Problem Set 7

JaeSeok Oh

March 19, 2024

Imputing Missing Data Practice

Q6.

- *logwages* are missing at approximately 25% $\left(= \frac{560}{1669} * 100\right)$ of the total number of observations. Since wages cannot be negative, it would not be the case of MNAR.

Table 1: Data Summary

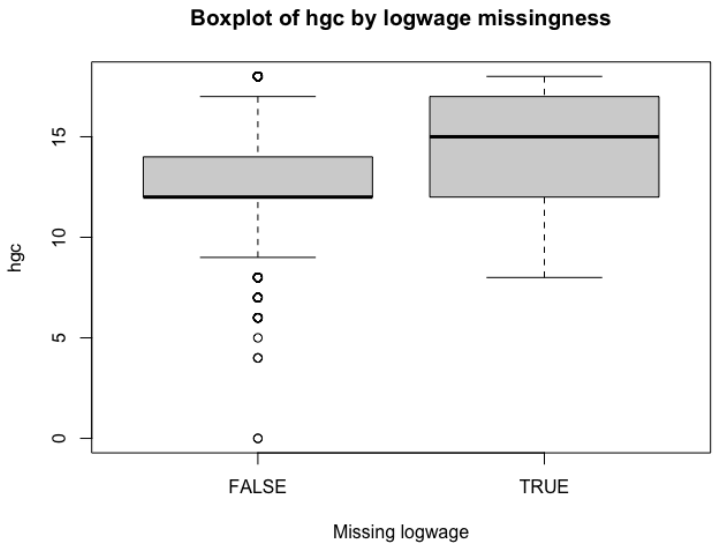
	mean	SD	Min	Max	Median	P0	P25	P50	P75	P100
logwage		0.39	0.00	2.26	1.66	0.00	1.36	1.66	1.94	2.26
hgc	13.10	2.52	0.00	18.00	12.00	0.00	12.00	12.00	15.00	18.00
tenure	5.97	5.51	0.00	25.92	3.75	0.00	1.58	3.75	9.33	25.92
age	39.15	3.06	34.00	46.00	39.00	34.00	36.00	39.00	42.00	46.00

Table 2: Missing Data Details

	college		married		Intersection			
	grad	ngrad	mar	sin	grd&mar	ngrd&mar	grd&sin	ngrd&sin
Missing	273	287	339	221	164	175	109	112
Complete	257	1412	1092	577	179	913	78	499

- In Table 2, we can see the ratio of individual characteristics(college and marriage) of *logwages* missing. Most wage values are from non-college graduate individuals, as well as married individuals. However, the ratio of missing values is high on graduate individuals and single individuals.
- In detail, I tabled the counts of the intersection terms between *married* and *college*. It is noticeable that those who graduated from college missed their wage values. Moreover, the ratio is much higher for the individuals who graduated and single.
- I also conducted the comparison of *hgc* between missing and complete *logwage*. As in Figure 1, *hgcs*, missing *logwages*(right side), have a higher mean and wider range of values. This result is consistent with the previous comment on that missings occurred more frequently in graduate individuals.
- Therefore, to make a conclusion that the missings are MAR cases is reasonable, even though the reason is not clear.

Figure 1:



Q7.

- Table 3 shows four different imputation method results by running linear regression:

$$logwage_i = \beta_0 + \beta_1 hgc_i + \beta_2 college_i + \beta_3 tenure_i + \beta_4 tenure^2 + \beta_5 age_i + \beta_6 married_i + \varepsilon_i \quad (1)$$

Table 3: Regression Results

	Raw	Mean Imp	Predicted Imp	Multiple Imp
(Intercept)	0.534***	0.708***	0.534***	0.606***
	(0.146)	(0.116)	(0.112)	(0.159)
hgc	0.062***	0.050***	0.062***	0.061***
	(0.005)	(0.004)	(0.004)	(0.006)
college	0.145***	0.168***	0.145***	0.125***
	(0.034)	(0.026)	(0.025)	(0.031)
tenure	0.050***	0.038***	0.050***	0.041***
	(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	0.000
	(0.003)	(0.002)	(0.002)	(0.003)
married	−0.022	−0.027*	−0.022+	−0.018
	(0.018)	(0.014)	(0.013)	(0.015)
Num.Obs.	1669	2229	2229	2229
R2	0.208	0.147	0.277	0.223

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

- Column 1 is the regression result from the data set after dropping the missing *logwages*, while the others are the estimates after filling out the missing values by the several imputation methods: (2) mean, (3) predicted from the result of Column 1, and (4) multiple imputation by using *mice* R package.

- Compared to the true value of $\widehat{\beta}_1 = 0.093$, all of them are down-ward estimated. It can apparently be explained by the pattern of missings: *logwages* of higher *hgc* are missing. This pattern should lead the estimates in terms of *hgc* to be lowered. We can see the 'mean imputation method' do the worst work because this might not be able to capture the *hgcs*' differences on the missing values.

- $\widehat{\beta}_1$ is estimated most closely in the first and the third regressions – these two are almost the same. This is evident because the data used in the third regression model is derived from the results of the first estimation with lower standard errors. Thus, the veracity of the predicted imputation depends on the validity of the first regression model. In contrast, the fourth regression result of β_1 seems a concrete reference to recover the true β_1 because it generates the values in random. However, in this case, we know that there is a skewness in the distribution of missing values among graduate and single individuals or higher mean *hgc*, which is not in completely random.

- Therefore, if one was able to figure out whether MCAR, MAR, or MNAR, one could choose the method. In the case of completely at random, multiple imputation (fourth estimates) might be better due to the uncertainty of missing pattern, while in the case of just at random, predicted imputation might reflect the differences(statistical logic or observed) better.

Q8.

- I am still using the data set I downloaded from **'kaggle'**. For me, it is required to look at firstly the literature of the sequential auction. And then, I could make a further decision to implement web-scrape or think of the other way to explore significant implications.

- Now, I am modeling the data as 'long-panel' data with a time trend to capture the price pattern overtime.