Econ-5253

Problem Set 7

Mieon Seong

Due by Mar 26th, 2024

Question 6-1 : Use modelsummary to produce a summary table of this data frame

	Mean	SD	Min	Median	Max
logwage	1.63	0.39	0.00	1.66	2.26
hgc	13.10	2.52	0.00	12.00	18.00
tenure	5.97	5.51	0.00	3.75	25.92
age	39.15	3.06	34.00	39.00	46.00

Question 6-2 : At what rate are log wages missing?

The number of variable is 2246. Also, the number of NA of log wages is 560. So, the rate is 24.9%.

Question 6-3 : Do you think the logwage variable is most likely to be MCAR, MAR, or MNAR?

If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR). In this data set, the probability of missing data about logwage variable is same (25%).

Question 7-1 : Perform the following imputation methods for missing logwage observations.

	Mean	SD	Min	Median	Max
logwage	1.63	0.39	0.00	1.66	2.26
hgc	13.10	2.52	0.00	12.00	18.00
tenure	5.97	5.51	0.00	3.75	25.92
age	39.15	3.06	34.00	39.00	46.00

Question 7-2 : The true value of $\hat{\beta}_1 = 0.093$. Comment on the differences of $\hat{\beta}_1$ across the models. What patterns do you see? What can you conclude about the veracity of the various imputation methods? Also discuss what the estimates of $\hat{\beta}_1$ are for the last two methods.

The first method is that we deleted all NA variables in the logwage and then calculated. However, in the second method, we performed mean imputation to fill in missing log wages, imputed missing log wages as the predicted values from the complete cases, and performed a multiple imputation.

Question 8 : Tell me about the progress you've made on your project. What data are you using? What kinds of modeling approaches do you think you're going to take?

In my project, I used movie rating data which includes some information such as name, rating, genre, year, score, year, budget, gross and so on. By using the data, I would like to correlation between budget and gross and so on.