Modeling of Missing Data

Focusing on Confidential Social Science Data

Jerry Chia-Rui Chang, Advisor: Jerome P. Reiter

Motivation & Goal

◣ The senior executive services (SES) program was established by the Office of Personnel Management (OPM) to select high-level executives within the federal government

◣ These executives are considered as "the backbone of Federal executive leadership" and are required to have leadership skills to lead strategic changes and achieve organizational goals

◣ Concerns have been raised on the effectiveness of the SES program. (e.g. lack of diversitity within the program, selection bias……)

◣ To understand what factors influence the promotion of the SES positions in terms of gender, race, and more through analyzing the federal government employee data .

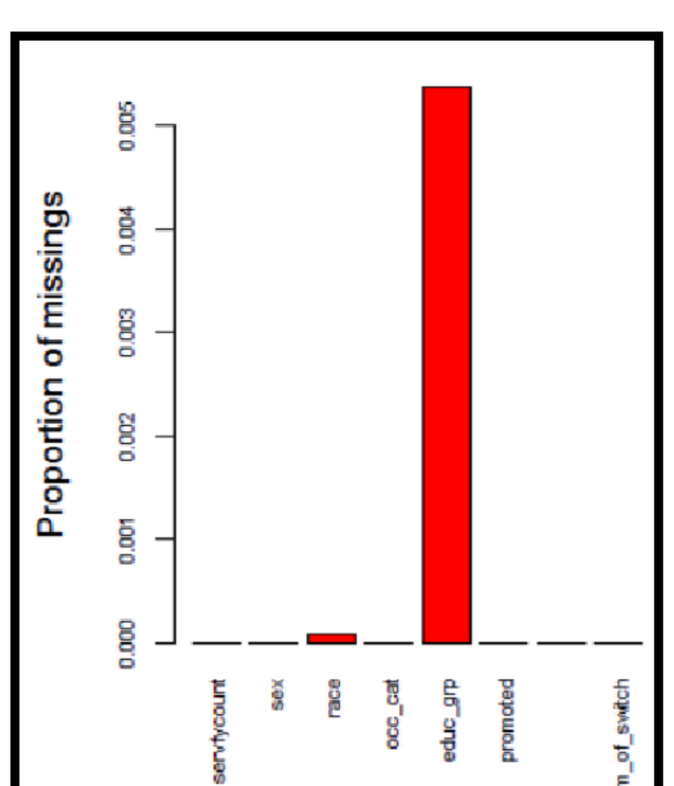
Missing Data

◣ Missing Mechanisms

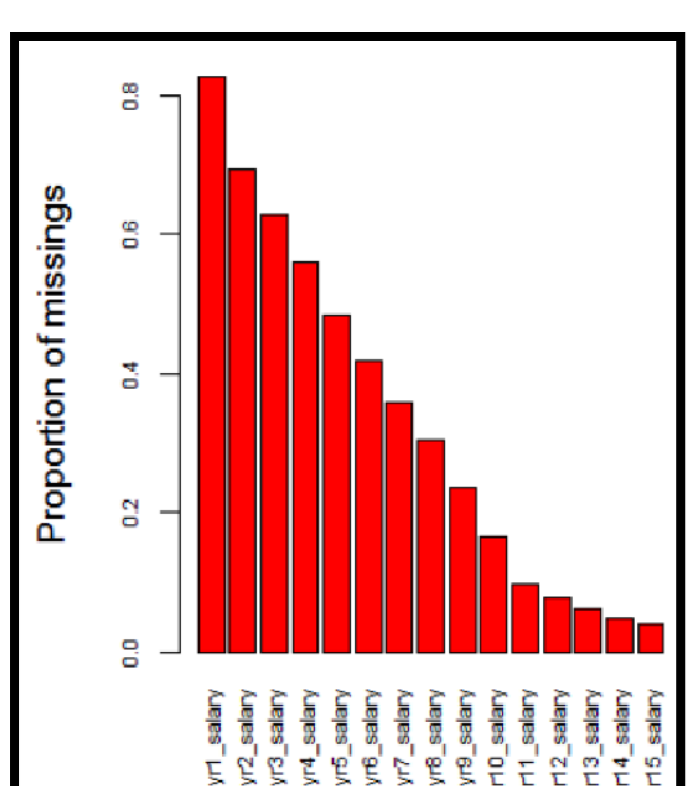
(1) Missing Completely at Random (**MCAR**): probability of observations being missing is unrelated to other subjects in the study
(2) Missing at Random (**MAR**): probability of missing only depends on observed values but not on unobserved values
(3) Not Missing at Random (**NMAR**): probability of missing depends on both observed and unobserved values.

◣ Two types of missingness in the OPM data

(1) Inherent Missingness: Race & Education Level
(2) Missingness due to Time Constraint: Pay Plan, Grade, Step Rate, Salary



Inherent Missingness



Missingness due to Time Constraint

Methodology

Multivariate Imputation by Chained Equations (MICE) with application of CART algorithm

◣ General Approach for MICE

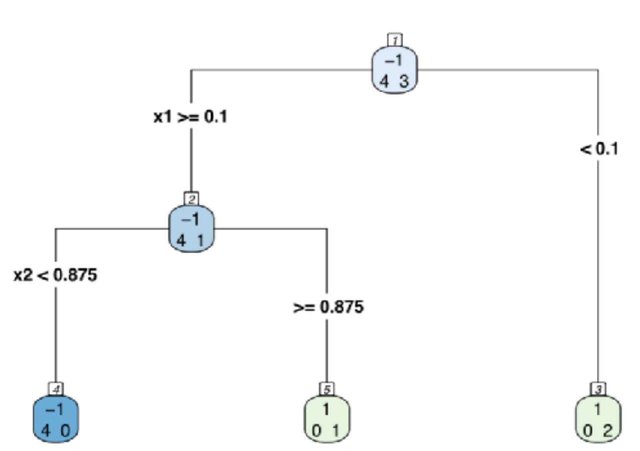
(1) Fill in the missing columns through drawing values from predictive conditional distribution to produce m complete datasets
(2) For each complete dataset, conduct analysis for parameters of interest
(3) Combine individual analysis to form final results

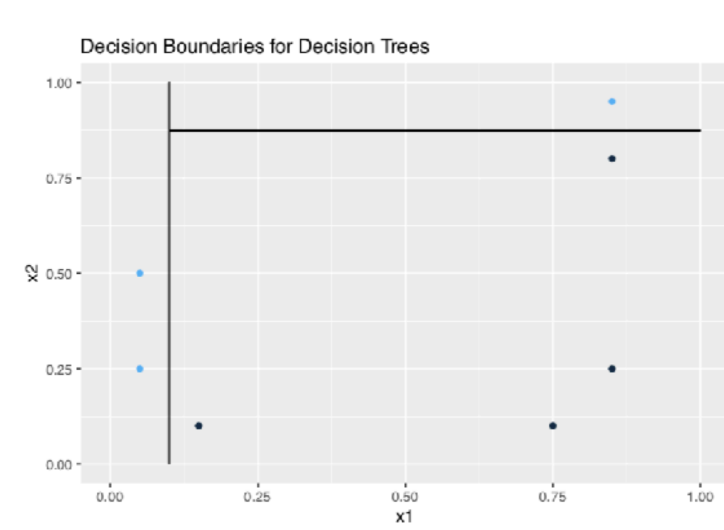
◣ Specifying Conditional Distribution - CART (Classification and Regression Tree)

◣ The CART algorithm performs binary splits of the predictors recursively to approximate the conditional distribution of a univariate outcome

◣ The partitions are found if the subsets of units have relatively homogeneous outcomes (Measured by Reduction in Gini Index)

◣ CART is a more flexible non-parametric modeling approach compared to standard generalize linear models (GLMs)

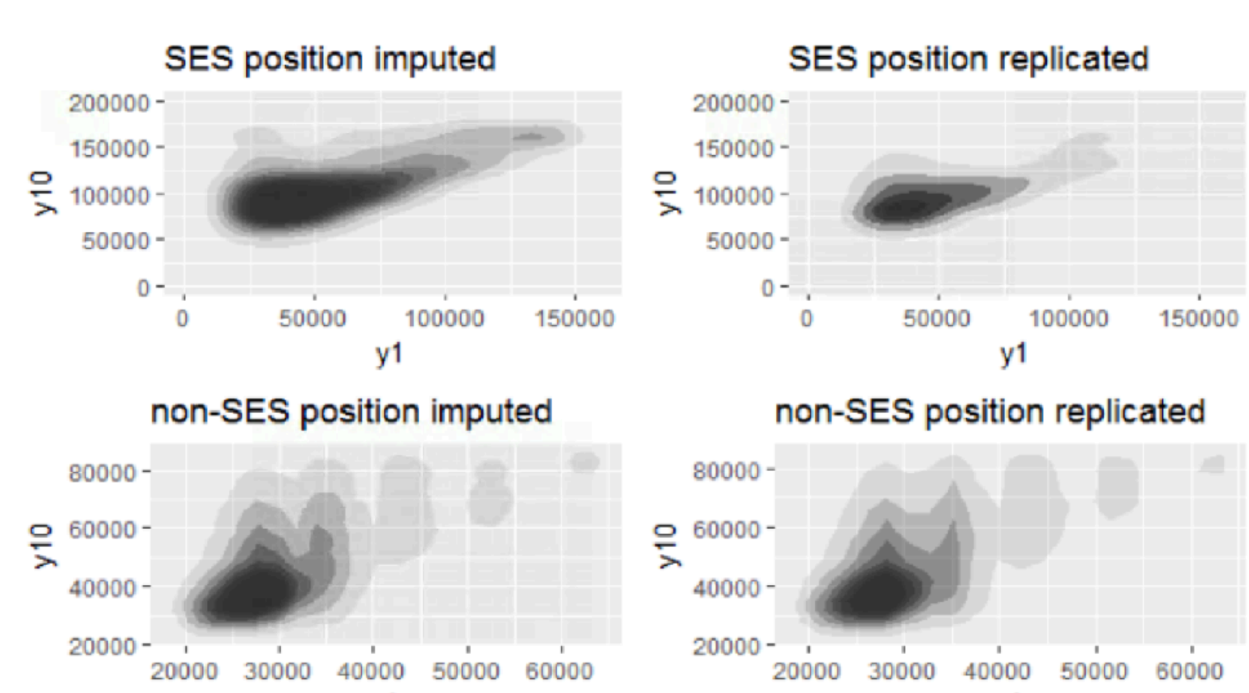




◣ Posterior Predictive Check (PPC)

◣ Check the robustness of the imputation model through re-imputation

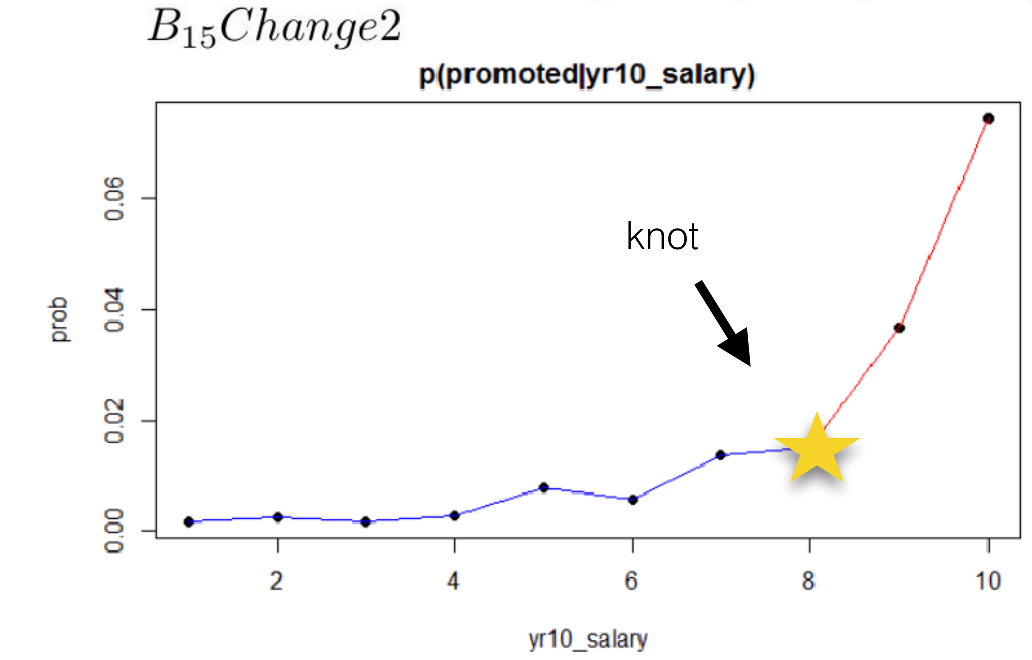
◣ Since our parameter of interest is the relationship between year 1 and year 10 salary, we compare the distribution between imputed datasets and replicated datasets.



Modeling of Complete Datasets

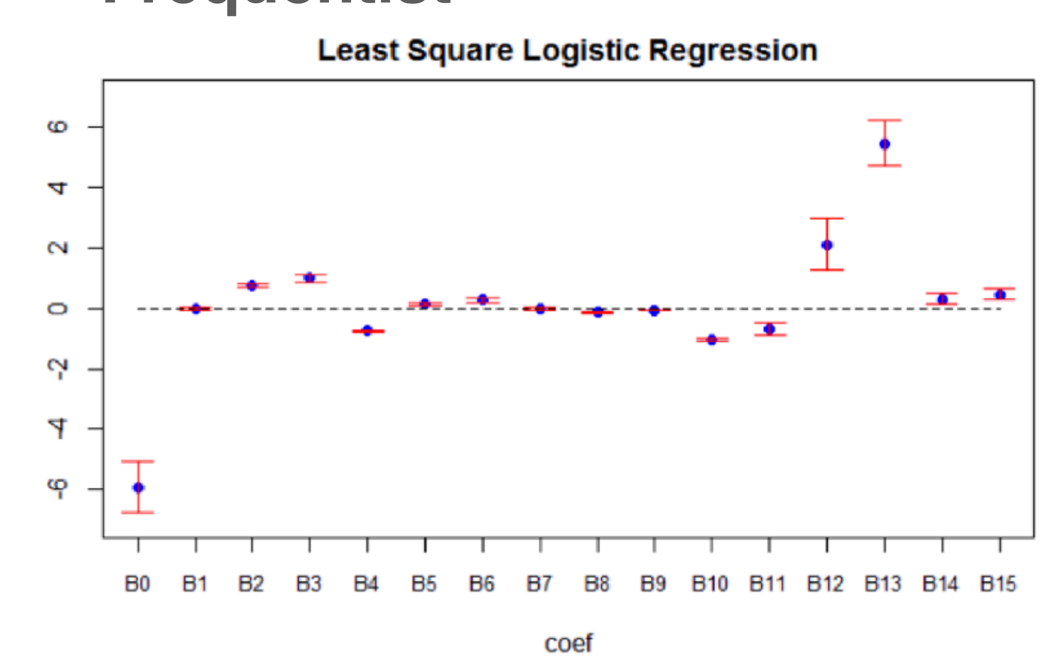
Logistic Regression with bspline application

$$logit(p) = B_0 + B_1 Female + B_2 OccA + B_3 OccO + B_4 EducA + B_5 EducC + B_6 EducD + B_7 EducE + B_8 NotWhite + B_9 NoSwitch + B_{10} NotSupervisor + B_{11} Grade_0 + B_{12} Salary + B_{13}(Salary - Knot) * I[Salary \geq Knot] + B_{14} Change1 + B_{15} Change2$$

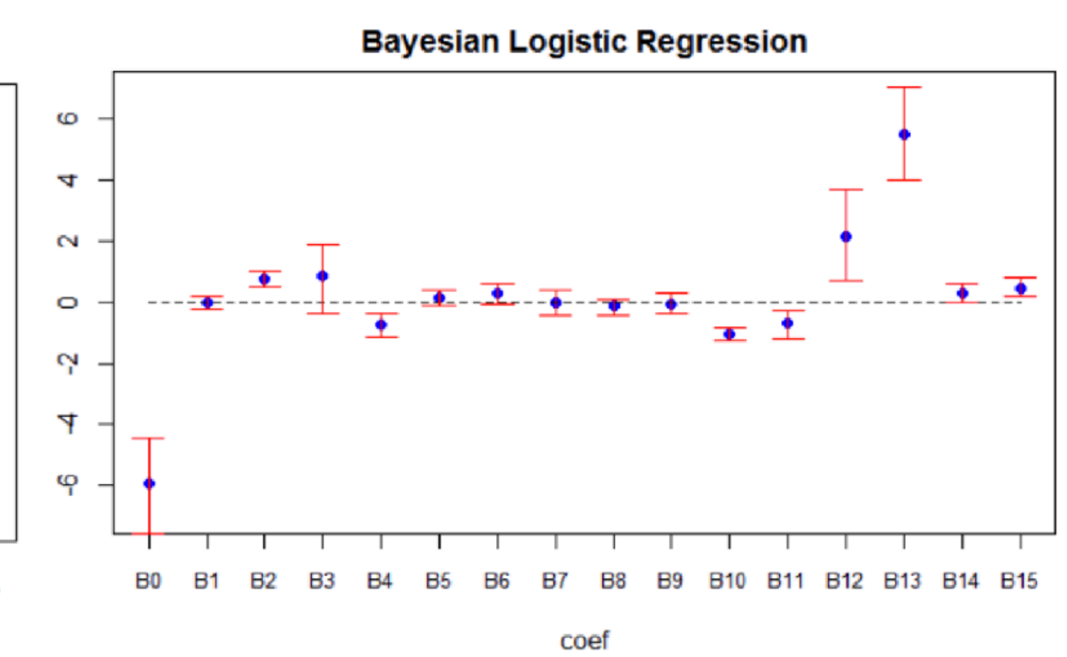


The relationship between the probability of promoted and year 10 salary is non-linear. To prevent underfitting, linear basis spline is applied. The knot is manually selected based on visualization.

Frequentist



Bayesian



Bayesian - Posterior Distribution

$$f(\beta_i) = N(0, 5), i = 0, ..., 15$$

$$f(y|\beta) = \binom{n}{y} logit^{-1}(\eta)^y (1 - logit^{-1}(\eta))^{n-y}$$

$$f(\beta|y, X) \propto f(\beta_0) \prod_{k=1}^{15} f(\beta_k) \prod_{i=1}^{N} logit^{-1}(\eta_i)^{y_i} (1 - logit^{-1}(\eta_i))^{n_i - y_i}$$

Combined Analysis

◣ Frequentist

◣ Parameter estimates through Maximum Likelihood approach
◣ Variable importance evaluating by p-value
◣ Combine m datasets results through averaging

◣ Bayesian

◣ Parameter estimates through Draws from Posterior Distribution
◣ Prior would affect the sensitivity of the posterior distribution
◣ Combine m datasets results through mixtrue draws from MCMC posterior outputs