Problem Set 7 - ECON 5253

Thomas Mondry

March 29, 2022

1 Wage data

1.1 Summary

The following table shows a summary of the numeric features in the wage data, after observations with missing schooling or tenure values have been omitted.

	NUnique	PercentMissing	Mean	SD	Min	Median	Max
logwage	670	25	1.63	0.39	0.00	1.66	2.26
hgc	16	0	13.10	2.52	0	12.00	18
tenure	259	0	5.97	5.51	0.00	3.75	25.92
age	13	0	39.15	3.06	34	39.00	46

The following table shows the frequencies of the two character features in the data.

		N
college	college grad	530
	not college grad	1699
married	married	1431
	single	798

1.2 Missing wage data

Log wages are missing for 25% of observations.

Comparing the observations with missing and non-missing wage information, respondents for whom wage values missing are 3 times more likely to be college graduates than not and 4.9 times more likely to be married than unmarried. One average, they have 2.1 more years of schooling and 3 more years of tenure, and there are no substantial outliers in the original data which could be causing these differences to be overstated. In practice, I don't think these discrepancies are prohibitive, but in theory I would call the data most likely MNAR for wages.

The following table shows the coefficients obtained when different methods are used to handle the observations with missing outcome variable:

	Complete cases	Mean imputation	OLS imputation	Multiple imputation
(Intercept)	0.534	0.708	0.534	0.618
	(0.146)	(0.116)	(0.112)	(0.156)
hgc	0.062	0.050	0.062	0.059
	(0.005)	(0.004)	(0.004)	(0.006)
collegenot college grad	0.145	0.168	0.145	0.123
	(0.034)	(0.026)	(0.025)	(0.031)
tenure	0.050	0.038	0.050	0.042
	(0.005)	(0.004)	(0.004)	(0.005)
I(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	0.000
	(0.003)	(0.002)	(0.002)	(0.003)
marriedsingle	-0.022	-0.027	-0.022	-0.014
	(0.018)	(0.014)	(0.013)	(0.016)
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	

Either (more likely) I have made an error somewhere which is escaping me, or the true value $\hat{\beta}_1 = 0.093$ came from a different dataset than the one I'm using.

However, we can observe that the coefficients from the first and third models are identical. This is no surprise, as the missing values in the third model were imputed from the model which was the "best" model (OLS) for the complete cases; therefore, exactly the same model will be "best" (perfect) for the missing cases, and thus also "best" for the full dataset after imputation. Further, we can see that mean imputation results in a model that is substantially different than the models obtained through any other method, which makes sense in the context of my observations about the differences between the average respondent who reported their wages and the average respondent who did not. Therefore, I would most likely select the multiple imputation model.

2 Project status update

As I've shown in previous problem sets, I'll be using data from *Jeopardy!* games. The outcome I'm thinking of modeling is a player's number of consecutive wins, given any available information about their first game. We attempted this problem as-is in a different class last semester and had very limited success; this time, in addition to being able to add more potentially informative features since my data scraping approach yields more information than the source we used previously, I am considering modeling a binary response for whether or not a player will win more than a specific number of games (e.g., more than 3 games) to see if I can get a more practically significant model. The big idea is to find a way to spot great players from their first game in a quantifiable way.

The problem itself is pretty straightforward, and I'll plan to use a variety of model types – for example, standard binomial GLM, tree-based approaches (random forest and/or gradient boosted trees), and support vector classifiers.