PS7-Onishi

Saryu Onishi

March 2023

1 wages.csv Summary

25% of 'logwage' data are missing (Table 1). Although it is difficult to say what type of missingness we see in the logwage variable, since we cannot account for the missingness using other variables, we can rule out MAR.
I would be inclined to suggest MCAR because the model using the complete cases-only dataset produced a beta estimate closest to the true value However, given the fact that MCAR almost never occurs in real life, I would guess that this missingness could be attributed to variables that are not measured, and hence is categorised at MNAR.

	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

Table 1: Data summary

2 Model Summary

$\hat{\beta_1}$ varies across the four models (Table 2). The model using likewise deletion to remove the missing data generated a $\hat{\beta_1}$ of 0.062. The model that predicted the missing values based on the complete cases model, has a $\hat{\beta_1}$ of 0.062. Both of these models had very close $\hat{\beta_1}$, and were the closest to the true value of 0.093. The next closest $\hat{\beta_1}$ was achieved through the multiple imputation regression model using the mice package. Here, the $\hat{\beta_1} = 0.058$.
The model with the $\hat{\beta_1}$ furthest from the true $\hat{\beta_1}$ was the model that used mean imputation, with a $\hat{\beta_1}$ value of 0.049. This does not surprise me as it is not generally recommended.
I think when we look at all the methods used here, it can be said that the simple list-wise deletion can work as well as multiple imputation regression models, in some circumstances.

3 Project

For the project, I was initially going to use the dataset I had been working with through the semester. However, I do not think the dataset will allow me to create a model, to produce a project that is up to the standard. I am currently looking at other data I may be able to use to generate a useful model.

	Complete Cases Only	Mean Imputation	Predicted from Compete Cases Model	Mice
(Intercept)	0.639	0.833	0.639	0.727
	(0.146)	(0.115)	(0.111)	(0.139)
hgc	0.062	0.049	0.062	0.058
	(0.005)	(0.004)	(0.004)	(0.005)
collegenot college grad	0.146	0.160	0.146	0.107
	(0.035)	(0.026)	(0.025)	(0.029)
tenure	0.023	0.015	0.023	0.022
	(0.002)	(0.001)	(0.001)	(0.002)
age	−0.001	−0.001	−0.001	−0.001
	(0.003)	(0.002)	(0.002)	(0.003)
marriedsingle	−0.024	−0.029	−0.024	−0.025
	(0.018)	(0.014)	(0.013)	(0.017)
Num.Obs.	1669	2229	2229	2229
Num.Imp.				5
R2	0.195	0.132	0.268	0.208
R2 Adj.	0.192	0.130	0.266	0.206
AIC	1206.1	1129.3	961.2	
BIC	1244.0	1169.3	1001.1	
Log.Lik.	−596.049	−557.651	−473.584	
RMSE	0.35	0.31	0.30	

Table 2: Model summary