PS7_Zilles

Andrew Zilles

March 2024

6. At what rate are log wages missing? Do you think the logwage variable is most likely to be MCAR, MAR, or MNAR?

The missing rate for logwage is: 0.2512337

I'm not sure. Scrolling through the missing values, it doesn't seem like any one of the other variables could explain it (e.g. there are plenty of missing values for high and low tenures, college eduction and no college education, etc.) That makes me think it's MCAR, but I'm not sure.

7. 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.

- Complete Cases Model: The estimated coefficient for hgc is 0.062, which is slightly lower than the true value of 0.093. This could indicate bias of only completed cases, especially if hgc is missing because of other variables in the model.
- Mean Imputation Model: Replacing missing values with the mean actually lowers the hgc coefficient to 0.050. This is quite a bit lower than both the true value and the estimates from the other models. Mean imputation usually underestimates the true coefficients so this makes sense.
- Regression Imputation Model: This model sort of calculates missing values of hgc using a regression model based on other variables. The estimated coefficient for hgc is 0.062, similar to the estimate from the complete cases model. The problem with this model is that it assumes a linear relationship with hgc and other variables, which isn't always true.

- Multiple Imputation Model: Multiple imputation generates multiple datasets with imputed values for missing data and combines results from analyses of each dataset. The estimated coefficient for hgc is 0.061, similar to the estimates from the other imputation methods. It's interesting that it decreased a little. Multiple imputation generally provides more reliable estimates compared to single imputation methods.

Since three of the four models are all quite similar to each other the best takeaway might be how much mean imputation underestimates coefficients. The other three methods have their own biases with relationships to other variables and don't seem to be capturing the complexity necessary to get us to the true 0.093.

8. 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?

Ah, yes the project. I haven't made much progress yet but hopefully I can get to it towards the end of spring break. My plan right now is to start working on my first year summer paper. I want to explore the relationship with public companies' reporting behaviors and threat of IRS audit. I'm going to use the reported 10-X data by companies but I haven't looked into how to do that yet and what I will need to do. I know there's also resources on $https://sraf.nd.edu/sec-edgar-data/lm_10x_summaries/$ that might already have the data I need but I haven't looked into it that much (it might just be code to obtain the data and I'll need to figure out how to run it and make it work.) Modeling wise, I'm really just trying to compare communication by two groups: companies facing certain audit and companies with audit uncertainty. The plan is to take their 10-X reports and find an average length/quality/tone/etc between these two groups and see if it's roughly the same or statistically different. There are interesting implications if the results go either way. If all of that is too ambitious before the end of the semester I might just settle with replicating "Tax Reporting Behavior Under Audit Certainty" by Ayers, Seidman, and Towery and extending it a little. They're the ones that created the metric to identify firms with certain audit/audit uncertainty.

	Complete Cases	Mean Imputation	Regression Imputation	Multiple Imputation
(Intercept)	0.534	0.708	0.534	0.616
	(0.146)	(0.116)	(0.112)	(0.129)
hgc	0.062	0.050	0.062	0.061
	(0.005)	(0.004)	(0.004)	(0.005)
collegenot college grad	0.145	0.168	0.145	0.138
	(0.034)	(0.026)	(0.025)	(0.029)
tenure	0.050	0.038	0.050	0.040
	(0.005)	(0.004)	(0.004)	(0.004)
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)	(0.002)
marriedsingle	−0.022	−0.027	−0.022	−0.021
	(0.018)	(0.014)	(0.013)	(0.015)
Num.Obs.	1669	2229	2229	2229
R2	0.208	0.147	0.277	0.215
R2 Adj.	0.206	0.145	0.275	0.213
AIC	1179.9	1091.2	925.5	1569.1
BIC	1223.2	1136.8	971.1	1614.8
Log.Lik.	−581.936	−537.580	−454.737	−776.542
RMSE	0.34	0.31	0.30	0.34