Problem Set 7

savannahjsimpson

March 26 2024

1 Wage Data

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
logwage	1,675	1.670	0.507	0.005	1.401	2.012	2.303
hgc	1,878	13.519	2.440	6	12	16	18
tenure	1,878	5.782	5.590	0.083	1.667	8.583	25.917
age	1,878	38.834	2.955	34	37	41	46

- The log wages are missing at a rate of 16.7 percent. This is most likely a Missing at Random (MAR) scenario. The missing values are likely related to other variables - for example, tenure may not apply to every entry.
- The computed hgc coefficient was slightly different across the different models. The results are listed here: Listwise deletion: 0.089, Mean imputation: 0.095, Predicted imputation: 0.090, and Multiple imputation: 0.094. Considering the true value of 0.093, the models did provide close predictions to the actual number. The least accurate, LD, is likely biased due to the randomness of the variables. The Mean Imputation did provide a close estimate, but ultimately overestimates the precision of this coefficient. Predicted imputation is close to the true value, but could be more accurate. Multiple imputation is the best measure as it accounts for the uncertain nature of these variables.

2 Project Progress

- My project focuses on the real estate market in my hometown, which I was luckily able to travel to this spring break. I was pleased to see that it is an increasingly discussed issue as there is more information available for me to work with. In the previous problem set I created 3 visuals with Texas housing data included in R. This allowed me to practice modeling

with a simpler data set, as I plan to create similar structures with my Palm Beach County data. I have been working with general housing data from the FED to use as a benchmark, as well as Zillow data specific to my research region.