Xiaojuan Zhu

June 22, 2015

Abstract

1 Introduction

Employee turnover is a topic that has drawn the attention of management researchers and practitioners for decades, because employee turnover is both costly and disruptive to the functioning of most organizations (**???**), and both private firms and governments spend billions of dollars every year managing the issue according to **?**. Therefore, understanding the causes of turnover: retirement and voluntary quit, examining the internal and external impacts, effectively forecasting the turnover by these two causes, and measuring the effectiveness and to what extent of the HR policy at firm and departmental levels are the key questions in this study for reducing it and for effective planning, budgeting, and recruiting in the human resource filed. As a funded research project, a large organizational secondary dataset including 12-year employees demographic information and records is transformed, analyzed and modeled by Cox proportional hazard regression models with a time dependent covariate using competing risks analysis to examine the statistically significant factors and to predict employees' conditional retiring and voluntary quitting probabilities. The dataset are also employed to logistic regression and time series models for compare the performance of cox proportional hazard model.This study also examines the forecasting capability of Cox proportional hazard model on the data with two kinds of bias (left truncation and right censor) by simulation.

2 Literature Review

3 Data Preparation

The turnover dataset is a large real world secondary dataset from a multipurpose research organization in the U.S. The dataset consists 4316 current active and 3782 terminated full-time employees' information including metrics such as payroll category, hired date, company start date, company credit service date, termination date, age at hired , years of service at hired (YCSH), gender, job classification (named as Cocscode), and department code (named as division). The company credit service date is the date that the organization starts to credit their retirement plan. Years of service (YCS) is the total years credit for employees' retirement plan. The employees are eligible to get full retirement or pension, when their age is at least 65 or the points is greater than 85, which is the sum of age and year of service. The window of time for the turnover dataset is from November 2000 to December 2012, i.e. the dataset consists the records only for the employees working in the organization from November 2000 to December 2012, indicating there is no records for employees leaving the organization before November 2000 and no termination date for 4316 current employees.

The turnover dataset is split into two datasets: training and holdout dataset. The training part is used to build the model and the holdout part is to validate the model performance. Two methods are used to slipt the dataset in order to validate the model performance: One is split data by a time point November 1 2010: training (November 1, 2000 - October 31, 2010) and holdout (November 1st, 2010 - December 31, 2012). The other on is to random split the turnover dataset into 2/3 of the dataset as training and 1/3 of the dataset as holdout. The covariates identified from the turnover dataset and used to build the models are payroll, gender, division, cocs code (Job category), age at hired, and year of service at hired:

- Payroll (PR): hourly, weekly, or monthly payroll,
- Gender: male, female
- division (ORG): ten divisions in the organization.
- Cocs code: Crafts(C), Engineers (E), General Administrative (G), Laborers (L), General Managers (M), Administrative (P), Operators (O), Scientists (S), Technicians (T))
- Age at hired: most recent age when an employee is hired.
- Years of service at hired (YCSH): the years of service which accounts for pension plan when employee is most recently hired.

Several economic indices are being considered and tested their as a variable impact on employee turnover. These indices include unemployment index, housing price index (HPI), investment index, and marketing index. Seasonal adjusted unemployment rate is published by Bureau of Labor Statics from United department of Labor (?). U.S housing price index, U.S. and southeastern monthly purchase-only index are considered as another economic indicator variables in the study (?). S&P 500 indices published from S&P Dow Jones Indices are also considered as investment index including S&P 500, Dividend, Earnings, Consumer index, Long Interest Rate, Real Price, Real Dividend, Real Earnings, P/E 10 ratio (?). Wilshire 5000 total market full cap index published by Wilshire Associates is considered as market index in the forecasting model(?). All these twelve indices are treated as variables using their twelve-month lag term in yearly data format. All these indices selected are indicators in various economic areas, such as job market, house market, and stock market, representing the fluctuation of these economic areas.

4 Model Development and Evaluation

Several questions have to be addressed by this study, such as when a specific employee will turnover, how many employees will turnover in a curtain department or a job category, what factors do affect their turnover, by what reason they will turnover: voluntary quit or retirement. And also the dataset has two kinds of unknown information, which are no records for employees leaving the organization before November 2000 and no termination date for 4316 current employees. These two kinds of unknown information cause two kinds of bias: right censor and left truncation. Besides, there is an intervention event implemented in year 2008 due to the downsizing policy in the organization. Therefore, how to model this incentive and to estimate its effect are another interesting questions. All these questions and problems can be solved by lifetime analysis, also called survival analysis. The survival analysis is to analyze the time duration for the occurrence of an events or certain events. The events can be the death of the patients, the failure of the machine, and the leaving of the employees by any reasons for this study. There are two kinds of survival statistic models: parametric survival models and Cox proportional hazards (PH) models. In this study, Cox PH model is employed to build the forecasting model, to generate a employees' working life baseline (distribution), and to identify significant factors for turnover. The parametric models are not appropriate for this study, because it is hard to fit the employees' working life distribution to any parametric distributions, such as Weibull or log-normal distribution. Time dependent covariates are incorporated for fitting the 2008 intervention event due to the downsize policy in the organization and for examining the effects of economic indicators. Competing risks analysis is applied for modeling employee retirement and voluntary quit. Besides, A simulation study is performed to examine the forecasting capability of cox proportional hazard model on the left truncation and right censor dataset.

4.1 Two data bias: right censor and left truncation

Right censor and left truncation are common in survival analysis. The right censor is that the event of interest (failure) occurs after the study window. Let T denotes the time of main event of interest to occur and let C denotes the end time of study. An observation is right censored when $T > C$, indicating the study do not have the failure time of the right censored observation. In this study, the study window is from November 2000 to December 2012 as shown in the figure **??**. Thus, the current active employees have unknown terminated date. They are treated as right censor. These right censored observations require special treatment

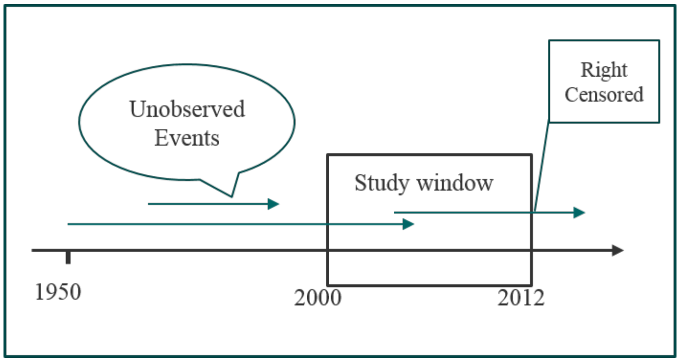

Figure 1: Right censor and left truncation

in survival analysis: a censor indicator variable is created:

$$\delta_i = \begin{cases} 1 & \text{if } t_i \leq c_i \text{ (uncensored)}, \\ 0 & \text{if } t_i > c_i \text{ (censored)}, \end{cases}$$

where, i denotes the ith observation, and the failure time of event for ith observation is minimum time between t_i and c_i, i.e., $min(t_i, c_i)$, that is when $c_i < t_i$, c_i is taken as end time of the ith observation in order to do next analysis.

Left truncation is that the occurrence of an intermediate event prior to the event of interest appear in the sample dataset. Let T denotes the time of event of interest to occur and let X denotes the time an individual enters the study, that is time of truncation events occurs. Only the individuals with $T \geq X$ are observed in the study window. Left truncation in this study occurs due to no records for employees leaving the organization before November 2000 as unobserved events shown in the figure **??**. The left truncation leads to another bias. As shown in the figure **??**, the longest arrow represents a life span for an employee hired in 1950 and left in 2006. Those employees who remain in the study window increase the apparent lifetimes. The existence of truncation in the data must be taken into account in order to overcome this bias and to achieve accurate estimation of survival analysis (**?**). Let t_{i0} denotes the start time of the ith observation, i.e., hired time or age at hired of ith employee, x_i denotes the entry time of the ith observation, i.e., the start time of study (November 1st, 2000) or age at November 1st, 2000. The start time of the observation is maximum value between t_{i0} and x_i, that is when $t_{i0} < x_i$, x_i is taken as start time of the ith observation in order to eliminate the left truncation bias (**?**). The number of failures in the t_j is redefined for left truncation. When $x_i < t_j \leq t_i$, the observation is in the risk set. When $t_j < x_i \leq t_i$, the ith observation has not entered study yet at t_j and it cannot be considered in the risk set. When $x_i \leq t_i < t_j$, it indicates the ith observation whose failure time before t_j, and it cannot be considered in the risk set at time t_j neither (**?**).

simulation study.

4.2 Cox PH regression model

Cox proportional hazards (PH) regression is a widely used method for estimating survival life events, introduced in a seminal paper by **?**. The Cox PH model is usually taken the form of hazard model formula as shown in the equation **??**:

$$h(t, x) = h_0(t)e^{(\sum_{i=1}^{k} \beta_i x_i)} \tag{1}$$

where $x = (x_1, x_2, \ldots, x_k)$, $h_0(t)$ is the baseline hazard occurring when $x = 0$, β is the coefficients of x. The model provides a hazard expression at time t for an individual with a given specification of a set of explanatory variables denoted by the x. The Cox PH formula is the product of quantities at hazard time t: $h_0(t)$ as the baseline hazard function and the exponential expression to the linear combination of $\beta_i x_i$, x does not involve time t, so it is time-independent covariates. x can also be time-dependent covariates, which named extended Cox PH regression as discussed in the section **??**. The Cox PH regression is "robust" and popular, because the baseline hazard function $h_0(t)$ is an unspecified function and its estimation can closely approximate correct parametric model (**?**). Taking the logarithm of both sides of the equation, the Cox PH model is rewritten in the equation **??**:

$$\log h(t,x) = \alpha(t) + \sum_{i=1}^{k} \beta_i x_i \tag{2}$$

where $\alpha(t) = \log h_0(t)$. If $\alpha(t) = \alpha$, the baseline is exponential distribution. In the Cox PH regression, $\alpha(t)$ do not limited on specific parametric distributions and it can take any form. The partial likelihood method is used to estimated β coefficients of the Cox model without having to specify the baseline (**?**). The Cox PH model is performed by SAS.

4.3 Time dependent covariate and counting process

A time dependent covariate is that a covariate is not constant through the whole study and its value changes over the course of the study. The extended Cox PH regression model incorporates both time-independent and time-dependent covariates as shown in the equation **??**:

$$h(t,x) = h_0(t)e^{(\sum_{i=1}^{k1} \beta_i x_i + \sum_{j=1}^{k2} \gamma_j x_j(t))} \tag{3}$$

where $x = (x_1, x_2, \ldots, x_{k_1}, x_1(t), x_2(t), \ldots, x_{k_2}(t))$, $h_0(t)$ is the baseline hazard occurring when $x = 0$, β and γ are the coefficients of x. There are two time dependent covariates in this study: policy and economic indicators. Policy is to handle the downsizing policy issued in January 2008 with three months response time window to accommodate a voluntary reduction in force from the organization. Policy is a dummy variable across years:

$$Policy = \begin{cases} 1 & \text{if employee works in year 2008,} \\ 0 & \text{if employee does not work in year 2008.} \end{cases}$$

Counting process method in SAS programming statements is used to handle time dependent covariates, which is each employee have multiple records. Each record is related to a time interval and the covariates in this record remain constant. Therefore, Each employee has up to 3 records: before 2008, in-between 2008, and after 2008. Two variables, age, year of service, are used for repenting two time terminals of each interval or record. For age, one time point is age at beginning of the certain period, named ”age at start”; and the other one is age at end of the curtain period, named ”age at end”. Two year of services points are also generated for each record: one is year of service at the beginning of the period, named ”YCS at start”; the other one is the year of service at the end of the period, named ”YCS at end”.

Economic indicators is another time dependent covariates. Because economic indicators are fluctuated across the year, all the employees have up to 12 years records based on the calender year, which interval starts from hired date or January 1, and ends at terminated date or December 31 of certain year during the study window as shown in equation . The economic indicators are taken the average value for each year into the optimal model identified from the internal covariates to examine their impacts on turnover.

$$\begin{aligned} (\text{start point}, \text{end point}) = (&max(\text{hired date}, \text{January 1 of a certain year}), \\ &min(\text{terminated date}, \text{December 31 of a certain year})) \end{aligned} \tag{4}$$

4.4 Competing risks

A competing risk is an event whose occurrence either precludes the occurrence of the event of interest or fundamentally alters the probability of occurrence of this event of interest (**?**). For example, turnover causes of an employee are exclusive and independent, i.e. an employee can experience only one event such as voluntary quit rather than retirement. This alters the probability of experiencing the event of interest, like retirement. Such events are known as competing risks events where one event of several different types of possible events can occur and hence the survival analysis for each event is calculated separately with the other events set as censored. Two mutually exclusive causes: retirement and voluntary quit are considered as the event of interests for each employees in this study, and the other events are treat as censored.

There are several reasons for selecting these two causes. One main reasean is because the organization are interested in forecasting the turnover of retirement and voluntary quit. There are 1/3 employees in that organization are over 50 years old who are eligible for retirement. The employee who voluntary quit usually is the one organization would like to keep (**?**). And also voluntary quit costs highly for the organizations and firms (**?**). Finally, the other reasons of turnover, such as layoff, transfer, death, or disability are caused by the factors which occurrence are random and hard to predict. The Cox PH regression for competing risks as shown in equation **??**:

$$h_j(t,x) = h_{j0}(t)e^{(\sum_{i=1}^{k}\beta_{ij}x_i)} \tag{5}$$

where, x_j is the covariate for a specific type of turnover. Note that the coefficient β is the effects of the covariates may be different from different turnover types. If β_{ij} is the same for all j, the model simplified to Cox PH model as shown in equation **??**.

4.5 Stratification model

I. what is stratification model;
II. how to select a stratify variable

4.6 Variable selection

All the covariates are putting into Cox PH regression model and selected by manually backwards selection method based on $P < 0.05$. The variable selection procedure is as follow: first, all the covariates are used to build the model. Second, remove the non-significant variable ($P > 0.05$) with the largest P value, and rerun the model with the other variables. Then, repeat the second step until there is no significant variable remaining in the model.

4.7 Model evaluation and comparison

The Cox PH model is evaluated by four statistics criteria: Akaikes information (AIC), Schwartzs Bayesian criterion (SBC), C-statistics, and mean absolute percentage value (MAPE). The optimal model should have low AIC, SBC, and MAPE value, and high C-statistics for both training and holdout dataset. In this study, the model performance on holdout dataset is considered more important than that on the training dataset. AIC and SBC are both information criteria using likelihood value. Usually, the best model comes with lowest AIC or SBC values. AIC, SBC values are automatically generated by the models.

C-statistics or the area under the receiver operating characteristic (ROC) curve is to test whether the probability of predicting the outcome is better than chance. It ranges from 0.5 to 1. Models are considered acceptable when the C-statistic is higher than 0.7 (?). C-statistics are calculated by using the predicted failure probability compared with the actual outcomes by SAS proc logistic. The predicted failure (retirement or voluntary quit) probability is actually the conditional failure probability for an employee at time t_j, given that the employee is active at time t_{j-1}. It is calculated based on the baseline and coefficients from Cox PH models for both training and holdout dataset as shown in equation **??**.

$$\begin{aligned} P\{t_{j-1} < T < t_j\} &= 1 - P\{T > t_j | T \geq t_j\} \\ &= 1 - \frac{S_{t_j}}{S_{t_{j-1}}} \\ &= 1 - \frac{S_0(t_j)^{(\sum_{i=1}^{k} \beta_i x_i)}}{S_0(t_{j-1})^{(\sum_{i=1}^{k} \beta_i x_i)}} \end{aligned} \tag{6}$$

where, T is survival time, t_j is a specific value for T, $S_0(t)$ is the baseline function generated by Cox PH model, x is the covariates, and β is the coefficient.

MAPE is another measure for comparing the accuracy of the model fitting between different forecast models since it measures relative performance (?) as shown in the equation **??**.

$$MAPE = \sum_{t=1}^{n} \left| \frac{y_t - \hat{y_t}}{y_t} \right| \frac{1}{n} \% \tag{7}$$

MAPE is calculated by using the yearly actual and predicted retirement or voluntary quit number as y_t and $\hat{y_t}$, respectively. The predicted retirement or voluntary quit number is the expected retirement or voluntary number summarized by aggregating all the failure probabilities for the active employees in the risk set at t_j as shown in **??**.

$$E(\text{turnover number at } t_j) = \sum_{i=1}^{k} P_i\{t_{j-1} < T < t_j\} \tag{8}$$

where, i denotes the ith employee. The logistic regression and time series moving average methods are also employed to compare with the performance of Cox PH regression model by MAPE value.

4.8 Results

4.8.1 Right censor and left truncation simulation results

4.8.2 Retirement model without external variables

1. four survival model have been generated for comparison. 2. significant variables. 3. stratfication variable deterimation. 4. model comparison based on validation way 1 and

way2 (one table show all the models)

4.8.3 Retirement model with external variables

best model and tested which variable does significantly impact on retirement.

4.8.4 Voluntary quit model without external variables

I. dependent variable are YCSH, because age is not able to predict well. II. shorten the length of risk set. iii. model comparison. (survival model, time seris model, and logsitic regression model)

4.8.5 Voluntary quit model with external variables

i. tested which variable does significantly impact on employee voluntary quit.

5 Conclusions and Managerial Implications