Problem Set 7

Dillon Beake

March 26, 2024

1 Summary Table

Table 1: Summary statistics of the dataset

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
logwage	1,669	1.625	0.386	0.005	1.362	1.655	1.936	2.261
hgc	2,229	13.101	2.524	0	12	12	15	18
tenure	2,229	5.971	5.507	0.000	1.583	3.750	9.333	25.917
age	2,229	39.152	3.062	34	36	39	42	46

2 Missing Logwage

2.1 Rate of missing Log Wages

The missing rate for logwage is 25.12

2.2 Explanation for Missing Log Wages - MNAR

I suspect that the missing log wages are Missing not at random (MNAR). The date being collected, e.g. wages, can be considered to be sensitive information. Some people may not want to report low wages, and also some people do not feel comfortable reporting high wages. This could be a likely cause for the missing data.

3 Regression Table

Table 2: Regression Results

Variable	Complete Cases	Mean Imputation	Predicted Values Imputation	Multiple Imputation
Intercept	0.534	0.708	0.534	0.534
hgc	0.062	0.050	0.062	0.062
college	0.145	0.168	0.145	0.145
tenure	0.050	0.038	0.050	0.050
I(tenure^2)	−0.002	−0.001	−0.002	−0.002
age	0.000	0.000	0.000	0.000
married	−0.022	−0.027	−0.022	−0.022

3.1 Analysis of coefficient b1 = hgc (years of schooling)

The complete cases estimated value = .062, which is lower than True Value = .093. This is likely due to the missing cases that were not included due to missing values.
The mean imputation predicted value = 0.50, which is even a lower estimate of the true value = .093. Mean imputations underestimate true errors, and it is hard to detect error that is not MCAR. Further, this could strengthen the case that the missing data was MNAR.
The predicted value estimation value = .062, which is lower than the true value = .093. The predicted value imputation includes predictions of missing values. However if missing values are not MAR, or MCAR, then it's hard to make a prediction as to why the values are missing. People may decide to just not anser that question because they don't want to share their level of education, or maybe a choice didn't apply to them.
Multiple imputation is typically the most robust because it takes complete cases, missing cases, and includes the prediction imputations. Typically, giving a more accurate measurement. HOwever, in this case the value = .062. Which is the same as the Predicted Values Imputation. This could again suggest that the missing items are not MCAR.

4 Project Progress

I need to speak with my mentor about the direction of my research. I still planning on using publicly available body measurements e.g. body composition, arm-length, wing-span, etc. And compare them to publicly available performance outcomes e.g. broad jump, vertical jump, pro agility drill, 40 yd dash splits, etc. The NFL combine is a great source for this data. Depending on how many predictor and outcome variables I decide to use. If I have one predictor variable and several outcome variables, I could use a Multivariate Analysis of Variance (MANOVA). If I have multiple predictor and outcome variables I could use a Multivariate Regression Analysis