Data Science
Ruichun Liu
Problem Set 7

1. **Question 6**

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
logwage	1,686	1.622	0.388	0.005	1.358	1.936	2.261
hgc	2,244	13.099	2.521	0.000	12.000	15.000	18.000
tenure	2,231	5.978	5.510	0.000	1.583	9.333	25.917
age	2,246	39.153	3.060	34	36	42	46

Table 2:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
logwage	1,669	1.625	0.386	0.005	1.362	1.936	2.261
hgc	2,229	13.101	2.524	0	12	15	18
tenure	2,229	5.971	5.507	0.000	1.583	9.333	25.917
age	2,229	39.152	3.062	34	36	42	46

From the table two, we can see that there are 1669 variables for logwage and 2229 for each others. Therefore, the missing rate is about 25 %. I think it is MNAR.

2. **Question 7**

Table 3 is on the next page. Since the true value is 0.093, we can see the results from complete cases and predicted value cases are quit close to the true value. The result from mean imputation is small.
Results in column 1 and 3 are the same. I think the reason might be that case 3 uses the predicated value from case 1, which would not change the value of coefficient. However, mean imputation will lead to a lower beta value. Missing Completely At Random and Missing At Random are good ways to predict the model.
Beta 1 in column 2 means that logwage would increase by 4.9% if there were 100% increase in hgc. Beta 1 in column 3 means that logwage would increase by 6.2% if there were 100% increase in hgc.

3. **Question 8**

Now I am working on a project about education in China. My topic is how education funding would affect the acceptance rate of high school in different regions and provinces in China. I have collected education funding data, the acceptance rate of high school, population, GDP, income, etc. during 1996-2017. I am going to use the panel data to run a FE Model (at province or city unit level). I would like to see how education funding affects the acceptance rate and if there are significant differences in different regions and provinces.

Table 3:

	Dependent variable:		
	logwage		
	(1)	(2)	(3)
hgc	0.062***	0.049***	0.062***
	(0.005)	(0.004)	(0.004)
collegenot college grad	0.146***	0.160***	0.146***
	(0.035)	(0.026)	(0.025)
tenure	0.023***	0.015***	0.023***
	(0.002)	(0.001)	(0.001)
age	−0.001	−0.001	−0.001
	(0.003)	(0.002)	(0.002)
marriedsingle	−0.024	−0.029**	−0.024*
	(0.018)	(0.014)	(0.013)
Constant	0.639***	0.833***	0.639***
	(0.146)	(0.115)	(0.111)
Observations	1,669	2,229	2,229
R^2	0.195	0.132	0.268
Adjusted R^2	0.192	0.130	0.266
Residual Std. Error	0.346 (df = 1663)	0.311 (df = 2223)	0.300 (df = 2223)
F Statistic	80.508*** (df = 5; 1663)	67.496*** (df = 5; 2223)	162.884*** (df = 5; 2223)
Note:			*p<0.1; **p<0.05; ***p<0.01