# Problem Set 7

Samuel Hunt

March 25, 2025

## 1 Imputation Methods

From our summary statistics table, we can see that log wages are missing at a rate of 25 percent. Additionally, because we can assume that this wages dataset is some form of survey data, we can assume the worst that the data is MNAR. It is possible that there is an attenuation bias in log wages, such as those with no wages reporting their value as NA, or people being embarrassed to report their log wages. This means that there could be an omitted variable explaining systemic differences in the true value of the omitted log wage variables, meaning the data would be MNAR.

Across the four different models, we see relatively similar coefficients for $\beta_1$. The first and third imputation method result in the same value of 0.062. Additionally, the multiple imputation method is very similar at a value of 0.059. The mean imputation method is the most different at 0.05. Although these values are relatively similar to each other, they are all quite different from the true value of $\beta_1$ which is 0.093. This tells us that MCAR and MAR are both unlikely to be accurate fits for our data, as neither model 1 or model 3 resulted in an accurate estimate of the coefficient. Mean imputation is generally unreliable, and that also held true in our test as it has the least accurate coefficient estimate. Finally, it looks like the multiple imputation model was also unable to be accurate as our data must be MNAR. This shows that imputation won't always be accurate when a attenuation bias occurs in the missing values of our data.

## 2 Project Update

My project for this class has been going well and I believe that I am making smooth progress. It is a research paper that I am doing for my honors thesis that looks at the impact of NAFTA on various mortality statistics in the United States. To make that paper my project for this class as well, I am hoping to create a fully replicable package of scripts that will be able to produce the models and images that are used in my paper. I have already started working with import and export data to create a NAFTA vulnerability metric for each county in the US, working off of a framework of an existing paper. For the models I plan to use, I know that I will end up utilizing an event-study model. There exists a general formula for this model that I will be able to add my chosen dependent variables of mortality statistics to, and this will show me the importance of NAFTA vulnerability for each mortality statistic before and after NAFTA was enacted.

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 |
| black | black | black | black | black | black | black | black |
| | | N | % | | | | |
| college | college grad | 530 | 23.8 | | | | |
| | not college grad | 1699 | 76.2 | | | | |
| married | married | 1431 | 64.2 | | | | |
| | single | 798 | 35.8 | | | | |

Table 2: Regression Outputs
Table 3:

| | Listwise Deletion | Mean Imputation | Predicted Imputation | Multiple Imputation |
|---|---|---|---|---|
| (Intercept) | 0.534 | 0.708 | 0.534 | 0.656 |
| | (0.146) | (0.116) | (0.112) | (0.161) |
| hgc | 0.062 | 0.050 | 0.062 | 0.059 |
| | (0.005) | (0.004) | (0.004) | (0.006) |
| collegenot college grad | 0.145 | 0.168 | 0.145 | 0.113 |
| | (0.034) | (0.026) | (0.025) | (0.031) |
| tenure | 0.050 | 0.038 | 0.050 | 0.042 |
| | (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 | -0.015 |
| | (0.018) | (0.014) | (0.013) | (0.016) |
| black | black | black | black | black |
| Num.Obs. | 1669 | 2229 | 2229 | |
| R2 | 0.208 | 0.147 | 0.277 | |
| R2 Adj. | 0.206 | 0.145 | 0.0.275 | |
| AIC | 1179.9 | 1091.2 | 925.5 | |
| BIC | 1223.2 | 1136.8 | 971.1 | |
| Log.Lik. | -581.936 | -537.580 | -454.737 | |
| RMSE | 0.34 | 0.31 | 0.30 | |