PS7 Felkner

Beth Felkner

March 2024

1 Data Summary of wages.csv

	Unique	Missing Pct.	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

25 percent of logwage observations are missing. I think the logwage variable is most likely to be MNAR. We know it can't be MCAR because that doesn't really exist in organically observed data (it could only really exist in a lab experiment type setting.) For it to be MAR, the missing logwages would need to be completely explained by the other variables we have such as education. This is not likely to be the case. Much more likely is that wages are missing because some people did not want to report their wages (likely low wages) which would be a case of MNAR.

2 Modelsummary with all four regressions

	Complete Cases	Mean Imp	Complete Imp	Mice
(Intercept)	0.639***	0.833***	0.639***	0.725***
	(0.146)	(0.115)	(0.111)	(0.129)
hgc	0.062***	0.049***	0.062***	0.059***
	(0.005)	(0.004)	(0.004)	(0.005)
collegenot college grad	0.146***	0.160***	0.146***	0.101**
	(0.035)	(0.026)	(0.025)	(0.031)
tenure	0.023***	0.015***	0.023***	0.023***
	(0.002)	(0.001)	(0.001)	(0.002)
age	−0.001	−0.001	−0.001	−0.001
	(0.003)	(0.002)	(0.002)	(0.002)
marriedsingle	−0.024	−0.029*	−0.024+	−0.018
	(0.018)	(0.014)	(0.013)	(0.016)
Num.Obs.	1669	2229	2229	
R2	0.195	0.132	0.268	
R2 Adj.	0.192	0.130	0.266	
AIC	1206.1	1129.3	961.2	
BIC	1244.0	1169.3	1001.1	
Log.Lik.	−596.049	−557.651	−473.584	
F	80.508	67.496	162.884	
RMSE	0.35	0.31	0.30	

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

The Beta1 estimation for complete cases only is 0.062, for mean imputation is 0.049, for imputation with predicted values from the complete cased model is also 0.062, and for multiple imputation regression is 0.059. Interesting, all of the values are significantly lower than the true value of 0.092. The clear pattern I see is that the complete case model and complete case imputation model yield exact the same Beta estimation values (for all Betas not just Beta1). This suggests to me that complete cases and complete case imputation methods have the most veracity, because their Beta estimations are close to the true Beta value. The multiple imputation Beta1 estimation is also close to the true Beta1, but I am suspicious of that method because the Beta2 (college grad) is so far off from the other Beta2 estimations (although I do not know the true Beta2 value so it could actually be the most accurate).

3 Progress on my project

In full transparency, I have not made very much progress on my project, but this is intentional. My mind is most productive through compartmentalization of tasks which I am aware of, and so my plan is to completely or mostly finish my

Masters Research project before diving into the data science project, because they are too similar for me to want to be doing both simultaneously. My Masters Research is coming along well with my data set compiled and clean and the bulk of my models ran, although I need to make some slight changes. My plan/aim is to finish adjusting the models and typing the paper part over sprig break.

Then when I return from spring break I will begin the data science project in earnest. I have made some incidental progress however, mostly through the problem sets that deal with it. I want to use Baseball Hall of Fame voting, which baseball-reference.com has available for all years (1936-present) in easily scrapable tables. My plan is to scrape each table and then compile them vertically into a train dataset(approximately 75 percent of total observations) and a test dataset(approximately 25 of total observations). I will then use machine learning techniques (not exactly sure what models I want to use but hoping the ML section of our class will help me decide) to predict HoF status in my test dataset. I will adjust it as necessary then once it does well on predicting HoF status in my test dataset, I will apply it to players who will be eligible for the upcoming classes and try to predict who will be inducted in 2025 and perhaps farther forward. Of course I will not know if my predictions are right for at least another year but I think it will be intersting still.