Problem Set Seven

Devin Williams

March 27, 2025

1 Data Summary Table

Table 1 shows the summary statistics for the dataset.

Table 1: Summary Statistics

	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 Missing Data Analysis

The rate of missing log wages is 25% seen from Table 1.

3 Is the logwage variable MCAR/MAR/MNAR?

It is very unlikely that the logwage variable is missing completely at random due to the high amount of missing data. 25% missing data demonstrates little chance of a purely random finding. It is possible that this data is MNAR, where people that are very low or very high on overall pay are less likely to report their wage.

4 Imputation methods and regression table

Table 2: Comparison of Returns to Schooling (β_1) Across Methods.

Method	Estimate	Std. Error	Bias from True Value
Complete Cases	0.0632	0.0054	-0.0298
Mean Imputation	0.0507	0.0043	-0.0423
Regression Imputation	0.0632	0.0042	-0.0298
Multiple Imputation	0.0640	0.0049	-0.0290
True Value	0.0930	NA	0.0000

5 Questions about table

- What patterns do you see?
  - All methods seem to underestimate the true value of 0.093. This is based on the bias from true value column.
  - Mean imputation performs the worst of these.
  - Multiple imputation performs the best.
- What can you conclude about the veracity of the various imputation methods?
  - Mean imputation is the least reliable, it may be distorting some of the relationships between variables as the single value is not the best way of imputing this large amount of missing values.
  - Complete cases perform well, but could create other biases by removing the cases entirely. This could make the data less representative.
  - Regression imputation has a smallest std. error in complete cases, demonstrating some efficiency gains. But could still have the same bias.
  - Multiple imputation has the best performance, it is able to best preserve the underlying relationships in the data while accounting for uncertainty.

- ◇ Discuss the Beta1 estimates for the last two methods.
    - ◇ Regression imputation shows the ability for predicated values to still have some efficiency. But really doesn't improve bias as mentioned. Multiple imputation demonstrates advantage for incorportating uncertainty. It comes closest to recovering the true parameter value.

6 Progress of Project

I am using worldwide soccer player data from transfermarkt to get an idea about potential biases from player nationality, while looking at what currently contributes most to a players transfer evaluation. I am hoping to create a multiple regression to get a better look at how things like goals, nationality, and age can impact this transfer evaluation.