Problem Set 7

Haocheng Yang

March 27, 2023

Abstract

This document summarizes and compares the regression results of female wages analysis. And the document also summarizes the progress of my final project.

1 Missing rate and the Type of Missingness

The missing rate of wages dataset is around 0.2512, In my opinion, the type of missing could either be Missing at Random or Missing not at random. The missing value of log wages could be relevant to the observed variables and also could be the result caused by unobserved variables. For example, a female worker with a graduate degree might also be a full-time housewife who does not work in industry, and a female worker who has a number of kids and senior citizen that need to be taken care of might not have a job in a company.

2 The Comparison Among s

The true value of ˆ1 = 0.093, which is different from all of the estimated values obtained from the imputation methods. You can get the information about s from the table listed on page two 1. The pattern we observe is that the estimates from the imputation methods are generally smaller than the true value of ˆ1.

Comparing the estimates, we can see that the mean imputation method has lowest of (= 0.043***), followed by the multiple imputation method (= 0.06). and the listwise deletion and missing log as predicted value methods have the same magnitude of s (= 0.065* and = 0.065***, respectively).

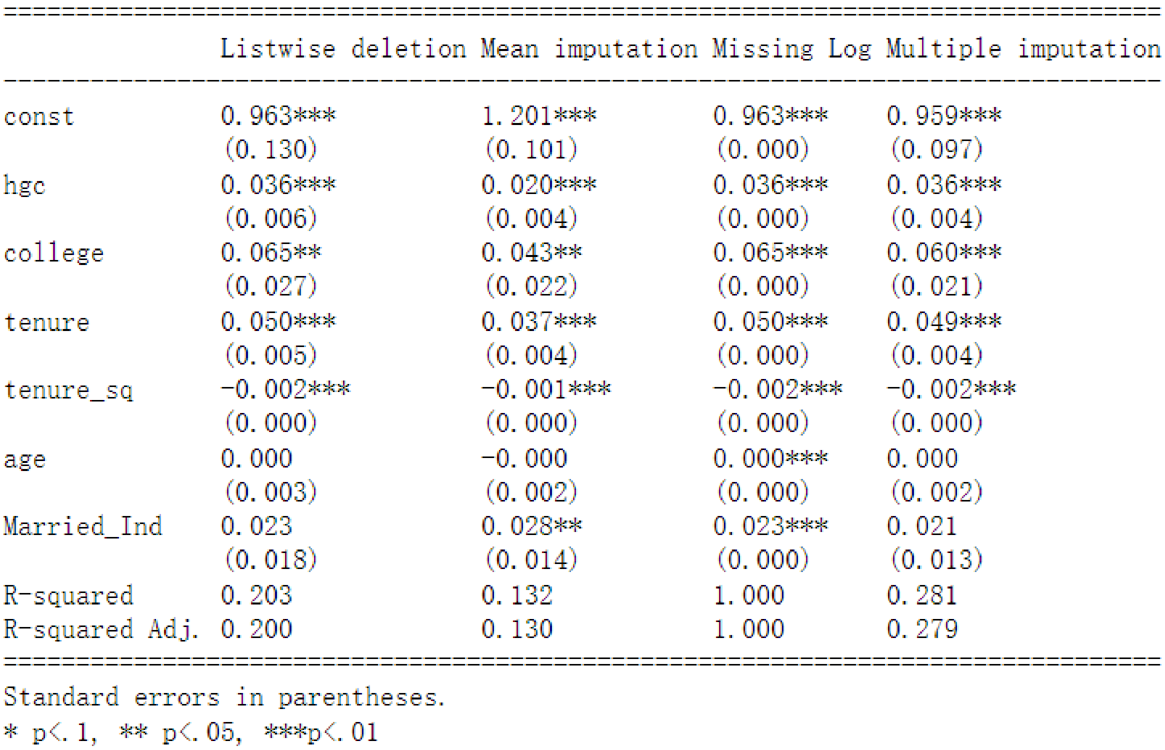
Mean imputation model gives the most biased value of 0.043, compared with the true beta of 0.093. The assumption for mean imputation is that the missing value is missing at random. If the missing values in our dataset is not missing at random or missing completely at random, the mean imputation model could be biased.

Regarding the estimates for the last two methods, the missing log as predicted value method produces a similar estimate (= 0.065***) to the listwise deletion method (= 0.065*). This is expected as both methods involve a form of deletion, either explicitly or implicitly, which may introduce similar biases. The missing log as predicted value method replaces the missing values with predicted values based on a regression model. This method may produce less biased estimates than mean imputation but can still be biased because it relies on a single imputed value.

The multiple imputation method produces a slightly smaller estimate (= 0.06**) than the other two methods. In my opinion, the multiple imputation method produces a more accurate and sophisticated results. beacuse it creates multiple imputed datasets by modeling the missing data based on observed data. Comparing among the four method, I consider that multiple imputation method generates more reliable correlation.

3 Final Project

The utilized model in my final project is OLS regression model. I am more familiar with this model but I might switch the model after we gain more knowledge in machine learning. I using financial data from WRDS database.



```
================================================================================
               Listwise deletion Mean imputation Missing Log Multiple imputation
--------------------------------------------------------------------------------
const          0.963***          1.201***        0.963***    0.959***
               (0.130)           (0.101)         (0.000)     (0.097)
hgc            0.036***          0.020***        0.036***    0.036***
               (0.006)           (0.004)         (0.000)     (0.004)
college        0.065**           0.043**         0.065***    0.060***
               (0.027)           (0.022)         (0.000)     (0.021)
tenure         0.050***          0.037***        0.050***    0.049***
               (0.005)           (0.004)         (0.000)     (0.004)
tenure_sq      -0.002***         -0.001***       -0.002***   -0.002***
               (0.000)           (0.000)         (0.000)     (0.000)
age            0.000             -0.000          0.000***    0.000
               (0.003)           (0.002)         (0.000)     (0.002)
Married_Ind    0.023             0.028**         0.023***    0.021
               (0.018)           (0.014)         (0.000)     (0.013)
R-squared      0.203             0.132           1.000       0.281
R-squared Adj. 0.200             0.130           1.000       0.279
================================================================================
Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01
```

Figure 1: The beta comparison table.