

PS7_Ryu

Junyeol Ryu

March 2024





1 Q6.

Answer 1. Rate of log wages missing: 25% (Figure1)

Answer 2. "logwage" variable is most likely to be MAR. This is because The higher the "hgc" value, the higher the missing value probability. (Figure2)

]

Figure 1: Summary Table

	Unique	Missing Pct.	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	0.0	12.0	18.0	
tenure	259	0	6.0	5.5	0.0	3.8	25.9	
age	13	0	39.2	3.1	34.0	39.0	46.0	

2 Q7.

Estimate the regression using only complete cases

hgc coefficient: 0.0623931

Perform mean imputation to fill in missing log wages

mean = 1.62519

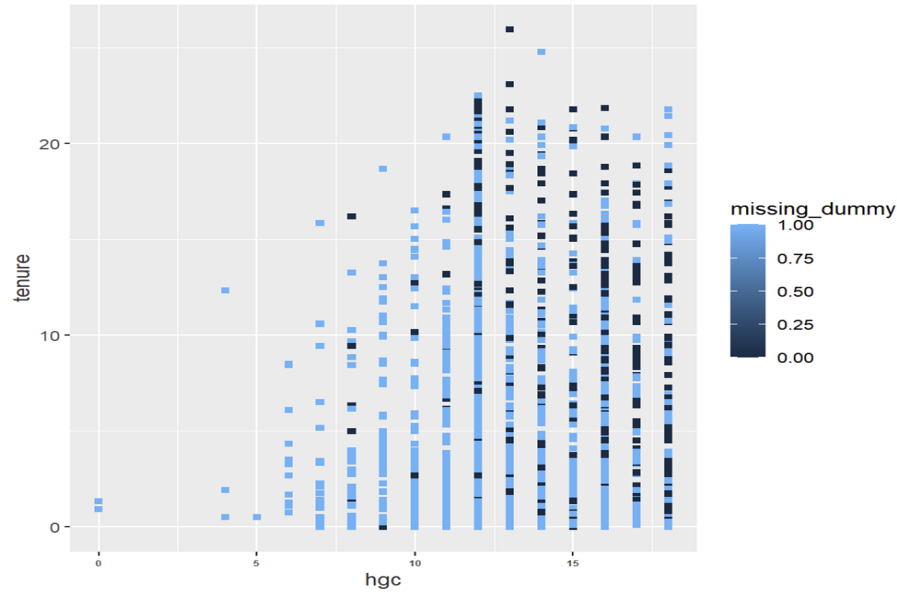
hgc coefficient: 0.0496877

Impute missing log wages as their predicted values from the complete cases regression above

hgc coefficient: 0.0623931

]

Figure 2: Scatter Plot hgc and tenure by logwage value



Answer about the comparison between true value: The estimated beta indicates 0.062, but the true value is known to be 0.093. which means underestimating the true effect of education on wages. We can think of possible reasons for the underestimation.

1) Missing value of wages: Missing values in the dependent variable(wage) lead to a reduced sample size, decreasing the degree of the accurate estimation. Moreover, if the missingness is not completely random(MCAR), it can cause bias into the estimated coefficient. This could be a possible reason for underestimation in complete case analysis with 1669 observations, but Mice and Amelia's results are also underestimated. Thus, we can think of other possible reasons. 2) Omitted Variable Bias: If there are other relevant variables that affect wages and are correlated with education but are not included in the model, this can lead to biased estimates of the education effect. 3) Measurement error: If there is a measurement error in the education variable, such as inaccurate information about the schooling year, this can reduce the estimated coefficient. 4) Sample selection: If the sample used in the analysis is not representative of the population, this could lead to differences in the estimated effect.

3 Q8.

My data set about the final project doesn't have missing value.

Figure 3: Mice Multiple Imputation Regression Model

	Listwise deletion	Mice	Amelia
(Intercept)	0.657	0.660	0.631
	(0.130)	(0.159)	(0.161)
hgc	0.062	0.061	0.063
	(0.005)	(0.006)	(0.005)
tenure	0.050	0.048	0.051
	(0.005)	(0.004)	(0.006)
tenure_sq	-0.002	-0.001	-0.002
	(0.000)	(0.000)	(0.000)
age	0.000	0.001	0.001
	(0.003)	(0.003)	(0.004)
married_dummy	0.022	0.026	0.024
	(0.018)	(0.023)	(0.018)
college_dummy	-0.145	-0.142	-0.144
	(0.034)	(0.032)	(0.047)
Num.Obs.	1669	2229	2229
Num.Imp.		5	5
R2	0.208	0.223	0.225
R2 Adj.	0.206	0.221	0.223
AIC	1179.9		
BIC	1223.2		
Log.Lik.	-581.936		
RMSE	0.34		