Data Science for Economists - Spring 21

Problem Set 7

Syed Waleed Mehmood Wasti

March 27, 2021

1. Done
2. Done
3. N/A
4. Done
5. Done
6. The share of missing values for logwage is about 0.249. I personally believe the logwage variable is Missing Not At Random (MNAR) since some individuals would deliberately choose not to report their income.

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
logwage	670	25	1.6	0.4	0.0	1.7	2.3
hgc	16	0	13.1	2.5	0	12.0	18
tenure	259	0	6.0	5.5	0.0	3.8	25.9
age	13	0	39.2	3.1	34	39.0	46

Table 1: Summary Table (after dropping missing values for hgc and tenure)

7. It can be seen from the table below that $\hat{\beta}_1$ varies significantly for the second regression while the values are relatively similar for the other 3 regressions. The estimates of the first and the last regressions are closest to the estimate therefore it seems like listwise deletion suits the data for this model.

   Seems like mean imputation is not such a good method. I think this should particularly be true when there is some variation in the data (as is the case for income).

   $\hat{\beta}_1$ for the last 2 methods are 0.059 and 0.062 respectively (although I believe the mice imputed one is not correct for some reason).

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.534	0.708	0.563	0.534
	(0.146)	(0.116)	(0.112)	(0.146)
hgc	0.062	0.050	0.059	0.062
	(0.005)	(0.004)	(0.004)	(0.005)
as.factor(college)not college grad	0.145	0.168	0.177	0.145
	(0.034)	(0.026)	(0.025)	(0.034)
poly(tenure, 2, raw = T)1	0.050	0.038	0.047	0.050
	(0.005)	(0.004)	(0.004)	(0.005)
poly(tenure, 2, raw = T)2	-0.002	-0.001	-0.002	-0.002
	(0.000)	(0.000)	(0.000)	(0.000)
age	0.000	0.000	0.000	0.000
	(0.003)	(0.002)	(0.002)	(0.003)
as.factor(married)single	-0.022	-0.027	-0.028	-0.022
	(0.018)	(0.014)	(0.013)	(0.018)
Num.Obs.	1669	2229	2229	1669
Num.Imp.				10
R2	0.208	0.147	0.223	0.208
R2 Adj.	0.206	0.145	0.221	0.206
AIC	1179.9	1091.2	956.8	
BIC	1223.2	1136.8	1002.4	
Log.Lik.	-581.936	-537.580	-470.382	
F	72.917	63.973	106.573	

Table 2: Estimates of the 4 regression models

8. I plan to work on estimating how entrepreneurship rate changes for people who opt for food stamps. For this I will be using the IPUMS CPS data for the years 1990-2015. I plan to conduct panel regression as my estimation model.