MAR.

Question 7

Table 2:

	<i>Dependent variable:</i>		
	logwage <i>OLS</i>	logwage_mean_imp <i>OLS</i>	logwage <i>Heckman selection</i>
	(1)	(2)	(3)
hgc	0.059*** (0.009)	0.036*** (0.006)	0.091*** (0.010)
union1	0.222** (0.087)	0.068 (0.047)	0.186** (0.084)
college1	-0.065 (0.106)	-0.126*** (0.048)	0.092 (0.100)
exper	0.050*** (0.013)	0.021*** (0.007)	0.054*** (0.012)
I(exper^2)	-0.004*** (0.001)	-0.001*** (0.0004)	-0.002* (0.001)
Constant	0.834*** (0.113)	1.149*** (0.078)	0.446*** (0.122)
Observations	1,545	2,229	2,229
R ²	0.038	0.020	0.092
Adjusted R ²	0.035	0.018	0.088
ρ			-0.998
Inverse Mills Ratio			-0.695*** (0.060)
Residual Std. Error	0.676 (df = 1539)	0.568 (df = 2223)	
F Statistic	12.106*** (df = 5; 1539)	9.207*** (df = 5; 2223)	

Note:

*p<0.1; **p<0.05; ***p<0.01

The true value of β_1 is 0.091, the coefficient we get based on our first method is 0.059 which is lower than our true value and by estimating second model our coefficient of interest is even lower than our true value but estimating the third model gives us the true value. heckit model would be the best fit here.