Modeling Federal Employees Data to Predict Promotion

Chia-Rui (Jerry) Chang

May 2018

# Abstract

The Senior Executive Service (SES) is a key leadership position within the Federal government. Federal employee with the SES position serves as the primary link between top political appointees and the rest of the Federal workforce. This study aims to understand factors that impact the SES promotion outcome, where we built a predictive model to predict promotion based on federal employees’ personal and career information. Due to recording error and time constraints, multiple variables have missing data, with some over 80% missing. To address potential data analysis challenges, we adopted multiple imputation technique to impute missing data. First, we generated 12 completed datasets through multivariate imputation by chained equations, with Classification and Regression Tree specified as the conditional distribution. Predictive check was then used to assess the robustness of imputation results. Second, we experimented various statistical models based on one completed dataset, where piecewise linear regression model provided appropriate fit to the data. Lastly, we estimated model coefficients with two approaches - maximum likelihood estimates and Bayesian posterior distribution. We then combined each completed dataset’s coefficient estimates to conduct inference.

# Introduction

The Senior Executive Service (SES) was established by the Civil Service Reform Act (CSRA) in 1978. The introduction of the SES program is to select and develop high-level executives corps with exceptional management skills. Granted with high authority, members of the SES not only ensure productivity and efficiency within the government but also provide leadership to agencies across administration. Each executive holds key position below the top political appointees, and serves as primary link between these appointees and government career workers (“Guide to the senior executive service,” 2017). Before the existence of the SES, the federal government was a fragmented bureaucracy. The creation of these centralized senior leaders therefore brought a “measure of coherence” to fulfil the larger corporate interests of the federal government (Huddleston & Boyer, 1996). With more than thirty years of establishment, the SES became “the backbone of Federal executive leadership”, where its members “play a crucial role in addressing unprecedented challenges facing our nation” (“Statement of Jeffrey D. Zients, U.S. Congress, Senate Committee on Homeland Security and Governmental Affairs, Subcommittee on Oversight of Government Management, the Federal Workforce, and the District of Columbia,” 2011).

The Office of Personnel Management (OPM), also established by the CSRA, is tasked to manage the overall SES program. There are four types of appointments in the SES. The first type is career appointment, where incumbents are selected through merit-based staffing and usually hold permanent position within the government. The other three types are either non-career appointment or appointment with limited terms. Since the majority of the SES positions are career appointments, the primary interest of this study focuses on career-track SES.

The selection of career-track SES is based on OPM’s Executive Core Qualifications (ECQ), which specifies a candidate’s competencies to build successful teams and bring out strategic integration within and outside the organization. Specifically, the OPM identifies various critical leadership skills for executives to succeed. For example, executives need to have the abilities to lead strategic changes and achieve organizational goals with high-quality results. Qualifications Review Boards (QRBs), consisting of three SES members from distinct agencies, further reviewed the approval of a candidate.

Tthe recruiting of the SES position consists of multiple phases, and we are interested in understanding factors that affect promotion outcomes. For example, promotion bias towards people with technical expertise existed in the early establishment of the SES program. Furthermore, some of the outside observers have raised concern about the diversity of the SES. As a result, various studies have been conducted on the current SES system. For example, Powell and Butterfield investigated the impact of gender on the SES promotion outcomes (Powell & Butterfield, 1994). Shafritz argued that the quality of the SES can be improved with more racial, ethnic and gender diversity (Carey, 2011). The motivation behind this project, therefore, is to build a statistical model in predicting promotion outcomes of the SES positions based on federal employees’ personal and career information.

# Data

The data of this study comes from the Office of Personal Management (OPM), consisting of the federal employees data from 1988 to 2011. It contains not only an employee’s personal information such as gender and race but also career-related information such as years of services, pay plan, and salary. The study used 10% samples of the OPM data, with 36,751 total employees.

*Personal Information*

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Short Description** | **Type** |
| Sex | gender of an employee | categorical |
| Race | race or national origin of an employee | categorical |
| Education | educational degree an employee obtained | categorical |

*Career-related Information*

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Short Description** | **Type** |
| Start Year | fiscal year an employee started working | continuous |
| Service Fiscal Year Count | number of years an employee has worked in the government excluding any break in services | continuous |
| Department Switch | number of times an employee switched department | continuous |
| Occupational Category | occupational category based on the nature of the work | categorical |
| Pay Plan | a two-dimensional matrix of basic pay rates for certain employees prescribed by law or other authoritative sources | categorical |
| Grade | hierarchical relationships among positions covered by the same pay plan or system | categorical |
| Step Rate | an indicator of salary range within certain pay plan and grade | categorical |
| Pay Basic | an employee’s basic pay rate adjusted for inflation based on 2011 standard | continuous |
| Promoted | an indicator variable of whether an employee has ever been promoted into the Senior Executive Service (SES) position | categorical |

The underlying goal of this study is to better understand the SES promotion outcomes without political sponsorship. Currently, the SES positions account for less than 1 percent of all positions in the government. With such low proportion, it is important to identify potential candidates before conducting statistical modeling and inference.

First, we defined our population to be career-track employees who started working in the federal government between 1978 to 1997 and had at least 15 years of working experiences. The reasons to choose this time period are twofold. First, it took around 22 years on average for a federal employee to be promoted in to the SES position (Powell & Butterfield, 1994). Because 2011 is the last year recorded in the OPM data, employees starting after 1997 would have less than 15 years of working experiences and are unlikely to be promoted within such a short period of time. We found it appropriate to remove them. Second, the OPM database was not established until 1988, so obtaining information of employees starting in the 1950s or 1960s periods would be challenging. Furthermore, the characteristics of employees in the past were much different from characteristics of employees nowadays. For example, the attainment of higher education and the proportion of women working are significantly higher nowadays. Given such constraints, we excluded employees who started before 1978.

Second, though it usually took more than 10 years for a federal employee to get promoted into the SES position, we recognized that very few employees achieved the SES position with less than 10 years of working experiences. Since we were more interested in understanding the long-term career promotion, we excluded employees who were promoted into the SES positions with less than 10 years of working experiences.

Third, we observed that those who got promoted usually made significant progress in their mid-career, which can be illustrated by features such as salary and grade achieved at working year 10. An employee’s grade demonstrates the hierarchical relationships among positions covered by the same pay plan or system. For example, the grades for General Pay Schedule (GS and GM pay plans) range from 1 to 15, with 15 being the highest grade. Since both pay plans account for approximately 75% of the employees in the data, we can gain some insights from the grade distribution of employees with GS or GM pay plan at working year 10. Specifically, most GS or GM pay plan employees who got promoted achieved grade 13 and above at working year 10, with few between grade 9 to 12, and none below grade 8. Based on this information, we believed employees with GS or GM pay plan had limited chance for promotion if they did not achieve grade 9 and above at working year 10; therefore, we found it suitable to exclude GS/GM pay plan employees whose grade were below 9 at working year 10.

*Grade achieved at yr10 for employees who were later promoted into the SES position with GS or GM pay plan*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Grade 1-8** | **Grade 9** | **Grade 10** | **Grade 11** | **Grade 12** | **Grade 13** | **Grade 14** | **Grade 15** |
| 0 | 5 | 2 | 16 | 50 | 96 | 126 | 146 |

Lastly, the federal government classifies each employee’s occupation into five occupational categories based on the nature of the work. Respectively, they are *Blue collar*, *Professional*, *Administrative*, *Technical*, *Clerical*, and *Other white collar*. The table below shows the detailed descriptions of each occupational category. We noticed that none of the SES positions contained *Blue collar* occupational category and less than one percent contained *Technical* and *Clerical*, we removed employees in these occupational categories.

*Occupational Categories*

|  |  |
| --- | --- |
| **Occupational Category** | **Description** |
| Blue collar | Occupations comprising the trades, crafts, and manual labor |
| Professional | White collar occupations that require knowledge in a field of science or learning characteristically acquired through education or training equivalent to a bachelor’s or higher degree with major study in or pertinent to the specialized field |
| Administrative | White collar occupations that involve the exercise of analytical ability, judgment, discretion, and personal responsibility, and the application of a substantial body of knowledge of principles, concepts, and practices applicable to one of more fields of administration or management |
| Technical | White collar occupations that involve work typically associated with and supportive of a professional or administrative field |
| Clerical | White collar occupations that involve structured work in support of office, business, or fiscal operations |
| Other white collar | White collar occupations that cannot be related to the above professional, administrative, technical, or clerical categories |

# Missing Data & Multiple Imputation

## Missing data

In social science research, survey item-nonresponse and data recording errors often lead to missing data. The OPM data is no exception. Rubin classified missing data into three properties: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Little & Rubin, 2014). First, data has MCAR property when the probability of observations being missing is unrelated to any variables. Second, data has MAR property when the probability of missing only depends on observed values but not on unobserved values. Though there are no formal ways to verify the MAR property, it is usually a reasonable assumption when we do not have comprehensive knowledge of the data. Third, data that cannot be attributed to either MCAR or MAR has NMAR property - the probability of missing depends on both observed and unobserved values.

We examined variables with missing data to assess the OPM data’s missing mechanism. In general, the missingness of the OPM data can be summarized into two categories - inherent missingness and missingness due to time constraint.

First, the OPM data contain more than 20 years of federal employees’ information. As rules and standards change overtime, recording errors would naturally lead to missing data. For example, *Race* variable was originally classified into 16 categories, but new standard established in 2016 simplified it into 6 categories. The main strategy to address the missingness of *Race* variable is to convert old recording standard to new standard. However, small percentage (less than 0.01%) of data is still missing even after conversion. *Education Level* is another variable that has missing data. The fact that employees might not fill in and update their personal information could lead to the missingness of *Education Level* (Figure 1).

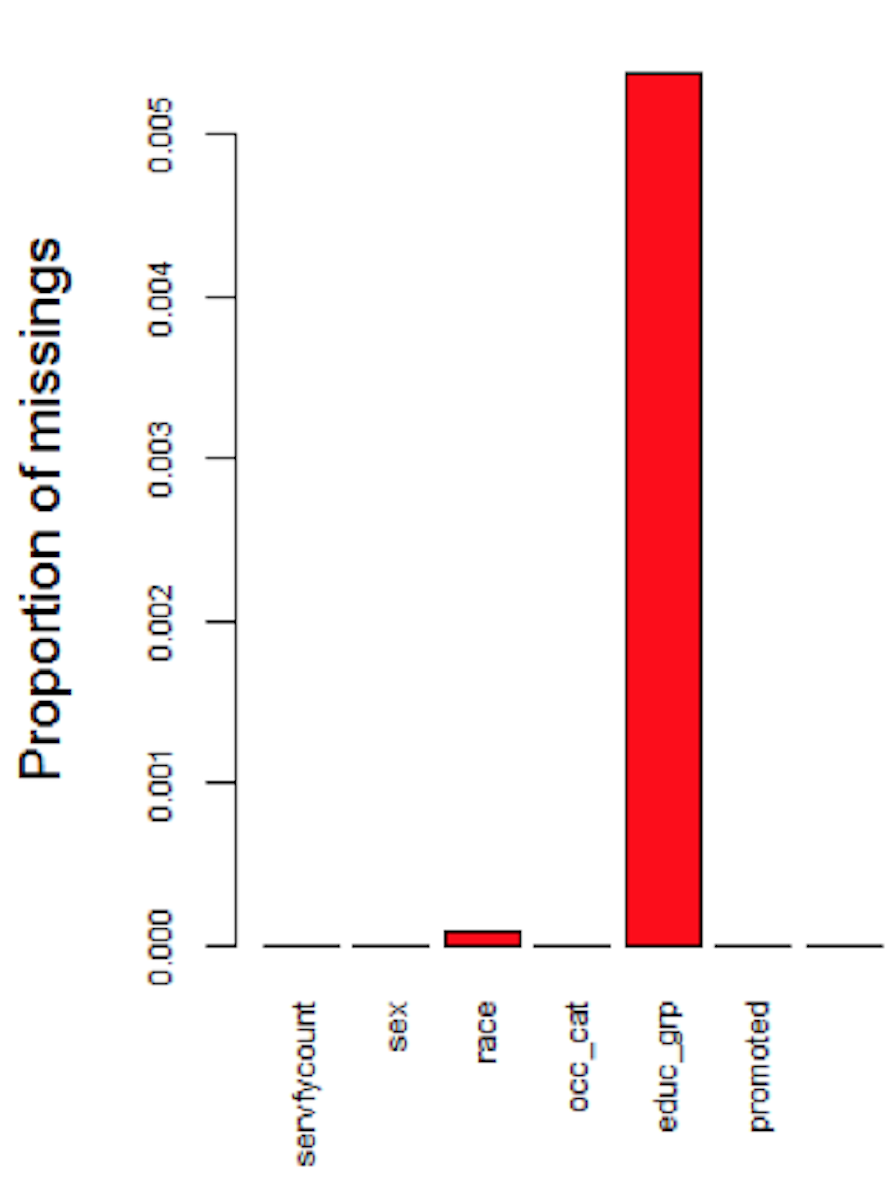


Figure 1 Inherent Missingness Independent of Time

Second, the OPM database was not created until 1988, so information regarding employees started before 1988 would not be recorded. For example, if an employee worked between 1980 and 1990, the OPM data would only record this employee’s information from 1988 to 1990, but data from 1980 to 1987 would be missing. Variables that change overtime belong to this category. As mentioned previously, a Federal employee’s salary is determined by particular grade, step rate and pay plan. Since *Pay plan*, *Grade*, *Step Rate*, and *Salary* are associated, these four variables have the same missing patterns. Specifically, the proportion of missing data decreases as year increases (Figure 2).

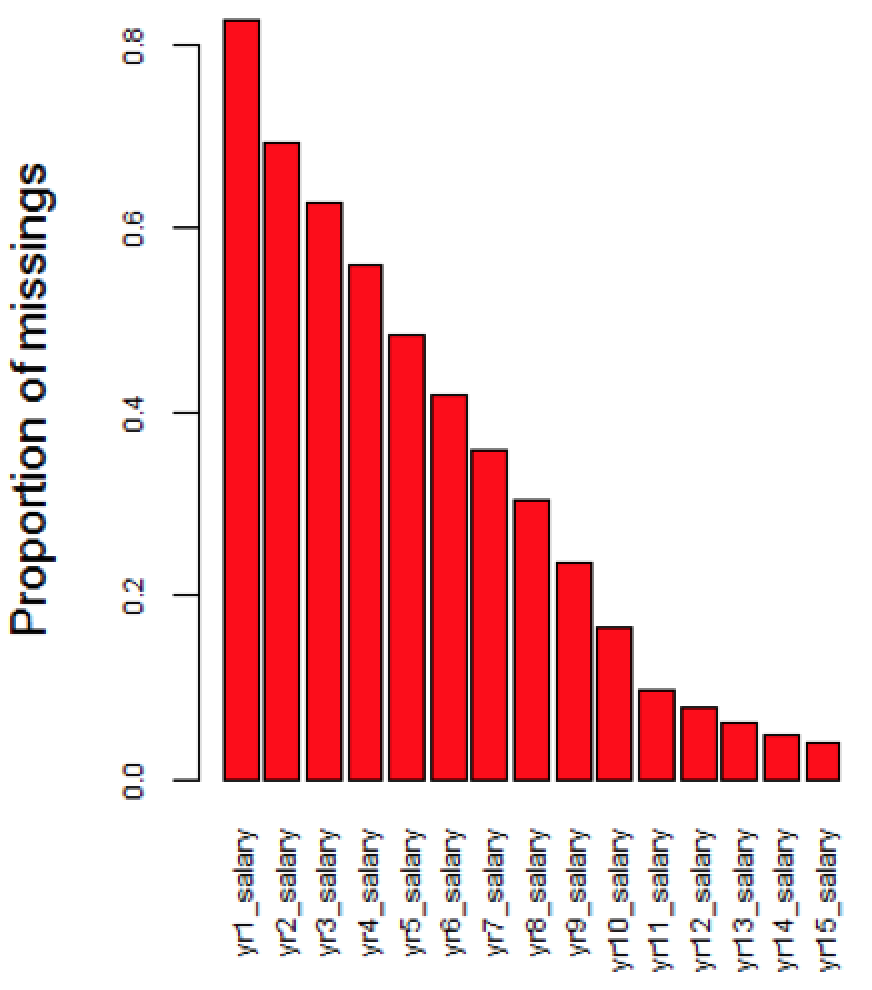


Figure 2 Missingness due to Time Constraint

We assumed the data has MAR property for two reasons. First, for inherent missingness, the proportion of missingness is small. Second, for time constraint missingness, we believed relationships for variables before 1988 and after 1988 are conditionally dependent. For example, salary information before 1988 is dependent on observed education levels, pay plan, and grade after 1988. The loss of information from missing data would produce bias and impact the robustness of statistical modeling and inference. We imputed the missing data to address this problem.

## 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 involves: (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 implementation of MICE involves 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 any “auxiliary variables” that are related to the missingness but not part of the covariates. The inclusion of “auxiliary variables” 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 parameters of interest to the final analysis. However, both variables provide information on the missingness of *Salary* and *Grade*, so we included them in the imputation model. Another auxiliary variable we included 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 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. 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. In the OPM data, 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.

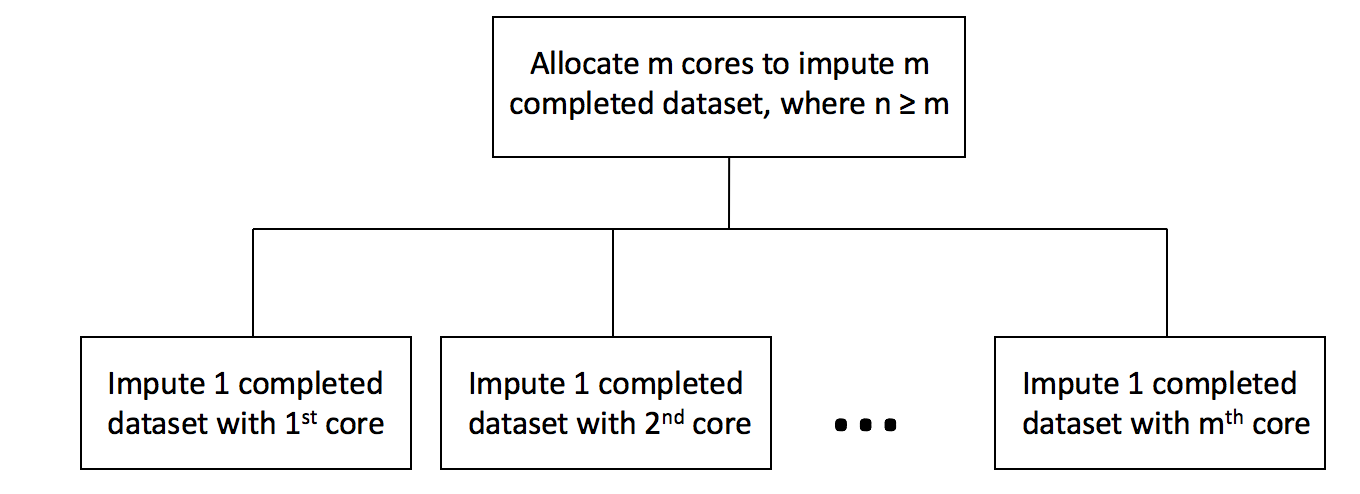
*Simplified Pay Plan*

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

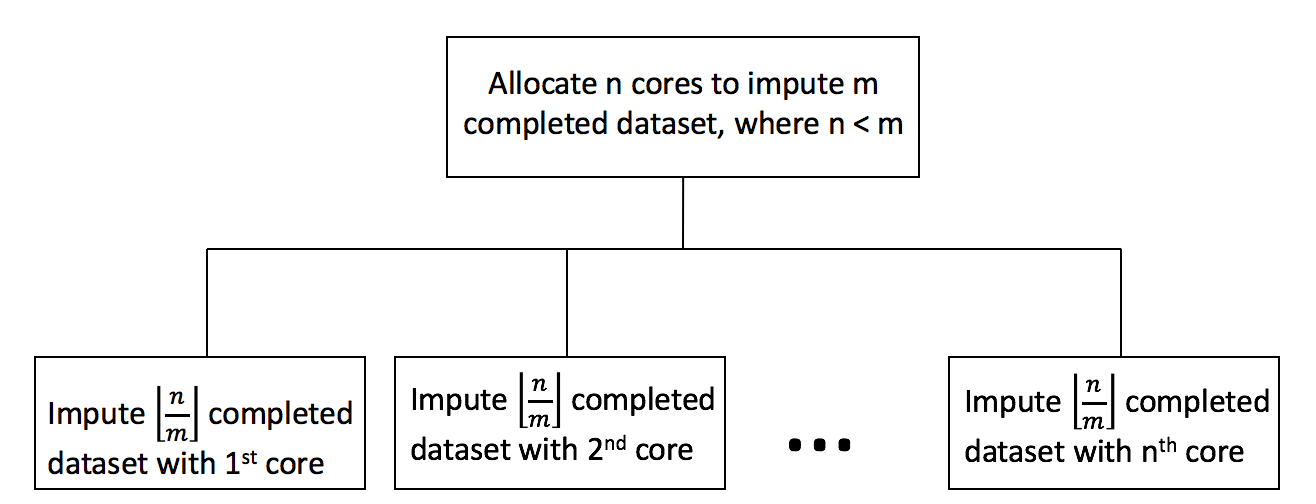
Another strategy used to increase the computational efficiency is through applying parallel computing technique on *Step 4* of the MICE. The implementation of the parallel computing includes the following steps.

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

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



**else**  
allocate cores  
conduct *Step 1* to *Step 3* of MICE 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 3 (He & Zaslavsky, 2012) shows the re-imputation process.

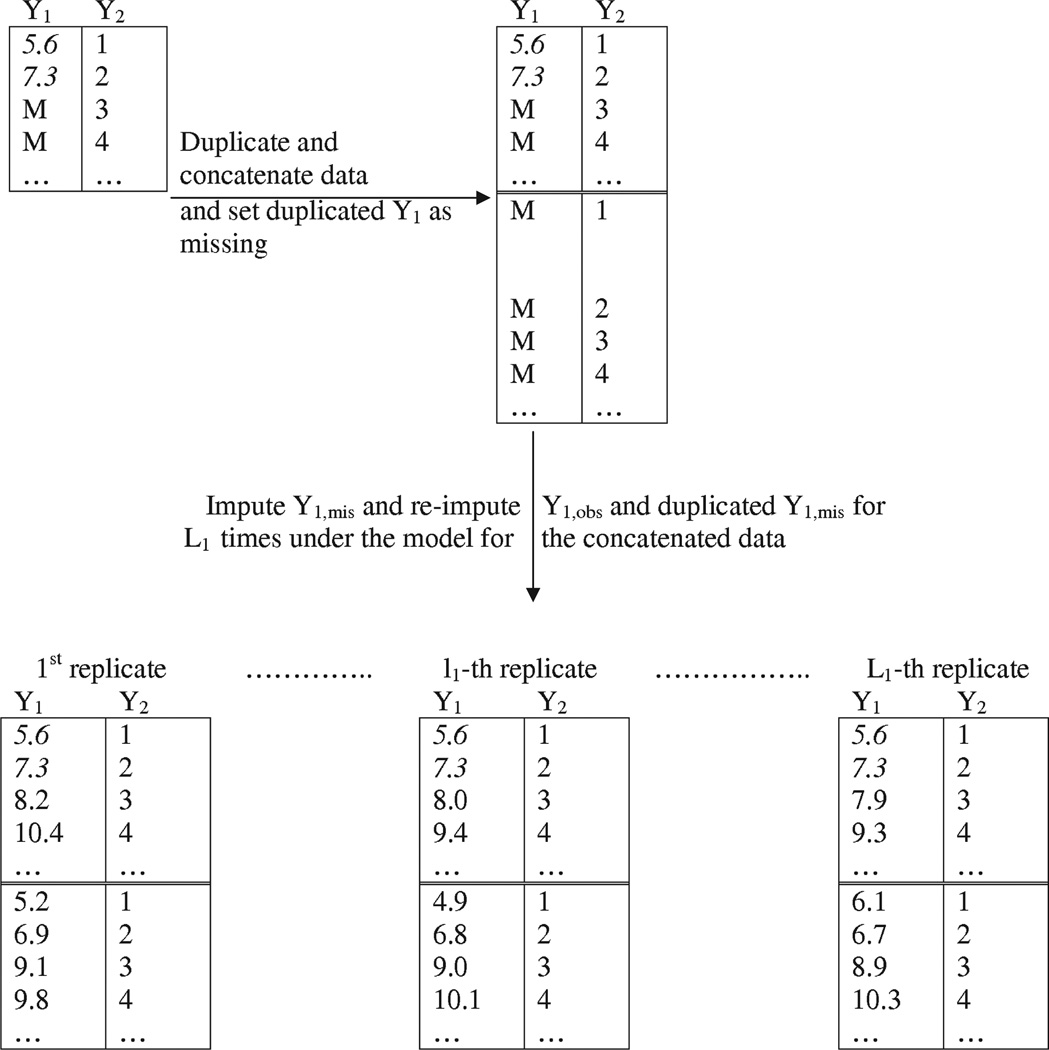


Figure 3 Generate replicated dataset

Our parameter of interest is the relationship between year 1 and year 10 salary. We assessed model adequecy through comparing the distribution of working year 1 and year 10 salary between imputed and replicated datasets. If the distribution of imputed datasets and the distribution of replicated datasets are similar, then imputation model provides plausible fit to the data. We generated replicated dataset with two methods. First, we conducted re-imputation 12 times by setting duplicated (variables with missing values) as completely missing. Second, we conducted re-imputation 12 times by only setting salary at working year 1 and 10 in duplicated as completely missing. Figure 4 shows the result of the first method, and Figure 5 shows the result of the second method.

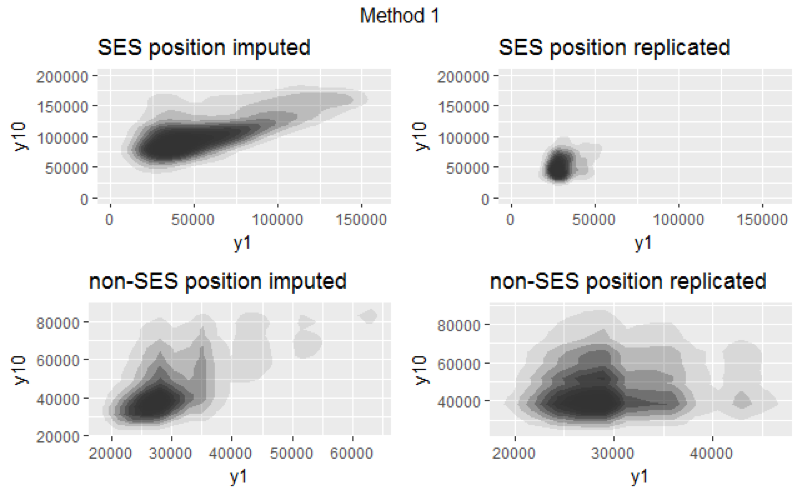


Figure 4 Salary distribution of method 1

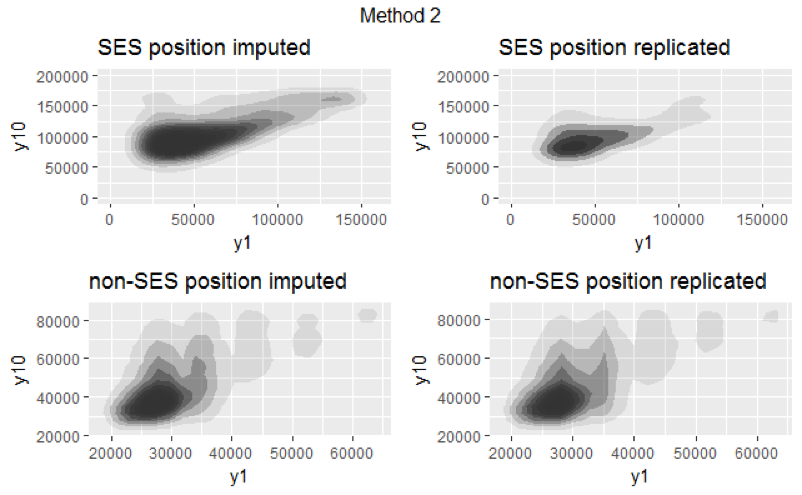


Figure 5 Salary distribution of method 2

For both figures, the x-axis represents salary at working year 1, and the y-axis represents salary at working year 10. Since salaries for employees with SES and non-SES position tend to vary a lot, we visualized the salary distribution separately for employees with SES position and non-SES position. Note that both figures are visualization of one replicated and one imputed datasets. (We compared all 12 replicated and 12 imputed datasets. Since they all yielded similiar results, we only included one here.)

Figure 4 shows that the distributions between imputed and replicated datasets for both positions are different, indicating 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 dependent on other partially missing variables. The second method provides a better model diagnostic. Figure 5 shows that the distributions between imputed and replicated datasets for both positions are similar, implying that our imputation model provides plausible fit to the data.

# Exploratory Data Analysis (EDA)

After checking the robustness of the imputation model, we conducted exploratory data analysis and model fitting based on one imputed dataset.

## Outcome variable and covariates selection

**Outcome Variable**

Variable *Promoted* is the outcome variable, indicating whether an employee was promoted into the SES position after working year 10 (1 = promoted, 0 = not promoted). In order for the analysis to be consistent, we only included employees with GS or GM pay plan. 482 out of 27,932 employees were promoted.

**Covariates**

The OPM data contain detailed information about the federal employees, but some variables are unrelated to the goal of the study. Based on past social science research on the selection of high-level executives within both private and public sectors, we logically chose variables relevant to our questions of interests as covariates.

First, *Gender* could be a predictor as past studies have shown that gender plays an important role in the promotion of high-level executives. For example, the glass ceiling phenomenon, the phenomenon that keeps women from reaching the top level of organizations, exists in both public and private sectors (Powell & Butterfield, 1994). Second, *Race* and *Education Level* are important predictors for federal employees’ salary information (Barrientos et al., 2017), and could potentially impact the promotion process. Both variables should be included as covariates.

Third, employees with the SES position are required to build coalitions both within and outside the organizations. We extrapolated that employees who had experiences working in various government agencies might possess this competency. *Department Switch*, a variable indicating whether an employee switched department between working year 1 and working year 10, could be a critical predictor. Fourth, senior executives are often tasked to lead people and make strategic changes, and variables such as *Occupational Category* and *Supervisory Status* might provide pertinent information.

Lastly, the promotion criteria depend on an employee’s performance at work. There is a formal appraisal system in the OPM, but such system was not established until 1996. Given the timeframe of the studies (employees starting between 1978 and 1997), it is not appropriate to include this variable. To address this problem, we decide to include *Salary* and *Grade* at working year 10 because these variables could potentially indicate an employee’s mid-career performance. Furthermore, we create an additional covariate *Change in Pay Rate*, which specifies an employee’s rate of change in salary between working year 1 and working year 10. Particularly, we posited that higher rate of change in salary would indicate better work performance.

**Assumptions**

The number of the SES position is limited in each agency and potential candidates tend to get promoted at different year. Though the promotion criteria might vary among years, we had to make one key assumption given the rarity of the total number of positions. That is, the covariates selected are general enough to capture the promotion criteria despite time differences. Furthermore, we adjusted salary information in terms of 2011 inflation rate to ensure we could compare variables dependent on time on the same basis.

## Inspection of each covariate

We examined the relationship between outcome variable and predictors with two strategies. First, for categorical predictors, we looked at the percentages of ones in outcome variable for each level. If a categorical predictor has many levels but the percentages of ones in outcome variable for certain levels is limited, we considered regrouping to reduce levels. Second, for continuous predictor, we broke predictor into equally sized, ordered groups, and computed the percentage of ones in outcome variable for each group. We then visualized the patterns between groups and explored the needs for transformation.

**Gender (Categorical)**

Variable *Gender* has 2 levels - Male and Female (M,F). There are more females than males, but each gender has similar proportion of ones.

*Gender*

|  |  |  |
| --- | --- | --- |
| Gender | population (%) | promoted (%) |
| Male (M) | 39.62 | 1.39 |
| Female (F) | 60.38 | 1.76 |

**Race (Categorical)**

Variable *Race* has 6 levels - Hispanic or Latino (1), American Indian or Alaska Native (2), Asian (3), Black or African American (4), Native Hawaiian or Other Pacific Islander (5), and White (6). Because there are significantly less people with race level (1) to (5) than people with race level (6), it is regrouped into two levels - Non-white (0), and White (1).

*Race before regrouping*

|  |  |  |
| --- | --- | --- |
| Race | population (%) | promoted (%) |
| Hispanic/Latino (1) | 7.65 | 0.91 |
| American Indian/Alaska Native (2) | 1.70 | 1.07 |
| Asian (3) | 4.46 | 0.82 |
| Black/African American (4) | 12.43 | 1.38 |
| Native Hawaiian/Other Pacific Islander (5) | 0.05 | 0 |
| White (6) | 73.69 | 1.79 |

*Race after regrouping*

|  |  |  |
| --- | --- | --- |
| Race | population (%) | promoted (%) |
| Non-white (0) | 26.30 | 1.13 |
| White (1) | 73.69 | 1.79 |

**Education Level (Categorical)**

Variable *Education Level* has 7 levels - High School Degree or less (0), More than High School, No Bachelor’s (1), Bachelor’s (2), Master’s (3), Professional Degree (4), Advanced Certification (5), and PhD (6). Because certain education levels have relatively small population size, it is regrouped from 7 to 5 levels based on each level’s characteristics: employees without Bachelor’s degree (level (0) and level (1)) are combined into one group, and employees with Professional Degree and Advanced Certification (Level (4) and (5)) are combined into another group.

*Education Level before regrouping*

|  |  |  |
| --- | --- | --- |
| Education Level | population (%) | promoted (%) |
| High School Degree or less (0) | 8.56 | 0.38 |
| More than High School, No Bachelor’s (1) | 18.57 | 0.59 |
| Bachelor’s (2) | 40.44 | 1.39 |
| Master’s (3) | 22.33 | 2.31 |
| Professional Degree (4) | 5.05 | 4.78 |
| Advanced Certification (5) | 0.25 | 1.45 |
| PhD (6) | 4.79 | 3.05 |

*Education Level after regrouping*

|  |  |  |
| --- | --- | --- |
| Education Level | population (%) | promoted (%) |
| Less than Bachelor’s (A) | 27.14 | 0.53 |
| Bachelor’s (B) | 40.44 | 1.39 |
| Master’s (C) | 22.33 | 2.31 |
| Professional Degree & Advanced Certification (D) | 5.30 | 4.62 |
| PhD (E) | 4.79 | 3.05 |

**Department Switch (Categorical)**

Variable *Department Switch* indicates whether employees switch department between working year 1 and working year 10. The majority did not switch department. Regardless of department switch, each group has similar proportion of people promoted into the SES position.

*Department Switch*

|  |  |  |
| --- | --- | --- |
| Department Switch | population (%) | promoted (%) |
| No Switch | 91.17 | 1.6 |
| Switch | 8.83 | 1.7 |

**Occupational Category (Categorical)**

Variable *Occupational Category* has 3 levels - Administrative (A), Professional (P), and Other White Collar (O).

*Occupational Category*

|  |  |  |
| --- | --- | --- |
| Occupational Category | population (%) | promoted (%) |
| Administrative (A) | 58.06 | 1.71 |
| Professional (P) | 38.97 | 1.56 |
| Other (O) | 2.97 | 0.49 |

**Supervisory Status (Categorical)**

Variable *Supervisory Status* indicates whether employees hold supervisory position at working year 10. Employees with supervisory positions have higher proportion of ones than people without supervisory positions.

*Supervisory Status*

|  |  |  |
| --- | --- | --- |
| Supervisory Status | population (%) | promoted (%) |
| Supervisor | 14.34 | 5.69 |
| Non-supervisor | 85.66 | 0.93 |

**Year 10 Grade (Categorical)**

Variable *Year 10 Grade* consists of 7 levels - grade 9 - 15. The proportion of ones for grade 9 to 12 is small, but increases significantly for grade 13 to 15. It is regrouped into three levels - 0 (Grade 9 - 12), 1 (Grade 13 & 14), and 2 (Grade 15).

*Year 10 Grade before regrouping*

|  |  |  |
| --- | --- | --- |
| Year 10 Grade | population (%) | promoted (%) |
| 9 | 16.46 | 0.11 |
| 10 | 2.44 | 0.30 |
| 11 | 22.88 | 0.26 |
| 12 | 24.35 | 0.75 |
| 13 | 20.81 | 1.69 |
| 14 | 9.66 | 4.77 |
| 15 | 3.39 | 15.73 |

*Year 10 Grade after regrouping*

|  |  |  |
| --- | --- | --- |
| Year 10 Grade | population (%) | promoted (%) |
| 0 | 66.13 | 0.4 |
| 1 | 30.48 | 2.66 |
| 2 | 3.39 | 15.73 |

**Year 10 Salary (Continuous)**

Variable *Year 10 Salary* specifies employees’ salary at working year 10. Based on OPM’s 2011 General Schedule pay table, employees with grade 9 has minimum salary of $41,563, and employees with grade 15 has maximum salary of $129,517. Due to random noises produced by the imputation model, the imputed datasets consist of grade 9 employees with salary less than 41,563, and grade 15 employees with salary greater than 129,517. Since these outliers account for less than 0.1% of the total observations for each imputed dataset, no concerns on the robustness of the analysis are raised. In addition, combining multiple imputed datasets to form final analysis could help averaging out the random noises.

The visualization graph (Figure 6) is produced based on the strategy mentioned before. The x-axis denotes 10 equally sized group, with each group representing certain salary ranges. (Year 10 Salary Table). The y-axis denotes the percentage of ones in each group. In general, employees with higher salary have higher percentage of ones. Group 1 to 8 have percentage of ones less than 2 %, whereas group 9 and 10 have percentage of ones larger than 2 %. The flat curve between group 1 and group 8 and the spike in group 9 and 10 indicate that the relationship between outcome variable and year 10 salary predictor is non-linear. To capture the non-linear relationship, we considered various transformation such as quadratic polynomial transformation, and semi-parametric piecewise spline function is the most effective in modeling this continuous predictor. The detailed descriptions of spline modeling will be mentioned in the next section.

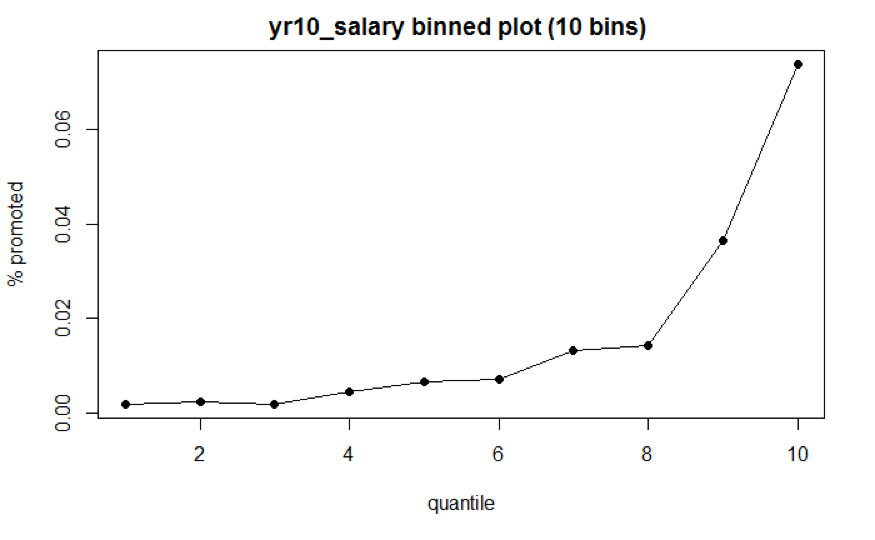


Figure 6 Year 10 Salary Binned Plot

*Year 10 Salary Table*

|  |  |  |
| --- | --- | --- |
| Group | Salary Range | % promoted |
| 1 | [25740,47744) | 0.18 |
| 2 | [47744,52653) | 0.22 |
| 3 | [52653,56801) | 0.18 |
| 4 | [56801,61726) | 0.44 |
| 5 | [61726,66339) | 0.66 |
| 6 | [66339,71136) | 0.69 |
| 7 | [71136,77180) | 1.3 |
| 8 | [77180,82160) | 1.43 |
| 9 | [82160,93591) | 3.65 |
| 10 | [93591,149505) | 7.39 |

Note that we recentered and rescaled the year 10 salary for interpretation purpose. First, year 10 salary is centered on the mean salary at grade 13 from OPM’s 2011 pay table (the mean is around $82,425). Second, it is also scaled by $10,000. Denote as employee ’s year 10 salary, .

**Change in Pay Rate (Categorical)**

Variable *Change in Pay Rate* specifies the change in pay rate between working year 1 and working year 10. The construction of the visualization graph for change in pay rate (Figure 7) is similar to that for year 10 salary. The x-axis denotes 10 equally sized group, with each group representing certain change in pay rate (Change in Pay Rate Table). The y-axis denotes the percentage of ones in each group. Though we believe that higher percentage of change in salary might indicate better performance at work, it is not reflected in the promotion of the SES position. Regardless of change rate, the graph does not indicate particular trend as the proportion of ones fluctuates among groups. This means that *Change in Pay Rate* probably does not provide predictive power if treated as continuous variable.



Figure 7 Change in Pay Rate Binned Plot

*Year 10 Pay Rate Change before regrouping*

|  |  |  |
| --- | --- | --- |
| Group | % Change Range | % promoted |
| 1 | [-10.3,40.4) | 1.39 |
| 2 | [40.4,57.2) | 1.79 |
| 3 | [57.2,69.8) | 1.24 |
| 4 | [69.8,81.1) | 0.10 |
| 5 | [81.1,92.5) | 1.35 |
| 6 | [92.5,105.5) | 1.39 |
| 7 | [105.5,120.5) | 1.57 |
| 8 | [120.5,140.7) | 1.83 |
| 9 | [140.7,174.8) | 1.87 |
| 10 | 174.8+ | 2.71 |

*Year 10 Pay Rate Change after regrouping*

|  |  |  |
| --- | --- | --- |
| Group | % Change Range | promoted (%) |
| Change1 | [-10.34,62) | 1.51 |
| Change2 | [62,160) | 1.42 |
| Change3 | 160+ | 2.66 |

To address the problem, we decide to model pay rate as a categorical variable with 3 levels. The first level consists of employees with rate changes below 62% (equals to around 5% change per year). This pay rate change range falls within the regular rate change of the federal government. The second level consists of employees with rate change between 62% and 160% (equals to around 5 - 10% change per year). This rate change range is higher than the regular rate change, indicating that employees with this rate change level might perform better than employees with regular rate change. The third level consists of employees with rate changes above 160%. This rate change is significantly higher than the regular change rate. Note that some of the very high rate change values are sensitive to the imputation models since more than 80% of the rate change data is missing.

# Model Fitting

## Logistic regression with piecewise linear spline application

In Generalized Linear Models (GLMs), the relationship between coefficients of predictors and univariate outcome variable is usually linear. The binnedplot of variable *Year 10 Salary* in the previous chapter, however, indicates that the relationship between outcome variable and the predictor is non-linear. Incorporating smoothing techniques such as the application of spline function is a common strategy to model data with non-linear relationship (Walther, 2017). Specifically, we applied piecewise linear spline function to model the data.

To recap the relationship between predictor *Year 10 Salary* and outcome variable, the proportion of ones below certain salary level is relatively flat, whereas the proportion of ones above certain salary level increases significantly (Figure 8). This indicates that the coefficient slopes above and below certain salary level are different.

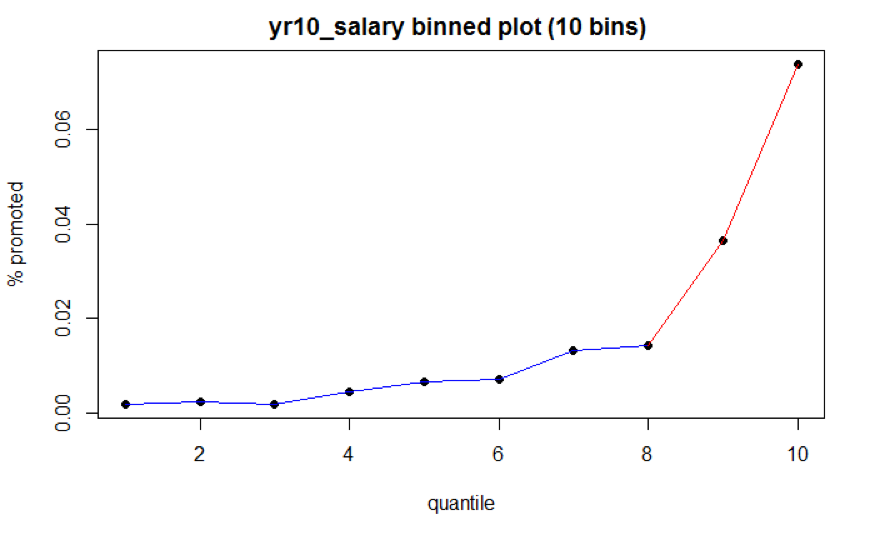


Figure 8 Year 10 Salary Binned Plot

A spline of degree is a continuous function formed by connecting polynomial segments. The points where the segments connect are called the knots of the spline. In general, a spline of degree D associated with a knot takes the form:

We created another binned plot with 25 bins (Fig 9) in order to select knots () granularly. The first knot is chosen as the cut point between black and blue segments, which takes the value of $71,674. Since year 10 salary is centered and scaled, this cut point should be . The second knot is chosen as the cut point between blue and red segments, which takes the value of $84,697 ().

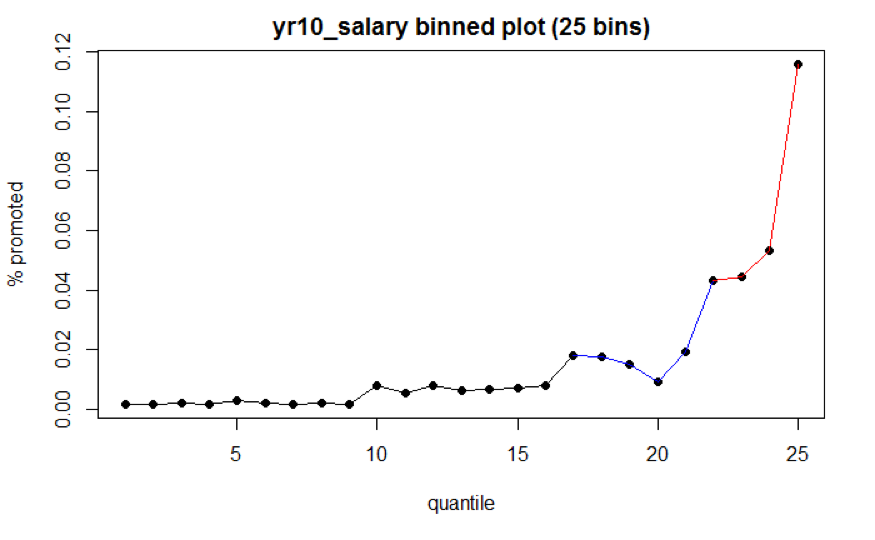


Figure 9 Year 10 Salary Binned Plot

Spline function with higher degree can better capture the non-linear relationship. However, we found it inappropriate to model the OPM data with higher degree due to the small proportion of ones in outcome variable. To avoid overfitting, we chose a degree of 1 for the spline function.

To sum up, denote as the year 10 salary after transformation. The prediction function for the piecewise linear spline is

Specifically, if year 10 salary is below the first cut point, then coefficient and equals to 0. If year 10 salary is between the first and second cut point, the coefficient can help capture the difference in slopes. Similarly, if year 10 salary is above the second cut point, the coefficient along with can help capture the difference in slopes. We can rewrite the prediction equation as

## Baseline

Other than covariate *Year 10 salary*, the rest of the covariates are categorical variables. Given that the proportion of ones in outcome variable is small, it is important to select reasonable baseline for each categorical variable. For variable *Gender*, we would like to learn whether women were put in disadvantages in promotion, so male is the baseline. For variable *Race*, since the population size for white is the largest, white is the baseline. For variable *Education Level*, Bachelor’s degree is the baseline because it is the most common degree federal employees earned. For variable *Department Switch*, switch is the baseline. For variable *Occupational Category*, Professional is the baseline. For variable *Supervisory Status*, supervisor is the baseline. For variable *Year 10 Grade*, regrouped grade 1 (grade 13 & 14) is the baseline. For variable *Change in Pay Rate*, change 2 (annual rate change between 5-10%) is the baseline.

## Model Form

To sum up, the model takes the following form.

# Pooled Analysis

We estimated the model coefficients of each completed dataset and conducted model diagnostic based on one completed dataset. We then combined model results of all datasets for inference.

## Estimates of model coefficients

In a frequentist setting, estimates of model coefficients are computed through maximum likelihood estimation. In a Bayesian setting, estimates of model coefficients are computed through posterior distribution, which requires the specification of prior distributions.

Prior to seeing the data, we assumed that the intercept () have normal distribution with mean -3 and variance 2. The coefficients of the predictors () have normal distribution with mean 0 and variance 5 - that are as likely to be positive as they are to be negative but unlikely to be far away from zero. After observing the data, we modeled the likelihood of each observation as a binomial distribution for ; is the probability of getting promoted, where . The posterior distribution is proportional to the product of the priors and the likelihood distribution. To sum up, the posterior distribution is derived as below.

**Prior Distribution**

**Likelihood**

**Posterior Distribution**

**Model Diagnostic**

Logistic regression does not require the assumptions of constant variance and normality of residuals. To check if the predictors of fitted logistic regression model are well specified, we applied binned residuals technique. The procedures to produce binned residuals include: first, we computed raw residuals of the fitted model and ordered observations by values of predicted probabilities; second, for continuous predictor, we formed equally sized, ordered bins using the ordered data and computed average residual of each bin. Next, we plotted average residual versus average predicted probability of each bin to check if there are any specific patterns that might cause problems; third, for categorical predictor, we computed average residual of each level.

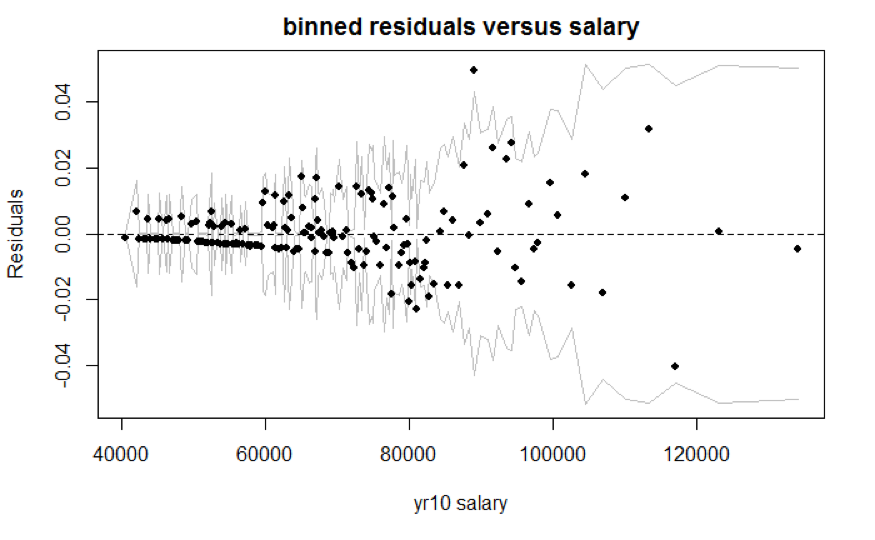


Figure 10 Residuals Binned Plot

We created the binned residuals plot (Figure 10) based on one imputed dataset’s model (Frequentist model) results. The average residuals are close to 0 for salary less than $82,425, but tend to fluctuate for salary greater than $82,425. That data with higher salary values are more sparse leads to the fluctuation. For categorical variable, the number of level is limited and majority of the time the prediction would yield 0 (not promoted), and thus, the average residuals for each bin are small. In general, we did not observe extreme residuals, but we should be wary of data with higher salary values.

Gender

|  |  |
| --- | --- |
| M | F |
|  |  |

Occupational Category Residuals

|  |  |  |
| --- | --- | --- |
| P | A | O |
|  |  |  |

Education Levels Residuals

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| B | A | C | D | E |
|  |  |  |  |  |

Race

|  |  |
| --- | --- |
| White | Non-white |
|  |  |

Switch

|  |  |
| --- | --- |
| Switch | No-switch |
|  |  |

Supervisory Position Residuals

|  |  |
| --- | --- |
| Supervisor | Non-supervisor |
|  |  |

Grade

|  |  |  |
| --- | --- | --- |
| Grade0 | Grade1 | Grade2 |
|  |  |  |

% Change Residuals

|  |  |  |
| --- | --- | --- |
| Change2 | Change1 | Change3 |
|  |  |  |

## Pooled results

We combined poole results for both Frequentist model and Bayesian model. For Frequentist model, we obtained the pooled results through (1) averaging out the estimated regression coefficients of each completed dataset (2) accounting for both within and across datasets variances. For Bayesian model, we obtained the pooled results through summarizing the mixture draws from posterior distribution.

### Frequentist model

For each completed dataset () and each regression coefficient (), let be the estimator of from dataset, and let be the estimator of the variance of from dataset. First, we computed the average of each dataset’s estimated slope to obtain the estimate of the model coefficient . Second, we computed the average of each dataset’s estimated slope variance to obtain the within-dataset variance . Thid, we computed the variance of the slopes across 12 datasets to obtain the across-datasets variance .

With average estimated regression coefficient (), within-dataset variance (), and across-datasets variance (), the point estimate of is , and its variance is . An approximate 95% confidence interval is . The table below shows the pooled estimated slope, standard error, and 95% confidence interval of each regression coefficient.

*Combined Coefficient Estimate, Standard Error, and 95% CI. (Frequentist Model)*

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficient | Estimate | Std. Error | 95% CI. |
| Intercept () | -6.018 | 0.697 | [-6.871, -5.164] |
| Female () | -0.035 | 0.012 | [-0.0815, 0.011] |
| OccA () | 0.744 | 0.015 | [0.709, 0.779] |
| OccO () | 1.005 | 0.299 | [0.845, 1.170] |
| EducA () | -0.712 | 0.036 | [-0.748, -0.675] |
| EducC () | 0.128 | 0.016 | [0.094, 0.161] |
| EducD () | 0.195 | 0.033 | [0.139, 0.252] |
| EducE () | -0.047 | 0.045 | [-0.129, 0.036] |
| Non-white () | -0.112 | 0.017 | [-0.151, -0.074] |
| No-switch () | -0.039 | 0.031 | [-0.057, -0.020] |
| Non-Supervisor () | -0.949 | 0.013 | [-1.008, -0.890] |
| Grade0 () | -0.450 | 0.066 | [-0.657, -0.244] |
| Grade2 () | 1.389 | 0.049 | [1.213,1.565] |
| ScaledSalary () | 1.939 | 0.618 | [1.157, 2.721] |
| (ScaledSalary + 1.0751) () | 2.919 | 0.642 | [2.077, 3.761] |
| (ScaledSalary - 0.2272) () | 3.328 | 0.902 | [2.362, 4.294] |
| Change1 () | -0.266 | 0.027 | [-0.457, -0.074] |
| Change3 () | 0.233 | 0.025 | [0.065, 0.401] |

### Bayesian Model

For Bayesian model, instead of averaging out the estimated regression coefficients, we obtained the pooled results through applying the approach proposed by Zhou and Reiter (2010). For each completed dataset and parameter of interest , denote as the MCMC draws from posterior distribution. First, simulate values of posterior draws (where is large) from each , with its distribution denoted as . Second, mix all to form . Third, sort total number of draws from , where the estimates of 95% posterior interval, posterior median, and other statistics can be obtained from the mixed posterior draws (Zhou & Reiter, 2010).

The pooled estimated slope, standard error, and 95% credible interval are generated through specifying to create a total number of draws. The table below shows the pooled estimated slope, standard error, and 95% credible interval of each regression coefficient.

*Combined Coefficient Estimate, Standard Error, and 95% CI. (Frequentist’s Model)*

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficient | Estimate | Std. Error | 95% CI. |
| Intercept () | -6.033 | 0.816 | [-7.732, -4.529] |
| Female () | -0.036 | 0.110 | [-0.254, 0.176] |
| OccA () | 0.745 | 0.124 | [0.502, 0.988] |
| OccO () | 0.877 | 0.571 | [-0.359, 1.887] |
| EducA () | -0.719 | 0.191 | [-1.103, -0.354] |
| EducC () | 0.127 | 0.127 | [-0.123, 0.375] |
| EducD () | 0.190 | 0.183 | [-0.174, 0.542] |
| EducE () | -0.056 | 0.212 | [-0.479, 0.355] |
| Non-white () | -0.116 | 0.132 | [-0.383, 0.140] |
| No-switch () | -0.029 | 0.178 | [-0.365, 0.326] |
| Non-Supervisor () | -0.950 | 0.112 | [-1.167, -0.729] |
| Grade0 () | -0.458 | 0.253 | [-0.966, 0.025] |
| Grade2 () | 1.395 | 0.218 | [0.969, 1.823] |
| ScaledSalary () | 1.933 | 0.770 | [0.530, 3.559] |
| (ScaledSalary + 1.0751) () | 2.915 | 0.781 | [1.476, 4.538] |
| (ScaledSalary - 0.2272) () | 3.316 | 0.927 | [1.573, 5.205] |
| Change1 () | -0.268 | 0.159 | [-0.586, 0.032] |
| Change3 () | 0.231 | 0.154 | [-0.078, 0.521] |

### Comparison between Frequentist and Bayesian Model

We compared the pooled results for both Frequentist model (Figure 11) and Bayesian model (Figure 12). The pooled coefficient means of both models are close to each other, but Bayesian model has higher pooled standard errors. There are certain requirements for Bayesian inference after multiple imputation. Our inherent constraints - prior parameter choices and number of available imputed datasets therefore lead to greater uncertainty. First, the posterior distribution depends on the choice of prior parameter, which usually requires knowledge from experts. Our choice of non-informative prior might not be the most ideal prior. Second, Zhou and Reiter (Zhou & Reiter, 2010) suggested that the number of imputed datasets should be large (at least 100 imputed datasets) in order for Bayesian inference to work well. Since there are only 12 available imputed datasets, Bayesian inference might not be valid. Thus, we conducted inference based on the Frequentist model.

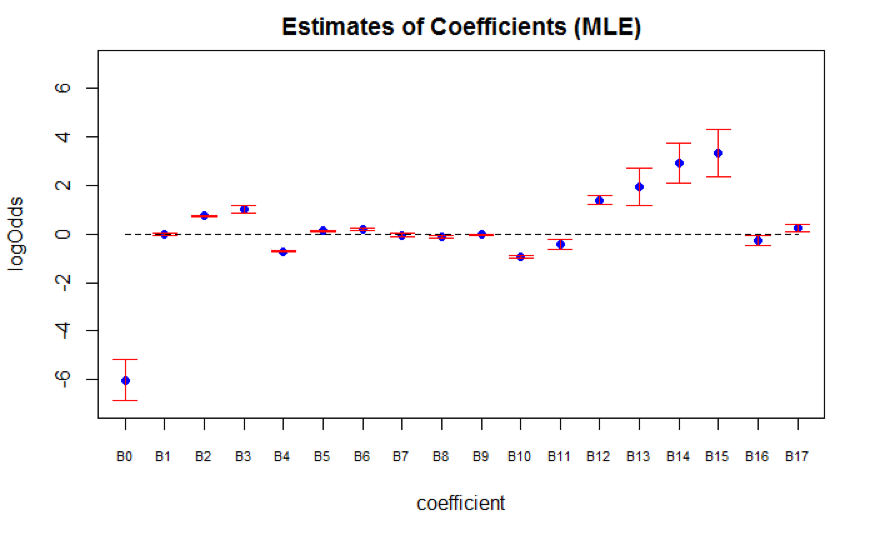


Figure 11 Frequentist Model

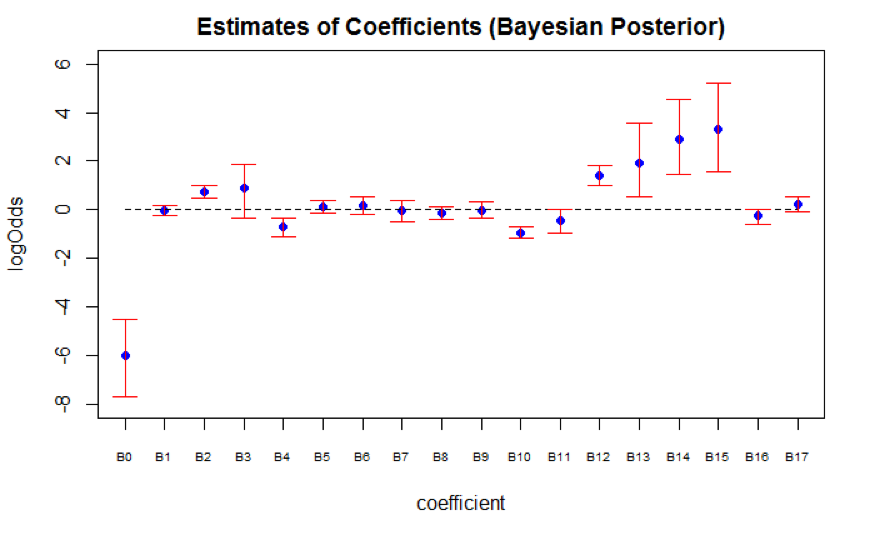


Figure 12 Bayesian Model

## Inference based on Frequentist model

In general, the selection of the SES positions is highly dependent on (1) personal skills (indicated by variable *Occupational Category* and *Supervisory Status*) (2) education level (indicated by variable *Education Group*) and (3) performance at work (indicated by variable *Salary*, *Grade*, and *Change in Pay Rate*)

First, for employees whose occupational category are *Administrative*, the odds of getting promoted are multiplied by a factor of 2.104 (95% CI: 2.031 to 2.180) compared to employees whose occupational category are *Professional*, holding all else constant. This implies that analytical ability, judgment, discretion, and personal responsibility (characteristics of occupational category *Administrative*) might be more important than having knowledge in a specific field of science (characteristics of occupational category *Professional*). Furthermore, for employees who do not serve in supervisory positions, the odds of getting promoted are multiplied by a factor of 0.387 (95% CI: 0.365 to 0.411) than employees who serve in supervisory positions, holding all else constant. Since leadership is one of the key competencies for the SES positions, employees without leadership experiences might be put into disadvantages.

Second, education level also plays an important role when it comes to predicting promotion outcome. For employees with degrees below Bachelor’s, the odds of getting promoted is multiplied by 0.491 (95% CI: 0.473 to 0.509) compared to employees with Bachelor’s degree, holding all else constant. Furthermore, though it is expected that the odds of getting promoted would increase with higher education level, there is certain limit. Specifically, compared to employees with only Bachelor’s degrees, the odds of getting promoted increases if employees acquired Master’s and Advanced degrees. However, the effect of PhD degrees on predicting promotion is not significant because its 95% confidence interval contains 0. Because PhD degrees focus on particular fields of studies, this result further confirms that having specific knowledge in a field of science might not be one of the important promotion criteria.

Third, compared to employees with grade 13 and 14 at working year 10, the odds of getting promoted are multiplied by a factor of 0.637 (95% CI: 0.519 to 0.784) for employees with grade below 13, but multiplied by a factor of 4.010 (95% CI: 3.364 to 4.781) for employees with grade 15. The visualization graphs further show the relationship between salary at given grade and % change in salary (Grade 15 for Figure 13, Grade 14 for Figure 14, Grade 13 for Figure 15, Grade < 13 for Figure 16). Regardless of Grade, the predicted probability of getting promoted increases as salary increases, and employees with bigger percentage change in salary also have higher odds of getting promoted. Note that the model yields predicted probability up to 0.25 at Grade 15. Given the small proportion of ones, we believe that the predicted probability at grade 15 is potentially overestimated due to sparsity of the data at higher salary values.

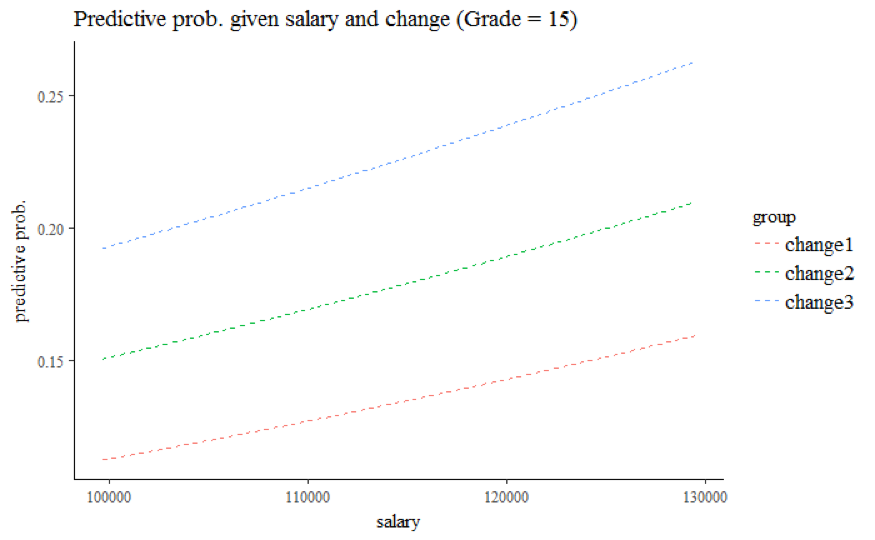


Figure 13 Grade 15

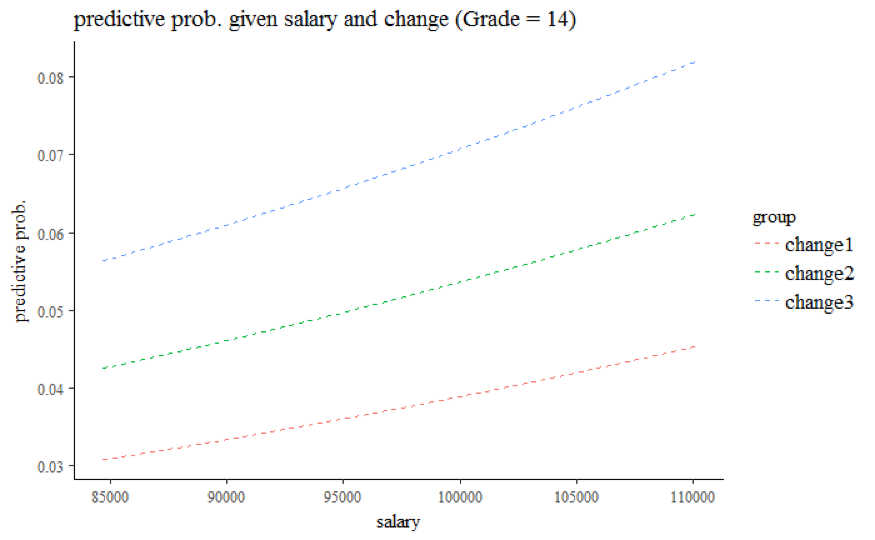


Figure 14 Grade 14

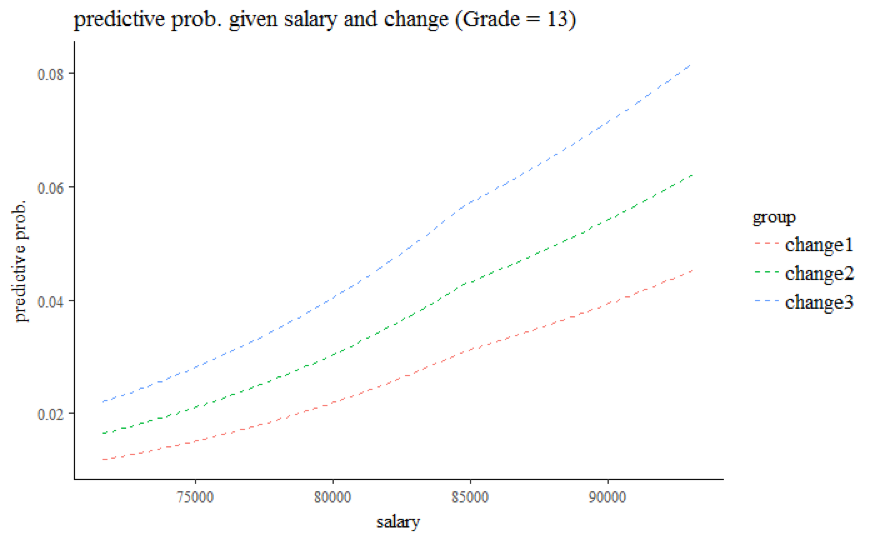


Figure 15 Grade 13



Figure 16 Grade < 13

Lastly, the 95% confidence interval of variable *Gender* contains value 0, suggesting that the effect of *Gender* is not significant in predicting promotion outcome. For employees whose race are *Non-white*, the odds of getting promoted are multiplied by a factor of 0.894 (95% CI: 0.868 to 0.929) compared to employees whose race are *White*, holding all else constant.

## Conclusion and future improvement

In conclusion, the model suggests an employee’s personal skills and performance at work are significant in predicting promotion outcomes, but certain limitations exist for the model. First, the analysis is based on 10% samples of the OPM data, and we would like to apply the current model to the full samples. Second, the predictive probability of employees with very high salary values is potentially overestimated because the piecewise linear model is not granular enough to account for small proportion of ones in outcome variables. In the future, we would like to experiment on various models to improve the predictive accuracy. Third, we did not have enough imputed datasets to account for random noises produced by the imputation models. This also affected the validity of the Bayesian inference after multiple imputation. In the future, we would like to impute more completed datasets through expanding computational capacity.

# References

Barrientos, A. F., Bolton, A., Balmat, T., Reiter, J. P., Figueiredo, J. M. de, Machanavajjhala, A., … DeLong, M. (2017). A framework for sharing confidential research data, applied to investigating differential pay by race in the us government. *arXiv Preprint arXiv:1705.07872*.

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

Carey, M. P. (2011). *Senior executive service: Background and options for reform*. DIANE Publishing.

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.

Guide to the senior executive service. (2017), 2–12.

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.

Huddleston, M. W., & Boyer, W. W. (1996). *The higher civil service in the united states: Quest for reform* (pp. 116–117). University of Pittsburgh Pre.

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/>

Powell, G. N., & Butterfield, D. A. (1994). Investigating the “glass ceiling” phenomenon: An empirical study of actual promotions to top management. *Academy of Management Journal*, *37*(1), 68–86.

Statement of Jeffrey D. Zients, U.S. Congress, Senate Committee on Homeland Security and Governmental Affairs, Subcommittee on Oversight of Government Management, the Federal Workforce, and the District of Columbia. (2011), 1.

Walther, G. (2017). Lecture 17: Smoothing splines, Local Regression, and GAMs.

Zhou, X., & Reiter, J. P. (2010). A note on bayesian inference after multiple imputation. *The American Statistician*, *64*(2), 159–163.