PS7_Wardwell

lwardwell

March 24th, 2025

Creating Tables and Imputing Missing Data Using R

Table 1: Summary Statistics - wages.csv

	Mean	Median	SD	Min	Max	N
logwage	NA	NA	NA	NA	NA	2229.0
hgc	13.1	12.0	2.5	0.0	18.0	2229.0
tenure	6.0	3.8	5.5	0.0	25.9	2229.0
age	39.2	39.0	3.1	34.0	46.0	2229.0
college_ind	0.2	0.0	0.4	0.0	1.0	2229.0
married_ind	0.6	1.0	0.5	0.0	1.0	2229.0
		N	%			
college	college grad	530	23.8			
	not college grad	1699	76.2			
married	married	1431	64.2			
	single	798	35.8			

1. At what rate are the log wages missing?
    - 25%
2. Do you think the logwage variable is most likely to be MCAR, MAR, or MNAR?
    - logwage is most likely missing at random, because the missingness of logwage is likely not related to the value of logwage or any other variables in the dataset.

Question 7

1. Comment on the differences of $\hat{\beta}_1$ between the different models.
    - $\hat{\beta}_1$ varies slightly across the models. In the mean imputation scenario, it decreases from .062 in the complete cases model to .050. This makes logical sense, as we have increased the total number of observations and replaced the missing values of the mean, which may produce less variation of logwage.
2. What patterns do you see?

- The Mean Imputation method has the highest deviation in coefficient magnitude from the Complete Cases scenario. It also seems to generate significance for the "married_ind" variable that is not present in any other scenario. The tenure_sq variable is the least affected by the choice of methodology. For all variables except the married_ind variable, significance on the coefficient does not change, while there are some slight deviations in magnitude. The adjusted R squared value is highest for the Single Imputation model at .229.

3. 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.

   - When concluding it is necessary and appropriate to impute missing values, it is important to consider how the imputation method may affect the model results. The Mean Imputation method has a notable effect on the coefficient of hgc, which may affect the estimates. The Mean Imputation method also has the lowest adjusted R squared value, .145, which generally indicates the model is not a good fit for the data. The coefficients on the last two models (Single and Multiple Imputation) are consistent with one another, with signage and magnitude on each coefficient being similar to the other model. The coefficients are similar, but the Multiple Imputation method generates 5 plausible datasets and incorporates this uncertainty into its standard errors, which may produce more reliable inferences than a single imputation.

Table 2: Regression Results

	Complete Cases	**Mean Imputation**	**Single Imputation**	**Multiple Imputation**
(Intercept)	0.657***	0.849***	0.705***	0.655***
	(0.130)	(0.103)	(0.112)	(0.120)
hgc	0.062***	0.050***	0.063***	0.064***
	(0.005)	(0.004)	(0.005)	(0.005)
college_ind	-0.145***	-0.168***	-0.163***	-0.156***
	(0.034)	(0.026)	(0.028)	(0.032)
tenure	0.050***	0.038***	0.045***	0.044***
	(0.005)	(0.004)	(0.004)	(0.005)
I(tenure^2)	-0.002***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
age	0.000	0.000	-0.001	0.000
	(0.003)	(0.002)	(0.002)	(0.003)
married_ind	0.022	0.027*	0.011	0.019
	(0.018)	(0.014)	(0.015)	(0.019)
Num.Obs.	1669	2229	2229	2229
Num.Imp.				5
R2	0.208	0.147	0.231	0.229
R2 Adj.	0.206	0.145	0.229	0.227
AIC	1179.9	1091.2	1463.9	
BIC	1223.2	1136.8	1509.6	
Log.Lik.	-581.936	-537.580	-723.973	
F	72.917	63.973	111.008	
RMSE	0.34	0.31	0.33	

Note: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

Question 8

1. What progress have you made on your project?

- I plan to take a look at whether having a PCAOB inspected auditor has an affect on a Bank's likelihood of receiving an FDIC enforcement action. This will require obtaining Bank auditor names from the FRY9 reports from the Federal Reserve, and combining that data with data on PCAOB inspections from the PCAOB website. I will also have to download enforcement action data and additional financial data for control variables from the FDIC.

2. What kinds of modeling approach do you think you're going to take?

    - Since we will be looking at the likelihood of an FDIC enforcement action, logistic regression seems like the best modeling tool to use, making sure to control for additional financial variables that may indicate a higher risk profile that would also increase the likelihood of regulatory enforcement.