ProblemSet 7

Yue Wu

March 2018

1 Type of missing values

The missing rate is 25.13

The missing value in logwage is more like MNAR. Because there is reasons that people keep their income secretly, and we can't use other variable to account the missing values.

2 Table

Table 1:

Statistic	N	Mean	St. Dev.	Min	Max
logwage	1,669	1.625	0.386	0.005	2.261
hgc	2,229	13.101	2.524	0	18
tenure	2,229	5.971	5.507	0.000	25.917
age	2,229	39.152	3.062	34	46

Table 2:

	logwage	hgc	college	tenure	age	married
1		12	not college grad	5.333	37	single
2	1.856	12	not college grad	5.250	37	single
3	1.613	12	not college grad	1.250	42	single
4	2.201	17	college grad	1.750	43	married
5	2.090	12	not college grad	17.750	42	married

Table 3:

	Dependent variable:		
	logwage		
	(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)
poly(tenure, 2)1	4.855***	3.799***	5.694***
	(0.346)	(0.312)	(0.301)
poly(tenure, 2)2	−1.836***	−1.977***	−2.318***
	(0.345)	(0.311)	(0.300)
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	0.709***	0.848***	0.726***
	(0.145)	(0.115)	(0.111)
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

3 Conclusion

According the table above, the b1 in regression (2) is 0.818. Compare to other models, b1 in model (2) is closer to the true value. Model (2) was estimated after missing values imputed by mean of logwages in complete cases. Due to the b1 in model (2) is closer to true, i believe that model (2) is better model than the other two models. Then, the missing values are more like MNAC.

The b1 in the model after multiple imputation is 0.79. If we increases "m" in the mice commend, the new b1 might be closer to true value. But imputation of mean is still better when "m" equal to 5.

4 Project Data

I will use 2 data sets in my project. One is about the information of previous games in past 20 years. The other data set is about the game players, what games they have spend money on. I have already clean the first data set, but I am still working on the finding of second data set.

The first model will be looks like

sales = b0+b1*price +b2*type +b3*online +b4*producer + e

I have to make sure there is no relationships between independent variables.

Also, if there is, I will find nonlinear relationship between dependent and independent variables.

For missing values, I will use multiple imputation. Because this method will give me most accurate result if "m" is large enough.