PS7

Colton Steele

March 28, 2023

1 Answers

Number 6: 1/4 of the observations do not have a value for logwage. In looking through the data, it does not seem like there is any relationship between the missing values and any of the other variables in our dataset. Given that in the question, you listed that the data is only for women that were working, the absence of zeroes in our data makes sense so I would say that this data is missing at random as I can't thing of an endogenous choice that would affect the missingness of our 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	0.0	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
tenure_sq	259	0	66.0	102.5	0.0	14.1	671.7

Number 7: It appears that my estimated beta coefficeint for hgc for all of my models is slightly less than the true value you stated the true value was. As we'd expect, the coefficient for the complete cases and predicted models are the same just with smaller standard errors as we have a larger sample size. The mean has the smallest coefficient which indicates that our predicted values are, on average, larger than the average. The Mice method gave us a value closer to the complete cases and predicted model, but slightly smaller in magnitude. Overall, this exercise shows that even though we can do our best to impute data, missing data is a problem and will oftentimes prevent us from getting the true values.

	Complete	Mean	Predicted	Mice
(Intercept)	0.534***	0.708***	0.534***	0.598***
	(0.146)	(0.116)	(0.112)	(0.156)
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***	0.108**
	(0.034)	(0.026)	(0.025)	(0.034)
tenure	0.050***	0.038***	0.050***	0.049***
	(0.005)	(0.004)	(0.004)	(0.005)
tenure_sq	−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+	−0.018
	(0.018)	(0.014)	(0.013)	(0.017)
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	
F	72.917	63.973	141.686	
RMSE	0.34	0.31	0.30	

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

Number 8: As of now, I have not made significant progress on my project. As I mentioned in the last problem set, for my project, I would like to create an elo system for NCAA softball using the softballR package. To do this, I plan to initialize my system at the start of the 2021-2022 season, then run the entire season's data and have the endpoints from last season be my starting point for this year and update each team's elo rating through the most recent time this year. Then, I want to use that rating system to predict which teams have the highest probabilities of making the Women's College World Series. I have searched and have not found anyone that has created a softball elo rating system so I think it would be a worthwhile and interesting endeavor.