Problem Set 7

Steven Plaisance

March 2023

1. Here we see the data summary output describing each numeric variable within the wages dataset. 25% of logwage observations are missing. I think the missing observations are likely to be missing at random.

	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	0.0	12.0	18.0	
tenure	259	0	6.0	5.5	0.0	3.8	25.9	
age	13	0	39.2	3.1	34.0	39.0	46.0	

2. Here we see the model summary output of each of the four models. Beta 1 seems to decrease as we account for missing logwage observations, indicating that the data is likely not missing completely at random. It seems that the list-wise deletion model and the predicted value model performed most closely to the true value of Beta 1. The last two methods use a modeling approach to generate predictions for the missing observations based on the other variables.

	(1)	(2)	(3)	(4)
(Intercept)	0.534	0.708	0.534	0.670
	(0.146)	(0.116)	(0.112)	(0.130)
hgc	0.062	0.050	0.062	0.058
	(0.005)	(0.004)	(0.004)	(0.005)
collegenot college grad	0.145	0.168	0.145	0.096
	(0.034)	(0.026)	(0.025)	(0.033)
tenure	0.050	0.038	0.050	0.050
	(0.005)	(0.004)	(0.004)	(0.005)
tenure2	−0.002	−0.001	−0.002	−0.002
	(0.000)	(0.000)	(0.000)	(0.000)
age	0.000	0.000	0.000	−0.001
	(0.003)	(0.002)	(0.002)	(0.003)
marriedsingle	−0.022	−0.027	−0.022	−0.014
	(0.018)	(0.014)	(0.013)	(0.016)
Num.Obs.	1669	2229	2229	
R2	0.208	0.147	0.277	
R2 Adj.	0.206	0.145	0.275	
AIC	1179.9	1091.2	925.5	
BIC	1223.2	1136.8	971.1	
Log.Lik.	−581.936	−537.580	−454.737	

3. My project is still in the early stages, but I am happy with my premise and excited to continue working. I am using GPS data that we collect at my job with the football team. Specifically, my project is focused on calculating longitudinal and lateral acceleration values using athlete orientation and angle of motion. In terms of modeling approaches, machine learning is a likely choice due to the vast amount of data available to me.