PS7$_T$*homasson*

Campbell Thomasson

March 25, 2024

1 PS7

At what rate are log wages missing? Do you think the logwage variable is most likely to be MCAR, MAR, or MNAR? After removing observations with missing values for hgc or tenure are missing, there are 2,229 observations remaining. Of those 2,229 observations, 2,229 - 1,669 = 560 logwage observations are missing. 560/2229 is a logwage observation missing rate of approximately 25.12 percent. The logwage variable is most likely to be MAR¡ as is standard across econoemtric research.

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.

First, See the table below:

Table 1: PS7 Models

Term	Statistic	Model 1	Model 2	Model 3	Model 4	Model 5
"(Intercept)"	"estimate"	"0.534***"	"0.534***"	"0.708***"	"0.534***"	"0.647***"
"(Intercept)"	"std.error"	"(0.146)"	"(0.146)"	"(0.116)"	"(0.112)"	"(0.149)"
"hgc"	"estimate"	"0.062***"	"0.062***"	"0.050***"	"0.062***"	"0.059***"
"hgc"	"std.error"	"(0.005)"	"(0.005)"	"(0.004)"	"(0.004)"	"(0.006)"
"collegenot college grad"	"estimate"	"0.145***"	"0.145***"	"0.168***"	"0.145***"	"0.112***"
"collegenot college grad"	"std.error"	"(0.034)"	"(0.034)"	"(0.026)"	"(0.025)"	"(0.032)"
"tenure"	"estimate"	"0.050***"	"0.050***"	"0.038***"	"0.050***"	"0.042***"
"tenure"	"std.error"	"(0.005)"	"(0.005)"	"(0.004)"	"(0.004)"	"(0.004)"
"I(tenure^2)"	"estimate"	"-0.002***"	"-0.002***"	"-0.001***"	"-0.002***"	"-0.001***"
"I(tenure^2)"	"std.error"	"(0.000)"	"(0.000)"	"(0.000)"	"(0.000)"	"(0.000)"
"age"	"estimate"	"0.000"	"0.000"	"0.000"	"0.000"	"0.000"
"age"	"std.error"	"(0.003)"	"(0.003)"	"(0.002)"	"(0.002)"	"(0.003)"
"marriedsingle"	"estimate"	"-0.022"	"-0.022"	"-0.027*"	"-0.022+"	"-0.016"
"marriedsingle"	"std.error"	"(0.018)"	"(0.018)"	"(0.014)"	"(0.013)"	"(0.016)"
"Num.Obs."	""	"1669"	"1669"	"2229"	"2229"	"2229"
"Num.Imp."	""	""	""	""	""	"5"
"R2"	""	"0.208"	"0.208"	"0.147"	"0.277"	"0.223"
"R2 Adj."	""	"0.206"	"0.206"	"0.145"	"0.275"	"0.221"
"AIC"	""	"1179.9"	"1179.9"	"1091.2"	"925.5"	""
"BIC"	""	"1223.2"	"1223.2"	"1136.8"	"971.1"	""
"Log.Lik."	""	"-581.936"	"-581.936"	"-537.580"	"-454.737"	""
"F"	""	"72.917"	"72.917"	"63.973"	"141.686"	""
"RMSE"	""	"0.34"	"0.34"	"0.31"	"0.30"	""

The $\hat{\beta}_1$ are all lower than the true 0.093. The lowest $\hat{\beta}_1$ value is Model 3, which is simply replacing missing values with the mean. The other variations are all in the 0.059-0.063 range. It seems that simply leaving out the missing observations gave a closer approximation to the true $\hat{\beta}_1$ than trying to use multiple imputations or replacing the missing values with the mean.

Model 4 uses the original model (Model 1) to predict the missing values of logwage, then uses those predicted values in the regression.

Model 5 performs multiple imputations (5 in this case) and pools the results together into one estimated model.

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? I have actually been meaning to ask you about this. I am an accounting Ph. D. student and obviously primarily conduct accounting research. I would like to use Compustat, CRSP, and I/B/E/S data to replicate a paper that I think would be useful for me to have a greater understanding of in order to write my Ph. D. program's first-year summer paper, but if you would prefer our research project be a completely original idea as opposed to a replication, I totally understand.