PS7$_{O}lubayode$

Olubayode Ebenezer

March 22, 2024

1 Introduction

This report presents a comprehensive analysis of a dataset containing information on approximately 2,250 women who were employed in the United States in 1988. A key variable of interest is logwage, the natural logarithm of wages, which facilitates the examination of factors influencing wage disparities. Specifically, the project aims to understand the impact of education (hgc), job tenure, age, marital status, and college education status on wage. The dataset includes variables such as logwage, years of education (hgc), tenure with the current employer, and age. A key focus of this analysis is to examine the missingness of log wage data and evaluate the efficacy of various imputation methods on the estimation of returns to schooling.

Data Description

The dataset, wages.csv, comprises several variables:
beginitemize
item logwage: The natural logarithm of wages, representing the dependent variable in our analysis.
item hgc: Years of education completed by an individual, serving as a proxy for human capital.
item tenure: The number of years an individual has been with their current employer.
item age: The age of the individual.
item married: Marital status of the individual, categorized as "married" or "single."

item college: Indicates whether the individual is a college graduate, with categories "college grad" and "not college grad."
enditemize Initially, the dataset contained missing values in hgc, tenure, and logwage. Observations with missing hgc or tenure were removed to ensure a complete analysis dataset.

2 Dataset Summary

The dataset provides a rich source of information on the employment characteristics of the sample population. The following table summarizes key statistics:

	logwage	hgc	tenure	age
count	1669.000000	2229.000000	2229.000000	2229.000000
mean	1.625190	13.101391	5.970615	39.151638
std	0.385534	2.524306	5.507216	3.061954
min	0.004940	0.000000	0.000000	34.000000
25%	1.362255	12.000000	1.583333	36.000000
50%	1.655079	12.000000	3.750000	39.000000
75%	1.936200	15.000000	9.333333	42.000000
max	2.261495	18.000000	25.916670	46.000000

3 Analysis of Missing Log Wages

The rate of missing log wages in the dataset is calculated to be 25.12%. To understand the nature of this missingness, an analysis was conducted comparing the mean values of other variables for observations with and without log wage data. The findings suggest a pattern that does not support the data being Missing Completely At Random (MCAR). Instead, the characteristics of the missing data indicate a potential Missing Not At Random (MNAR) scenario, particularly due to the sensitive nature of wage information and the selective non-disclosure by participants.

4 Regression Analysis: Model Summary

The regression analysis results, including all coefficients for each imputation method, are presented in the table below. This comprehensive view helps to understand the impact of different imputation techniques on the estimated parameters of the model, particularly focusing on the returns to education and other factors influencing wage determination.

Table 1: Regression Model Coefficients Across Different Imputation Methods

	Complete Cases	Mean Imputation	Predicted Imputation	Multiple Imputation
Intercept	0.533569	0.707596	0.533569	0.567189
C(college)[T.not college grad]	0.145168	0.168228	0.145168	0.139741
C(married)[T.single]	-0.022046	-0.026833	-0.022046	-0.021299
hgc	0.062393	0.049688	0.062393	0.061235
tenure	0.049525	0.038168	0.049525	0.041061
np.power(tenure, 2)	-0.001560	-0.001330	-0.001560	-0.001000
age	0.000441	0.000200	0.000441	0.000435

Analytical Approach and Modeling

The analytical approach adopted for this project involves regression models to estimate the impact of the independent variables on logwage. Four distinct modeling strategies were employed to address missing data in the logwage variable:

beginenumerate

item Complete Cases Analysis: This model utilized only observations with no missing logwage data, serving as a baseline for comparison with other imputation methods.

item Mean Imputation: Missing logwage values were replaced with the mean of observed logwage values, and the model was re-estimated.

item Predicted Imputation: Missing logwage values were imputed using predicted values from a regression model estimated on the complete cases. This approach assumes that the missingness is related to observed data (MAR).

item Multiple Imputation: A more sophisticated method, multiple imputation by chained equations (MICE), was simulated to handle missing logwage values, creating several imputed datasets that were analyzed separately. The results were then pooled to provide a comprehensive estimate.

endenumerate

Table

This table below highlights the differences in coefficient estimates across the four models: Complete Cases, Mean Imputation, Predicted Imputation, and Multiple Imputation. Such differences underscore the varying impacts of imputation strategies on the analysis, particularly in terms of interpreting the influence of education (hgc), tenure, and other variables on wage outcomes.

To assess the impact of different imputation methods on the estimated returns to education (hgc), four regression models were analyzed. The results are summarized as follows:

Method	$\hat{\beta}_1$
Complete Cases	0.062393
Mean Imputation	0.049688
Predicted Imputation	0.062393
Multiple Imputation	0.061235

5 Regression Analysis

The analysis of $\hat{\beta}_1$ across these models highlights the limitations and biases associated with each imputation method. The complete cases and predicted imputation methods provide estimates closer to the true effect of education on wages, while mean imputation significantly underestimates this effect. Multiple imputation offers a more nuanced estimate but still does not fully capture the true value of 0.093.

This report underscores the critical importance of selecting appropriate imputation methods for handling missing data in wage analysis. The findings from the regression analysis illuminate the biases inherent in simpler imputation techniques and advocate for the use of more sophisticated methods like multiple imputation. However, it also acknowledges the challenges in fully compensating for the missingness, especially when the data is MNAR. Future research should continue to explore and refine methods for dealing with missing data to enhance the accuracy and reliability of economic analyses.

Conclusion and Next Steps

The preliminary analysis underscores the importance of carefully selecting imputation methods when handling missing data. The next steps in this project will involve a more in-depth examination of the differences in model estimates, further exploration of multiple imputation techniques, and potentially incorporating additional variables or interaction terms to better understand the dynamics influencing wage disparities.