

PS7

Alex Skipper

March 2018

1 Part 6

Table 1:

Statistic	N	Mean	St. Dev.	Min	Max
logwage	1,669	1.625	0.386	0.005	2.261
hgc	1,669	12.556	2.322	0	18
tenure	1,669	5.225	5.095	0.000	24.750
age	1,669	39.171	3.085	34	45

The rate at which logs are missing is .2512337. The log wages are MNAR due to certain personal reasons from the sample. Some people might of chosen to disclose their wages or skipped it. Due to wages being a personal matter, these missing observations are not due to any sampling errors or bias, but do to personal reasons.

2 Part 7

Table 2:

	<i>Dependent variable:</i>	
	logwage	
	(1)	(2)
hgc	0.062*** (0.005)	0.049*** (0.004)
collegenot college grad	0.146*** (0.035)	0.160*** (0.026)
tenure	0.023*** (0.002)	0.015*** (0.001)
age	-0.001 (0.003)	-0.001 (0.002)
marriedsingle	-0.024 (0.018)	-0.029** (0.014)
Constant	0.639*** (0.146)	0.833*** (0.115)
Observations	1,669	2,229
R ²	0.195	0.132
Adjusted R ²	0.192	0.130
Residual Std. Error	0.346 (df = 1663)	0.311 (df = 2223)
F Statistic	80.508*** (df = 5; 1663)	67.496*** (df = 5; 2223)

Note:

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

Figure 1: Fitted Regression

Coefficients:

(Intercept)	hgc	collegenot college grad
0.785609	0.059265	0.103506
tenure	age	marriedsingle
0.022591	-0.002495	-0.021843

The B1's across the table all differ due to the ways log wage as manipulated or imputed. B1 decreased with the first mean imputation, while increasing during multiple mean imputations. A pattern can be seen more more mean imputation can lead to a closer true value when dealing with missing data observations. The accuracy of the various imputation methods are not very accurate due to the missing values being filled in by the mean of logwages. The true values of the missing logwages could be much higher or lower than the mean, resulting the data results to be skewed. However, it is as close as we can get to the true value.

The estimates for B1 in the last two methods differ. The first mean imputation finds the mean of the available logwages and replaces the missing values with the mean. Which is easy to use but consistency of the data is reduced. So, this impacts other estimates, such as variance. The last imputation method is more accurate than the first one due to the fact that it performs multiple imputations to fill in the missing values. From the multiple imputations, the mean from the many models is taken and filled into the missing values, which gives more accurate results. Leaving you close to the true value.

3 Part 8

My project will be over a the ranking of education systems based on economics factors. It will be a cross section analysis. I will be using the average of the reading, math and science scores on the PISA test. While looking at many different factors that might affect the rankings from two different years, 2000 and 2010. The data will examine those countries that typically do well. All of the data can be found in the World Data Bank, United Nations Development Programme: Human Development Reports, World Prison Brief, United Nations Educational, Scientific and Cultural Organization, PISA 2000:Overview of the Study Design, Method and Results written by Stanat · Artelt · Baumert · Klieme · Neubrand · Prenzel · Schiefele · Schneider · Schümer · Tillmann · Weiß, and Organisation for Economic Co-operation and Development (OECD). I will look at enrollment rate of both sexes as a percentage from the Data World Bank, the Education Index which is a number between 0 and 1 which accounts for the mean

years of schooling in a country as percentage of the expected years of schooling for an individual from Human Development Report, School Expenditure as percentage of GDP from the Data World Bank, and the unemployment Rate as a percentage from the Data World Bank. I will also include health expenditure per capita in USD from the Data World Bank, the ratio of pupils to teacher from United Nations Educational, Scientific and Cultural Organization, the prison population rate in per hundred thousand of the national population from World Prison Brief, and gross domestic saving as a percentage of GDP from the Data World Bank. I will also include other variables as I progress with the project. So, far ill be looking at doing OLS, but I will look into different models as well.