PS7

Becky Crouse

March 28, 2023

Imputation of Missing Values

For Problem Set 7, I ran regression models using the wages data set. The model used for each regression is:

$$logwage_i = \beta_0 + \beta_1 hgc_i + \beta_2 college_i + \beta_3 tenure_i + \beta_4 tenure_i^2 + \beta_5 age_i + \beta_6 married_i + \epsilon_i$$

Each model differs in the method used for imputing missing observation data for the dependent variable $logwage$. The following sections describe the data set and the regression models. Tables are generated using the modelsummary package in R.

Summary Table

Table 1: Summary 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

- Values of logwage are missing for 25 percent of the sample.
- Given the observations in the data set all relate to women in the late 1980s, I would expect any missing wages are missing not at random (MNAR).

Regression Summaries

The table below shows the coefficients obtained from running the above regression function for each method of imputation of the missing wage data.

Given the true value of $\hat{\beta_1} = .093$, the models above all underestimate this coefficient. The imputation techniques resulting in the closes estimates are the Complete Cases and Fitted Value Imputation, which both show a coefficient of

Table 2: Regression results

	Complete Cases	Mean Imputation	Fitted Value Imputation	Multiple Imputation
(Intercept)	0.534***	0.708***	0.534***	0.560***
	(0.146)	(0.116)	(0.112)	(0.143)
hgc	0.062***	0.050***	0.062***	0.061***
	(0.005)	(0.004)	(0.004)	(0.006)
college	0.145***	0.168***	0.145***	0.126**
	(0.034)	(0.026)	(0.025)	(0.035)
tenure	0.050***	0.038***	0.050***	0.042***
	(0.005)	(0.004)	(0.004)	(0.006)
$tenure^2$	−0.002***	−0.001***	−0.002***	−0.001**
	(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)
married	−0.022	−0.027*	−0.022+	−0.019
	(0.018)	(0.014)	(0.013)	(0.019)
Num.Obs.	1669	2229	2229	2229
Num.Imp.				5
R2	0.208	0.147	0.277	0.224
R2 Adj.	0.206	0.145	0.275	0.222
AIC	1179.9	1091.2	925.5	
BIC	1223.2	1136.8	971.1	
Log.Lik.	−581.936	−537.580	−454.737	
RMSE	0.34	0.31	0.30	

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

.062. The Multiple Imputation method is also nearly the same with a coefficient estimate of .063. The Mean Imputation technique, however, was furthest from the true value at .05.

It is apparent that none of the imputation techniques helps get us to an answer that mirrors the true value. However, both the Fitted Value and the Multiple Imputation techniques lead us to an answer that is very close to the Complete Cases technique.

It makes sense that the Fitted Value Imputation method provides us with the same coefficient estimate as the Completed Cases method because the imputed y variables all fall on the fitted line. The key difference between these models is that the Fitted Value coefficients have smaller standard errors, which mechanically results from using a larger amount of data.

The Multiple Imputation method runs separate models to calculate values for the missing data points, which are then pooled resulting in output similar to regression coefficients. I used the default arguments, so the number of models used in my imputation was 5 and the method for imputing is predictive mean matching (pmm). At a high-level, the pmm method

uses multiple instances of random sampling of the posterior predictive distribution of the complete cases regression to calculate values for the missing y observations, and then pools these into one value.

Final Project

I plan to use the ISORA tax authority survey data (used in PS6), along with other available data on tax rates and country-level information to understand what factors lead to the best enforcement outcomes. I am most interested in using machine learning for prediction and classification to work through this, and would like to better understand if or how unsupervised models could be used (so I am anxious to learn more about these techniques).