ECON 5253: PS7

William Lorton

March 25, 2021

1 Question 6: Missing Values

As indicated in Table 1, approximately 25 percent of the logwage variable observations are missing.

The logwage variable is most likely to be Missing Not At Random (MNAR). Whether someone is working or not (i.e., has or does not have a wage to be observe) is an endogenous issue; one's work status (and any resulting earnings or lack thereof) may come about for a variety of reasons that we don't have any real insight into here. Also, there is no indication that the values are missing, say, because the person who compiled the data set randomly forgot to input values for these particular observations.

2 Question 7: Imputation

We now impute the missing logwage observations in four different ways and estimate the following linear regression model for each case:

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

The coefficient of interest is β_1 which indicates the return to schooling in terms of earnings from work. Table 2 gives the four linear model estimates. Model 1 performed imputation of the logwage variable via listwise deletion (and in doing so assumed the missing observations were MCAR). Model 2 performed mean imputation. Model 3 imputed the missing values by inserting their predicted values from the listwise deletion regression in Model 1. This is the "matching" form of imputation whereby we find a "donor" for the observation with missing logwage data (specifically one with "similar enough" other values); note also that this method assumes the data are MAR. Model 4 performed multiple imputation via the R package mice. The default of 5 imputation methods was chosen and the resulting consolidated model is derived from the "pooled" regression results after running the linear model with the data from each imputation method.

	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	12.0	18
tenure	259	0	6.0	5.5	0.0	3.8	25.9
age	13	0	39.2	3.1	34	39.0	46

Table 1: Summary Table for Wage Data (NAs under hgc and tenure dropped)

We are given that the true value of β_1 is 0.093, indicating that the true return to schooling for women in the U.S. is an approximate 9.75 percent increase in one's wage for every one year increase in education level, all else equal.[1]

It is thus clear that all of the models significantly underestimate the return to schooling, the model that used mean imputation (Model 2) being the worst. We can also notice that there is virtually no difference in the β_1 estimate across models 1, 3, and 4. This may be a result of the fact that none of the corresponding imputation methods match up with the most reasonable assumption of MNAR. Model 1 assumes MCAR and Model 3 optimally assumes MAR. Model 4 may also assume MAR in the case that the mice package primarily works around some form of "matching" imputation; this would also perhaps explain why Model 3 and Model 4 are so similar with regard to their β_1 estimates.

As an additional note, we might guess that mean imputation results in the hgc coefficient falling so significantly because we are assigning the sample average logwage to women who aren't working despite being relatively highly educated (e.g. at least completed college) as a result of, say, their decision to be stay at home mothers. On the other hand, we are also assigning this value to some women who have significantly less education, so perhaps the return-to-schooling understatement effect balances out a bit.

The last two methods (matching via OLS and multiple imputation) both result in estimates of an approximate 6.40 percent increase in earnings for every one year increase in education level, all else equal.

3 Question 8: Project

I'm using division I NCAA football and men's basketball data to estimate the effect of team success on an institution's applicant pool size and the quality of their admit pool. This of course also requires data on applicant numbers, admit test scores, and so on. I'm planning to construct panel data sets for each sport and use them to estimate fixed effects models. I'm currently in the stage of constructing the data sets.

[1] We interpret the slope coefficient as a $(exp(coefficient) - 1) * 100$ percent increase in the dependent variable per every one unit increase in the associated explanatory variable.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.534	0.708	0.534	0.577
	(0.146)	(0.116)	(0.112)	(0.138)
hgc	0.062	0.050	0.062	0.062
	(0.005)	(0.004)	(0.004)	(0.005)
collegenot college grad	0.145	0.168	0.145	0.134
	(0.034)	(0.026)	(0.025)	(0.031)
tenure	0.050	0.038	0.050	0.043
	(0.005)	(0.004)	(0.004)	(0.005)
I(tenureˆ2)	-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	-0.023
	(0.018)	(0.014)	(0.013)	(0.017)
Num.Obs.	1669	2229	2229	2229
Num.Imp.				5
R2	0.208	0.147	0.277	0.226
R2 Adj.	0.206	0.145	0.275	0.224
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	

Table 2: Output from Linear Models Using Data Imputed in Four Different Ways