PS7 Li

Donald Li

March 2018

1 Part 6

After looking through the data set it seems to be MNAR. Some people may not feel comfortable in revealing their incomes. Between characteristics and missingness there are 560 missing obs out of 2246.

2 Part 7

The imputation methods will never be the exact beta value of .093. However, with each level of imputation we get a little bit closer. Casewise deletion gives the worst, followed by Mean then finally Predictive mean matching gives us the best.

3 Part 8

For my final project we are using twitter data for sentiment analysis! I have already worked on a practice data set of fake/real text messages to cut my teeth at natural language processing! I am very excited to see the results. Right now we are in the gather data phase.

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

Table 2:

	Dependent variable:
	logwage
hgc	0.062***
	(0.005)
collegenot college grad	0.146***
	(0.035)
tenure	0.023***
	(0.002)
age	−0.001
	(0.003)
marriedsingle	−0.024
	(0.018)
Constant	0.639***
	(0.146)
Observations	1,669
R^2	0.195
Adjusted R^2	0.192
Residual Std. Error	0.346 (df = 1663)
F Statistic	80.508*** (df = 5; 1663)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3:

	Dependent variable:
	logwage
hgc	0.062***
	(0.005)
collegenot college grad	0.146***
	(0.035)
tenure	0.023***
	(0.002)
age	−0.001
	(0.003)
marriedsingle	−0.024
	(0.018)
Constant	0.639***
	(0.146)
Observations	1,669
R^2	0.195
Adjusted R^2	0.192
Residual Std. Error	0.346 (df = 1663)
F Statistic	80.508*** (df = 5; 1663)
Note:	*p<0.1; **p<0.05; ***p<0.01

	est	se	t	df	Pr(>|t|)	lo 95	hi 95	nmis	fmi	lambda
(Intercept)	1	0	5	94	0	0	1	NA	0	0
hgc	0	0	12	85	0	0	0	2	0	0
college2	0	0	5	360	0	0	0	NA	0	0
tenure	0	0	17	1232	0	0	0	15	0	0
age	0	0	-1	60	0	0	0	0	0	0
married2	0	0	-2	344	0	0	0	NA	0	0