Problem Set 7

Karley Nadolski

March 2021

1 Summary Table of Wages Data

	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

The log wages are missing at a rate of 25 percent. Without further examining the data, I am unsure whether the missing information is MCAR, MAR, or MNAR. Without being able to rule it out completely, I would guess that the missing data is MNAR. It seems difficult to prove that data could ever be missing completely at random, but I also didn't see any obvious relationships between the missing data spots and the data that was available for a given observation.

2 Comparing Imputation Techniques

(Table included on the second page of this document.)

Knowing that the true value of the β_1 is 0.093, it's easier to compare the four models for accuracy. At first glance, it's easy to see that the estimates for Models 1, 3, and 4 are all the same across the board to the hundredths place. They are also the three models that are closest to the true β_1 value of 0.093.

It is telling though that these three methods, though they provided consistent estimates across methodology, are not that close to the true β value. This may be related to the nature of the missing data as MNAR (which is the hardest context to draw plausible imputed values from). Model 2, where I substituted the mean in for the missing values, was the least accurate at least as it pertained to estimating β_1.

3 Project Update

To be honest, I haven't made a ton of progress on my project for the end of this semester. But I have given it some thought and I've already downloaded and

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.534	0.708	0.534	0.534
	(0.146)	(0.116)	(0.112)	(0.146)
hgc	0.062	0.050	0.062	0.062
	(0.005)	(0.004)	(0.004)	(0.005)
as.factor(college)not college grad	0.145	0.168	0.145	0.145
	(0.034)	(0.026)	(0.025)	(0.034)
poly(tenure, 2, raw = T)1	0.050	0.038	0.050	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	-0.022	-0.022
	(0.018)	(0.014)	(0.013)	(0.018)
Num.Obs.	1669	2229	2229	1669
Num.Imp.				20
R2	0.208	0.147	0.277	0.208
R2 Adj.	0.206	0.145	0.275	0.206
AIC	1179.9	1091.2	925.5	
BIC	1223.2	1136.8	971.1	
Log.Lik.	-581.936	-537.580	-454.737	
F	72.917	63.973	141.686	

cleaned the main data that I'll be using. I am interested in doing something with the Australian Rain data that I used for visualization on the last problem set, but I haven't quite decided what approach I want to take. Last semester, I was in a spatial statistics class that taught me about spatial analysis that would be interesting if integrated with some of the skills we've been working on in this class.

If I'm able to find the data, I would be interesting in looking at the relationships between rain totals and agricultural/economic indices throughout Australia. At this point, I'm looking for additional economics data and other applications that would be interesting for the original data set.