# EMPLOYEE TURNOVER PROBABILITY PREDICTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
INDUSTRIAL ENGINEERING

By
Hüsameddin Deniz Barın
September 2022

EMPLOYEE TURNOVER PROBABILITY PREDICTION
By Hüsameddin Deniz Barın
September 2022

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Savaş Dayanık(Advisor)

Firdevs Ulus

Kemal Kılıç

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

# ABSTRACT

## EMPLOYEE TURNOVER PROBABILITY PREDICTION

Hüsameddin Deniz Barın
M.S. in Industrial Engineering
Advisor: Savaş Dayanık
September 2022

Employee turnover prediction is crucial for the companies in the sense that the precautionary action by the employers can be made in advance. A turnover data provided by a company was examined throughout the thesis. Firstly, the missing data were imputed. Then a hierarchical model aiming to explain the attrition heterogeneity among the employees and preventing separation was fitted to the data set. Finally, the results of the implementation were analyzed along with the benchmark models. Based on the results, the proposed hierarchical model had a higher performance on the target metric and the heterogeneity across the units was inferred through the hierarchical model which outperformed the benchmark models.

*Keywords:* Employee turnover, Bayesian hierarchical models, Missing data imputation.

# ÖZET

# PERSONEL KAYBI OLASILIĞI TAHMİNLEME

Hüsameddin Deniz Barın
Endüstri Mühendisliği, Yüksek Lisans
Tez Danışmanı: Savaş Dayanık
Eylül 2022

Personel kaybı tahmini, işverenler için önlem alınabilmesi adına önem arz etmektedir. Bu tezde, personel kaybı verisi incelenmiştir. Veri setindeki eksik veriler doldurulmuştur. Verideki ayrılabilirlik ve personelin ayrılma eğilimine yönelik heterojenlik, hiyerarşik modelleme kullanılarak açıklanmaya çalışılmıştır. Son olarak; uygulama sonuçları alternatif modeller eşliğinde incelenmiştir. Sonuçlar ışığında, önerilen hiyerarşik model alternatif modellerden hedeflenen ölçü cinsiden daha iyi performans sergilemiş ve ayrılma eğilimine yönelik heterojenlik, alternatif modellerden daha iyi performans sergileyen hiyerarşik model sayesinde görülmüştür.

*Anahtar sözcükler*: Personel kaybı, Hiyerarşik modelleme, Eksik veri doldurma.

# Acknowledgement

I would like to thank to my advisor Prof. Dr. Savaş Dayanık who supported me and guided me through his wisdom and knowledge. It was a privilige to work along with him and I appreciate him for giving me this chance. I would like to thank to my mother, Ayşegül Barın, and father, Zafer Barın, for always supporting me in my education and everything. I could not have been the person I am today and I could have not achieved without them. I would also like thank to Pınar Çep for her patient support and for convincing me that I can succeed everytime I faced with an issue.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This thesis introduces a hierarchical logistic regression model to estimate the turnover probabilities of employees working for a company manufacturing agricultural machinery. The name of the company will not be provided due to privacy concerns. The main purpose of the study is to come up with a robust model estimating the employees apt to leave the company so that the company can necessarily take precautionary actions beforehand. The company's Human Resources Department suffers from high turnover rates and expects to predict the employment status on a yearly basis.

The data consisted of the rows corresponding to the employment status of the employees during a calendar year and some explanatory variables so the data are in unbalanced panel format since each employee might have multiple observations and, the observations start at the beginning of the employment or the data collection process and end with the termination of the employment or last update of the data. The data are grouped by the employees causing heterogeneity because employees might have certain sensitivities for attrition which separates them from the others. The heterogeneity is expected to exist due to employees' varying expectations from their jobs. Therefore, the individual differences among the employees should be accounted for by the proposed model. These differences

can be explained thorough the variation in the demographic features of the employees which will base the grouping strategy of the observations. In addition, the variables are mostly categorical which brings the problem of separability so the model should also prevent overfitting. At this point, Rossi et al. [1, Chapter 5] proposes a hierarchical logistic regression model where regression coefficients of units are determined by demographic information which enables to have heterogeneous regression coefficients. Also, using a Bayesian model allows to use a prior shrinking the coefficients towards 0 to prevent overfitting. In addition to the estimation problem, some of the variables are partially missing which should be handled appropriately to hinder the estimation model leading to biased results.

The data contain 3282 observations of employees from 2015 to 2019. There are 1012 employees worked for the company in total during this period of time. The number of predictors used within the models is 17.

Second chapter briefly covers the literature review regarding the models proposed for employee turnover problem and some other models that might be appropriate for the data at hand, and the missing data imputation problem. Third chapter of the thesis includes an introductory part to the data at hand, and the relationship between the predictors and response are discussed using some exploratory analysis tools. Forth chapter focuses on how to handle partially missing data. Fifth and sixth chapter introduces the proposed Bayesian hierarchical model and the diagnostics regarding the model. In seventh chapter, a couple of benchmark models are introduced which might help validating the proposed model. In eighth chapter, the Bayesian hierarchical model and the benchmark models are compared in terms of the target metric and the inferences regarding the master model is made. Finally, the overall conclusion is discussed.

# Chapter 2

# Literature Review

In literature, there are various models aiming to predict employee turnover. Gao et al. [2] uses WQRF (Weighted Quadratic Random Forest) in order to classify employment status. The proposed approach aims to handle the problem of unbalanced classes of response which is common in employee turnover prediction problems because only a small fraction of employees leave the company during a short period of time such as a year or a month. In addition, the model attempts to overcome the problem of overfitting by variable selection. Ajit [3] argues that the turnover data are tended to be noisy due to under investment of Human Resources Analytics Divisions or Departments. To overcome this issue, Extreme Gradient Boosting (XGBoost), introduced by Chen and Guestrin [4], is proposed due to its capability of preventing overfitting in noisy data. Since both WQRF and XGBoost models are shown to outperform benchmark models such as Naive Bayes, Random Forest and Logistic Regression on employee turnover problems, these models can be considered as alternatives for the data at hand. In addition to those models, Gelman et al. [5] proposes a weakly automatic prior on generic data sets which shrinks the regression coefficients towards 0.

For the missing values, Van Buuren [6], Buuren and Groothuis-Oudshoorn [7], Van Buuren et al. [8] and Van Buuren et al. [9] introduce a missing data imputation model which can handle the missing values by using the observed

data.

For modelling the heterogeneity across the units of a data set, Rossi et al. [1] proposes a logistic Bayesian hierarchical model using demographic variables to explain the differences across the units.

Gelman et al. [10, Chapter 11] discusses the diagnostics helping to evaluate the parameters of the hierarchical model.

# Chapter 3

# A Glimpse of Data and Exploratory Analysis

The data contain demographic variables which are observed and assumed to remain as it is for each employee during the employee's lifetime. Time dependent variables are expected to change year by year and are expected to be available at the beginning of each calendar year unless they are missing.

A glimpse of the data is available in Table 3.1. Each employee has its own ID and the observations of employees are on a yearly basis. The data is in panel format since the groups are defined by the employees and each group contains at least one row which are observed at different time periods. Occasionally, there are missing values for some of the variables.

Table 3.1: A Glimpse of the Data

| Observation | ID | Working.Status | Year | Number.of.Years.Worked | College.Department | Alma.Mater | Highest.Degree | First.Year | Gender | Child | Child.Born | Level.of.English.Proficiency | Job.Switch | Marital.Status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10000 | Working | 2015 | Medium | Support | 20 | Over Bachelor | 0 | Man | 1 | 0 | Insufficient | 1 | Married |
| 2 | 10000 | Working | 2016 | Medium | Support | 20 | Over Bachelor | 0 | Man | 1 | 0 | Insufficient | 1 | Married |
| 3 | 10000 | Working | 2017 | Medium | Support | 20 | Over Bachelor | 0 | Man | 1 | 0 | Insufficient | 1 | Married |
| 4 | 10000 | Left | 2018 | Medium | Support | 20 | Over Bachelor | 0 | Man | 1 | 0 | Insufficient | 1 | Married |
| 5 | 10001 | Working | 2015 | Medium | NA | NA | NA | 0 | Woman | 0 | 0 | Insufficient | 1 | Married |
| 6 | 10001 | Left | 2016 | Medium | NA | NA | NA | 0 | Woman | 0 | 0 | Insufficient | 1 | Married |

## 3.1 Variables

The list of the predictors and their definitions are provided below.

**Time Dependent Variables**

- Age
- Number of Years Worked
- Child: It is equal to '1' if the employee parents a child. Otherwise, it is '0'.
- Marital Status: The marital status of the employee at the beginning of the year.
- Child Born: It is equal to '1' if the employee had a new born child during the year. Otherwise, it is '0'.
- Performance Score: The performance score given by employee's manager showing how well the employee performed during the previous year. It might be labeled as bad, okay, good or "none" if it is the employee's first year in the company. The previous year's performance score is used as a predictor in current year's employment status since it will be yet available at the end of the year.
- Appreciation Score: The appreciation score given by employee's manager showing how compatible the employee worked with its environment during the previous year. It might be labeled as bad, okay, good or "none" if it is the employee's first year in the company. The previous year's appreciation score is used as a predictor in current year's employment status since it will be yet available at the end of the year.
- First Year: It is equal to '1' if it is the employee's first year at the company. Otherwise, it is '0'.
- Salary: Salary of an employee during the year. This variable was masked by a linear transformation of the true values so that the salary information is kept private.

- Salary Raise: Percentage of increase in salary normalized by average of percentage of increase in salaries among all employees during the year. For instance, if an employee had a raise of 20% and the average raise among the all employees during the year were 25%, the raise would be $\frac{1.20}{1.25} = 0.96$. The employees having their first year at the company is labeled as "none".
- Highest Degree: The degree of the highest education. For the employees who did not go to the college, the groups were combined since these groups contained a small number of observations.
- Job Switch: It is equal to '1' if the employee worked for another company before.

**Demographics**

- Alma Mater: The cluster of the university that the employee studied. The clusters were formed manually considering the reputation of the universities.
- College Department: The cluster of the department that the employee graduated from. The clusters were formed manually considering the similarity of the majors.
- Gender
- Workplace: The location of the workplace (The company owns a factory and agencies in other locations).
- Level of English Proficiency
- No School: It is equal to '1' if the employee's highest graduation is 'High School' or below. This variable is perfectly collinear with one of the categories in Alma Mater and College Department. The categories in Alma Mater and College Department are dropped from both benchmark models and the original model.

- Education Level: A combination of English Proficiency and Alma Mater demonstrating how qualified the employee is with respect to others (This variable was replaced with English Proficiency and Alma Mater in hierarchical model due to giving better results on the validation set).

It is essential to note that some categories and variables were perfectly collinear. For instance, the Performance Score variable has a category as 'none' which is perfectly collinear with First Year variable. To have stable results in models, perfectly collinear categories were removed.

## 3.2 Response

Response variable, Working Status, is a binary variable demonstrating if the attrition occurred or not during any time of the year.

Figure 3.1 shows an overall look to what the categories are like across the variables. There are certain variables such as Age and Number.of.Years.Worked having flat distributions across the categories since these variables were transformed from the raw data based on their percentiles. On the other hand, the variables such as Workplace do not contain much information due to being sparse other than their explanatory power. Furthermore, the variables having minor categories contribute to the overfitting problem.

## 3.3 Exploratory Analysis

Exploratory analysis is made to examine the tentative relationship between the predictors and the response. Although some of the predictors are numerical, they were transformed into categorical form to capture possibly non-linear relationship between the parameters and the response for the main model since the logistic regression models are a type of linear models. The transformation was

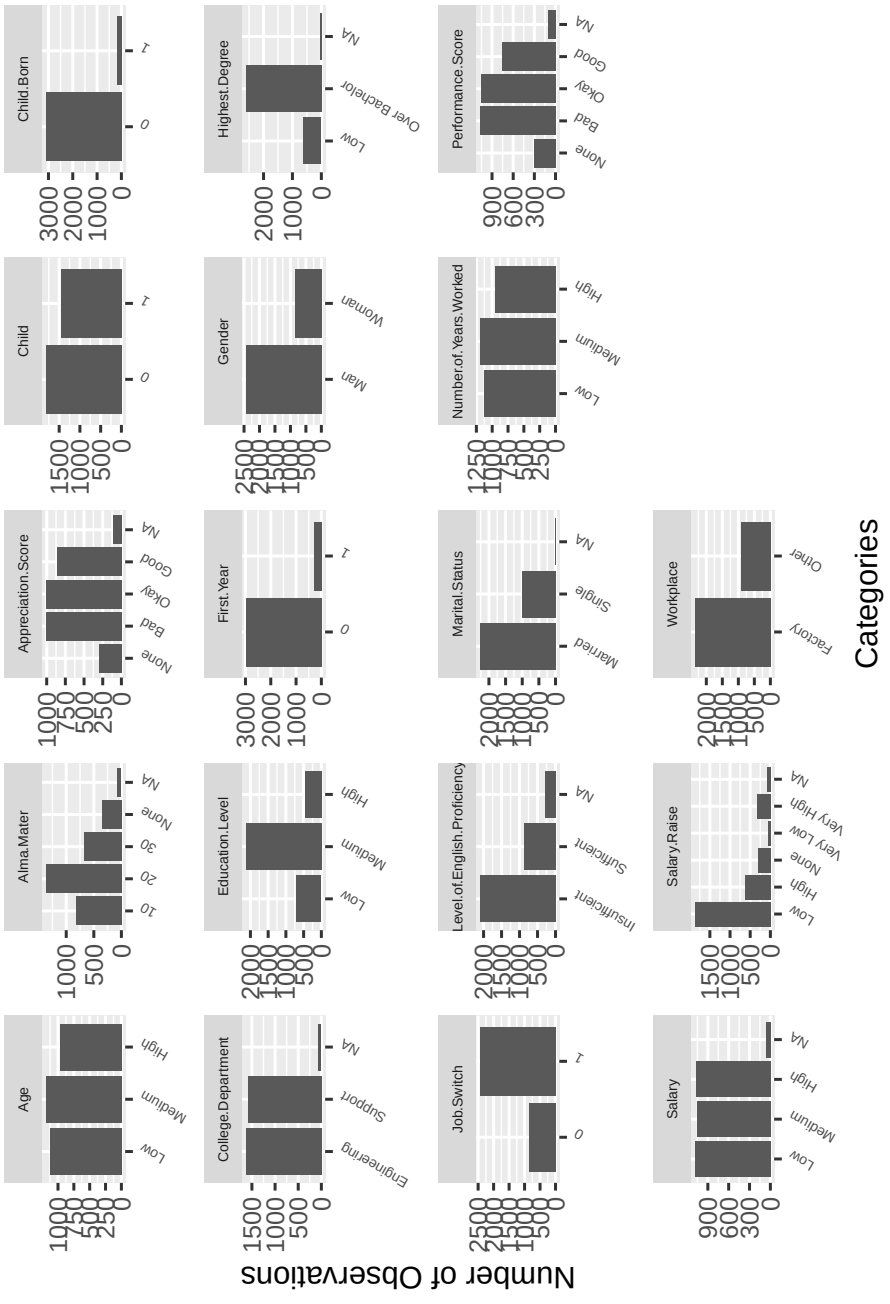


Figure 3.1: Visualization of the Data

not applied for the benchmark models (Weighted Quadratic Random Forest and XGBoost) that can capture the non-linear relationship between the outcome and the variables.

Binning the continuous variables leads to have a separable data and lose a certain amount of information caused by aggregation. For instance, Figure 3.2 illustrates that there is not a clear distinction between the age values for different values of the outcome. This might lead to an inference such that age might not have an impact on explaining the outcome which might be contradicted from Table 3.2 showing the conditional distributions of Working Status given Age and it suggests that people at early ages might be more inclined to leave as similar to experienced workers. Then, using these types of variables as continuous in linear models might prevent capturing the information available. On the other hand, variables such as raise might have monotonic relationship with the outcome which does not have to be linear enforcing variables to transform to categorical form.

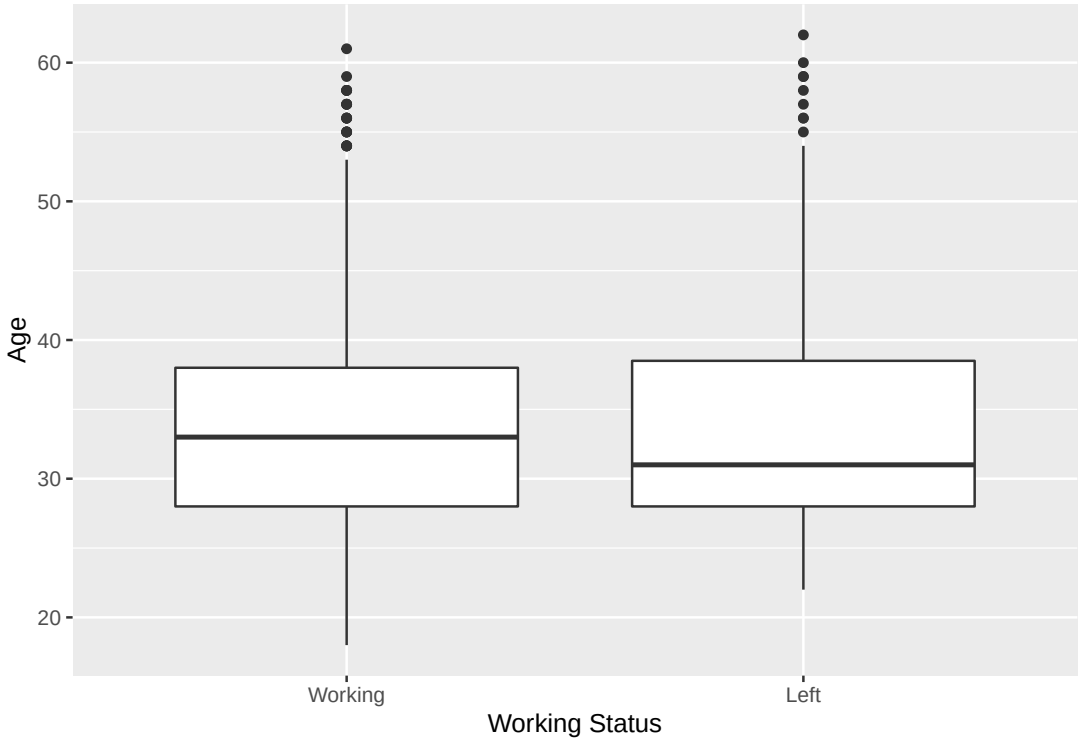


Figure 3.2: Effect of Continuous Age on Working Status

Table 3.2: Distribution of Working Status Conditional on Categorical Age

| | Working | Left |
|---|---|---|
| (18,30] | 88.68 | 11.32 |
| (30,50] | 92.30 | 7.70 |
| (50,65] | 75.81 | 24.19 |

## 3.4 Chi-Square Test of Independence for Categorical Variables

Chi square test of independence is an appropriate tool to compare categorical variables to measure if there might exist a statistical dependence. The hypothesis test is as the following:

$$\begin{aligned} H_0 &: \text{The variables are statistically independent.} \\ H_A &: \text{It is rejected that the variables are statistically independent.} \end{aligned} \tag{3.1}$$

This test can be used to check if there might be dependence between the response and the predictors.

Table 3.3: Results of Chi-Square Test of Independence

| Categorical Variables | p-value |
|---|---|
| Child | 0.00 |
| Marital.Status | 0.00 |
| Appreciation.Score | 0.00 |
| Salary.Raise | 0.00 |
| Number.of.Years.Worked | 0.00 |
| Age | 0.00 |
| Level.of.English.Proficiency | 0.00 |
| Performance.Score | 0.00 |
| Child.Born | 0.00 |
| Job.Switch | 0.00 |
| Alma.Mater | 0.02 |
| First.Year | 0.03 |
| College.Department | 0.07 |
| Salary | 0.11 |
| Gender | 0.22 |
| Highest.Degree | 0.33 |
| Workplace | 0.90 |

According to Table 3.3, at the confidence level of 95%, the null hypothesis is rejected for some of the variables, such as Child, Marital Status, Salary Raise,

Table 3.4: Missing Percanteges of Variables

| Variable | Missing Percentage |
|---|---|
| Level.of.English.Proficiency | 10.28 |
| Alma.Mater | 5.73 |
| College.Department | 5.34 |
| Highest.Degree | 5.04 |
| Appreciation.Score | 3.38 |
| Performance.Score | 3.35 |
| Salary | 1.89 |
| Marital.Status | 0.70 |

which may suggest some variables might help explaining the variance in the response. For instance, single employees seem to have more tendency for attrition or employees having the lowest raise category might be more likely to leave the company. The test was applied more than once which might be associated with making the Bonferroni correction for confidence level but most variables' p-value are already practically 0 where Bonferroni correction is not needed.

## 3.5 Missingness

Some variables are partially missing. The percentages of the missing variables are presented in Table 3.4.

# Chapter 4

# Handling Missing Data

As shown in Table 3.4, there are partially missing variables and the missing values must be handled appropriately which was retrospectively discussed in Rubin [11] highlighting under what conditions the missingness, a matrix showing the missingness in the data, can be ignored in the likelihood. It is straightforward to have a desire to ignore the missing data since ignoring the missing data corresponds to ignoring the parameters leading to missingness which are not of the interest of the study [6, Chapter 1,2]. To ignore the missingness in inference of the parameters of interest, the reason leading to missingness or the matrix showing the missingness, the missing mechanisms are defined as in Van Buuren [6, Chapter 1,2] which are Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR).

## 4.1 Notation

The following notation used in Buuren and Groothuis-Oudshoorn [7] will be used for the rest of this section. Let $Y$, $Y_{miss}$ and $Y_{obs}$ denote the data, missing part and observed part of the data respectively. $Y_{miss,j}$, $Y_{obs,j}$ and $Y_{-j}$ denote the missing and observed parts of variable $j$ and variables other than $j$ respectively. $\theta$ and $\phi$

denote the parameter of interest and parameters explaining the missingness. The parameter of interest corresponds to the main model's parameters which provides the employee turnover probability predictions. The parameters explaining the missingness depends on the missing mechanisms that will be further described in the next section. For instance, it might be a constant reflecting the missingness probability of a partially missing variable as shown in Van Buuren [6, Chapter 2]. $\psi_j^t$ corresponds to the parameters of the parametric model that will be used for the estimation of missing values of the $jth$ variable at iteration $t$ and $\dot{\psi}_j^t$ is the sample drawn out of the conditional distribution specified for $\psi_j^t$. Lastly, $R$ denotes the non-response matrix of data $Y$ where '1' and '0' denote the observed and missing values respectively.

## 4.2 Missing Data Mechanisms

Van Buuren [6, Chapter 1,2] describes the mechanisms leading to missingness as the following:

MCAR: The data are Missing Completely at Random if the missingness depends on parameters $\phi$, but are conditionally independent of observed and missing values. This implies that the process leading to missingness is not related to the information in the data, but it could be explained by some parameters $\phi$

$$p(R = 0|Y_{miss}, Y_{obs}, \phi) = p(R = 0|\phi). \tag{4.1}$$

MAR: The data are Missing at Random if the distribution of missingness can be explained by parameters $\phi$ and $Y_{obs}$ so the missingness still depends on the observed data conditional on $\phi$. It is essential to note that $R$ is conditionally independent of $Y_{miss}$ when $Y_{obs}$ and $\phi$ are given. Then, the assumption is that the observed part of the data is sufficient to explain the missingness

$$p(R = 0|Y_{miss}, Y_{obs}, \phi) = p(R = 0|\phi, Y_{obs}). \tag{4.2}$$

MNAR: The data are Missing Not Completely at Random if the distribution

of missingness still depends on unobserved values, $Y_{miss}$, given $\phi$ and $Y_{obs}$

$$p(R = 0|Y_{miss}, Y_{obs}, \phi). \tag{4.3}$$

The expression above cannot be simplified. The importance of missing mechanisms root on the need to ignore the missing mechanism leading to missingness, i.e. $\phi$ [6, Chapter 1,2].

## 4.3 Ignorability

The observed data likelihood can be defined as the combination of $R$ and $Y$, i.e. $p(Y_{obs}, R|\phi, \theta) = \int p(Y, R|\phi, \theta)dY_{miss}$ [10, Chapter 8.2]. Since $R$ matrix contains information which might affect the inference of $\theta$, $\phi$ is also contained in the likelihood equation [10, Chapter 8.2]. However, the inference of parameters $\phi$ is not of interest of this study since the aim is to infer $\theta$. Then determining the conditions to ignore the missing mechanism $\phi$ would simplify the analysis [6, Chapter 1,2].

Gelman et al. [10, Chapter 8.2] presents the conditions to assume that $\phi$ is ignorable and proves that the ignorability can be ensured under the given conditions. Let $p(\theta, \phi|Y_{obs}, R)$ be the posterior distribution of the parameters. Then

$$p(\theta, \phi|R, Y_{obs}) = \frac{p(\theta, \phi)p(R, Y_{obs}|\theta, \phi)}{p(R, Y_{obs})}. \tag{4.4}$$

where the first and second term on the nominator are the prior and likelihood. Then

$$\begin{aligned} p(\theta|R, Y_{obs}) &= \frac{\int p(\theta, \phi)p(R, Y_{obs}|\theta, \phi)d\phi}{p(R, Y_{obs})} \\ &= \frac{\int p(\theta, \phi) \int p(R, Y|\theta, \phi)dY_{miss}d\phi}{p(R, Y_{obs})} \\ &= \frac{\int p(\theta, \phi) \int p(Y|\theta, \phi)p(R|Y, \phi, \theta)dY_{miss}d\phi}{p(R, Y_{obs})} \\ &= \frac{\int p(\theta, \phi) \int p(Y|\theta)p(R|Y, \phi)dY_{miss}d\phi}{p(R, Y_{obs})}. \end{aligned} \tag{4.5}$$

Let's assume the MAR assumption holds and $p(\phi, \theta) = p(\phi)p(\theta)$. Then the missing data mechanism is said to be ignorable for Bayesian inference [12]. In many cases, the latter condition holds as $\theta$ provides little information about $\phi$ or vice versa [13]. Then

$$\begin{aligned} p(\theta|R, Y_{obs}) &= \frac{\int p(\theta,\phi) \int p(Y|\theta)p(R|Y,\phi)dY_{miss}d\phi}{p(R, Y_{obs})} \\ &= \frac{\int p(\theta,\phi)p(R|Y_{obs},\phi) \int p(Y|\theta)dY_{miss}d\phi}{p(R, Y_{obs})} \\ &= \frac{\int p(\theta,\phi)p(R|Y_{obs},\phi)p(Y_{obs}|\theta)d\phi}{p(R, Y_{obs})} \\ &= \frac{\int p(R,\phi|Y_{obs})p(Y_{obs},\theta)d\phi}{p(R, Y_{obs})} \\ &= \frac{p(R|Y_{obs})p(Y_{obs},\theta)}{p(R|Y_{obs})p(Y_{obs})} = p(\theta|Y_{obs}). \end{aligned} \tag{4.6}$$

The second equation is by $p(R|Y, \phi) = p(R|Y_{obs}, \phi)$ and the forth equation is by $p(\phi, \theta) = p(\phi)p(\theta)$. It can be concluded that the inference regarding $\theta|Y_{obs}$ does not depend on $R$ so missing data mechanism can be ignored [10, Chapter 8.2].

## 4.4 Missing Mechanism of the Data at Hand

If $Y_{obs}$ depends on $R$, it might be assumed that missing mechanism is not MCAR as demonstrated in Buuren and Groothuis-Oudshoorn [7]. Since the non-response cannot be explained by a set of constants $\phi$, where observed values are still helpful in explaining the distribution of $R$. Therefore, the dependence between any partially missing variable's missingness and any other observed variable's values might point out that MCAR is unlikely to be the mechanism leading to missingness. Although the dependence between missingness and observed values might be helpful to decide if the missing mechanism is MCAR or not, it is not feasible to differentiate MAR and MNAR since the necessary information to distinguish them is already missing [6, Chapter 1,2]. There are two suggestions proposed in Van Buuren [6, Chapter 6] when the missing mechanism is assumed to be MNAR. Either expanding the data so that the mechanism is expected to get

Table 4.1: Dependence between Missingness and Observed Variables

| Missingness versus Variable | p-value |
|---|---|
| College.Department-Level.of.English.Proficiency | 0.01 |
| Alma.Mater-Level.of.English.Proficiency | 0.00 |
| Highest.Degree-Level.of.English.Proficiency | 0.01 |
| Level.of.English.Proficiency-Workplace | 0.00 |
| Performance.Score-Workplace | 0.00 |
| Appreciation.Score-Workplace | 0.00 |

closer to MAR or setting a model handling MNAR and making sensitivity analysis on parameters of interest as illustrated in Van Buuren et al. [8] are plausible to handle MNAR. Setting a model which is sensitive to MNAR is based on the finding, $p(Y|Y_{obs}, R = 1) \neq p(Y|Y_{obs}, R = 0)$ [6, Chapter 1,2]. For the turnover data set, assigning distinct distributions on missing and observed values is not plausible due to the number of missing variables and the complexity of the distributions the missing values might have as opposed to the case in Van Buuren et al. [8] where the missing blood pressures' distribution are modeled by only shifting the mean of the observed values' distribution. Also, adding new variables to the missing imputation model is not a viable option due to the pre designed data collection process of the company. At this point, MAR assumption might be seen as an inevitable starting point for imputation.

To see if the missing mechanism might be MCAR, the dependence between each missing variable's missingness and other observed variables' values should be investigated. Then, Chi-Square Test of Independence might be useful for this purpose.

Table 4.1 demonstrates the p-values of Chi-Square Test of Independence between the missing indicators of partially missing variables and observed values of missing variables/fully observed variables. Since the missing indicator of each partially missing variable depend on at least one of the variable's observed values at any reasonable amount of significance level, the missing mechanism is not expected to be MCAR. As an instance to the exploratory analysis regarding missingness, employees working in places other than the factory are more likely to have missing values in English Proficiency which might be justified by job descriptions

since employees working outside the factory are less likely to use English in their day to day tasks where they are mostly required to do local tasks. Then, there might have been less desire to follow up with these employees' qualification of English.

## 4.5 Ad-hoc Methods

Before discussing the proposed imputation model, it is necessary to mention some of the ad-hoc methods presented by Van Buuren [6, Chapter 1] and Gelman and Hill [14, Chapter 25] to handle the missing data. These methods are attractive due to their easiness in terms of applicability.

One of the most popular approach among ad-hoc methods is complete case analysis/ pairwise deletion/ available case analysis where partially missing observations are removed and the inferences are made on complete data which may lead to waste the available information at hand [6, Chapter 1]. Also, this type of analysis is only suitable under MCAR assumption as suggested by Van Buuren [6, Chapter 1] which does not seem to be the missing mechanism leading to missingness for the data at hand. Listwise deletion is a method that treats data as if the observations have a multivariate normal distribution and make the imputation based on the covariances and means estimated out of the observed data. This type of treatment is unsuitable for the data containing categorical variables. Another method is to fill in the missing values by the observed values' mean or mode depending on the variable type. This method might bring biased parameter inferences unless the missing mechanism is MCAR [6, Chapter 1]. Deterministic or Stochastic Regression are also available choices to impute the data which might provide unbiased mean estimates of parameter of interest under MAR [6, Chapter 1]. However, only fully observed variables can be used to impute the partially missing variables which might weaken the MAR assumption since the other partially missing variables can still help explaining the missingness.

Although these methods can be easily applied, they are either not plausible

to apply for the data at hand or might lead to biased inferences due to their incapability of explaining the missing mechanism.

## 4.6 Multivariate Imputation by Chained Equations

Buuren and Groothuis-Oudshoorn [7] notes that one way of imputing missing data is Joint Modeling where a multivariate distribution is assigned to the variables and the other way is to impute each partially missing variable in a univariate manner using Fully Conditional Specification (MICE). The latter's feature of univariate modeling is what makes MICE desirable since it allows the user to specify conditional distribution for each incomplete variable without being attached to the multivariate structure [7]. MICE iteratively imputes each missing variable one by one conditional on others, imputed and not missing, after the initial imputation where the algorithm usually converges after 10-20 imputations [7]. The imputation is repeated $n \in \mathbb{Z}^+$ times so that the difference in the imputed values reflect the uncertainty of the values to be imputed and the main model, which would have been implemented if the data was complete, is implemented for each replication [7]. Then, the estimates are pooled for final estimates as described in Van Buuren et al. [8]. The steps of the model is presented in Van Buuren [6, Chapter 4.5] as the following:

1. A conditional model, $Y_{miss,j}|Y_{obs,j}, Y_{-j}, R$, is specified for each variable $j$.
2. Missing values are initially imputed for each variable $j$ by taking random draws from $Y_{obs,j}$.
3. $\dot{\psi}_j^t$ is drawn out of $p(\psi_j^t|Y_{obs,j}, \dot{Y}_{-j}^t, R)$ where $t$ denotes the iteration number and $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \ldots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \ldots, \dot{Y}_p^{t-1})$ where $\dot{Y}_j^t$ denotes the imputed $jth$ variable for $tth$ iteration.
4. $\dot{Y}_{miss,j}^t$ is drawn out of $p(Y_{miss,j}|Y_{obs,j}, \dot{Y}_{-j}^t, R, \dot{\psi}_j^t)$.
5. Third and forth steps are consecutively repeated for each variable which corresponds to one iteration when each incomplete variable is imputed once.

At this point, $R$ can be ignored when making the imputations by the assumption of MAR. MICE algorithm can be considered as an MCMC(Markov Chain Monte Carlo) algorithm where the parameter space is the collection of missing values and if all the conditionals are compatible, i.e. a joint distribution exists, the algorithm yields to Gibbs Sampler [6, Chapter 4]. Although the joint distribution may not exist due to incompatibility of conditionals, incompatibility issues is not all that concerning if the conditionals impute the data coherent with the missing mechanism [6, Chapter 4].

## 4.7 Conditional Distributions of Partially Missing Variables

Van Buuren [6, Chapter 3] introduces some of the conditional distributions that could be incorporated into MICE Algorithm which were implemented in *mice* function of *mice* package in *R*.

For numerical variables to be imputed, Predictive Mean Matching (PMM) was set as the default choice of the function *mice*. The main idea behind this conditional specification is to fill in the missing values by observed values. The steps of the model is as the following:

1. For partially missing variable $j$, $S = X^\top X$ is computed where $X$ is consisted of the observations where the $jth$ variable is observed and the columns that will be used to predict $jth$ variable.
2. $V = (S + diag(S)\kappa)^{-1}$ is calculated for some small $\kappa$ so that the covariance matrix is ensured to be positive definite.
3. The least square is calculated as $\hat{\beta} = VX^\top y$ where $y$ is the response of the regression.
4. By assuming the normality of the error term in linear regression equation and constancy of the error term's standard deviation, $\dot{\sigma}^2 = (y - X\hat{\beta})^\top (y - X\hat{\beta})/\dot{g}$ can be randomized where $g \sim \chi^2_{n-k}$. This is due to $(y - X\hat{\beta})^\top (y -$

$X\hat{\beta})/\sigma^2 \sim \chi^2_{n-k}$ where $k$ is the dimension of $\beta$.

5. $\dot{B}$ is drawn out of $Normal(\hat{\beta}, \dot{\sigma}^2 V)$ by calculating $\dot{B} = \hat{\beta} + \dot{\sigma}\dot{z}V^{1/2}$ where $\dot{z}$ and $V^{1/2}$ are standard multivariate normal sample and Cholesky Decomposition of the unnormalized covariance matrix.
6. For $ith$ missing observation of variable $j$, $|X_i\hat{\beta} - X_k\dot{\beta}|$ is calculated for each $k$ indexing the rows of $X$.
7. 5 of $k$s having minimum $|X_i\hat{\beta} - X_k\dot{\beta}|$ are selected and one of those are randomly chosen as the imputed value. The selected candidates, which are expected to reflect the distribution of $ith$ missing value, are called the donors within the context of PMM.
8. The sixth and seventh steps are repeated for each missing observation of variable $j$.

As Van Buuren [6, Chapter 3] suggests, instead of using $|X_i\hat{\beta} - X_k\hat{\beta}|$ or $|X_i\dot{\beta} - X_k\dot{\beta}|$, $|X_i\hat{\beta} - X_k\dot{\beta}|$ might be more suitable in terms of selecting the donors since $|X_i\hat{\beta} - X_k\hat{\beta}|$ does not reflect the uncertainty of the values to be imputed because the parameter estimates are assumed to be certain [6, Chapter 3]. On the other hand, for $|X_i\dot{\beta} - X_k\dot{\beta}|$, $\dot{\beta}$ cancels out especially when the dimension of $X$ is small so the same donors are chosen at each iteration which might also block to reflect the uncertainty of the values to be imputed [6, Chapter 3]. The number of donors was set to 5 based on the simulation results mentioned by the author and some other distance metrics as an alternative to Euclidian distance are available to be used to select the donors. It is also suggested that PMM has the advantage of eliminating post processing the imputed values which might occur when the imputed values do not make sense such as negative salary etc. Therefore, PMM remained as the choice to impute the numeric variables.

In a similar manner, the binary categorical variables are suggested to be imputed using a Bayesian Logistic Regression as the following:

- For variable $j$, $\hat{\beta}$ and $V$ are calculated similar to PMM.
- Similarly, $\dot{B}$ is drawn out of $Normal(\hat{\beta}, V)$ by calculating $\dot{B} = \hat{\beta} + \dot{z}V^{1/2}$ where $\dot{z}$ and $V^{1/2}$ are standard multivariate normal sample and Cholesky

Decomposition of the unnormalized covariance matrix.

- $\dot{p} = (1 + e^{-X_i^\top \dot{\beta}})^{-1}$ is calculated and if $\dot{u}$ sampled from $Uniform(0, 1)$ is greater than $\dot{p}$, the value is set to 1, otherwise, it is set to 0. Sampling $\dot{\beta}$ reflects the uncertainty of the outcome which will allow unlikely values to appear in some of the replications.

Similarly, for variables with more than two categories, the algorithm described above is followed in the form of Multinomial Regression [6, Chapter 3]. There are other methods that *mice* package provides to impute the categorical variables such as Linear Discriminant Analysis. However, *polyreg* is superior to *lda* which are the function calls for Bayesian Polytomous Regression and Linear Discriminant Analysis [6, Chapter 6]. In addition to superiority, it seemed favorable to reflect the uncertainty of the missing values in the replications as algorithms such as linear discriminant analysis cannot reflect the uncertainty in missing values as suggested in Van Buuren [6, Chapter 6]. In addition, PMM seemed superior to the Bayesian Normal Linear Regression since it annihilates the post processing that might be needed.

## 4.8 Tuning Mice

There are a couple of choices that must be made before building the model as pointed out in Buuren and Groothuis-Oudshoorn [7]. The first choice to make is to determine the conditional distribution that will be assigned to each partially missing variable which was already discussed in the previous section. Another choice is to select the variables that will be used in the imputation of each partially missing variable. All variables appearing in the main model should be added to the imputation model to avoid biased inferences [7]. However, since some of the variables contained many categories which triggered perfect prediction, as discussed in Buuren and Groothuis-Oudshoorn [7], each incomplete variable was assigned predictors considering explaining the missingness and the predictive power. Another important choice to be made is to select the number of iterations

where convergence is reached which will be further discussed in the following section based on the diagnostics of the generated data sets. Lastly, the number of replicated data sets is crucial to reflect uncertainty in missing values [6, Chapter 3]. Then, the larger the number of replications are generated, the better the uncertainty is reflected. Therefore, the number of replicated data sets will be decided based on the computation time of the algorithm since the main algorithm will be run for each data set. More replications of imputed data will better reflect uncertainty in missing values so using replications as many as possible is favorable by accounting for the computation time.

The imputation model is expected to deal with both demographic and time-dependent data. It is infeasible to impute both demographic and time-dependent variables simultaneously at the same iteration because the algorithm might impute different values for demographics of employees. For instance, an employee might have different College Department values within different time periods. To prevent that, the data was processed accordingly at each iteration where each observation belongs to an employee in estimating demographics and the data is reverted back to panel format for time dependent variables. Then, average and mode of numeric and categorical variables per employee for time-dependent variables were computed so that time-dependent variables can represent "demographic like" information and finally, these, so called "demographic like", variables were used all together with the demographic variables to predict demographic variables. These steps were repeated as many as the number of iterations.

Lastly, the imputation was made with respect to the monotonicity of the missingness meaning observed values are used firstly to impute the missing as suggested in Buuren and Groothuis-Oudshoorn [7]. For instance, if a value in Alma Mater is missing, it is guaranteed that the English Proficiency is also missing. Then, fitting a model for English Proficiency regressed on Alma Mater at first rather than fitting a model for Alma Mater regressed on English Proficiency, which will lead to fitting observed values on imputed values, is more reasonable.

## 4.9 Diagnostics

Although there is no definite way to determine if the algorithm converged, imputed values' statistics might be plotted against iteration number [7]. For categorical variables, each category's fraction was plotted against iterations. For variable Salary, the mean and standard deviation of imputed values were also examined against iteration number as in Buuren and Groothuis-Oudshoorn [7]. Although this way of monitoring convergence may not provide any idea about each missing value's convergence because it is assumed that each missing value has a distinct distribution, a picture of lack of convergence might indicate that the values might not have converged.

By the plots 4.1 and 4.2, the imputed values do not exhibit a lack of convergence case.

## 4.10 Notes on Implementation of Mice

- The numeric variables such as Number of Years Worked and Age were converted into categorical form since there might not be linearly increasing or decreasing relationship between these variables and the others. The Salary variable was kept continuous. The reason being is that Salary Raise is computed using Salary and it was not feasible to transform Salary to categorical form and preserve the Salary Raise's, which is a strong predictor of the outcome, information within the categories of Salary.

- To estimate Salary, average of observed Salary across years was computed per employee so that the samples drawn out of conditional distribution are reasonable. For instance, some employees' salary column contained both observed and missing values depending on the year. For some of these employees, a decrease in salary for future periods was observed in the imputed data sets which did not make sense because HR Department noted that lowering wage was not a used practice within the company. Therefore,

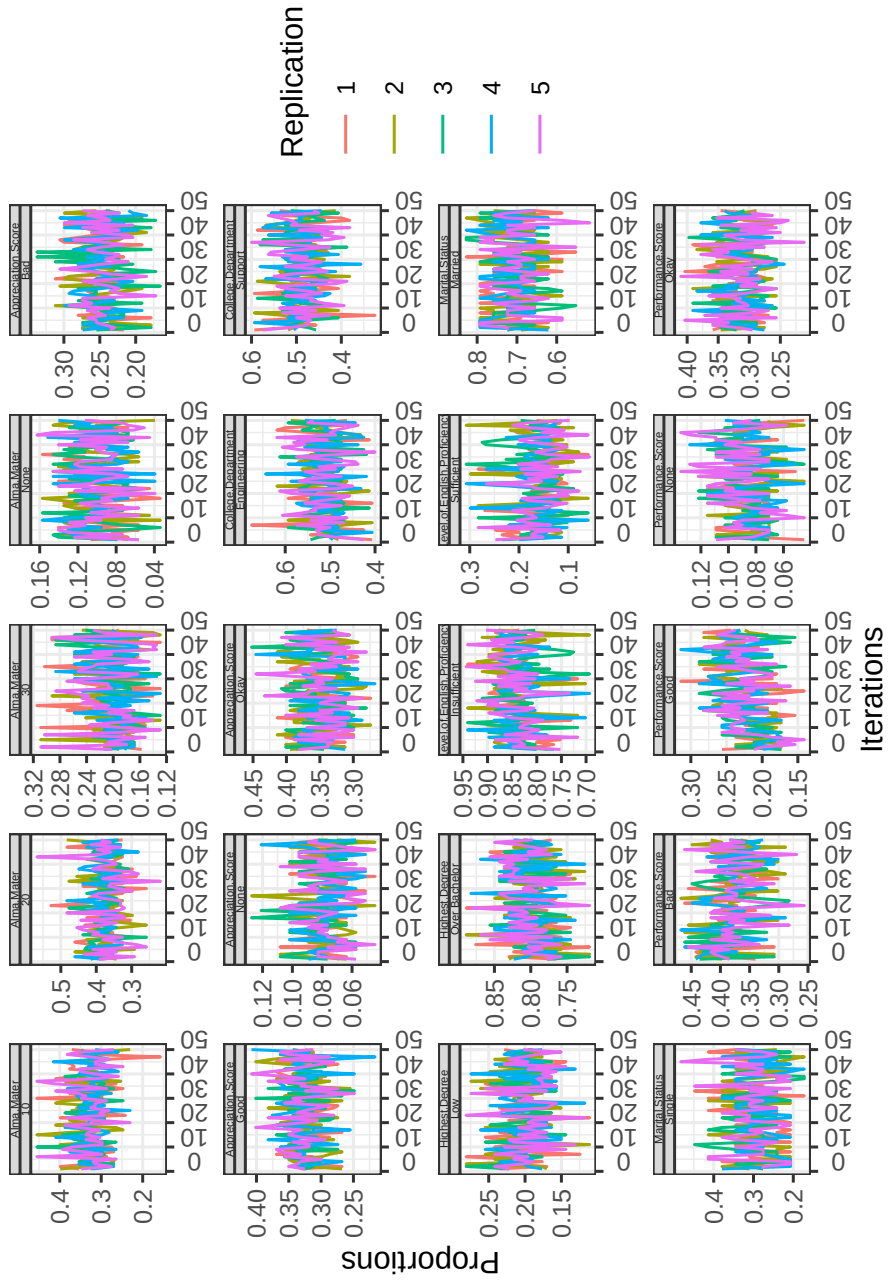


Figure 4.1: Categorical Variables Fractions versus Iterations

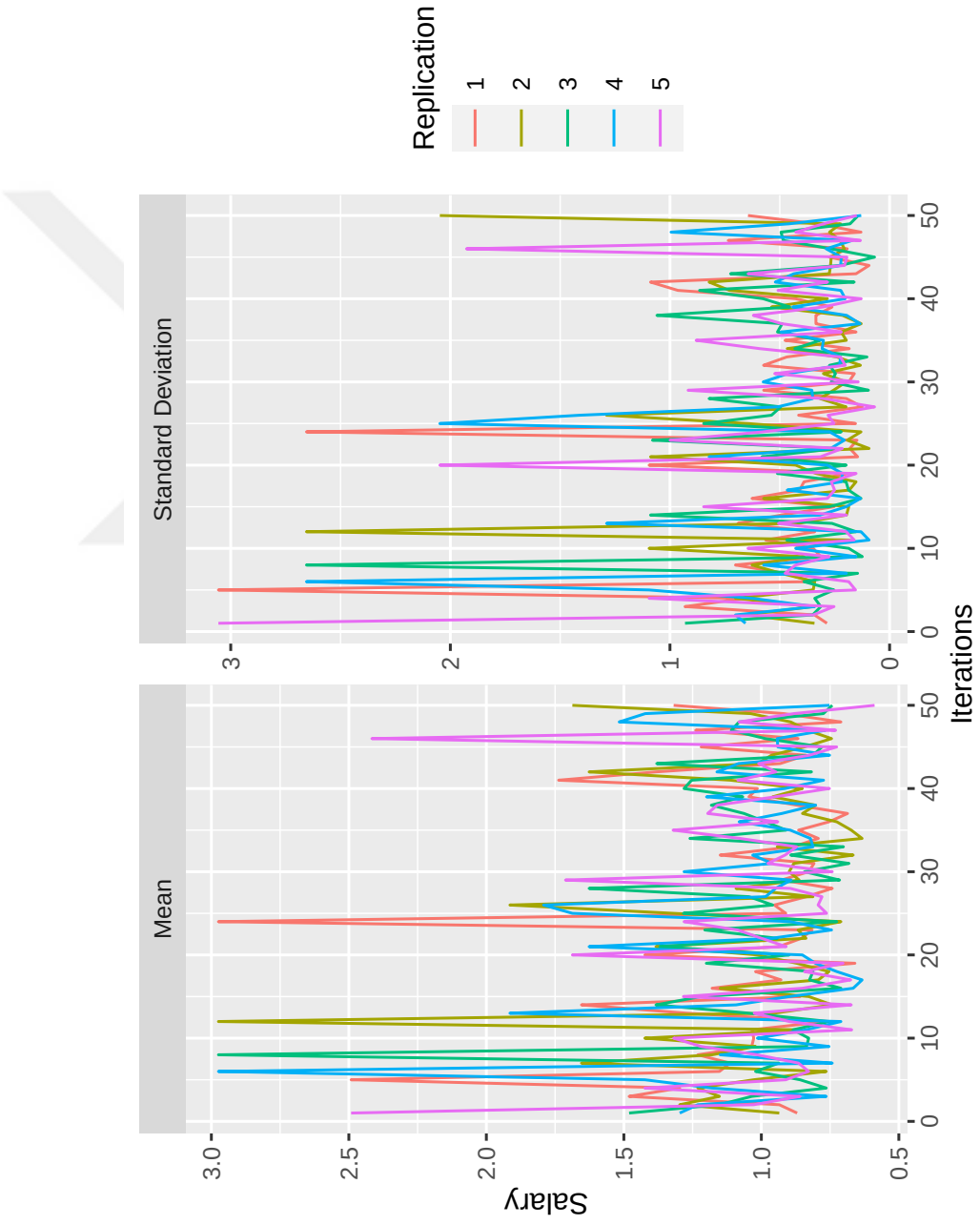


Figure 4.2: Mean and Standard Deviation of Salary versus Iterations

Average Salary was used in the estimation of Salary.

- After imputing the data for 2015, the records belonging to this year were removed. Since, there are variables such as Salary Raise using the previous year's information and the data for 2014 is not available.

# Chapter 5

# Bayesian Hierarchical Model

In this section, an introductory part concerning the notions in Bayesian Hierarchical Models Literature will be covered. The claims and arguments in this section are based on Rossi et al. [1, Chapter 2] where the introductory mechanics of Bayesian statistics are discussed. The primary focus of Bayesian study is to treat parameters as random as opposed to the methods in classical statistics and similar to classical approaches, the study is concentrated on making inferences regarding the parameters [1, Chapter 2]. The inferences regarding parameters are based on the information provided by the data and apriori knowledge or assumptions on the distribution of the parameters. Then, the posterior distribution of the set of parameters, $\theta$, given the prior assumptions and the data is as indicated below by Bayes Rule

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \propto p(\theta)p(y|\theta), \tag{5.1}$$

where $p(\theta|y)$, $p(\theta)$ and $p(y|\theta)$ are posterior and prior distribution and likelihood respectively. $p(\theta|y) \propto p(\theta)p(y|\theta)$ is mostly the interest of the study because $p(y)$ is a constant normalizing the posterior distribution so that it integrates out to 1. In other words, if $p(\theta|y) \propto f(\theta|y)$ and $f(\theta|y)$ is the density of a known distribution, it means the ignored constants cancel out to 1, i.e. $p(\theta|y) = f(\theta|y)$. The main interest is to find the expectation of a function, such as moments, quantiles or

marginals, of $\theta|y$ [1, Chapter 2], i.e.

$$E_{\theta|y}(h(\theta)) = \frac{\int h(\theta)p(\theta)p(y|\theta)d\theta}{\int p(\theta)p(y|\theta)d\theta}. \tag{5.2}$$

In particular, it is not feasible to integrate out the parameters and come up with a closed form solution for "sophisticated" models due to the dimension of parameters and complex dependence structure between the components of $\theta$. Then, sampling from $p(\theta|y)$ might make sense to approximate $E_{\theta|y}(h(\theta)) \approx \frac{\sum_{t=1}^{n} h(\theta^t)}{n}$ by the Law of Large Numbers [1, Chapter 3].

$E_{\theta|y}(h(\theta))$ is feasible to be approximated by using the importance sampling which produces independent samples drawn out of the target distribution (See Rossi et al. [1, Chapter 2] for the importance sampling). However, it is likely to come up with an importance density not capturing the support of the posterior density which might lead to estimate the posterior parameter of interest with great precision although the massive region of the posterior is missed by the importance density which might be impossible to detect for problems with high dimensional parameter space [1, Chapter 2]. In hierarchical models, the dimension of parameters is too large to handle the inference problem with importance sampling as Rossi et al. [1, Chapter 2] suggests using importance sampling when the number of parameters is up to 20.

Another approach to sample from the posterior distribution is to lay down each axis of the parameter space into grids and compute the posterior density at each of those points in the parameters space and sample based on the density values of the points on the intersection of the grids as suggested in Rossi et al. [1, Chapter 3]. However, this approach suffers from the curse of dimensionality, since the number of points to be evaluated increases drastically as the number of dimensions increases [1, Chapter 3]. For instance, let's assume the parameter space is 10 dimensional and each axis is layed down into 10 grids. Then, the posterior density to be evaluated would be $10^{10}$. Due to incapability of these approaches to sample from the posterior when the parameters space is large, MCMC (Monte Carlo Markov Chain) are used to sample from the posterior distribution. MCMC samplers are constructed on the parameter space of $\theta$ to sample from the posterior distribution and if it is assured that MCMC sampler converges to

posterior distribution, then, the samples drawn out of posterior distribution can be attained via MCMC [1, Chapter 3]. Then, these samples could be used to approximate the function of interest of $\theta$.

## 5.1 Monte Carlo Markov Chains (MCMC)

The arguments for this section and the sections regarding the sampling methods are based on the arguments in Rossi et al. [1, Chapter 2,3]. MCMC samplers are Markov Chains constructed on the parameter space and are used to sample from the concerned distribution to approximate the aforementioned integrals. By initially drawing $\theta_0$ from any point in the parameters space, $\theta_r|F(\theta_{r-1})$ is drawn consecutively where $F$ denotes the conditional distribution of the parameters given the last state of the sampler [1, Chapter 3].

As suggested in Rossi et al. [1, Chapter 3], let $\pi$ and $\Theta$ denote the posterior distribution and the parameter space of $\theta$ and $K(\theta, A)$ is defined as the Kernel that maps the last state of the sampler, $\theta$, to $A \in \Theta$ such that $\pi(A) > 0$, i.e. $K(\theta, A) = \int_A p(\theta, v)dv$. Then, if it is assumed that $K^n(\theta, A) > 0$ for $n \in Z^+$ , i.e the chain is likely to move to the set $A$ after finite number of draws from any state in the parameter state, then the chain, $K$, is said to be irreducible with respect to $\pi$ which implies the chain converges to some stationary distribution $\omega$ [1, Chapter 3]. Furthermore, if $\pi(\theta)p(\theta, v) = \pi(v)p(v, \theta)$, i.e. the chain is time-reversible with respect to $\pi$, the stationary distribution of the chain, $\omega$, is equivalent to the posterior distribution, $\pi$ [1, Chapter 3]. Then, under the given conditions, it is implied that the sampler converges to the posterior distribution as $r$ tends to $\infty$ [1, Chapter 3]. This is actually what makes MCMC desirable since it enables to sample from the posterior distribution so that the expectations of functions of parameters can be approximated but the problem with MCMC chains is that the consequent iterations might depend even after hundreds of thousands of iterations which violates Week Law of Large Numbers assuming independence of samples [1, Chapter 3]. However, the sample averages are still expected to converge to the expectation by assuming the sampler's ergodicity [1,

Chapter 3].

Using MCMC methods is problematic in the sense that consequent draws from the sampler will produce correlated samples so the dependence between the draws makes the samples less informative of the posterior. Then, the amount of dependency between consequent draws will reveal the efficiency of the sampler which will be enlarged upon in diagnostics section. In addition, the initial values of the sampler might not reflect the posterior distribution if the initial samples of the sampler is far away from the substantial mass of the posterior distribution. These initial draws are called burnin iterations and a considerable amount of iterations might be disposed depending on the pace of the convergence of MCMC.

## 5.2 Gibbs Sampler

Gibbs Sampler is one of the sampling algorithms enabling to sample from the target distribution by using Markov Chains. As suggested in Rossi et al. [1, Chapter 3], let's assume that it is desired to sample from a multi-dimensional random variable $\theta$ and it is difficult to directly sample from $\theta \sim p(\theta)$ but it is available to sample from closed form conditional distributions of $k$ sets of parameters partitioning $\theta$, i.e. $\theta = (\theta_1, \ldots, \theta_k)$. By iteratively drawing $\theta_i \quad i \in \{1, \ldots, k\}$ from the conditional distributions, i.e.

$$\begin{aligned}
\theta_1^{t+1} &\sim p(\theta_1|\theta_2^t, \ldots, \theta_k^t) \\
\theta_2^{t+1} &\sim p(\theta_2|\theta_1^{t+1}, \theta_3^t, \ldots, \theta_k^t) \\
&\vdots \\
\theta_k^{t+1} &\sim p(\theta_k|\theta_1^{t+1}, \ldots, \theta_{k-1}^{t+1}),
\end{aligned} \tag{5.3}$$

where $t$ denotes the iteration number, MCMC converges to $p(\theta)$ as the sampler is irreducible [1, Chapter 3]. It is suggested that if the parameter space is the Cartesian product of intervals and the posterior density is strictly positive at any point in the parameter space, the chain is irreducible which implies the

convergence of the sampler to the invariant distribution for any starting point in the parameter space. Then, the sampler converge to $p(\theta)$ regardless of the starting point of the chain [1, Chapter 3].

## 5.3 Metropolis Algorithms

Although Gibbs Sampler is easy to sample from, since the conditionals are of known type, the conditionals might not always be available to directly sample from [1, Chapter 3]. Then Metropolis Algorithm provides a generic tool to sample from any form of posterior distribution [1, Chapter 3]. As suggested by Rossi et al. [1, Chapter 3], let $q(\theta, v)$ denote the transition function of the sampler given $\theta$. This transition function is specified by the user and determines the efficiency of the sampler which will be discussed after the model is provided. The steps of the continuous state space Metropolis Algorithm is presented as the following. A new sample is drawn from the conditional distribution $\theta_1 \sim q(\theta_0, v)$ where $\theta_0$ is the initial value of the parameter vector. With probability $\alpha$, the chain moves to the state $\theta_1$, otherwise with probability $1 - \alpha$, the chain remains at the state $\theta_0$ where $\alpha(\theta_0, \theta_1) = min\{1, \pi(\theta_1)q(\theta_1, \theta_0)/\pi(\theta_0)q(\theta_0, \theta_1)\}$. These steps correspond to one sample and are repeated until the the specified number of iterations is reached. Then, with the given configuration of the sampler, the algorithm has a stationary distribution if the chain is irreducible, i.e. $q(\theta, v) > 0$, and aperiodic, $\alpha(\theta, v) > 0$ [1, Chapter 3]. In addition, the kernel of Metropolis chains assure time-reversibility with respect to $\pi$ so if it is ensured that the chain converges to a stationary distribution, the chain converges to the posterior distribution by the construction of $\alpha$ which ensures the time-reversibility [1, Chapter 3].

## 5.4 Asymptotic Approximation of the Posterior Distribution

Rossi et al. [1, Chapter 2] presents another way of approximating the integrals as the following:

$$\begin{aligned} I &= \int h(\theta)p(\theta)p(D|\theta)d\theta \\ &= \int h(\theta)p(\theta)e^{L(\hat{\theta})+G^\top(\theta-\hat{\theta})+(\theta-\hat{\theta})^\top H(\theta-\hat{\theta})}d\theta \\ &\cong \int h(\theta)p(\theta)e^{L(\hat{\theta})+\frac{1}{2}(\theta-\hat{\theta})^\top H(\theta-\hat{\theta})}d\theta, \end{aligned} \tag{5.4}$$

where $\hat{\theta}$ is the maximum likelihood estimator, $G = \frac{\partial L(\hat{\theta})}{\partial \theta}$,$H = \frac{\partial^2 L(\hat{\theta})}{\partial \theta \partial \theta^\top}$ and $L$ denotes the $log-likelihood$ function. To use the Taylor Approximation in the second equation, the likelihood is assumed to be continuously differentiable. Third equation is by the gradient being 0 at the maximum. By using a Normal Prior on $\theta$,

$$I = e^{L(\hat{\theta})} \int h(\theta)(2\pi)^{-k/2}|A|^{1/2}e^{-\frac{1}{2}\left(\theta-\bar{\theta}\right)^\top A\left(\theta-\bar{\theta}\right)}e^{-\frac{1}{2}\left(\theta-\hat{\theta}\right)^\top H^*\left(\theta-\hat{\theta}\right)}d\theta, \tag{5.5}$$

where the $\bar{\theta}$ and $A$ are the mean vector and precision matrix of the normal prior and $H^* = -H$. By $(x-\mu_1)^\top A_1(x-\mu_1) + (x-\mu_2)^\top A_2(x-\mu_2) = (x-\tilde{\mu})^\top(A_1+A_2)(x-\tilde{\mu}) + (\mu_1-\tilde{\mu})^\top A_1(\mu_1-\tilde{\mu}) + (\mu_2-\tilde{\mu})^\top A_2(\mu_2-\tilde{\mu})$ where $\tilde{\mu} = (A_1+A_2)^{-1}(A_1\mu_1 + A_2\mu_2)$,

$$\begin{aligned} I = e^{L(\hat{\theta})-\frac{1}{2}\left(\tilde{\theta}-\bar{\theta}\right)^\top A\left(\tilde{\theta}-\bar{\theta}\right)-\frac{1}{2}\left(\tilde{\theta}-\hat{\theta}\right)^\top H^*\left(\tilde{\theta}-\hat{\theta}\right)}|A|^{1/2}|A+H^*|^{-1/2} \\ \int h(\theta)(2\pi)^{-k/2}|A+H^*|^{1/2}e^{-\frac{1}{2}\left(\theta-\tilde{\theta}\right)^\top (A+H^*)\left(\theta-\tilde{\theta}\right)}d\theta. \end{aligned} \tag{5.6}$$

Then, when $h(\theta) = 1$, $I = \int p(\theta)p(D|\theta)d\theta$ would be equal to the the normalizing constant before the integral in the Equation (5.6) since the integral integrates out to 1. Then, $\theta|D \sim Normal(\tilde{\theta}, (A+H^*)^{-1})$. This approximation suffers from the curse of dimensionality and is not useful if the dimension of the parameters is more than a few [1, Chapter 2].

## 5.5 Heterogenous Modeling for Units

Employees have different demographic attributes such as alma mater, gender etc. So as expected, they might have heterogeneous sensitivity concerning attrition. Then, assigning separate parameters for employees might make sense to be able to model their characteristic differences. As suggested by Rossi et al. [1, Chapter 5], a general way of specifying the likelihood is to model each unit conditionally independent given the unit specific parameters, i.e. $p(y_1, \ldots, y_m | \prod_{i=1}^{m} \theta_i) = \prod_{i=1}^{m} p(y_i|\theta_i)$ where $m$ is the number of units so the posterior distribution of the parameters induced under a joint prior on unit parameters is as the following [1, Chapter 5]

$$p(\theta_1, \ldots, \theta_m | y_1, \ldots, y_m) \propto p(\theta_1, \ldots, \theta_m | \tau) \prod_{i=1}^{m} p(y_i|\theta_i), \tag{5.7}$$

where the first term is the joint prior on unit parameters conditional on hyperparameters, $\tau$ which are assumed to be constant at this point. It is substantial to note that the number of parameters is massive so the joint prior might be further simplified as $p(\theta_1, \ldots, \theta_m|\tau) = \prod_{i=1}^{m} p(\theta_i|\tau)$ for convenience [1, Chapter 5]. Then, the problem reduces to the evaluation of the hyperparameters of unit level parameters, $\tau$ which might be estimated as in the case of Zellner [15]. On the other hand, $\tau$ might also be assigned a distribution to preserve uncertainty in parameters. Then, the posterior becomes as the following which is presented in Rossi et al. [1, Chapter 5]

$$p(\theta_1, \ldots, \theta_m, \tau | y_1, \ldots, y_m, h) \propto p(\tau|h) \prod_{i=1}^{m} p(y_i|\theta_i) p(\theta_i|\tau). \tag{5.8}$$

This specification on the distribution of hyperparameters forms a hierarchical structure on parameters where first stage and the second stage prior are specified on unit level parameters and hyperparameters respectively [1, Chapter 5].

## 5.6 Multivariate Regression

Before diving into the model, Multivariate Regression Model(MRM) and the derivation of the natural conjugate priors for its parameters demonstrated in Rossi et al. [1, Chapter 2] must be mentioned. MRM natural conjugate prior structure is used in modeling the distribution of the second stage parameters.

Let's assume that the response of the regression is multivariate normal and the multivariate normal error of the response shares the same correlation structure across the observations. Let $\beta_i$ and $u_i$ denote the $k$ dimensional response and error vector for $ith$ observation. $z_i$ is the $n_z$ dimensional explanatory variables for $ith$ observation and $\Delta$ is the $n_z \times k$ dimensional matrix of regression coefficients. Lastly, $V_\beta$ denotes the correlation matrix shared by all the observations. In other words,

$$\beta_i = \Delta^\top z_i + u_i \quad u_i \sim N(0, V_\beta) \quad i \in \{1, \dots, m\}. \tag{5.9}$$

Then, the likelihood can be written as the following:

$$\begin{aligned} p(\beta_1, \dots, \beta_m | V_\beta, \Delta) &= \prod_{i=1}^{m} (2\pi)^{-\frac{k}{2}} |V_\beta|^{-\frac{1}{2}} e^{-\frac{1}{2}(\beta_i - \Delta^\top z_i)^\top V_\beta^{-1} (\beta_i - \Delta^\top z_i)} \\ &\propto |V_\beta|^{-\frac{m}{2}} e^{-\frac{1}{2}\sum_{i=1}^{m}(\beta_i - \Delta^\top z_i)^\top V_\beta^{-1}(\beta_i - \Delta^\top z_i)} \\ &= |V_\beta|^{-\frac{m}{2}} e^{-\frac{1}{2} tr\left((B - Z\Delta) V_\beta^{-1} (B - Z\Delta)^\top\right)} \\ &= |V_\beta|^{-\frac{m}{2}} e^{-\frac{1}{2} tr\left((B - Z\Delta)^\top (B - Z\Delta) V_\beta^{-1}\right)}, \end{aligned} \tag{5.10}$$

where $Z_{m \times n_z}$ and $B_{m \times k}$ denote the covariate matrix and response matrix and $tr$ denotes the trace of a matrix. Third equation is by the cyclic property of the trace. Let $\hat{\Delta} = (Z^\top Z)^{-1} Z^\top B$ which yields to the least square solution for given $B$ and $Z$. Then,

$$\begin{aligned} (Z\Delta)^\top (B - Z\hat{\Delta}) &= \Delta^\top Z^\top B - \Delta^\top Z^\top Z \hat{\Delta} \\ &= \Delta^\top Z^\top B - \Delta^\top Z^\top Z (Z^\top Z)^{-1} Z^\top B \\ &= 0, \end{aligned} \tag{5.11}$$

and

$$\begin{aligned}(Z\hat{\Delta})^\top(B - Z\hat{\Delta}) &= \hat{\Delta}^\top Z^\top B - \hat{\Delta}^\top Z^\top Z\hat{\Delta} \\ &= \hat{\Delta}^\top Z^\top B - \hat{\Delta}^\top Z^\top Z(Z^\top Z)^{-1}Z^\top B \\ &= 0,\end{aligned} \tag{5.12}$$

hold. Therefore, by considering $(B - Z\Delta)^\top(B - Z\Delta) = (Z\hat{\Delta} + (B - Z\hat{\Delta}) - Z\Delta)^\top(Z\hat{\Delta} + (B - Z\hat{\Delta}) - Z\Delta)$, the following would be obtained by Equations (5.11) and (5.12):

$$\begin{aligned}&(B - Z\Delta)^\top(B - Z\Delta) = (Z\hat{\Delta} + (B - Z\hat{\Delta}) - Z\Delta)^\top(Z\hat{\Delta} + (B - Z\hat{\Delta}) - Z\Delta) \\ &= \hat{\Delta}^\top Z^\top Z\hat{\Delta} - \hat{\Delta}^\top Z^\top Z\Delta + (B - Z\hat{\Delta})^\top(B - Z\hat{\Delta}) - \Delta^\top Z^\top Z\hat{\Delta} + \Delta^\top Z^\top Z\Delta \\ &= \underbrace{(B - Z\hat{\Delta})^\top(B - Z\hat{\Delta})}_{S} + (\hat{\Delta} - \Delta)^\top Z^\top Z(\hat{\Delta} - \Delta).\end{aligned} \tag{5.13}$$

Then, the likelihood would be

$$p(B|\Delta, V_\beta) \;\propto\; |V_\beta|^{-\frac{m-n_z}{2}} e^{tr\left(-\frac{1}{2}SV_\beta^{-1}\right)} |V_\beta|^{-\frac{n_z}{2}} e^{-\frac{1}{2}tr\left((\hat{\Delta}-\Delta)^\top Z^\top Z(\hat{\Delta}-\Delta)V_\beta^{-1}\right)}. \tag{5.14}$$

Then, by $tr(A^\top B) = vec(A)^\top vec(B)$ and $vec(ABC) = (C^\top \otimes A)vec(B)$, the exponential term on the right hand side of the likelihood would come to multivariate normal density form as follows:

$$\begin{aligned}tr\big((\hat{\Delta} - \Delta)^\top Z^\top Z(\hat{\Delta} - \Delta)V_\beta^{-1}\big) &= vec(\hat{\Delta} - \Delta)^\top vec\big(Z^\top Z(\hat{\Delta} - \Delta)V_\beta^{-1}\big) \\ &= vec(\hat{\Delta} - \Delta)^\top V_\beta^{-1} \otimes Z^\top Z vec(\hat{\Delta} - \Delta).\end{aligned} \tag{5.15}$$

Finally, the likelihood would be

$$p(B|\Delta, V_\beta) \;\propto\; \underbrace{|V_\beta|^{-\frac{m-n_z}{2}} e^{tr\left(-\frac{1}{2}SV_\beta^{-1}\right)}}_{I} \underbrace{|V_\beta|^{-\frac{n_z}{2}} e^{-\frac{1}{2}\left(vec(\hat{\Delta}-\Delta)^\top V_\beta^{-1} \otimes Z^\top Z vec(\hat{\Delta}-\Delta)\right)}}_{II}, \tag{5.16}$$

where $I$ depends on only $V_\beta$ so prior on $V_\beta$ should be proportional to $I$ to obtain a conjugate prior. Also, realize that $II$ is a Normal Kernel because it has the form of $vec(\Delta)|V_\beta$. Then, the joint conjugate prior would simplify as the following:

$$p(V_\beta, \Delta) = p(V_\beta)p(\Delta|V_\beta). \tag{5.17}$$

Finally, substituting the following priors for the likelihood will result in a posterior in the prior's form which is equivalent to say that the the prior is conjugate to the likelihood. In addition, since the likelihood and the prior has the same form, the conjugate prior is natural.

$$\begin{aligned} V_\beta &\sim IW(v, V) \\ vec(\Delta)|V_\beta &\sim N(vec(\bar{\Delta}), V_\beta \otimes A^{-1}). \end{aligned} \tag{5.18}$$

Then, the posterior distribution of the parameters would be

$$\begin{aligned} p(V_\beta, \Delta|B) &\propto p(B|\Delta, V_\beta)p(\Delta, V_\beta) = p(B|\Delta, V_\beta)p(V_\beta)p(\Delta|V_\beta) \\ &\propto |V_\beta|^{-\frac{m}{2}} e^{-\frac{1}{2}tr\left((B-Z\Delta)^\top (B-Z\Delta)V_\beta^{-1}\right)} \times |V_\beta|^{-\frac{(v+k+1)}{2}} e^{tr(-\frac{1}{2}VV_\beta^{-1})} \\ &\times |V_\beta|^{-\frac{n_z}{2}} e^{-\frac{1}{2}tr\left((\Delta-\bar{\Delta})^\top A(\Delta-\bar{\Delta})V_\beta^{-1}\right)}. \end{aligned} \tag{5.19}$$

Assume $A$, which is the precision matrix for $\Delta$, is positive definite so $A = U^\top U$ for some upper triangular matrix by Cholesky Decomposition. Then,

$$(\Delta - \bar{\Delta})^\top A(\Delta - \bar{\Delta}) + (B - Z\Delta)^\top (B - Z\Delta) = (P - R\Delta)^\top (P - R\Delta), \tag{5.20}$$

where

$$P = \begin{bmatrix} B \\ U\bar{\Delta} \end{bmatrix} \quad R = \begin{bmatrix} Z \\ U \end{bmatrix}. \tag{5.21}$$

Then, the least square equation for Equation (5.20) would be

$$\begin{aligned} \tilde{\Delta} &= (R^\top R)^{-1} R^\top P \\ &= (Z^\top Z + U^\top U)^{-1}(Z^\top B + U^\top U\bar{\Delta}) \\ &= (Z^\top Z + A)^{-1}(Z^\top Z \underbrace{(Z^\top Z)^{-1} Z^\top B}_{\hat{\Delta}} + U^\top U\bar{\Delta}) \\ &= (Z^\top Z + A)^{-1}(Z^\top Z(Z^\top Z)^{-1} Z^\top B + A\bar{\Delta}) \\ &= (Z^\top Z + A)^{-1}(Z^\top Z\hat{\Delta} + A\bar{\Delta}). \end{aligned} \tag{5.22}$$

Finally,

$$
\begin{aligned}
&(\Delta - \bar{\Delta})^\top A(\Delta - \bar{\Delta}) + (B - Z\Delta)^\top (B - Z\Delta) \\
&= (P - R\Delta)^\top (P - R\Delta) \\
&= (P - R\tilde{\Delta})^\top (P - R\tilde{\Delta}) + (\tilde{\Delta} - \Delta)^\top R^\top R(\tilde{\Delta} - \Delta) \\
&= \underbrace{(\tilde{\Delta} - \bar{\Delta})^\top A(\tilde{\Delta} - \bar{\Delta}) + (B - Z\tilde{\Delta})^\top (B - Z\tilde{\Delta})}_{S} + (\tilde{\Delta} - \Delta)^\top (Z^\top Z + A)(\tilde{\Delta} - \Delta),
\end{aligned}
\tag{5.23}
$$

where the second equation is similar to the expansion of the least square in Equation (5.13). Then, the posterior distribution for the parameters would be

$$
\begin{aligned}
V_\beta | B &\sim IW(v + m, V + S) \\
vec(\Delta) | V_\beta, B &\sim N(vec(\tilde{\Delta}), V_\beta \otimes (Z^\top Z + A)^{-1}),
\end{aligned}
\tag{5.24}
$$

where the joint posterior density of the hyperparameters would be

$$
p(V_\beta, \Delta | B) \;\propto\; |V_\beta|^{-\frac{(v+m+k+1)}{2}} e^{tr(-\frac{1}{2}(V+S)V_\beta^{-1})} \times |V_\beta|^{-\frac{n_z}{2}} e^{-\frac{1}{2}tr\left((\Delta-\tilde{\Delta})^\top (Z^\top Z+A)(\Delta-\tilde{\Delta})V_\beta^{-1}\right)}.
\tag{5.25}
$$

The properties of the natural conjugate prior setting will be discussed after the proposition of the model.

## 5.7 Hierarchical Logit Model

As outlined in the previous sections, the model should account for the individual differences of employees regarding attrition. In this context, the hierarchical logit model and the sampling strategy proposed in Rossi et al. [1, Chapter 5], which is expected to explain the heterogeneity among the employees, will be presented throughout the rest of the chapter. Furthermore, inducing a prior distribution shrinking the regression coefficients towards 0 is expected to overcome the problem of overfitting.

An important remark to be made is that there are many demographic variables available and the demographic information might actually help explaining heterogeneity. For instance, highly educated employees might have higher tendency

towards salary raise since they might have more job opportunities compared to their colleagues due to their qualification. Then, let $z_i$ and $\beta_i$ denote the demographic covariates and regression coefficients for $ith$ unit. The setting presented in MRM can be applied to the problem at hand by Rossi et al. [1, Chapter 5] as the following:

$$\beta_i = \Delta^\top z_i + u_i \quad u_i \sim N(0, V_\beta) \quad i \in \{1, \ldots, m\}, \tag{5.26}$$

where $m$ is the total number of units and $\Delta$ determines the effects of the demographics on time-varying regression coefficients. The idea is to use the demographic variables to estimate $\{\beta_i\}$ via $\Delta, Z$. This corresponds to having prior belief about $\{\beta_i\}$ such that they will depend on the demographic information. The model in Equation (5.26) can be written in the matrix form as the following:

$$B = Z\Delta + U, \quad u_i \sim N(0, V_\beta), \tag{5.27}$$

where the rows of $B$, $U$ and $Z$ are $\beta_i$, $u_i$ and $z_i$ respectively. The likelihood for the hyperparameters given the outcome, $\{\beta_i\}$, can be written as demonstrated in Section 5.6 [1, Chapter 5]. Then, the priors for $V_\beta$ and $\Delta$ can be set as the natural conjugate priors to take advantage of the Gibbs Sampler [1, Chapter 5]. Since, setting conjugate priors provides closed form fully conditional distributions so that the parameters can be sampled in sequence directly given the dependent parameters [1, Chapter 5]. The natural conjugate priors are as the following:

$$\begin{aligned} &vec(\Delta | V_\beta) \sim N(vec(\bar{\Delta}), V_\beta \otimes A^{-1}) \\ &V_\beta \sim IW(v, V), \end{aligned} \tag{5.28}$$

where $\bar{\Delta}$ and $A$ denote the location and precision prior for $\Delta$. $v$ and $V$ are degrees of freedom and location parameters of the Inverse Wishart prior. Fully conditional distributions and the sampling method will be further discussed in the following sections. Let $y_i$ and $x_i$ denote the outcomes and time dependent data of the $ith$ group. Then, Directed Acyclic Graph(DAG) would be as in the Figure 5.1. for the given setting.

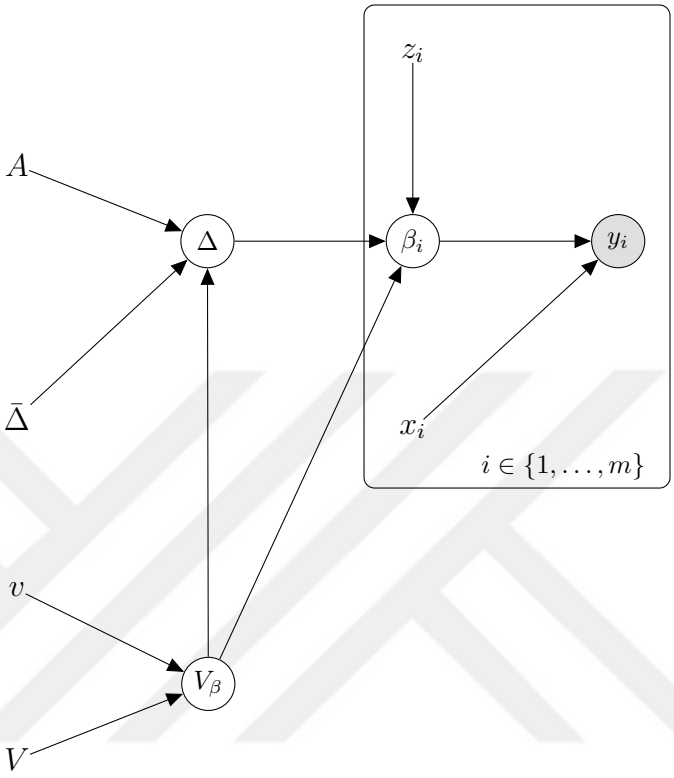


Figure 5.1: The Graphical Model or Directed Acyclic Graph(DAG)

## 5.8 Sampling from the Posterior and Analyis of the Conditional Distributions

The purpose is to obtain samples from $p(\{\beta_i\}, \Delta, V_\beta | A, \bar{\Delta}, v, V, X, Z, Y)$. Then, the posterior distribution can be simplified as

$$p(\{\beta_i\}, \Delta, V_\beta | A, \bar{\Delta}, v, V, X, Z, Y) \propto p(\Delta, V_\beta | \{\beta_i\}, A, \bar{\Delta}, v, V, Z) \prod_{i=1}^{m} p(\beta_i | X, Z, Y, \Delta, V_\beta). \tag{5.29}$$

Given $\{\beta_i\}$, $p(\Delta, V_\beta | \{\beta_i\}, A, \bar{\Delta}, v, V, Z)$ can be drawn using MRM outlined in the Section 5.6 and Gibbs Sampler can be used to draw these hyperparameters consecutively [1, Chapter 5]. Given $\Delta, V_\beta$, there is no suitable way of sampling from $\{\beta_i\}$ via the conditional closed form distributions given that the regression model is not linear [1, Chapter 2].

By modelling the likelihood of $\{\beta_i\}$ using the inverse logit transformation, Metropolis Sampler can be used to sample from $p(\beta_i|X, Z, Y, \Delta, V_\beta)$ since the posterior of $\beta_i$ cannot be expressed in a closed form as similar to $\Delta, V_\beta$ [1, Chapter 5]. As mentioned in the previous sections, since $\{\beta_i\}'s$ are assumed to be conditionally independent given $\Delta, V_\beta$, it is possible to draw from $\beta_i$ one by one [1, Chapter 5]. For the given priors in the previous section, the fully conditional distributions are as the following as demonstrated in Rossi et al. [1, Chapter 5]

$$\begin{aligned}
&L(\beta_i|x_i, y_i) \\
&V_\beta|B, Z, V, v, A, \bar{\Delta} \sim IW(v + m, V + S) \\
&vec(\Delta)|B, Z, V_\beta, A, \bar{\Delta} \sim N(vec(\tilde{\Delta}), V_\beta \otimes (Z^\top Z + A)^{-1}) \\
&\hat{\Delta} = (Z^\top Z)^{-1} Z^\top B \\
&\tilde{\Delta} = (Z^\top Z + A)^{-1}(Z^\top Z\hat{\Delta} + A\bar{\Delta}) \\
&S = (\tilde{\Delta} - \bar{\Delta})^\top A(\tilde{\Delta} - \bar{\Delta}) + (B - Z\tilde{\Delta})^\top (B - Z\tilde{\Delta}).
\end{aligned} \tag{5.30}$$

where $L(\beta_i|x_i, y_i)$ denotes the likelihood for $\beta_i$.

An important decision to be made is to group the units. An option is to assign a distinct $\beta_i$ on different employees. However, each employee has at most one success and the employees that are in the interest of the study do not have a successful outcome afterall. In that case, based on the experiments, a small prior on $V_\beta$ has to be selected such that $\beta_i's$ have insignificant difference so that the model does not overfit. Since, in case of not inducing a very small prior on $V_\beta$, the prior becomes weak over the likelihood and the effect of the likelihood over the posterior leads to unbounded posteriors. This requirement undermines the purpose of the study since the likelihood of each group should have a nontrivial impact on $\beta_i's$. The solution to this problem was to assign the employees having the same demographics to the same group which is essentially compromises with the purpose of the study which is based on modelling the heterogeneity based on the different demographic information. Therefore, if two employees share the same demographic information, there is no reason to have a prior belief such that they might have different tendency for attrition.

It is worth mentioning that the posterior mean of $\Delta$ is a weighted average of

least square estimator, $\hat{\Delta}$, and the prior mean, $\bar{\Delta}$ where the weights are prior precision and information matrix of MRM [1, Chapter 2]. In addition, even if the densities of the prior and likelihood are located far away from each other, the posterior of $\Delta$ is more concentrated than both the likelihood and the prior [1, Chapter 2]. Since the precision of the posterior is the sum of the information matrix and the precision prior. This might be undesirable since if the information provided by the prior and likelihood do not support each other, it might make sense to have a less concentrated posterior distribution of $\Delta$ and also, this property is specific to natural conjugate priors where it might be troublesome to check if there is such disagreement between the likelihood and prior for high dimensional parameter space [1, Chapter 2]. Other than this, as the number of units, $m$, increase, the information is also most likely to increase so the likelihood might dominate posterior and Bayes estimator converges to the least square estimator as the number of units tends to $\infty$ [1, Chapter 2].

The normal form of the prior distribution on $\{\beta_i\}$ brings a great deal of shrinkage due to the thin tails of normal distribution and this characteristic of the normal prior setting might not be desirable in case of having a considerable amount of information in units [1, Chapter 5]. Since, in those cases, the model should have the flexibility to allow the likelihood to dominate the posterior. However, since some of the units only contain a couple of observations, the thin tails of the normal distribution is desired.

There are two samplers given to sample from $p(\beta_i|X, Z, Y, \Delta, V_\beta)$ which are Independence and Normal Random Walk Metropolis. These both require to use a proposal distribution to approximate the posterior distribution of $\beta_i|\Delta, V_\beta$. Although most proposal densities are driven by the maximum likelihood, a maximum does not exist for units not choosing all types of the response and even some with all the responses existing [1, Chapter 5]. For the data at hand, not having a sufficient amount of data for each unit is more extreme due to the fact that the outcome is unbalanced so each unit is less likely to have a bounded likelihood. To overcome the issue, the following likelihood calculation is proposed by the authors

to approximate the posteriors and guarantee the existence of a maximum

$$\ell_i^*(\beta) = \ell_i(\beta)\bar{\ell}(\beta)^\tau. \tag{5.31}$$

where $\ell_i$ and $\bar{\ell}$ denote the unit level likelihood and pooled likelihood, which is the likelihood of the aggregate data, respectively and $\tau = \frac{n_i}{cN}$. $c$, $n_i$ and $N$ denote the tuning constant, the number of unit level observations and aggregate observations. $\tau$ is designed to "normalize" the pooled likelihood to the individual level likelihood. $\bar{\ell}(\beta)$ would dominate $\ell_i^*(\beta)$ when $\tau = 1$ since the number of observations used in the pooled likelihood will be much greater than the unit level likelihood's. Then, it is proposed that the posterior distribution can be approximated by the asymptotic approximation previously discussed in Section 5.4 and the proposal density for independence Metropolis can be set to $Normal(\beta_i^*, (H_i + V_\beta^{-1})^{-1})$ where $\beta_i^* = (H_i + V_\beta^{-1})^{-1}(H_i\hat{\beta}_i + V_\beta^{-1}\bar{\beta})$. $\hat{\beta}_i$ and $H_i$ denote the maximum likelihood estimator and negative hessian at the maximum of Equation (5.31). $\bar{\beta}$ and $V_\beta$ denote the mean and variance of the Normal prior. In a similar manner, $s(H_i + V_\beta^{-1})^{-1}$ was set to the variance of the Normal Random Walk where $s$ was proposed to be set to $2.93/\sqrt{dim(\beta)}$ based on the experimental results. Independence Metropolis might have the disadvantage of missing the location of the posterior which might result in higher rates of rejection [1, Chapter 2]. On the other hand, $V_\beta$ can be directly used as the covariance matrix of Normal Random Walk as suggested in Rossi et al. [1, Chapter 5] but this proposal density will ignore the information coming from the likelihood and since some of the units have a decent amount of observations, the rejection rates for these units might be relatively low which reduces the efficiency of the algorithm [1, Chapter 5]. Therefore, $s(H_i + V_\beta^{-1})^{-1}$ was used as the covariance matrix of Normal Random Walk.

At each iteration, $\beta_i^{can} \sim Normal(\beta_i^{cur}, s(H_i + V_\beta^{-1})^{-1})$ is primarily sampled for each $ith$ unit. Then, for given $\Delta$ and $V_\beta$, $\{\beta_i\}$ are sampled using the following

transition probability

$$\begin{aligned}\alpha_i &= \frac{\pi_i^{can}}{\pi_i^{cur}} \\ &= \frac{e^{-\frac{1}{2}(\beta_i^{can}-\Delta^\top z_i)^\top V_\beta^{-1}(\beta_i^{can}-\Delta^\top z_i)}}{e^{-\frac{1}{2}(\beta_i^{cur}-\Delta^\top z_i)^\top V_\beta^{-1}(\beta_i^{cur}-\Delta^\top z_i)}} \times \frac{\prod_{j=1}^{n_i}\left(\sigma(X_j.\beta_i^{can})\right)^{y_j}\left(1-\sigma(X_j.\beta_i^{can})\right)^{1-y_j}}{\prod_{j=1}^{n_i}\left(\sigma(X_j.\beta_i^{cur})\right)^{y_j}\left(1-\sigma(X_j.\beta_i^{cur})\right)^{1-y_j}} \\ &= \frac{e^{-\frac{1}{2}(\beta_i^{can}-\Delta^\top z_i)^\top V_\beta^{-1}(\beta_i^{can}-\Delta^\top z_i)}}{e^{-\frac{1}{2}(\beta_i^{cur}-\Delta^\top z_i)^\top V_\beta^{-1}(\beta_i^{cur}-\Delta^\top z_i)}} \times \prod_{j=1}^{n_i}\frac{\left(\sigma(X_j.\beta_i^{can})\right)^{y_j}\left(1-\sigma(X_j.\beta_i^{can})\right)^{1-y_j}}{\left(\sigma(X_j.\beta_i^{cur})\right)^{y_j}\left(1-\sigma(X_j.\beta_i^{cur})\right)^{1-y_j}}.\end{aligned} \tag{5.32}$$

where *can* and *cur* denote the candidate and current $\beta_i$ and $\sigma$ denotes the inverse logit. Lastly, $x_j$ denotes the $jth$ observation for the respective $ith$ unit. It is necessary to note that the left and right hand term denote the proportions of the prior and likelihood respectively. Furthermore, the transition probability does not depend on the proposal density of Metropolis. Since, $q(\beta_0, \beta_1)/q(\beta_1, \beta_0)$ cancel out to 1 when $q$ is the density of Normal [1, Chapter 3].

## 5.9 Manager's Effect on Employees

The variable Manager could not be imputed since it was suspected that the missing mechanism leading to the missingness of the variable Manager could not be MAR without variable Working Status since the missingness in variable Manager seems to depend on Working Status. Then, using Working Status in MICE might leak information into the missing values in Manager which results in using a function of the response as a predictor. Then, the model might lead to biased inferences.

Although using Manager variable is problematic to be used, the managers might have impact on employees' employment status. To measure that, a binomial regression model on managers was built by assuming that missing mechanism is MCAR and all the observations having missing values were removed. Each observation corresponded to manager's annual information and the response was set to the number of employees working and left which were supervised by that manager. The model did not turn out to be helpful in explaining the effect of Manager

on the turnover. The full model containing all the variables was compared to the model containing only the intercept to find out if some variables might be likely to explain the variance in the response. The p-value of the likelihood ratio test for the null model and full model turned out to be 0.087. Then, it cannot be rejected that the null model is as good as the full model at the confidence level of 95%. Then, by assuming the missing mechanism is MAR, manager's information does not seem to have a significant impact on the attrition of the employees. However, it should still be noted that the MCAR assumption is quite naive and complete case analysis might lead to biased results if the missing mechanism is not MCAR.

# Chapter 6

# Diagnostics

## 6.1 Effective Sample Size

The ideas and arguments presented in Rossi et al. [1, Chapter 3] and Gelman et al. [10, Chapter 11] regarding sampler's convergence were used in this chapter. MCMC sampler's sequential iterations have correlation due to the nature of Markov Chains since each new sample depends on the state of the last sample. As a matter of fact, there can be significant amount of correlation even after more than hundreds of thousands of iterations [1, Chapter 3]. The dependence among the iterations reduces the efficiency of the sampler. Since, autocorrelated consecutive draws mostly represent the same region of the distribution and it is desired to have samples from different regions of the support in consecutive draws. Then, a measure quantifying how fast the sampler drifts around the support of the posterior can show the efficiency of the sampler as suggested by Rossi et al.

[1, Chapter 3] which introduces a measure based on the following formula:

$$\begin{aligned}
\hat{\mu} &= \frac{\sum_r \mu^r}{R} \\
var(\hat{\mu}) &= \frac{1}{R^2}(var(\mu^1) + cov(\mu^1, \mu^2) + \ldots + cov(\mu^1, \mu^R) + \ldots \\
&\quad + var(\mu^R) + \ldots, + cov(\mu^{R-1}, \mu^R)) \\
&= \frac{var(\mu)}{R}(1 + 2\sum_{j=1}^{R-1}(\frac{R-j}{R}\rho_j)) = \frac{var(\mu)}{R} f_R,
\end{aligned} \tag{6.1}$$

where $\mu$ is the parameter sampled using the methods discussed and $\mu^j$ is the $jth$ sample and $\rho_j$ is the correlation between $rth$ and $(r-j)th$ term. For the third equation to hold, it should be assumed that the sampler have converged to the stationary distribution [1, Chapter 3]. For large values of $j$, $\rho_j$ is too noisy so instead, $\sum_{j=1}^{J} \rho_j$ can be bounded on $T$ which is the first lag where two consecutive lags are negative [10, Chapter 11]. Then, $f_T$ denotes the term showing the total correlation and it resembles how efficient the sampler is. Rossi et al. [1, Chapter 3] notes that the parameter $s$ used in the covariance matrix of the Normal Random Walk can be tuned using $f_T$ value so that the sampler becomes more efficient. However, it must be assumed that the sampler converges to the posterior distribution of the parameter for different values of $s$. Based on the experiments, for smaller values of $s$, the sampler navigates slowly and does not navigate all the regions of the support of the posterior. For higher values of $s$, the samples get rejected very frequently and the sampler cannot drift around the support of the posterior. Then, this method can only be used for the $s$ values where the converge is assured. However, this process is more time consuming considering the fact that tuning $s$ should bring time efficiency. Therefore, the value proposed by Rossi et al. [1, Chapter 3], $s = 2.93/sqrt(k)$, was used where $k$ denotes the number of time dependent variables of $X$. Nonetheless, this idea can still be useful to compute how much information the sampler extracts out of the posterior distribution as suggested by Gelman et al. [10, Chapter 11] where $\frac{T}{f_T}$ is essentially called the Effective Sample Size(ESS). This statistic is suggested to be handy to compare with the actual sample size so that it can measure how close the sampler is to an independent sampler because $f_T$ should ideally be 1 when the sampler draws independent samples out of the posterior distribution.

To compute the ESS, it must be ensured that the values drawn out of the sampler closely resemble the posterior distribution. However, the initial draws from the parameter space might lie out of the support of the posterior and they might not be representing the posterior when the number of iterations are finite as suggested in Gelman et al. [10, Chapter 11] and Rossi et al. [1, Chapter 3]. That idea brings the notion of burnin period where the initial draws of the samplers are filtered out when making the inference. The sampler was run for 100000 iterations with 10000 iterations of burnin. The sampler seems to converge after a couple of thousand of draws based on the trace plots. Based on the experiments and suggestion of Rossi et al. [1, Chapter 3], as the parameter size of the posterior distribution increases, the samples have more autocorrelation and the sampler becomes more inefficient. The model at hand suffers from the same issue as well since there are more than 1000 parameters drawn at each iteration. ESS was computed for each replication of MICE and these values were averaged across the replications.

Table 6.1: Effective Sample Size for $\{\beta_i\}$

| | X.Intercept. | Number.of.Years.WorkedMedium | Number.of.Years.WorkedHigh | Highest.DegreeOver.Bachelor | First.Year1 | Child1 | Child.Born1 | Marital.StatusSingle | WorkplaceOther | Performance.ScoreBad | Performance.ScoreOkay | Appreciation.ScoreBad | Appreciation.ScoreGood | SalaryMedium | SalaryHigh | AgeMedium | AgeHigh | Salary.RaiseVery.High | Salary.RaiseHigh | Salary.RaiseVery.Low |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unit 1 | 1028 | 553 | 441 | 782 | 600 | 753 | 915 | 685 | 1121 | 617 | 635 | 722 | 722 | 907 | 505 | 775 | 436 | 899 | 717 | 960 |
| Unit 2 | 1007 | 635 | 367 | 415 | 514 | 394 | 736 | 605 | 651 | 536 | 532 | 653 | 599 | 720 | 531 | 492 | 376 | 535 | 725 | 919 |
| Unit 3 | 1048 | 656 | 458 | 459 | 500 | 356 | 353 | 474 | 667 | 644 | 501 | 512 | 562 | 594 | 472 | 469 | 328 | 492 | 740 | 717 |
| Unit 4 | 740 | 660 | 624 | 470 | 566 | 514 | 722 | 608 | 1027 | 658 | 630 | 740 | 776 | 810 | 563 | 538 | 397 | 743 | 823 | 611 |
| Unit 5 | 1387 | 706 | 349 | 610 | 641 | 380 | 575 | 575 | 931 | 777 | 594 | 609 | 641 | 772 | 544 | 682 | 410 | 598 | 735 | 1020 |
| Unit 6 | 891 | 680 | 333 | 662 | 738 | 443 | 451 | 548 | 622 | 615 | 510 | 556 | 414 | 755 | 518 | 470 | 288 | 487 | 754 | 957 |
| Unit 7 | 884 | 530 | 483 | 591 | 457 | 498 | 468 | 480 | 794 | 769 | 533 | 784 | 685 | 716 | 537 | 409 | 395 | 763 | 754 | 1060 |
| Unit 8 | 401 | 776 | 494 | 426 | 673 | 490 | 669 | 495 | 904 | 572 | 651 | 820 | 637 | 763 | 403 | 631 | 454 | 758 | 681 | 1023 |
| Unit 9 | 970 | 799 | 527 | 581 | 694 | 652 | 422 | 594 | 810 | 759 | 648 | 525 | 628 | 660 | 641 | 856 | 480 | 705 | 737 | 992 |
| Unit 10 | 850 | 663 | 600 | 410 | 452 | 451 | 625 | 611 | 953 | 612 | 502 | 594 | 704 | 750 | 269 | 529 | 331 | 704 | 741 | 1159 |
| Unit 11 | 1064 | 862 | 620 | 779 | 563 | 660 | 719 | 633 | 995 | 973 | 964 | 897 | 735 | 838 | 735 | 870 | 496 | 755 | 817 | 952 |
| Unit 12 | 1117 | 702 | 529 | 528 | 573 | 507 | 575 | 566 | 786 | 773 | 532 | 598 | 724 | 531 | 381 | 596 | 336 | 517 | 637 | 1025 |
| Unit 13 | 1164 | 647 | 513 | 495 | 543 | 515 | 473 | 699 | 557 | 507 | 532 | 660 | 489 | 708 | 351 | 449 | 380 | 520 | 583 | 853 |
| Unit 14 | 1176 | 731 | 471 | 420 | 529 | 395 | 620 | 668 | 499 | 681 | 560 | 574 | 764 | 653 | 389 | 460 | 407 | 711 | 820 | 749 |
| Unit 15 | 1157 | 806 | 649 | 672 | 642 | 824 | 463 | 657 | 873 | 748 | 516 | 661 | 591 | 778 | 389 | 636 | 371 | 1054 | 766 | 1031 |
| Unit 16 | 1101 | 618 | 400 | 558 | 675 | 416 | 697 | 596 | 767 | 575 | 535 | 590 | 605 | 750 | 525 | 768 | 387 | 610 | 666 | 721 |
| Unit 17 | 1171 | 537 | 545 | 699 | 770 | 559 | 611 | 676 | 736 | 989 | 809 | 891 | 785 | 668 | 665 | 612 | 404 | 597 | 859 | 674 |
| Unit 18 | 1257 | 836 | 539 | 645 | 520 | 313 | 645 | 604 | 733 | 523 | 393 | 656 | 652 | 760 | 524 | 471 | 273 | 490 | 718 | 879 |
| Unit 19 | 1352 | 724 | 408 | 759 | 549 | 725 | 818 | 786 | 565 | 668 | 634 | 760 | 845 | 993 | 604 | 619 | 398 | 867 | 789 | 1119 |
| Unit 20 | 865 | 973 | 506 | 569 | 683 | 551 | 998 | 743 | 655 | 834 | 759 | 863 | 667 | 810 | 535 | 728 | 421 | 785 | 761 | 1056 |
| Unit 21 | 1261 | 591 | 474 | 592 | 552 | 507 | 665 | 612 | 894 | 627 | 609 | 884 | 693 | 741 | 573 | 785 | 496 | 527 | 843 | 801 |
| Unit 22 | 1050 | 793 | 638 | 620 | 584 | 610 | 653 | 794 | 775 | 693 | 590 | 728 | 792 | 643 | 445 | 670 | 466 | 788 | 793 | 1002 |
| Unit 23 | 1435 | 603 | 463 | 781 | 593 | 352 | 566 | 480 | 805 | 570 | 641 | 901 | 687 | 767 | 559 | 465 | 244 | 614 | 556 | 963 |
| Unit 24 | 973 | 734 | 516 | 609 | 649 | 580 | 386 | 676 | 617 | 649 | 464 | 782 | 522 | 649 | 340 | 745 | 518 | 707 | 521 | 717 |

Table 6.2: Effective Sample Size for $V_\beta$

| | X.Intercept. | Number.of.Years.WorkedMedium | Number.of.Years.WorkedHigh | Highest.DegreeOver.Bachelor | First.Year1 | Child1 | Child.Born1 | Marital.StatusSingle | WorkplaceOther | Performance.ScoreBad | Performance.ScoreOkay | Appreciation.ScoreBad | Appreciation.ScoreGood | SalaryMedium | SalaryHigh | AgeMedium | AgeHigh | Salary.RaiseVery.High | Salary.RaiseHigh | Salary.RaiseVery.Low |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X.Intercept. | 2615 | 380 | 255 | 349 | 365 | 245 | 347 | 390 | 600 | 481 | 316 | 473 | 362 | 562 | 238 | 311 | 202 | 385 | 651 | 1045 |
| Number.of.Years.WorkedMedium | 380 | 910 | 1433 | 1114 | 1351 | 671 | 1258 | 1680 | 3604 | 1815 | 935 | 1118 | 1617 | 2150 | 1217 | 1233 | 852 | 1195 | 668 | 1957 |
| Number.of.Years.WorkedHigh | 255 | 1433 | 2724 | 2352 | 657 | 552 | 1707 | 4526 | 5207 | 6171 | 1622 | 1433 | 1131 | 5734 | 2944 | 2731 | 2344 | 1783 | 636 | 1363 |
| Highest.DegreeOver.Bachelor | 349 | 1114 | 2352 | 1619 | 878 | 629 | 1379 | 3200 | 3725 | 3495 | 1263 | 1540 | 1964 | 2917 | 2468 | 1778 | 1733 | 1401 | 632 | 1699 |
| First.Year1 | 365 | 1351 | 657 | 878 | 634 | 402 | 957 | 1323 | 1786 | 1780 | 805 | 1871 | 653 | 1722 | 716 | 1148 | 615 | 772 | 681 | 1822 |
| Child1 | 245 | 671 | 552 | 629 | 402 | 458 | 688 | 772 | 1401 | 1065 | 531 | 975 | 510 | 1238 | 566 | 796 | 431 | 659 | 427 | 857 |
| Child.Born1 | 347 | 1258 | 1707 | 1379 | 957 | 688 | 1783 | 2326 | 2756 | 2362 | 1118 | 1600 | 1644 | 3387 | 1669 | 1321 | 1223 | 1384 | 792 | 1665 |
| Marital.StatusSingle | 390 | 1680 | 4526 | 3200 | 1323 | 772 | 2326 | 6906 | 5705 | 7552 | 1916 | 1824 | 2290 | 5462 | 4042 | 2313 | 3860 | 3240 | 947 | 1853 |
| WorkplaceOther | 600 | 3604 | 5207 | 3725 | 1786 | 1401 | 2756 | 5705 | 5774 | 6953 | 3516 | 2952 | 2600 | 5879 | 5309 | 4047 | 6195 | 3899 | 1839 | 2558 |
| Performance.ScoreBad | 481 | 1815 | 6171 | 3495 | 1780 | 1065 | 2362 | 7552 | 6953 | 8036 | 2681 | 2270 | 2293 | 6082 | 2937 | 3215 | 4271 | 2682 | 1226 | 2419 |
| Performance.ScoreOkay | 316 | 935 | 1622 | 1263 | 805 | 531 | 1118 | 1916 | 3516 | 2681 | 1088 | 1205 | 1303 | 2469 | 1376 | 1606 | 1215 | 1415 | 773 | 1489 |
| Appreciation.ScoreBad | 473 | 1118 | 1433 | 1540 | 1871 | 975 | 1600 | 1824 | 2952 | 2270 | 1205 | 1461 | 2186 | 2461 | 936 | 928 | 698 | 1437 | 897 | 2228 |
| Appreciation.ScoreGood | 362 | 1617 | 1131 | 1964 | 653 | 510 | 1644 | 2290 | 2600 | 2293 | 1303 | 2186 | 1324 | 3022 | 1369 | 1794 | 1185 | 1463 | 783 | 1875 |
| SalaryMedium | 562 | 2150 | 5734 | 2917 | 1722 | 1238 | 3387 | 5462 | 5879 | 6082 | 2469 | 2461 | 3022 | 6388 | 4140 | 2884 | 5495 | 2783 | 1276 | 2015 |
| SalaryHigh | 238 | 1217 | 2944 | 2468 | 716 | 566 | 1669 | 4042 | 5309 | 2937 | 1376 | 936 | 1369 | 4140 | 2948 | 1452 | 3197 | 1408 | 485 | 1005 |
| AgeMedium | 311 | 1233 | 2731 | 1778 | 1148 | 796 | 1321 | 2313 | 4047 | 3215 | 1606 | 928 | 1794 | 2884 | 1452 | 2426 | 1702 | 1855 | 646 | 1992 |
| AgeHigh | 202 | 852 | 2344 | 1733 | 615 | 431 | 1223 | 3860 | 6195 | 4271 | 1215 | 698 | 1185 | 5495 | 3197 | 1702 | 4013 | 1016 | 445 | 943 |
| Salary.RaiseVery.High | 385 | 1195 | 1783 | 1401 | 772 | 659 | 1384 | 3240 | 3899 | 2682 | 1415 | 1437 | 1463 | 2783 | 1408 | 1855 | 1016 | 1186 | 1015 | 1692 |
| Salary.RaiseHigh | 651 | 668 | 636 | 632 | 681 | 427 | 792 | 947 | 1839 | 1226 | 773 | 897 | 783 | 1276 | 485 | 646 | 445 | 1015 | 877 | 2024 |
| Salary.RaiseVery.Low | 1045 | 1957 | 1363 | 1699 | 1822 | 857 | 1665 | 1853 | 2558 | 2419 | 1489 | 2228 | 1875 | 2015 | 1005 | 1992 | 943 | 1692 | 2024 | 2232 |

Table 6.3: Effective Sample Size for $\Delta$

| | X.Intercept. | Number.of.Years.WorkedMedium | Number.of.Years.WorkedHigh | Highest.DegreeOver.Bachelor | First.Year1 | Child1 | Child.Born1 | Marital.StatusSingle | WorkplaceOther | Performance.ScoreBad | Performance.ScoreOkay | Appreciation.ScoreBad | Appreciation.ScoreGood | SalaryMedium | SalaryHigh | AgeMedium | AgeHigh | Salary.RaiseVery.High | Salary.RaiseHigh | Salary.RaiseVery.Low |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X.Intercept. | 2114 | 267 | 203 | 261 | 269 | 164 | 258 | 281 | 433 | 344 | 229 | 321 | 263 | 417 | 188 | 225 | 161 | 287 | 404 | 682 |
| GenderWoman | 3106 | 1075 | 762 | 644 | 1288 | 1055 | 831 | 1193 | 755 | 1106 | 972 | 1243 | 916 | 919 | 769 | 1095 | 831 | 894 | 1234 | 993 |
| JobSwitch1 | 3085 | 1117 | 811 | 993 | 1066 | 1036 | 820 | 1099 | 975 | 976 | 665 | 1420 | 1124 | 1104 | 1172 | 1340 | 934 | 953 | 978 | 1495 |
| College.DepartmentSupport | 2547 | 1226 | 858 | 944 | 1026 | 1027 | 976 | 1007 | 1320 | 1030 | 1048 | 1106 | 1015 | 1087 | 1019 | 1080 | 779 | 1283 | 1264 | 1305 |
| Education.LevelLow | 2996 | 1218 | 1080 | 1083 | 1441 | 1341 | 1025 | 1357 | 935 | 1463 | 1074 | 1391 | 1446 | 1355 | 945 | 920 | 1122 | 1497 | 1142 | 1632 |
| Education.LevelMedium | 4320 | 1750 | 923 | 1259 | 975 | 1308 | 1280 | 1488 | 1475 | 1284 | 1186 | 1727 | 1418 | 1557 | 1351 | 1034 | 988 | 1125 | 1311 | 1375 |

It is sufficient to have 100 or even 10 draws [10, Chapter 11]. Then, by Table 6.1, 6.2 and 6.3, ESS values seem to be sufficient.

## 6.2 Converge of MCMC

After the simulation draws were obtained, it is necessary to check if all the draws truly resemble the posterior distribution. Since it is assumed that MCMC sampler is expected to converge to the posterior distribution as the invariant distribution of the sampler as the iterations tend to $\infty$, initial iterations might not resemble the posterior distribution. Therefore, a certain number of iterations might be thrown away as burnin. Gelman et al. [10, Chapter 11] suggests disposing half of the iterations as a safe choice.

As suggested in Gelman et al. [10, Chapter 11], although the burnin iterations were disposed, MCMC sampler might have not reached the support of the distribution which might imply that MCMC sampler did not converge to the posterior distribution. On the other hand, MCMC might have not been simulated long enough to fully resemble the posterior distribution such that the regions of the posterior are fully travelled [10, Chapter 11]. For instance, drawn samples might not cover all the modes of a multi-modal posterior. Gelman et al. [10, Chapter 11] name these two conditions that must be satisfied as stationarity and mixing and proposes a diagnostic inspecting these concurrently by using a quantitative measure. This test compares the chains to have an idea if they are sampled from the same distribution. To run this test, $m$ chains having scattered initial values are run in parallel and the burnin/warm-up period is disposed for each run. By running chains with scattered initial points, mixing of the chain is tested since chains with scattered initial points should visit the same regions of the posterior density [10, Chapter 11]. Secondly, each chain's iterations are divided into two (Each part is treated as a distinct chain). By partitioning the chains into two, the stationarity of the chain is tested since first and second half of the chain should reflect the same distribution [10, Chapter 11]. Then, the following calculations

are made

$$\bar{\psi} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_j$$
$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\psi}_j - \bar{\psi})^2,\ where\ \bar{\psi}_j = \frac{1}{n}\sum_{j=1}^{n}\bar{\psi}_{ij}$$
$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2,\ where\ s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - \bar{\psi}_j)^2 \tag{6.2}$$
$$\widehat{var}(\psi) = \frac{n-1}{n}W + \frac{1}{n}B$$
$$\hat{R} = \sqrt{\frac{\widehat{var}(\psi)}{W}},$$

where $\psi$ is any one dimensional parameter, $n$ and $m$ are the total number of iterations per chain and the number of chains. $\widehat{var}(\psi)$ is the unbiased estimator of the variance of $\psi$ and $\hat{R}$ is the potential scale reduction factor which is expected to decrease to 1 as iterations tend to $\infty$ [10, Chapter 11]. To obtain scattered initial values for parameters, the posterior is approximated by an overdispersed mixture of multivariate-t distribution as pointed out in Gelman and Rubin [16]. However, since the dimension of the posterior distribution is quite large, trying to approximate the posterior by a mixture of multivariate-t distribution is not quite feasible. Also, the first and second stage parameters depend on each other so approximating their distribution is not trivial. The initial values of $\beta_i's$ were sampled from $Uniform(-1, 1)$. Since the columns of the data matrix were scaled after one-hot encoding, $Uniform(-1, 1)$ distribution is assumed to be sufficiently overdispersed for regression coefficients.

To compare the chains, the variance between the chains, $B$ and variance within the chains, $W$ are calculated. The variance between the chains is expected to converge to 0 as the iterations tend to $\infty$ if the chains sample from the same distribution [10, Chapter 11]. Within variance is expected to converge to the variance of the parameter of interest [10, Chapter 11]. Then, if $\hat{R}$ is close to 1, the chains are near the target distribution [16].

Table 6.4: Potential Scale Reduction Factors for $\{\beta_i\}$

| | X.Intercept. | Number.of.Years.WorkedMedium | Number.of.Years.WorkedHigh | Highest.DegreeOver.Bachelor | First.Year1 | Child1 | Child.Born1 | Marital.StatusSingle | WorkplaceOther | Performance.ScoreBad | Performance.ScoreOkay | Appreciation.ScoreBad | Appreciation.ScoreGood | SalaryMedium | SalaryHigh | AgeMedium | AgeHigh | Salary.RaiseVery.High | Salary.RaiseHigh | Salary.RaiseVery.Low |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unit 1 | 1.002 | 1.005 | 1.007 | 1.005 | 1.006 | 1.002 | 1.002 | 1.001 | 1.009 | 1.003 | 1.002 | 1.004 | 1.003 | 1.006 | 1.008 | 1.001 | 1.005 | 1.004 | 1.001 | 1.002 |
| Unit 2 | 1.015 | 1.008 | 1.010 | 1.015 | 1.014 | 1.003 | 1.004 | 1.006 | 1.002 | 1.013 | 1.009 | 1.042 | 1.006 | 1.007 | 1.014 | 1.002 | 1.023 | 1.017 | 1.010 | 1.035 |
| Unit 3 | 1.002 | 1.006 | 1.013 | 1.005 | 1.005 | 1.011 | 1.011 | 1.004 | 1.013 | 1.005 | 1.001 | 1.006 | 1.007 | 1.008 | 1.029 | 1.004 | 1.003 | 1.001 | 1.003 | 1.009 |
| Unit 4 | 1.004 | 1.005 | 1.011 | 1.006 | 1.006 | 1.002 | 1.003 | 1.004 | 1.011 | 1.004 | 1.002 | 1.003 | 1.006 | 1.007 | 1.005 | 1.001 | 1.011 | 1.001 | 1.001 | 1.004 |
| Unit 5 | 1.003 | 1.004 | 1.015 | 1.007 | 1.002 | 1.011 | 1.008 | 1.002 | 1.002 | 1.005 | 1.008 | 1.008 | 1.004 | 1.006 | 1.037 | 1.003 | 1.005 | 1.002 | 1.001 | 1.002 |
| Unit 6 | 1.002 | 1.006 | 1.006 | 1.004 | 1.006 | 1.006 | 1.004 | 1.005 | 1.028 | 1.003 | 1.004 | 1.008 | 1.003 | 1.006 | 1.007 | 1.004 | 1.015 | 1.008 | 1.004 | 1.011 |
| Unit 7 | 1.039 | 1.003 | 1.021 | 1.009 | 1.005 | 1.009 | 1.002 | 1.017 | 1.006 | 1.009 | 1.004 | 1.021 | 1.024 | 1.002 | 1.003 | 1.011 | 1.007 | 1.027 | 1.002 | 1.010 |
| Unit 8 | 1.010 | 1.005 | 1.007 | 1.012 | 1.005 | 1.001 | 1.006 | 1.007 | 1.006 | 1.003 | 1.001 | 1.005 | 1.011 | 1.001 | 1.017 | 1.010 | 1.002 | 1.003 | 1.006 | 1.002 |
| Unit 9 | 1.010 | 1.003 | 1.027 | 1.005 | 1.039 | 1.005 | 1.004 | 1.002 | 1.034 | 1.007 | 1.010 | 1.009 | 1.002 | 1.001 | 1.005 | 1.007 | 1.007 | 1.026 | 1.004 | 1.004 |
| Unit 10 | 1.002 | 1.005 | 1.008 | 1.006 | 1.019 | 1.003 | 1.008 | 1.001 | 1.004 | 1.004 | 1.010 | 1.003 | 1.004 | 1.001 | 1.003 | 1.002 | 1.005 | 1.004 | 1.003 | 1.009 |
| Unit 11 | 1.003 | 1.001 | 1.010 | 1.022 | 1.004 | 1.002 | 1.009 | 1.005 | 1.061 | 1.003 | 1.007 | 1.001 | 1.012 | 1.011 | 1.010 | 1.003 | 1.002 | 1.003 | 1.002 | 1.023 |
| Unit 12 | 1.002 | 1.005 | 1.004 | 1.003 | 1.003 | 1.004 | 1.007 | 1.003 | 1.020 | 1.001 | 1.013 | 1.011 | 1.012 | 1.003 | 1.003 | 1.002 | 1.005 | 1.002 | 1.008 | 1.007 |
| Unit 13 | 1.003 | 1.009 | 1.012 | 1.006 | 1.006 | 1.011 | 1.010 | 1.004 | 1.028 | 1.011 | 1.002 | 1.076 | 1.022 | 1.002 | 1.007 | 1.010 | 1.037 | 1.004 | 1.004 | 1.038 |
| Unit 14 | 1.003 | 1.013 | 1.006 | 1.011 | 1.024 | 1.030 | 1.017 | 1.004 | 1.004 | 1.017 | 1.024 | 1.004 | 1.011 | 1.009 | 1.046 | 1.022 | 1.002 | 1.006 | 1.002 | 1.021 |
| Unit 15 | 1.003 | 1.006 | 1.005 | 1.005 | 1.007 | 1.002 | 1.010 | 1.003 | 1.046 | 1.002 | 1.000 | 1.003 | 1.004 | 1.006 | 1.012 | 1.005 | 1.002 | 1.003 | 1.006 | 1.023 |
| Unit 16 | 1.002 | 1.004 | 1.005 | 1.006 | 1.004 | 1.003 | 1.003 | 1.003 | 1.005 | 1.005 | 1.003 | 1.006 | 1.005 | 1.005 | 1.010 | 1.004 | 1.008 | 1.002 | 1.001 | 1.008 |
| Unit 17 | 1.003 | 1.008 | 1.006 | 1.001 | 1.001 | 1.009 | 1.004 | 1.003 | 1.014 | 1.003 | 1.005 | 1.011 | 1.013 | 1.003 | 1.004 | 1.003 | 1.004 | 1.005 | 1.003 | 1.003 |
| Unit 18 | 1.002 | 1.004 | 1.010 | 1.006 | 1.005 | 1.008 | 1.007 | 1.006 | 1.007 | 1.007 | 1.001 | 1.003 | 1.004 | 1.004 | 1.003 | 1.004 | 1.004 | 1.003 | 1.002 | 1.002 |
| Unit 19 | 1.004 | 1.007 | 1.009 | 1.006 | 1.005 | 1.005 | 1.003 | 1.005 | 1.011 | 1.004 | 1.003 | 1.009 | 1.008 | 1.003 | 1.008 | 1.005 | 1.009 | 1.004 | 1.005 | 1.002 |
| Unit 20 | 1.002 | 1.003 | 1.005 | 1.002 | 1.003 | 1.003 | 1.006 | 1.002 | 1.014 | 1.005 | 1.001 | 1.006 | 1.003 | 1.004 | 1.013 | 1.004 | 1.008 | 1.003 | 1.001 | 1.004 |
| Unit 21 | 1.001 | 1.009 | 1.005 | 1.007 | 1.001 | 1.009 | 1.007 | 1.002 | 1.016 | 1.003 | 1.010 | 1.012 | 1.008 | 1.008 | 1.040 | 1.003 | 1.004 | 1.006 | 1.003 | 1.013 |
| Unit 22 | 1.006 | 1.007 | 1.023 | 1.005 | 1.011 | 1.006 | 1.009 | 1.006 | 1.031 | 1.038 | 1.006 | 1.063 | 1.024 | 1.003 | 1.011 | 1.013 | 1.046 | 1.013 | 1.003 | 1.079 |
| Unit 23 | 1.004 | 1.005 | 1.009 | 1.012 | 1.008 | 1.003 | 1.004 | 1.006 | 1.024 | 1.009 | 1.016 | 1.021 | 1.009 | 1.002 | 1.010 | 1.015 | 1.013 | 1.014 | 1.007 | 1.058 |
| Unit 24 | 1.005 | 1.005 | 1.004 | 1.007 | 1.007 | 1.008 | 1.011 | 1.010 | 1.026 | 1.007 | 1.003 | 1.012 | 1.004 | 1.006 | 1.013 | 1.005 | 1.008 | 1.017 | 1.010 | 1.046 |

Table 6.5: Potential Scale Reduction Factors for $V_\beta$

| | X.Intercept. | Number.of.Years.WorkedMedium | Number.of.Years.WorkedHigh | Highest.DegreeOver.Bachelor | First.Year1 | Child1 | Child.Born1 | Marital.StatusSingle | WorkplaceOther | Performance.ScoreBad | Performance.ScoreOkay | Appreciation.ScoreBad | Appreciation.ScoreGood | SalaryMedium | SalaryHigh | AgeMedium | AgeHigh | Salary.RaiseVery.High | Salary.RaiseHigh | Salary.RaiseVery.Low |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X.Intercept. | 1.001 | 1.002 | 1.006 | 1.003 | 1.001 | 1.003 | 1.005 | 1.002 | 1.006 | 1.004 | 1.003 | 1.002 | 1.004 | 1.002 | 1.002 | 1.002 | 1.003 | 1.004 | 1.003 | 1.002 |
| Number.of.Years.WorkedMedium | 1.002 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.002 | 1.000 | 1.002 | 1.000 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 |
| Number.of.Years.WorkedHigh | 1.006 | 1.001 | 1.000 | 1.001 | 1.002 | 1.001 | 1.001 | 1.000 | 1.000 | 1.001 | 1.001 | 1.002 | 1.001 | 1.000 | 1.001 | 1.001 | 1.000 | 1.001 | 1.002 | 1.001 |
| Highest.DegreeOver.Bachelor | 1.003 | 1.001 | 1.001 | 1.001 | 1.001 | 1.002 | 1.001 | 1.000 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 |
| First.Year1 | 1.001 | 1.000 | 1.002 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.002 | 1.002 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.001 | 1.001 |
| Child1 | 1.003 | 1.001 | 1.001 | 1.002 | 1.001 | 1.002 | 1.004 | 1.001 | 1.003 | 1.002 | 1.001 | 1.002 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 |
| Child.Born1 | 1.005 | 1.001 | 1.001 | 1.001 | 1.001 | 1.004 | 1.002 | 1.000 | 1.000 | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.000 | 1.002 | 1.004 | 1.002 |
| Marital.StatusSingle | 1.002 | 1.000 | 1.000 | 1.000 | 1.001 | 1.001 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.000 | 1.000 | 1.001 | 1.001 | 1.000 | 1.000 | 1.001 | 1.001 |
| WorkplaceOther | 1.006 | 1.001 | 1.000 | 1.001 | 1.002 | 1.003 | 1.000 | 1.000 | 1.000 | 1.000 | 1.002 | 1.001 | 1.001 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.002 | 1.002 |
| Performance.ScoreBad | 1.004 | 1.001 | 1.001 | 1.001 | 1.002 | 1.002 | 1.001 | 1.000 | 1.000 | 1.000 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.000 | 1.001 | 1.002 | 1.001 |
| Performance.ScoreOkay | 1.003 | 1.002 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 | 1.002 | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.000 | 1.001 | 1.001 | 1.001 | 1.001 |
| Appreciation.ScoreBad | 1.002 | 1.000 | 1.002 | 1.000 | 1.001 | 1.002 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 |
| Appreciation.ScoreGood | 1.004 | 1.002 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.002 | 1.001 |
| SalaryMedium | 1.002 | 1.000 | 1.000 | 1.001 | 1.001 | 1.001 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.001 |
| SalaryHigh | 1.002 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 | 1.000 | 1.001 | 1.000 | 1.001 | 1.001 | 1.001 |
| AgeMedium | 1.002 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.000 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.001 |
| AgeHigh | 1.003 | 1.001 | 1.000 | 1.001 | 1.001 | 1.001 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.001 | 1.001 | 1.000 | 1.000 | 1.000 | 1.001 | 1.000 | 1.002 | 1.001 |
| Salary.RaiseVery.High | 1.004 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.002 | 1.000 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 | 1.001 | 1.001 | 1.000 | 1.002 | 1.001 | 1.001 |
| Salary.RaiseHigh | 1.003 | 1.001 | 1.002 | 1.001 | 1.001 | 1.001 | 1.004 | 1.001 | 1.002 | 1.002 | 1.001 | 1.001 | 1.002 | 1.001 | 1.001 | 1.001 | 1.002 | 1.001 | 1.002 | 1.001 |
| Salary.RaiseVery.Low | 1.002 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.002 | 1.001 | 1.002 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 |

Table 6.6: Potential Scale Reduction Factors for $\Delta$

| | X.Intercept. | Number.of.Years.WorkedMedium | Number.of.Years.WorkedHigh | Highest.DegreeOver.Bachelor | First.Year1 | Child1 | Child.Born1 | Marital.StatusSingle | WorkplaceOther | Performance.ScoreBad | Performance.ScoreOkay | Appreciation.ScoreBad | Appreciation.ScoreGood | SalaryMedium | SalaryHigh | AgeMedium | AgeHigh | Salary.RaiseVery.High | Salary.RaiseHigh | Salary.RaiseVery.Low |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X.Intercept. | 1.001 | 1.005 | 1.016 | 1.008 | 1.004 | 1.003 | 1.012 | 1.005 | 1.022 | 1.009 | 1.009 | 1.004 | 1.002 | 1.004 | 1.005 | 1.006 | 1.008 | 1.002 | 1.011 | 1.004 |
| GenderWoman | 1.000 | 1.006 | 1.002 | 1.007 | 1.005 | 1.005 | 1.002 | 1.001 | 1.067 | 1.030 | 1.014 | 1.003 | 1.003 | 1.002 | 1.002 | 1.002 | 1.001 | 1.003 | 1.003 | 1.071 |
| JobSwitch1 | 1.043 | 1.052 | 1.021 | 1.019 | 1.013 | 1.079 | 1.013 | 1.002 | 1.053 | 1.001 | 1.038 | 1.105 | 1.067 | 1.008 | 1.085 | 1.005 | 1.021 | 1.019 | 1.019 | 1.030 |
| College.DepartmentSupport | 1.019 | 1.012 | 1.025 | 1.031 | 1.020 | 1.009 | 1.009 | 1.014 | 1.055 | 1.013 | 1.008 | 1.031 | 1.052 | 1.003 | 1.011 | 1.016 | 1.008 | 1.018 | 1.010 | 1.015 |
| Education.LevelLow | 1.015 | 1.021 | 1.015 | 1.012 | 1.022 | 1.084 | 1.062 | 1.036 | 1.011 | 1.013 | 1.007 | 1.024 | 1.021 | 1.014 | 1.092 | 1.011 | 1.006 | 1.008 | 1.028 | 1.032 |
| Education.LevelMedium | 1.001 | 1.009 | 1.001 | 1.003 | 1.000 | 1.003 | 1.019 | 1.001 | 1.173 | 1.003 | 1.006 | 1.005 | 1.001 | 1.037 | 1.081 | 1.013 | 1.001 | 1.002 | 1.002 | 1.029 |

For each component of $\beta_i$, $\Delta$ and $V_\beta$, Gelman-Rubin diagnostic was computed and it turned out that MCMC seemed to mix well and reached stationarity for 90000 iterations after a 10000 iterations of burnin phase. As suggested in Brooks and Gelman [17], if the potential scale reduction factor is smaller than 1.2, it is ensured that the convergence is reached. As pointed out in Table 6.4, 6.5 and 6.6, the potential scale reduction factor, $\hat{R}$, for all the parameters are well under the threshold.

# Chapter 7

# Benchmark Models

A couple of benchmark models were considered to validate the accuracy of the hierarchical model. The hierarchical model and the benchmark models will be compared based on the Receiver Operating Characteristic (ROC) and Area Under Curve (AUC). ROC is a plot showing true positive and false positive rates for all possible threshold values for classifying the predicted probabilities of a classification model. AUC is the area under the ROC plot and can be used to compare different classification models.

## 7.1 A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models

Gelman et al. [5] proposes an automatic and weakly informative prior distribution on regression coefficients so that it generalizes to any data set suffering from the problem of separation in the regression context. The proposed idea is to induce weak priors on each coefficient independently to prevent separation by shrinking the coefficients towards 0.

The first step of the model is to induce the same prior distribution on the regression coefficients. To do that, the regression inputs are normalized at first. A common approach is to standardize the regression inputs and standardizing binary inputs have the issue of interpretability as demonstrated in Gelman [18] which suggests solely centering for the binary inputs. By assuming that each binary variable has a balanced population rate, the standard deviation would be around 0.5. To make them on the same scale with continuous variables, the binary inputs are centered and the continuous inputs are centered and divided by two times the standard deviation. Hence, the numeric variables are expected to have the same scale as the binary variables. This way of scaling might still be problematic in terms of making the variables scale invariant if the proportion of any binary input is extremely unbalanced [18]. Since, in that case, the standard deviation of the binary input would be much lower than 0.5 which might be problematic as Gelman [18] points out.

The second step is to select the prior distribution such that the prior allows occasionally higher values of the regression coefficients which might not be plausible due to thin tails in case of a Normal prior [5]. As suggested by Gelman et al. [5], let us consider a logistic regression with one parameter $\theta$ and assume that the success and failure probability of the outcome are equal. Then, the likelihood function would be

$$L(\theta|y) = (\frac{e^{\theta}}{1+e^{\theta}})^{y}(\frac{1}{1+e^{\theta}})^{1-y} = \frac{e^{\theta y}}{1+e^{\theta}}. \tag{7.1}$$

Then, the likelihood would be $\frac{e^{\frac{\theta}{2}}}{e^{\theta}+1}$ by assuming equal chances of success and failure [5]. Finally, this function is assumed to constitute a prior on $\theta$. It is suggested that a similar prior to this function is a scaled-t distribution with $d.o.f. = 7$ and $s = 2.5$. The author considers a more conservative prior with $d.o.f. = 1$ by allowing coefficients to have higher values occasionally.

## 7.2 Naive Bayes

Let $Y$ and $X_i$ denote the response and $ith$ variable among n predictors. The Naive Bayes classifies an unseen instance by finding the class corresponding to maximum posterior density, i.e. $c^* = argmax_{c \in C}\{p(Y = c|X_1, \ldots, X_n)\}$ where $C$ is the set of categorical outcomes. The posterior density can be simplified as

$$\begin{aligned} p(Y|X_1, \ldots, X_n) &= \frac{p(Y, X_1, \ldots, X_n)}{p(X_1, \ldots, X_n)} = \frac{p(X_1, \ldots, X_n|Y)p(Y)}{p(X_1, \ldots, X_n)} \\ &= \frac{p(X_1, \ldots, X_n|Y)p(Y)}{\sum_{c \in C} p(X_1, \ldots, X_n|Y = c)p(Y = c)}. \end{aligned} \tag{7.2}$$

The first term and the second term in the numerator denote the likelihood and prior respectively. By assuming $X_i's$ are conditionally independent given $Y$,

$$p(X_1, \ldots, X_n|Y) = \prod_{i=1}^{n} p(X_i|Y), \tag{7.3}$$

which might be unreasonable since $X_i's$ might still depend even in the presence of $Y$. For categorical variables, $p(X_i|Y)$ can be estimated for an unseen instance by calculating $p(X_i = k|Y = c)$ using the training data where $k$ denotes any of the categories of $ith$ variable. For numeric variables, a continuous distribution is fit to the variable and the parameters of the distribution are estimated for each class separately by using the training data. Then, the density value for an unseen instance can be computed by using the parameter estimates for each class. The prior distribution, $p(Y)$, might be based on the prior knowledge of the user on the class probabilities or it might be automatically set to the fractions of classes of $Y$ in the training data.

## 7.3 Weighted Quadratic Random Forest (WQRF)

Gao et al. [2] uses a random forest algorithm to classify the employee turnover data. The data belongs to a communication company trying to detect the employees that are likely to quit. By doing this, they aim to prevent the turnover

of employees by enhancing their working conditions or hiring new employees in advance.

Proposed WQRF algorithm involves three steps namely forming training, validation and test set, variable selection and calculating the weights of the ensemble voting.

The data sets are firstly set. Training data is composed of $n$ data sets resampled out of the original data as much as the number of observations in the original data set. Each decision tree is built using only a portion of the predictors to avoid overfitting. Then, the classification error is calculated per decision tree by using out of bag data of each tree which is the remainder of the observations not chosen in the resampled data. It is suggested that the number of observations in the out of bag data set is expected to approximately amount one third of the original data set. Then, for each predictor, the values are updated by adding some noise and the classification error is calculated using out of bag data once again per decision tree. The average of the difference in the prediction error with and without noise among trees corresponds to variable importance. Since, if the predictor is significant, adding noise to the variable should increase the error in a relatively greater amount. At each iteration, the variable with the least importance is removed from the model and the steps above is repeated until the algorithm reaches the predefined number of variables. Instead of adding noise to the variables for the data at hand, the orders of the values were shuffled.

The advantage of the variable selection is to diminish the classification error and computation time [2]. After selecting the variables, the model is trained on the training set. Herein, the aim of the model is to detect the employees that are likely to leave. By the nature of the data, the response is unbalanced and the proportion of the outcomes is one to nine in favor of the working employees. Then, the classification error may not reflect the accuracy of the model. Since, labeling each employee as "not leaving" yields to approximately 90% of accuracy where labeling all samples as negative is worthless. It is proposed that F-measure of decision trees, $F_{measure} = \frac{2*recall*precision}{recall+precision}$, should be used as the weights of voting instead of using simple voting used in the random forest. F-measure considers the

precision as well as the recall so F-measure regards both capturing the positive outcomes as well as being precise in the estimation of them [2]. F-measure of each tree is estimated using the validation set which was also generated by resampling. The outcome estimates for the unseen observations are calculated based on the weighted ensemble.

## 7.4 XGBoost Algorithm for Prediction of Employee Turnover

Ajit [3] considers that Human Resources Information Systems are mostly underfunded which leads to noise in the data causing overfitting and it is noted that using Extreme Gradient Boosting (XGBoost) Algorithm might prevent the overfitting by adding a regularization term in the objective function. This might be the case for the turnover data at hand since there is only one employee trying to collect the data and the missingness is also seen in the recent data which might indicate the insufficiency of the data collection process. Even if this is not the case, an algorithm attacking overfitting might be useful yet.

XGBoost algorithm, which is proposed in Chen and Guestrin [4], is a tree ensemble method where each tree's prediction for an outcome are added up to predict the response.

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \tag{7.4}$$

where $F = \{f(x) = w_{q(x)}\}(q : R^m \to T, w \in R^T)$, $m$ is the number of variables, $T$ is the number of leaves and $q$ represents the tree mapping the predictors to $q(x)$ index of $w_{m\times 1}$. The proposed regularized objective is as the following:

$$\begin{aligned} \mathcal{L}(\phi) &= \sum_i \ell(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\ \text{such that } \Omega(f) &= \gamma T + \frac{1}{2}\lambda ||w||^2, \end{aligned} \tag{7.5}$$

where $\ell$ is the loss function and $k$ denotes the index of the $kth$ regression tree. The penalization term increases as the number of leafs and regression weights increase.

With the given setting, it is not feasible to minimize the given objective using the conventional optimization methods [4]. Instead, the objective is minimized in an iterative manner

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t), \tag{7.6}$$

which is approximated by Equation (7.7) for a given $q$ (See Chen and Guestrin [4] for the transition from $\mathcal{L}^{(t)}$ to $\tilde{\mathcal{L}}^{(t)}(q)$).

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T, \tag{7.7}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} \ell(y_i, \hat{y}_i^{(t-1)})$, $h_i = \partial^2_{\hat{y}^{(t-1)}} \ell(y_i, \hat{y}_i^{(t-1)})$ and $I_j$ is the set of the index of observations in $jth$ leaf. This corresponds to minimizing the objective at iteration $t$ with respect to $w$ for a given tree $q$. However, it is not likely to find out the tree giving the best possible reduction in the objective by evaluating them all [4]. Therefore, the following measure is used for finding the best split at each node:

$$\mathcal{L}_{split} = \frac{1}{2} \left[ \frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma, \tag{7.8}$$

where $I_L$, $I_R$ and $I$ are left and right child nodes after the split and the parent node.

A brief of how the algorithm works is as the following. An initial tree is firstly set. At each iteration, $q$ is constructed based on $\{g_i\}$, $\{h_i\}$, $\lambda$, $\gamma$ and $\mathcal{L}_{split}$ which only depends on the hyperparameters and the trees which were set yet. Then, for the given tree, the tree weights can be set and the tree construction is repeated accordingly.

Most favorable feature of XGBoost apart from the regularization term is the heuristic methods used in finding the best split for decreasing computation time. However, using a heuristic is not needed.

# Chapter 8

# Results and Comparison of the Models

## 8.1 Hieararhical Model Training

The hierarchical model was compared to the benchmark models to validate the overall accuracy of the model. The performance measure was selected as AUC since the accuracy as a validation measure is not useful due to the imbalanced response. The model was trained on from 2016 to 2017, validated on 2018 to tune the parameters and tested on 2019. For each replication of MICE, the model was trained and the success probabilities were computed for each iteration by taking the inverse logit transformation. Then, the estimates across the replications and iterations of the hierarchical model were pooled by taking the average. These steps were also incorporated for the other models which required tuning.

The model was ran for grid of values of $A$, $V$ and $v$. The results were validated on the validation set and the hyperparameter setting giving the highest AUC value was selected for the training set which was used to estimate the success probabilities of the outcome in the test set. $A = 10$, $v = 100$ and $V = 0.025 \times I_{k\times k}$, where $I_{k\times k}$ denotes the identity matrix, was selected as the values of the

hyperparameters based on the grid search. It should be noted that $v$ affects the degrees of freedom as well as the mean of the $V_\beta$ so the value being over 30 is not trivial. In addition, $\bar{\Delta}$ was set to 0 matrix so that $\{\beta_i\}$ can be shrunk towards 0 to avoid overfitting.

100000 iterations with a burnin period of 10000 were drawn for each 5 replication of MICE. Although the Hybrid Sampler combining the Gibbs Sampler and Metropolis Hastings is not efficient considering the high autocorrelation, it took approximately 3 minutes for model to generate the results. Since the data is based on the annual information of the employees, the model is expected to be ran a couple of times a year, especially when there is an update in the data. It does not seem to be feasible to convert the data into monthly based format so that the estimations can be made on a monthly basis. Since, the response is already imbalanced and narrowing the observation of outcome period would make imbalance more severe. In a different context, the models sampling from more parameters might require more efficient sampling techniques such as NUTS sampler.

The model was ran using a framework called *Rcpp* in *R* which allows to use some packages in $C++$. *Armadillo* package increased the pace of the algorithm. Especially, the loops in $C++$ were handled faster compared to *R*'s. Instead of running the model serially, independent chains were ran to fully use the CPU of the machine. 10 chains were ran for 5 replications of MICE where two chains were ran for each replication by keeping the burnin period as 10000 and having the same amount of actual samples so 55000 iterations with a burnin period of 10000 iterations were ran for each chain.

## 8.2 Inference About the Model

The purpose of the model was to segregate the regression parameters of the units so that the heterogeneity of the employees tendency towards the attrition can be explained. There might be two cause of the difference in the $\{\beta_i\}$. The first

reason is that their prior mean is different which is driven by the difference in the demographic information. Obviously, the second reason is due to the difference of the information in the likelihood. $V_\beta$ balances the effect of these two mechanisms which are the prior mean, driven by demographics, and the likelihood. As $V_\beta$ gets smaller, the prior becomes more flat and the likelihood becomes more dominant on the posterior. It should also be realized that as $V_\beta$ gets smaller, the prior on $\Delta$, which has 0 mean, becomes less dense and $\Delta$ values increase. This seems as a defect of using conjugate prior because tuning one of the parameters might have an undesired effect on the other parameters. However, the effect of changing $V_\beta$ considering $\{\beta_i\}$ can be suppressed by $A$ where increasing precision parameter yields to shrink $\Delta$ towards 0.

Furthermore, based on the experiments on the validation set, the models built up on using the Alma Mater and English Proficiency, which are demographic variables, lead to having many units not having sufficient amount of observations and this lead to a lower performance on the target metric. On the other hand, combining these two variables based on the content of the categories into one variable, which is called to be Education Level, provided higher AUC values on the validation set. Specifically, the employees having a High Alma Mater and Sufficient English Proficiency were assigned to the High category and the employees having a Low Alma Mater and Insufficient English Proficiency were assigned to the Low category and the rest were labeled as Medium.

To confirm if the model works as intended, the very first check that should be made is to see if $\{\beta_i\}$ are different. Table 8.1 shows the $\{\beta_i\}$ input's coefficients across the different units. It should be noted that the intercept values are more likely to deviate because the intercept values include the effects of raw demographic variables. Then, the intercept term can be thought as a combination of multiple demographic values. To compare $\{\beta_i\}$ values, the statistical tests can be applied but in case of applying statistical tests with this much of samples, small differences in the distributions lead to rejections. This can be handled by thinning the samples. However, there is also the problem of assuming a distribution on the samples coming from. Although $\{\beta_i\}$ have a Normal Prior, the posterior might not be reflective of it. Bimodal marginal distributions were observed in

some of $\{\beta_i\}$ so it is not trivial to make assumptions regarding $\{\beta_i\}$.

Table 8.1: Sample Mean of $\{\beta_i\}$

| | X.Intercept. | Number.of.Years.WorkedMedium | Number.of.Years.WorkedHigh | Highest.DegreeOver.Bachelor | First.Year1 | Child1 | Child.Born1 | Marital.StatusSingle | WorkplaceOther | Performance.ScoreBad | Performance.ScoreOkay | Appreciation.ScoreBad | Appreciation.ScoreGood | SalaryMedium | SalaryHigh | AgeMedium | AgeHigh | Salary.RaiseVery.High | Salary.RaiseHigh | Salary.RaiseLow |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unit 1 | -2.690 | 0.253 | -0.056 | -0.044 | -0.170 | -0.265 | -0.128 | -0.011 | -0.056 | -0.158 | 0.043 | 0.212 | -0.216 | -0.047 | 0.067 | -0.324 | 0.047 | -0.349 | -0.350 | 0.133 |
| Unit 2 | -2.680 | 0.262 | 0.004 | -0.234 | -0.317 | -0.336 | -0.051 | -0.122 | 0.221 | -0.056 | -0.109 | 0.217 | -0.233 | -0.104 | -0.072 | -0.116 | -0.033 | -0.311 | -0.360 | 0.263 |
| Unit 3 | -2.605 | 0.172 | -0.095 | -0.169 | -0.232 | -0.349 | -0.113 | -0.032 | -0.005 | -0.012 | -0.205 | 0.230 | -0.225 | -0.020 | -0.046 | -0.170 | -0.080 | -0.177 | -0.364 | 0.165 |
| Unit 4 | -2.565 | 0.272 | -0.103 | -0.268 | -0.282 | -0.383 | -0.163 | 0.035 | -0.067 | -0.036 | -0.056 | 0.217 | -0.227 | 0.014 | 0.098 | -0.191 | -0.127 | -0.291 | -0.532 | -0.108 |
| Unit 5 | -2.505 | 0.186 | 0.017 | 0.066 | -0.214 | -0.383 | -0.099 | 0.097 | 0.057 | -0.017 | -0.215 | 0.225 | -0.125 | -0.039 | 0.069 | -0.028 | -0.109 | -0.123 | -0.429 | 0.128 |
| Unit 6 | -2.416 | 0.168 | -0.208 | -0.123 | -0.135 | -0.306 | -0.119 | -0.105 | -0.053 | 0.107 | -0.210 | 0.401 | -0.322 | -0.044 | 0.004 | -0.106 | -0.187 | -0.249 | -0.331 | 0.018 |
| Unit 7 | -1.989 | 0.212 | 0.030 | 0.016 | -0.316 | -0.260 | -0.153 | 0.001 | -0.231 | -0.101 | -0.244 | 0.144 | -0.087 | 0.025 | 0.093 | -0.135 | 0.028 | -0.083 | -0.320 | -0.023 |
| Unit 8 | -1.886 | 0.356 | -0.141 | 0.042 | -0.270 | -0.377 | -0.243 | -0.016 | 0.058 | -0.089 | -0.183 | 0.237 | -0.203 | 0.104 | 0.187 | -0.146 | -0.078 | -0.006 | -0.305 | 0.008 |
| Unit 9 | -2.281 | 0.106 | -0.007 | -0.094 | -0.307 | -0.378 | -0.126 | 0.056 | 0.004 | -0.049 | -0.228 | 0.027 | -0.053 | -0.072 | 0.000 | -0.082 | 0.010 | -0.256 | -0.360 | 0.290 |
| Unit 10 | -2.682 | 0.280 | -0.079 | -0.282 | -0.348 | -0.594 | -0.266 | 0.054 | -0.025 | -0.012 | -0.095 | 0.231 | -0.192 | 0.012 | 0.267 | -0.142 | -0.048 | -0.261 | -0.443 | 0.326 |
| Unit 11 | -1.590 | 0.173 | 0.019 | -0.021 | -0.234 | -0.278 | -0.091 | -0.053 | -0.022 | 0.023 | -0.158 | -0.040 | 0.007 | 0.123 | 0.043 | -0.093 | 0.052 | -0.177 | -0.205 | 0.147 |
| Unit 12 | -1.966 | 0.142 | 0.063 | -0.084 | -0.206 | -0.445 | -0.158 | 0.121 | 0.052 | 0.136 | -0.209 | -0.011 | -0.039 | -0.028 | 0.078 | -0.092 | 0.174 | -0.205 | -0.416 | 0.271 |
| Unit 13 | -2.127 | 0.203 | -0.084 | -0.173 | -0.269 | -0.346 | -0.158 | -0.055 | 0.086 | 0.017 | -0.128 | 0.272 | -0.234 | 0.000 | 0.139 | -0.174 | -0.056 | -0.046 | -0.366 | 0.302 |
| Unit 14 | -2.126 | 0.219 | -0.066 | -0.196 | -0.264 | -0.295 | -0.157 | -0.049 | 0.068 | 0.063 | -0.116 | 0.158 | -0.127 | 0.006 | 0.112 | -0.183 | 0.028 | -0.194 | -0.286 | 0.431 |
| Unit 15 | -1.534 | 0.141 | 0.018 | -0.051 | -0.174 | -0.271 | -0.116 | 0.018 | 0.028 | 0.086 | -0.219 | 0.182 | -0.025 | 0.087 | -0.017 | -0.016 | -0.053 | 0.034 | -0.317 | 0.034 |
| Unit 16 | -2.148 | 0.267 | -0.142 | -0.149 | -0.310 | -0.403 | -0.142 | -0.093 | 0.031 | -0.049 | -0.140 | 0.190 | -0.254 | 0.162 | -0.046 | -0.297 | -0.142 | -0.139 | -0.248 | 0.239 |
| Unit 17 | -1.446 | 0.141 | 0.012 | -0.080 | -0.178 | -0.257 | -0.062 | 0.003 | 0.021 | 0.106 | -0.172 | 0.119 | -0.074 | -0.006 | -0.016 | -0.083 | 0.005 | -0.155 | -0.128 | 0.199 |
| Unit 18 | -1.938 | 0.147 | 0.015 | -0.115 | -0.209 | -0.260 | -0.100 | -0.006 | -0.041 | 0.030 | -0.201 | 0.202 | -0.093 | 0.088 | -0.030 | -0.092 | -0.053 | -0.076 | -0.257 | 0.242 |
| Unit 19 | -1.748 | 0.202 | -0.222 | -0.117 | -0.371 | -0.382 | -0.191 | 0.067 | -0.121 | -0.059 | -0.026 | 0.398 | -0.133 | -0.053 | 0.095 | -0.076 | -0.262 | -0.111 | -0.235 | -0.063 |
| Unit 20 | -2.035 | 0.191 | -0.032 | -0.022 | -0.195 | -0.378 | -0.149 | -0.085 | -0.011 | -0.026 | -0.108 | 0.136 | -0.123 | -0.005 | 0.157 | -0.142 | 0.033 | -0.186 | -0.273 | 0.184 |
| Unit 21 | -1.905 | 0.149 | 0.074 | -0.167 | -0.227 | -0.189 | -0.100 | -0.231 | 0.047 | -0.076 | -0.019 | -0.016 | -0.050 | 0.010 | -0.055 | -0.107 | 0.049 | -0.229 | -0.230 | 0.142 |
| Unit 22 | -1.247 | 0.088 | 0.049 | 0.023 | -0.127 | -0.190 | -0.069 | -0.087 | 0.084 | 0.053 | -0.168 | 0.045 | -0.016 | 0.044 | 0.021 | -0.052 | 0.080 | -0.058 | -0.160 | 0.127 |
| Unit 23 | -1.880 | 0.208 | -0.154 | -0.097 | -0.276 | -0.297 | -0.111 | -0.073 | -0.006 | -0.028 | -0.146 | 0.338 | -0.249 | -0.039 | 0.035 | -0.093 | -0.157 | -0.149 | -0.305 | 0.109 |
| Unit 24 | -1.705 | 0.147 | -0.023 | -0.120 | -0.191 | -0.258 | -0.095 | -0.063 | -0.066 | 0.043 | -0.203 | 0.135 | -0.114 | 0.041 | 0.077 | -0.143 | -0.028 | 0.070 | -0.291 | 0.259 |

To make inference regarding the difference in $\{\beta_i\}$, for each unit $i$, the regression coefficients were ranked in the ascending order. For each rank and regression coefficient pair, the occurrences of ranks were counted across the units. The first and last three ranks versus regression coefficients are provided in Table 8.2. One can realize that the intercept term consistently has the lowest value compared to other coefficients for all the units. This is expected considering the imbalanced response so the intercept is significantly lower than 0. Also, $Marital.StatusSingle$ input has $2nd$ for one of the units. Contrarily, it has $18th$ rank for another unit. Based on the distribution of the ranks across different units leads to the conclusion that $\{\beta_i\}$ might have heterogeneity.

Table 8.2: Inputs versus Ranks across Units

| Inputs | Rank:1 | Rank:2 | Rank:3 | Rank:18 | Rank:19 | Rank:20 |
|---|---|---|---|---|---|---|
| AgeMedium | 0 | 0 | 0 | 0 | 0 | 0 |
| AgeHigh | 0 | 0 | 0 | 0 | 1 | 0 |
| Appreciation.ScoreBad | 0 | 0 | 0 | 4 | 6 | 6 |
| Appreciation.ScoreGood | 0 | 0 | 1 | 0 | 0 | 0 |
| Child.Born1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Child1 | 0 | 12 | 8 | 0 | 0 | 0 |
| First.Year1 | 0 | 0 | 5 | 0 | 0 | 0 |
| Highest.DegreeOver.Bachelor | 0 | 0 | 0 | 0 | 0 | 0 |
| Marital.StatusSingle | 0 | 1 | 0 | 1 | 0 | 0 |
| Number.of.Years.WorkedMedium | 0 | 0 | 0 | 3 | 13 | 8 |
| Number.of.Years.WorkedHigh | 0 | 0 | 0 | 1 | 0 | 0 |
| Performance.ScoreBad | 0 | 0 | 0 | 1 | 0 | 0 |
| Performance.ScoreOkay | 0 | 0 | 1 | 0 | 0 | 0 |
| SalaryMedium | 0 | 0 | 0 | 2 | 0 | 0 |
| SalaryHigh | 0 | 0 | 0 | 6 | 0 | 0 |
| Salary.Raise.Low | 0 | 0 | 0 | 4 | 4 | 10 |
| Salary.Raise.High | 0 | 11 | 8 | 0 | 0 | 0 |
| Salary.RaiseVery.High | 0 | 0 | 1 | 0 | 0 | 0 |
| WorkplaceOther | 0 | 0 | 0 | 2 | 0 | 0 |
| X.Intercept. | 24 | 0 | 0 | 0 | 0 | 0 |

Finally, the other purpose of the model was to shrink regression coefficients towards 0 so that the model does not overfit. This can be ensured by the model's performance on the test set which will be discussed in the next sections.

## 8.3 Becnhmark Models Application

XGBoost algorithm was tuned based on $\gamma$ and $\lambda$ in the objective function which prevent overfitting. In addition, $\eta$ is used to scale the weights, $w$, at each iteration which corresponds to learning rate. The subsample variable was used to avoid overfitting which chooses a ratio of the samples in the tree to be grown. Lastly, the positive instances' gradients were scaled to deal with imbalanced response. By scaling the gradients, positive instances have greater impact in the loss function so the loss of positive instances become more costly. XGBoost algorithm also allows many other parameters to be set but since these parameters raise the number of calculations exponentially considering the curse of dimensionality, they were excluded from the grid search.

WQRF algorithm involves variable selection which restrict the number of variables to 15 as in Gao et al. [2]. However, this number is irrelevant to the data at hand so the number of variables corresponding to the highest AUC value in the validation set was picked.

For both WQRF and XGBoost, Salary and Salary Raise variables were not converted into categorical form and $Age$ and $Number.of.Years.Worked$ variables were transformed into continuous form after MICE. Since, both algorithms promise to capture the nonlinearity of the response with variables. Therefore, the concerned variables were converted into continuous form to fully use the information within the variables.

The other models did not require tuning so they were applied directly. The prior value in Naive Bayes was set based on the outcome ratios in the training sets.

## 8.4 Comparison of the Models

The ROC curves for each model were computed and AUC values were calculated based on them. The Figure 8.1 shows that the hierarchical model outperforms the other models on AUC and performs only slightly better than Weakly Automated Prior Model.

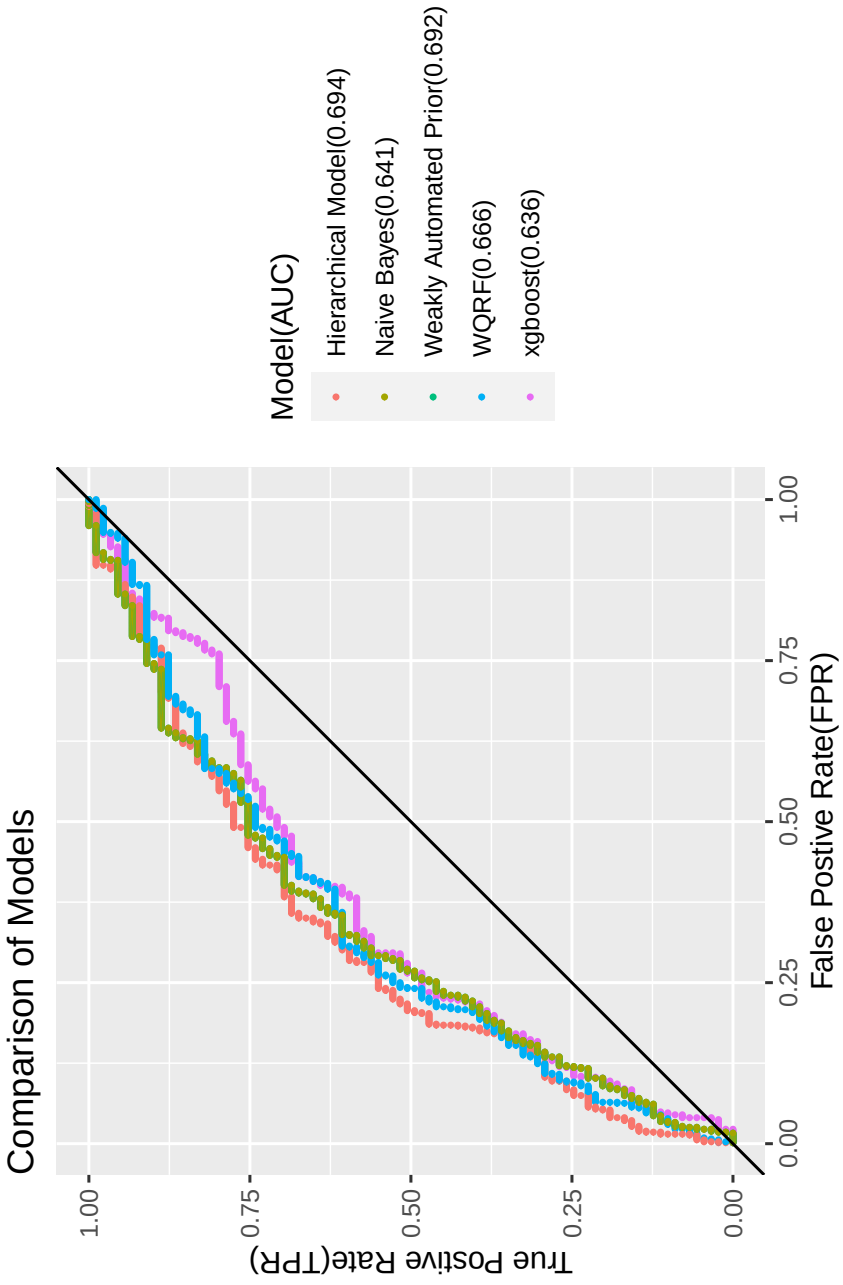


Figure 8.1: ROC and AUC Values across Models

# Chapter 9

# Conclusion

In this study, the employee turnover problem was discussed along with the data provided by a company manufacturing agricultural machinery. The purpose of the study was to estimate the turnover probabilities on a yearly basis. The provided data contained missing values which were imputed by MICE. A hierarchical model targeting to explain the heterogeneity in attrition tendency was applied. The model was designed to explain the heterogeneity by demographic features of the units. In addition, since the data contained many categorical variables, it was open to overfitting which was also expected to be taken care of by the hierarchical model.

Based on the literature search, WQRF, Weakly Automated Prior, XGBoost and Naive Bayes algorithms were used for modelling the response to provide a benchmark to the main model. The main model seemed to outperform all the models other than Weakly Automated Prior Model and most importantly, this was achieved by unleashing the heterogeneity of the employees towards attrition. The model is expected to be used by the company which provided the data and it is anticipated that the employee turnovers can be predicted in advance to make the proactive actions.

# Bibliography


[1] P. E. Rossi, G. M. Allenby, and R. McCulloch, *Bayesian statistics and marketing*. John Wiley & Sons, 2012.

[2] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," *Mathematical Problems in Engineering*, vol. 2019, 2019.

[3] P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms," *algorithms*, vol. 4, no. 5, p. C5, 2016.

[4] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[5] A. Gelman, A. Jakulin, M. G. Pittau, Y.-S. Su, *et al.*, "A weakly informative default prior distribution for logistic and other regression models," *Annals of applied Statistics*, vol. 2, no. 4, pp. 1360–1383, 2008.

[6] S. Van Buuren, *Flexible imputation of missing data*. CRC press, 2018.

[7] S. v. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, pp. 1–68, 2010.

[8] S. Van Buuren, H. C. Boshuizen, and D. L. Knook, "Multiple imputation of missing blood pressure covariates in survival analysis," *Statistics in medicine*, vol. 18, no. 6, pp. 681–694, 1999.

[9] S. Van Buuren, J. P. Brand, C. G. Groothuis-Oudshoorn, and D. B. Rubin, "Fully conditional specification in multivariate imputation," *Journal of statistical computation and simulation*, vol. 76, no. 12, pp. 1049–1064, 2006.

[10] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[11] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[12] R. J. Little and D. B. Rubin, "Bayes and multiple imputation," *Statistical analysis with missing data*, p. 120, 2002.

[13] J. L. Schafer, *Analysis of incomplete multivariate data*. CRC press, 1997.

[14] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[15] A. Zellner, "On assessing prior distributions and bayesian regression analysis with g-prior distributions," *Bayesian inference and decision techniques*, 1986.

[16] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical science*, vol. 7, no. 4, pp. 457–472, 1992.

[17] S. P. Brooks and A. Gelman, "General methods for monitoring convergence of iterative simulations," *Journal of computational and graphical statistics*, vol. 7, no. 4, pp. 434–455, 1998.

[18] A. Gelman, "Scaling regression inputs by dividing by two standard deviations," *Statistics in medicine*, vol. 27, no. 15, pp. 2865–2873, 2008.