Modeling Federal employees data to predict promotions

Chia-Rui (Jerry) Chang

May 2018

# If you are creating a PDF you’ll need to write your preliminary content here or

Placeholder

# Introduction

Placeholder

# This chunk ensures that the thesisdowndss package is

Placeholder

## Missing Data

# Imputation of Missing Data

## Overview of Multiple Imputation

Multiple Imputation (MI) is a flexible approach for researchers to address the problem of missing data. The method for carrying out multiple imputation involve: (1) fill in the missing values through drawing values from predictive conditional distribution to produce completed datasets; (2) for each completed dataset, conduct analysis for parameters of interest; and (3) combine individual analysis to form final results (Little & Rubin, 2014). Multivariate imputation by chained equation (MICE), where multiple imputed datasets are created through drawing samples sequentially from predictive conditional distribution, is often applied to generate multiple completed datasets at step (1). MICE is one of the most popular approach for imputing missing data. It provides flexible imputation results and can handle both continuous and categorical data. The implementations of MICE involve the following steps.

Here, we assume we have a data matrix , where are completely observed and are partially observed.

**Step 1**  
For , fill in initial missing values of through draws from predictive conditional distribution .  
**Step 2**  
For , (a) construct a predictive distribution for fully conditional on .  
(b) Draw values from the conditional distribution and update the original missing values of column . **Step 3** Perform step 2 for times.  
**Step 4** Perform step 1-3 for times to generate completed datasets.

We applied MICE to impute missing values in the OPM data, where and . Since the robustness of MICE is dependent on the predictive conditional distribution, one strategy to improve imputation result is to include “auxiliary variable” that are related to the missingness but not part of the covariates. Indeed, the inclusion of “auxiliary variable” can make the MAR assumption more reasonable (Collins, Schafer, & Kam, 2001). In the OPM data, variables such as *Pay Plan* and *Step Rate* are not of interest to the final analysis, but they provide information on the missingness of *Salary* and *Grade*. Another auxiliary variable we include is salary between working year 11 and 15. Though we are only interested in the rate of change in salary between working year 1 and 10, salary between working year 11 and 15 can help predict past salary change rate.

## Specifying Predictive Conditional Distribution - the CART method

In the practice of MICE, one of the most common models for specifying predictive conditional distribution is Generalized Linear Models (GLMs). GLMs such as multiple linear regressions often produce consistent imputation results. However, if the data to be imputed contain hundreds of variables, GLMs might be too simple to capture the true distribution. For example, relationships among variables might be interactive and non-linear. Specifying parametric models for data with great complexity is therefore inappropriate (Burgette & Reiter, 2010). The OPM data contains 65 variables. Eeach categorical variable has various levels. Since non-linear relationships might exist among multiple variables and levels, specifying the standard GLMs on the conditional distribution could lead to biased parameter estimates and produce inconsistent results. To address this challenge, non-parametric model can be more appropriate; specifically, we used Classification and Regression Trees (CART) to impute missing data.

The CART algorithm performs binary splits of the predictors recursively to approximate the conditional distribution of an univariate outcome. The partitions are found if the subsets of units have relatively homogeneous outcomes. The leaf would be reached after multiple partitions, with values in each leaf representing the conditional distribution of the outcome. If the outcome variable is categorical, Classification Tree would be adopted; on the other hand, Regression Tree would be implemented if the outcome variable is continuous. For its application in MICE, we would use CART to derive the conditional distribution for each on the completely observed variables in step 1 and each given in step 2 (Burgette & Reiter, 2010) (Doove, Van Buuren, & Dusseldorp, 2014).

Though one of the disadvantages of the CART method is its difficulty for model interpretations when the number of tree level is high, it should not be our major concern because the goal for adopting CART method is to plausibly impute the missing data. Indeed, the application of non-parametric CART models in MICE can result in “more reliable inferences compared with naive applications of MICE” (Burgette & Reiter, 2010).

## Improving Computational Efficiency

The CART method is computationally efficient if predictors to be split are continuous variables. Specifically, continuous data will be sorted in ascending order before partitions, which reduces the computational complexity. On the other hand, if predictors to be split are categorical variables, the CART method might encounter computational difficulties when the variable has multiple levels. If a categorical variable has levels, the CART method would examine every possible splits, which results to possible partitions. Indeed, variable *Grade* has more than 20 levels, which means there are more than million possible partitions.

One solution to increase the computational efficiency of the CART imputation method is to reduce the number of levels for categorical variables. For example, we reduced *Pay Plan* from 173 categories into 7 categories. The OPM defines pay plan as “a two-digit alphabetical code used to identify Federal civilian pay systems” (Office of Personal Management (OPM), n.d.). The most common pay plan in the OPM data is the General Schedule pay system, which covers around 78 percent of white-collar Federal employees. Other pay plans cover employees who have unique occupations or serve for particular agency. For example, the *AL* pay plan applies to administrative law judges, and *SV* refers to pay plan in the Transportation Security Administration. Because some pay plan codes are only applicable to relatively small subset of population, we decide to merge pay plans with small sample size into “others” category. The table below is the simplified pay plan after merging.

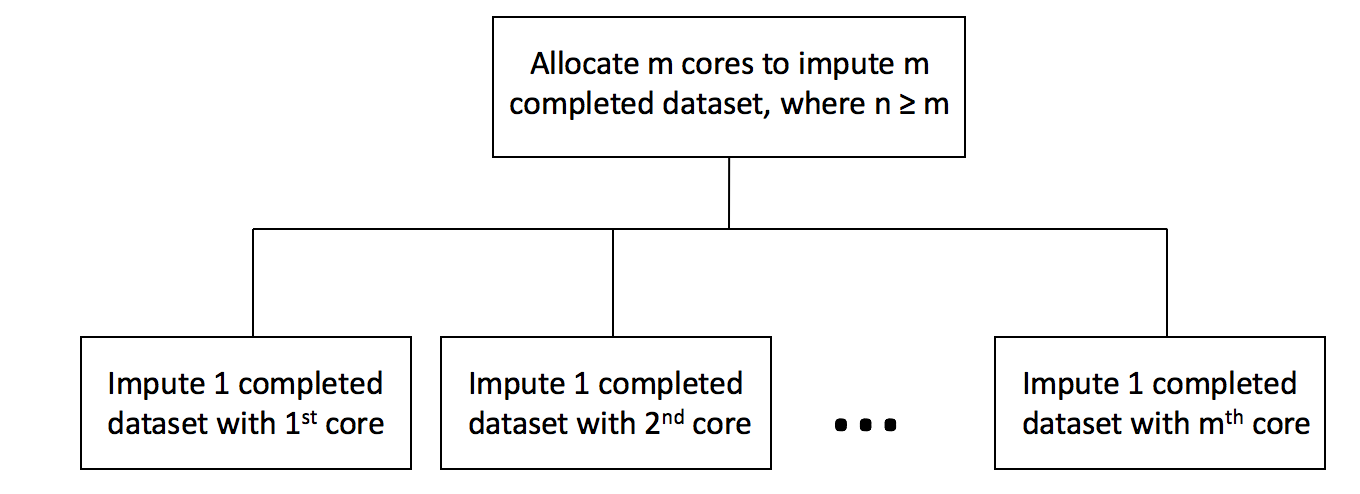
*Simplified Pay Plan*

|  |  |
| --- | --- |
| **Type** | **Codes** |
| General Schedule | GS, GM |
| Non General Schedule | AD, ES, SV, VN, Others |

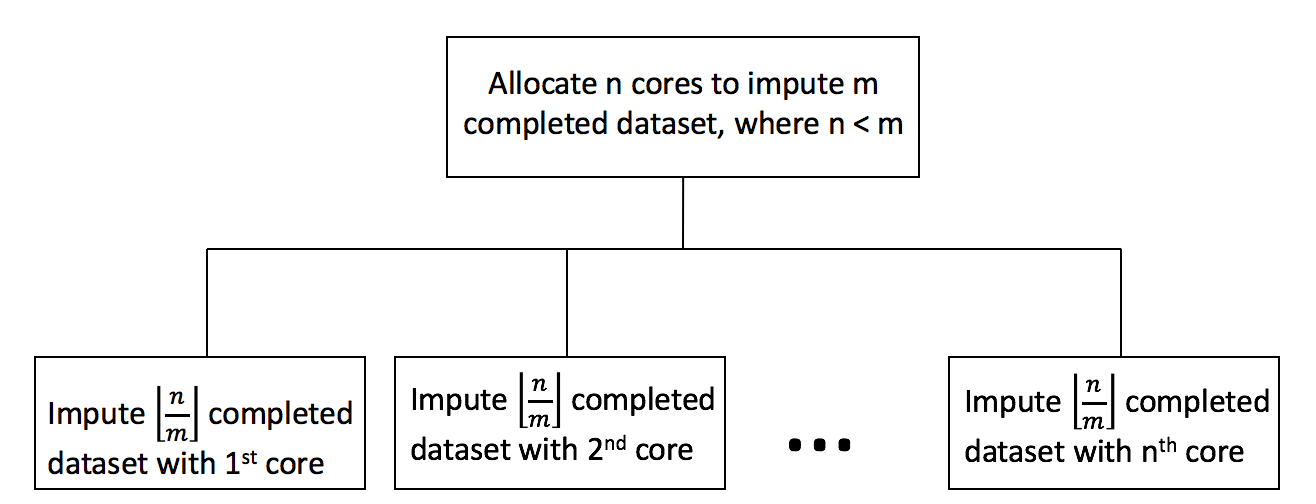
Another strategy used to increase the computational efficiency is through applying parallel computing technique on *Step 4* of the MICE. The implementation of the parallel computing includes the following steps.

Here, we assume we want to impute completed datasets with cores.

**If ()**  
allocate cores  
conduct *Step 1* to *Step 3* once for each core in parallel



**else**  
allocate cores  
conduct *Step 1* to *Step 3* times for each core in parallel



## Imputation Results and Predictive Checks

We generated 12 completed datasets with the CART imputation model. The validity of the imputed datasets depends on the use of an appropriate imputation model, so it is important to check whether the model yields reasonable results. We applied predictive checks to assess imputation model adequacy. Predictive checks is a Bayesian model checking technique designed to investigate the potential model inadequacy between imputed and replicated datasets. Though our imputation model is not fully Bayesian, predictive checks could still be used to measure the predictive differences between imputed and replicated data (He & Zaslavsky, 2012).

Denote , where is partially observed and is completely observed. To generate replicated data , the steps involve: (1) create duplicates of ; (2) set duplicated as completely missing; (3) combine and to form concatenated dataset; and (4) re-impute concatenated dataset with original imputation model. Figure 1 shows the re-imputation process (He & Zaslavsky, 2012). Since our parameter of interest is the relationship between year 1 and year 10 salary, we compare the distribution of these variables between imputed and replicated datasets. If the distribution of these two variables are similar between imputed and replicated datasets, then we can conclude that the imputation model provides good fit to the data.

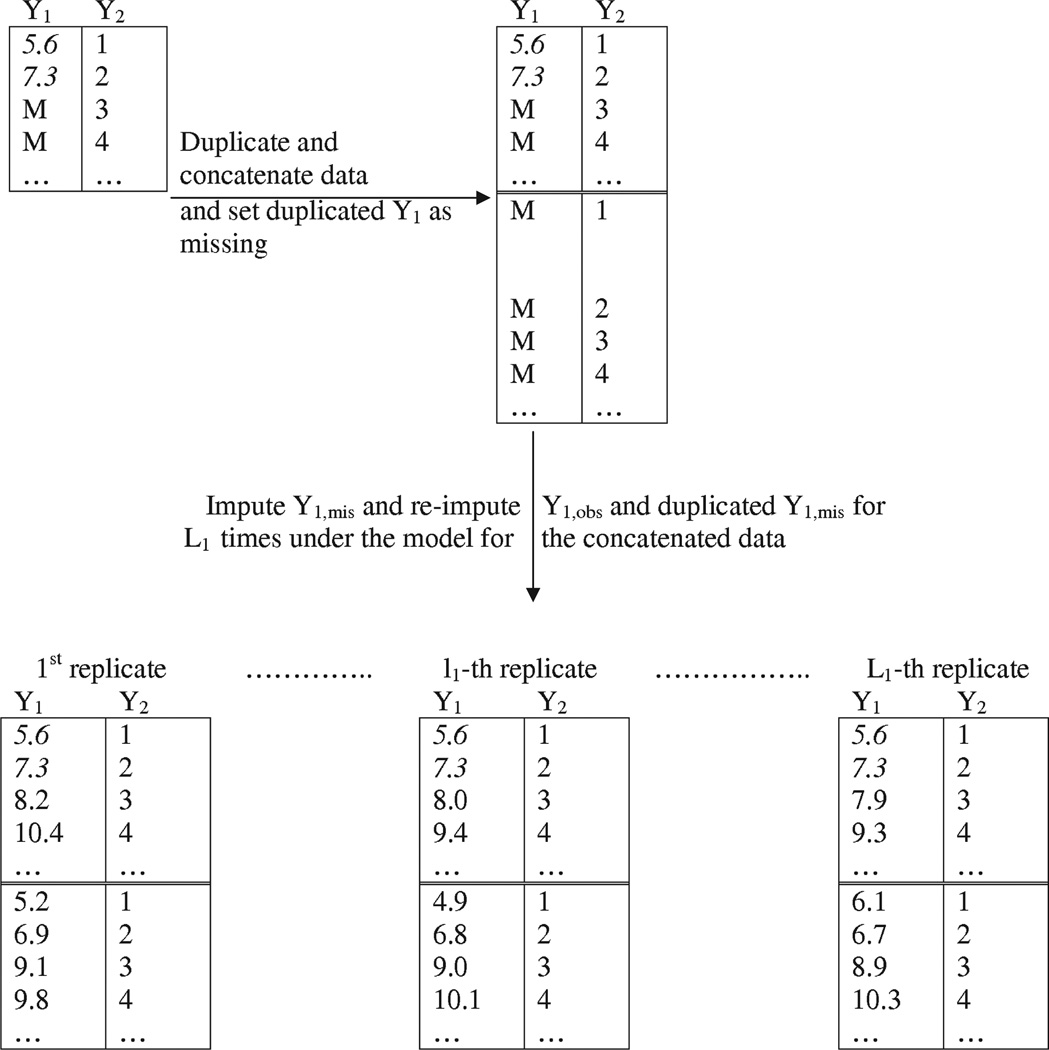


Figure 1 Generate replicated dataset

Two different methods are used to generate replicated dataset. First, conduct re-imputation by setting duplicated (variables with missing values) as completely missing. Second, conduct re-imputation by only setting salary at working year 1 and 10 in duplicated as completely missing. The distribution between imputed and replicated dataset is shown by Figure 2 for the first method, and Figure 3 for the second method. The first method indicates that our imputation model is not a good fit to the data. However, this method might not be ideal because the imputed values of year 1 and year 10 salary are highly dependent on other partially missing variables. The second method provides a better model diagnostic. The distribution between imputed and replicated datasets are similar, which implies that our imputation model provides plausible fit to the data.

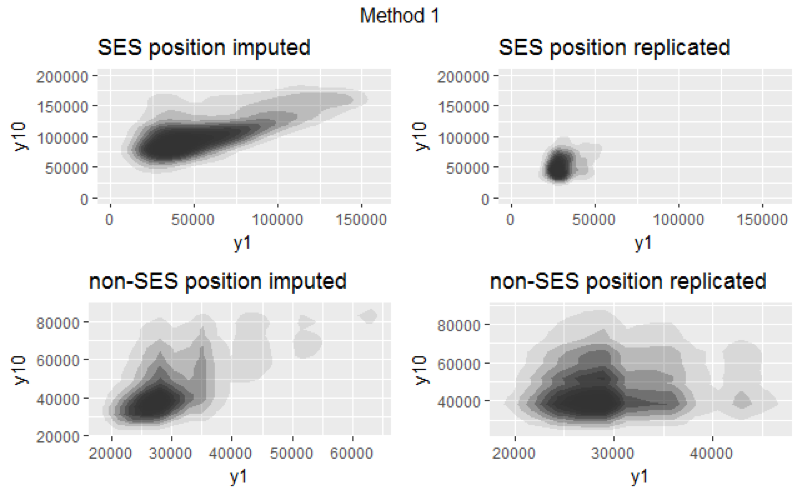


Figure 2 Method 1

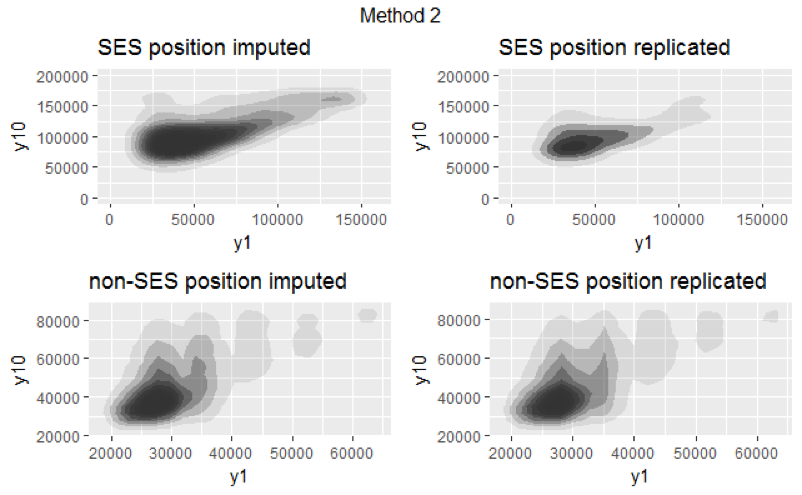


Figure 3 Method 2

# This chunk ensures that the thesisdowndss package is

Placeholder

## EDA

# This chunk ensures that the thesisdowndss package is

Placeholder

## Frequentist - Maximum Likelihood Approach

## Bayesian - Posterior Distribution

# This chunk ensures that the thesisdowndss package is

Placeholder

# References

Placeholder

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070–1076.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330.

Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, *72*, 92–104.

He, Y., & Zaslavsky, A. M. (2012). Diagnosing imputation models by applying target analyses to posterior replicates of completed data. *Statistics in Medicine*, *31*(1), 1–18.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Office of Personal Management (OPM). (n.d.). Pay & leave. Retrieved from <https://www.opm.gov/policy-data-oversight/pay-leave/pay-administration/fact-sheets/pay-plans/>