Problem Set 7- Questions 6,7 and 8

Yeganeh Karbalaei

1 Summary Table

Table 1: Summary Statistics of the Dataset

Variable	Unique	Missing Pct.	Mean	SD	Min	Max
logwage	670	25.0%	1.6	0.4	0.0	2.3
hgc	16	0.0%	13.1	2.5	0.0	18.0
tenure	259	0.0%	6.0	5.5	0.0	25.9
age	13	0.0%	39.2	3.1	34.0	46.0

Category	N	%
college: college grad	530	23.8%
college: not college grad	1699	76.2%
married: married	1431	64.2%
married: single	798	35.8%

There are 560 missing logwage values in your dataset after removing observations with missing hgc or tenure. This represents 25% of the total observations. If the data has been missed completely at random, then the missingness is unrelated to any observed or non-observed variables but if it is MAR we can realize it based on variables we have information about. In MNAR, values are missing but not completely random and it is based on unobserved variables. For this question, we have to compare the average values of variables age, hgc and tenure with and without missing data. There is almost 2 units difference between hgc with missing data and without missing data. Also, there is 3 units difference between average values of tenure variable with and without considering missing data. So, we see that the missingness is related to observed variables, so it is MAR.

2 Regression Table

In "Complete Cases" method, the coefficient of variable hgc is 0.062, in "Mean Imputation" it is 0.050, in "Predicted Imputation" it is 0.062 and based on "Multiple Imputation" method, the coefficient is 0.060. All imputation methods underestimate the returns of length of schooling. However, mean imputation method is the worst. Both "Complete Cases" and "Predicted Imputation" gave us the same number which means one of them reproduce the other method result. The result we got for the coefficient of hgc suggests that the process of imputation is more complicated and we have to consider other imputation methods.

	Complete Cases	Mean Imputation	Predicted Imputation	Multiple Imputation
(Intercept)	0.534***	0.708***	0.534***	0.593***
	(0.146)	(0.116)	(0.112)	(0.138)
hgc	0.062***	0.050***	0.062***	0.060***
	(0.005)	(0.004)	(0.004)	(0.005)
college: not college grad	0.145***	0.168***	0.145***	0.121**
	(0.034)	(0.026)	(0.025)	(0.035)
tenure	0.050***	0.038***	0.050***	0.044***
	(0.005)	(0.004)	(0.004)	(0.005)
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)	(0.003)
married: single	-0.022	-0.027*	-0.022^{+}	-0.015
	(0.018)	(0.014)	(0.013)	(0.017)
Num.Obs.	1669	2229	2229	2229
Num.Imp.				5
R^2	0.208	0.147	0.277	0.228
R^2 Adj.	0.206	0.145	0.275	0.226
AIC	1179.9	1091.2	925.5	
BIC	1223.2	1136.8	971.1	
Log.Lik.	-581.936	-537.580	-454.737	
RMSE	0.34	0.31	0.30	

$^{+}$ p ¡ 0.1, * p ¡ 0.05, ** p ¡ 0.01, *** p ¡ 0.001

Table 2: Comparison of Regression Models with Different Imputation Methods

3 Research Progress

Since I want to see the effect of Violence Against Women Act on Native American Women who are living in reservations, I could get the related data from the reports of the office of Violence Against Women. I want to take other Native Women who live off-reservation as the control group and women from tribes as the treatment group, and based on their age group, relationship with the offenders, and type of victimization, see whether the policy was helpful or not. I am not sure is it possible when the data is not individual-level or not. This is an initial step of the further research in which I want to study the effect of domestic violence on infant mortality.