# PS7$_{Z}hang$

Haotian Zhang

March 2025

## 1 Pictures and Discussions I

Table 1: Summary Statistics of Numeric Variables

| Variable | Unique | Missing Pct. | Mean | SD | Min | Median | Max |
|---|---|---|---|---|---|---|---|
| logwage | 671.0 | 25.0 | 1.6 | 0.4 | 0.0 | 1.7 | 2.3 |
| hgc | 17.0 | 1.0 | 13.1 | 2.5 | 0.0 | 12.0 | 18.0 |
| tenure | 259.0 | 1.0 | 6.0 | 5.5 | 0.0 | 3.8 | 25.9 |
| age | 13.0 | 0.0 | 39.2 | 3.1 | 34.0 | 39.0 | 46.0 |

Table 2: Summary Statistics of Categorical Variables

| Variable | N | % |
|---|---|---|
| college grad | 532.0 | 23.7 |
| not college grad | 1714.0 | 76.3 |
| married | 1442.0 | 64.2 |
| single | 804.0 | 35.8 |

Approximately 24.93% of the `logwage` values are missing in the dataset. The most reasonable assumption based on the data would be that logwage is MAR (Missing at Random), for the missingness is related to observed variables (e.g., education, tenure).

## 2 Pictures and Discussions II

We can observe that all methods yield estimates below the true value of 0.093.

The complete case and predicted imputation methods yield identical estimates for $\beta_1$, both at 0.062. The complete case approach regresses only those observations that do not have missing values for any variable, including `logwage`. The predicted imputation approach fills in the missing `logwage` values with the predicted values from a regression model trained on the complete cases. This means that the imputed values are deterministic functions of the other variables in the complete case data. Since both approaches use essentially the same regression structure and the same underlying data to estimate $\beta_1$, they produce the same estimate for the `hgc` coefficient.

Mean imputation results in the lowest estimate of $\beta_1$ (0.050). This method replaces all missing `logwage` values with the mean of the observed `logwage` values. This artificially reduces the variability of the `logwage` variable because each imputed value is exactly the same. As a result, the relationship between `logwage` and other predictors becomes weaker, since the imputed values do not reflect the natural variation present in the observed data. Therefore, the $\beta_1$ produced by this method is the lowest among the four approaches.

Table 3: Estimates of the Four Regression Models

| | **Complete Cases** | **Mean Imputation** | **Predicted Imputation** | **Multiple Imputation** |
|---|---|---|---|---|
| (Intercept) | 0.534 | 0.708 | 0.534 | 0.658 |
| | (0.146) | (0.116) | (0.112) | (0.141) |
| hgc | 0.062 | 0.050 | 0.062 | 0.058 |
| | (0.005) | (0.004) | (0.004) | (0.005) |
| collegenot college grad | 0.145 | 0.169 | 0.145 | 0.103 |
| | (0.034) | (0.026) | (0.025) | (0.029) |
| 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) |
| marriedsingle | -0.022 | -0.027 | -0.022 | -0.018 |
| | (0.018) | (0.014) | (0.013) | (0.016) |
| Num. Obs. | 1669 | 2229 | 2229 | 2246 |
| Num. Imputations | – | – | – | 5 |
| $R^2$ | 0.208 | 0.146 | 0.277 | 0.227 |
| Adj. $R^2$ | 0.206 | 0.144 | 0.275 | 0.225 |
| AIC | 1179.9 | 1093.8 | 925.5 | – |
| BIC | 1223.2 | 1139.5 | 971.1 | – |
| Log Likelihood | -581.936 | -538.912 | -454.737 | – |
| RMSE | 0.34 | 0.31 | 0.30 | – |

Multiple imputation produces an estimate of $\beta_1$ equal to 0.058. This method creates multiple versions of the dataset, each containing different plausible values for the missing data. These values are drawn from statistical distributions, so they incorporate random variation that reflects the uncertainty associated with the missing values. The final regression results are obtained by pooling the estimates across all imputations, which tends to produce more realistic coefficient estimates and standard errors. Although the multiple imputation estimate of $\beta_1$ (0.058) is still lower than the true value (0.093), it is closer than the estimate produced by mean imputation (0.050), and it better captures the inherent uncertainty of the missing data.

By comparison, we can conclude that mean imputation is the least reliable because it artificially lowers standard errors by reducing variability in the imputed values. The predictive imputation method preserves the original data structure but may reinforce any existing biases present in the complete case data, as it relies entirely on the same regression model. Multiple imputation, while still producing a slightly underestimated $\hat{\beta}_1$, is arguably the most statistically sound method because it incorporates the uncertainty and variability associated with the imputation process.

Thus, among the four methods evaluated, multiple imputation offers the best balance of realism and statistical rigor, making it the preferred approach for handling missing data in this context.

# 3 Progress on my project

I'm going to use the loan data I used last time to finish my fianl project. In last problem set I finish the data clean and data visualization part. During the spring break I made progress on Literature review part. I want to use the machine learning models to finish the project, but it will depend on the following class we take.