PS7

Amir Tayebi

March 12, 2019

1 Summary Table

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
logwage	1,669	1.625	0.386	0.005	1.362	1.936	2.261
hgc	2,229	13.101	2.524	0	12	15	18
tenure	2,229	5.971	5.507	0.000	1.583	9.333	25.917
age	2,229	39.152	3.062	34	36	42	46

2 Answer to Question 6

The rate at which logwages are missing is 0.2512337.

Tables 2 represents the summary statistics on individuals for whom we have the data on wage. On the other hand, table 3 shows the summary statistics on the missing observations. It would be ideal to perform the t-test to see if there is any significant differences between variables of tables two and three, but we can make some conclusions without doing so. Other than age, the differences between the variables of the two tables look to be significant. The data on individuals who are either married or uneducated tend to be not reported. In addition, the difference between hgc in table two and three seems to be significant. This also the case for tenure, but I don't know what they really are. I think missing values are kind of MAR since he values which are missing can be completely explained by the data we already have.

Table 2: Summary Statistics without Missing Observations

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
logwage	1,669	1.625	0.386	0.005	1.362	1.936	2.261
hgc	1,669	12.556	2.322	0	12	14	18
college	1,669	0.846	0.361	0	1	1	1
tenure	1,669	5.225	5.095	0.000	1.417	7.917	24.750
age	1,669	39.171	3.085	34	36	42	45
married	1,669	0.346	0.476	0	0	1	1

Table 3: Summary Statistics on Missing Observations

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
hgc	560	14.727	2.403	8	12	17	18
college	560	0.512	0.500	0	0	1	1
tenure	560	8.192	6.069	0.000	2.917	12.833	25.917
age	560	39.093	2.995	34	37	41	46
married	560	0.395	0.489	0	0	1	1

3 Comment on Beta

All the models underestimate the coefficient of interest, but the estimated coefficient by the linear regression is a little bit closer to the real value in comparison to others. The fact that the coefficient of interest in all the models is much lower than the real value is another sign that data are missing completely at random. As for the interpretation of the coefficient of interest in the mice package, mice assumes that the missing data are Missing at Random (MAR), which means that the probability that a value is missing depends only on observed value and can be predicted using them. It imputes data on a variable by variable basis by specifying an imputation model per variable.

4 Information on My Research

I am currently working on the data gathered from VoteSmart project and some other data sets to clean them. In this project, I'm investigation the effects of limitations on politicians on various outcomes such as their quality, their voting behavior, etc. I am also working on another paper studying the degree to which democratic governments distort media. However, at this point I am not sure which one I will be presenting for this class.

Table 4: Regression Results

	Dependent variable:		
		logwage	
	Complete	Mean	Prediction
	(1)	(2)	(3)
hgc	0.062***	0.050***	0.062***
	(0.005)	(0.004)	(0.004)
collegenot college grad	0.145***	0.168***	0.145***
	(0.034)	(0.026)	(0.025)
tenure	0.050***	0.038***	0.050***
	(0.005)	(0.004)	(0.004)
tenure2	−0.002***	−0.001***	−0.002***
	(0.0003)	(0.0002)	(0.0002)
age	0.0004	0.0002	0.0004
	(0.003)	(0.002)	(0.002)
marriedsingle	−0.022	−0.027**	−0.022*
	(0.018)	(0.014)	(0.013)
Constant	0.534***	0.708***	0.534***
	(0.146)	(0.116)	(0.112)
Observations	1,669	2,229	2,229
R^2	0.208	0.147	0.277
Adjusted R^2	0.206	0.145	0.275
Residual Std. Error	0.344 (df = 1662)	0.308 (df = 2222)	0.297 (df = 2222)
F Statistic	72.917*** (df = 6; 1662)	63.973*** (df = 6; 2222)	141.686*** (df = 6; 2222)
Note:			*p<0.1; **p<0.05; ***p<0.01