

Xiaojuan Zhu

June 15, 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 (Staw, 1980; Mueller and Price, 1989; Kacmar et al., 2006), and both private firms and governments spend billions of dollars every year managing the issue according to Leonard (2001). 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 field. 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 Cocscod), 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.

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 Bureau of Labor Statistics, 2015). U.S housing price index, U.S. and southeastern monthly purchase-only index are considered as another economic indicator variables in the study (Federal Housing Finance Agency, 2015). 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 (S&P Dow Jones Indices, 2015). Wilshire 5000 total market full cap index published by Wilshire Associates is considered as market index in the forecasting model(Wilshire Associates, 2015). 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 Description

4.1 How to model employee turnover

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, for survival analysis and forecasting. In this study, Cox PH model is employed, because it is hard to fit the employees' working life distribution in that organization to any parametric distributions, such as Weibull or log-normal distribution. Two bias in the dataset are handled by right censor and left truncation during model fitting. 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.

4.1.1 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 Cox (1972). The Cox PH model is usually taken the form of hazard model formula as shown in the equation 1:

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

where $x = (x_1, x_2, \dots, x_k)$, $h_0(t)$ 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 4.1.3. 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 (Kleinbaum, 1998). Taking the logarithm of both sides of the equation, the Cox PH model is rewritten in the equation 2:

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

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 (Allison, 1995).

4.1.2 Two data bias: right censor and left truncation

Right censoring and left truncation are common in survival analysis. The right censoring is that the main 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. Thus, the current active employees have unknown terminated date. They are treated as right censor. These right censored observations require special treatment 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 i th observation, and the failure time of event for i th 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 i th observation in order to do next analysis.

Left truncation is that the occurrence of an intermediate event prior to the main event of interest appear in the sample dataset. Left truncation in this study occurs when employees enter to the organization at random time, that is, the hired time precedes the origin of the study window. Then, the employees are followed from this entry time until the turnover events occur or until they are right-censored. So, the employees turnover before the entry time of the study will not be aware or not be in the sample dataset. Let T denotes the time of main 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. The left truncation leads to the bias in the analysis. For

example, employees who entered the company in 1950s, appear to stay longer as compared the employees entered in 1960s or later, causing by the fact that the sample dataset does not include the observation recorders for the employees who left the organization before November, 2000. 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 (Carrión et al., 2010). Let t_{i0} denotes the start time of the i th observation, i.e., hired time or age at hired of i th employee, x_i denotes the entry time of the i th 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 i th observation in order to eliminate the left truncation bias (Allison, 1995). 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 i th 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 i th observation whose failure time before t_j , and it cannot be considered in the risk set at time t_j neither (Carrión et al., 2010). simulation study.

4.1.3 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 3:

$$h(t, x) = h_0(t) e^{(\sum_{i=1}^{k_1} \beta_i x_i + \sum_{j=1}^{k_2} \gamma_j x_j(t))} \quad (3)$$

where $x = (x_1, x_2, \dots, x_{k_1}, x_1(t), x_2(t), \dots, x_{k_2}(t))$, $h_0(t)$ is the baseline hazard occurring when $x = 0$, β and γ are the coefficients of x .

I. due to 2008 intervention, each employee have different age at 2008, we have use counting process. Each employee has up to 3 records related to the time, which is before 2008, in-between 2008, and after 2008. The variables related to time are changed based on years, such as policy, age, year of service. Policy is a dummy variable in the form as shown in equations. Two age points are calculated for each record at each period: one 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 variables are also generated for each record at each period: one is year of service at the beginning at the period, named "YCS at start"; the other one is the year of service at the end of the period, named "YCS at end".

II. economic indicators, is time dependent external variables, we have split each employee's the record into yearly records, so each employee has up to 12 records. What is counting process.

4.1.4 Competing risks

what is competing risks. competing risks can help forecasting employee retirement and voluntary quit.

why select these two reasons to model.

4.1.5 Stratification model

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

4.1.6 Variable selection

all the variables are putting into model and selected by backwards selection methods based on P-value.

4.2 Model validation and comparison

4.2.1 Measurements

AIC, BIC, MAPE, c statistic.

4.2.2 Model validation

1. Forecast employee turnover how to forecast and calculate employee turnover number
2. conditional probability (The prediction is conditional probability): given the employee is survival at last year, what is the probability they survival or quit for this year.
3. the total number of employee turnover is the aggregate all turnover probabilities of employees at each year

4. split data into two ways: both training and holdout, training to build the model and holdout is to validate the model, and compare the actual vs forecasting, calculated MAPE and C statistics.

WAY1: training (10 years) and holdout (2 years). forecasting the total number of employee leaving the organization, compared to the actual number to calculate MAPE and C statistics. compare to logistic regression and time series methods to shown survival is better.

WAY2: random split the data into training (2/3) and holdout (1/3), using c statistics to select one best survival model.

4.3 Results

4.3.1 Right censor and left truncation simulation results

4.3.2 Retirement model without external variables

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

way2 (one table show all the models)

4.3.3 Retirement model with external variables

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

4.3.4 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 series model, and logit regression model)

4.3.5 Voluntary quit model with external variables

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

5 Conclusions and Managerial Implications

References

- P. D. Allison. Survival analysis using the sas system: A practical guide. Cary, North Carolina: Sas institute, 1995.
- A. Carrión, H. Solano, M. L. Gamiz, and A. Debón. Evaluation of the reliability of a water supply network from right-censored and left-truncated break data. *Water resources management*, 24(12):2917–2935, 2010.
- P. R. Cox. *Life Tables*. Wiley Online Library, 1972.
- Federal Housing Finance Agency. Monthly purchase-only indexes. <http://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx>, 2015. Accessed: 2015-01-30.
- K. M. Kacmar, M. C. Andrews, D. L. Van Rooy, R. C. Steilberg, and S. Cerrone. Sure everyone can be replaced but at what cost? turnover as a predictor of unit-level performance. *Academy of Management Journal*, 49(1):133–144, 2006.
- D. G. Kleinbaum. Survival analysis, a self-learning text. *Biometrical Journal*, 40(1):107–108, 1998.
- B. Leonard. Turnover at the top, May 2001.
- C. W. Mueller and J. L. Price. Some consequences of turnover: A work unit analysis. *Human Relations*, 42(5):389–402, 1989.

S&P Dow Jones Indices. Standard and poor's (s&p) 500 index data including dividend, earnings and p/e ratio. <http://data.okfn.org/data/core/s-and-p-500>, 2015. Accessed: 2015-01-30.

B. M. Staw. The consequences of turnover. *Journal of Occupational Behaviour*, pages 253–273, 1980.

U.S Bureau of Labor Statistics. (seas) unemployment rate. <http://data.bls.gov/timeseries/LNS14000000>, 2015. Accessed: 2015-01-30.

Wilshire Associates. Wilshire 5000 total market index. <https://research.stlouisfed.org/fred2/series/WILL5000INDFC/downloaddata>, 2015. Accessed: 2015-01-30.