Xiaojuan Zhu

August 7, 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 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 Cocs code), and Organization level (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' pension plan. The employees are eligible to get a full pension, when their age is at least 65 or their points is greater than 85, which is the sum of age and year of service. Employees have different YCS when they are hired because their YCS can be transferred from their previous job if their previous job also accounts for the pension plan. Common Occupational Classification System (COCS) code is a standardized code used to describe the job category by the organization for reporting to Common Occupational Classification System. In this study, COCS code is highly correlated with payroll category: managers, engineers, administrative, and scientists are monthly payroll, general administrative and technicians are weekly payroll, the other categories are hourly payroll. Organization level code is used to distinguish the departments. In this study, the division in the organization do not stabilize like COCS code for an employee, because the division can be renamed, reduced, or dismissed by the change of production plan or organization's budget. The division is considered as time independent variable for employees due to no historical record for divisions provided by HR department. 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. These two kinds of unknown information cause two kinds of bias: right censor and left truncation. The right censor is due to the no termiation date for current employees, and the left truncation is due to no records for employee leaving before November 2000.

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 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. The economic indices are originally in the dayly or monthly form. The average values by twelve month for each year are used into the model fitting.

4 Model Development and Evaluation

Several questions have to be addressed by this study: Can turnover in term of retirement or voluntary quit be predicted? When will a employees turnover? Who will retire or voluntary quit in term of job categories or divisions? what age groups are more likely to retire or voluntary quit? What economic conditions related to retirement or voluntary quit? What is the magnitude or impact of buyout program? How do the tenure and age impact the retirement or voluntary quit? Besides, how to deal with the data biases: right censor and left truncation existing in the dataset. 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

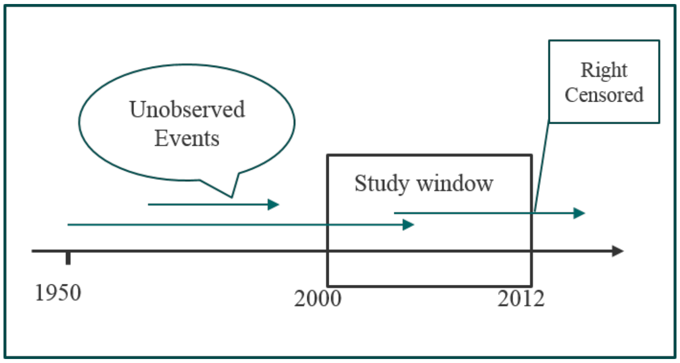


Figure 1: Right censor and left truncation

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 1. 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 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 1. The left truncation leads to another bias. As shown in the figure 1, 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 (Carrión et al., 2010). 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 (all). 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 (Carrión et al., 2010).

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

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

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 4.3. The key assumption for Cox PH regression model is proportion hazards. However, Cox regression can handle non proportional hazards using time-dependent covariate or stratification. 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:

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

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 (all). 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 3:

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

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 Stratification model

An alternative for handling nonproportional hazards is stratification. A stratified model allows each subgroup of data as defined by a grouping variable to have its own baseline hazard while sharing parameters for other covariates across. If the proportional hazards assumption holds within these subgroups then this model allows us to get valid common estimates of covariate effects using all of the observations. Equation 5 below represents the hazard function for strata z;

$$h(t, x, z) = h_0^z(t)e^{(\sum_{i=1}^{k} \beta_i x_i)} \tag{5}$$

where z represents the grouping variable, and $h^z\sigma_0(t)$ is a baseline hazard based for stratam z and β_i are common effects of covariates ac. Note that the strata covariates cannot be the covariates in the Cox PH model.

The proportional hazard assumption can be tested using Schoenfeld residuals which works even if the model includes time-dependent covariates; see Allison (2010); **?**. An alternative is to test the interaction between time-dependent and time-independent covariates in the Cox PH model. The assumption is valid if the interaction is not statistically significant ($P > 0.05$). Including a stratified covariate, when appropriate, can improve the Cox model's performance. The C-statistic is used to compare models with and without stratification with a higher C value indicating a better model (Lemke, 2012).

4.5 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

(Tableman and Kim, 2003). 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 (Allen et al., 2010). And also voluntary quit costs highly for the organizations and firms (Selden and Moynihan, 2000). 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 6:

$$h_j(t,x) = h_{j0}(t)e^{(\sum_{i=1}^{k}\beta_{ij}x_i)} \tag{6}$$

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 1.

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 (Hosmer et al., 2013). 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 7.

$$\begin{aligned} P\{t_{j-1} < T < t_j\} &= 1 - P\{T > t_j | T \geq t_j\} \\ &= 1 - \frac{S_{t_j}}{S_{t_{j-1}}} \\ &= 1 - \frac{S_0(t_j)^{(\sum_{i=1}^{k} \beta_i x_i)}}{S_0(t_{j-1})^{(\sum_{i=1}^{k} \beta_i x_i)}} \end{aligned} \tag{7}$$

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 (Chu, 1998) as shown in the equation 8.

$$MAPE = \sum_{t=1}^{n} \left| \frac{y_t - \hat{y_t}}{y_t} \right| \frac{1}{n} \% \tag{8}$$

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 9.

$$E(\text{turnover number at } t_j) = \sum_{i=1}^{k} P_i\{t_{j-1} < T < t_j\} \tag{9}$$

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 Simulation on right censor and left truncation

In order to understand the performance and efficiency of the Cox PH model in right censored and left truncated data we perform a simulation study.

Generated $n = 100, 200, 500, 1000, 2000,$ and 4000 observations from a a Weibull regression model with one covariate which we referred to as age.

Age is uniformly distributed from 22 to 70 years of age, which is chosen to mimic the actual distribution of workers ages in our sample.

In the regression model, the coefficients for $\beta_age = -.025$ (Why?) and the coefficient for $\beta_0 = 1.5$.

The survival times T_i are randomly generated from a Weibull distribution with shape parameter α and scale parameter λ, where $\alpha = 1.5$ and $\lambda = exp(-0.025age + \beta_0)^{\frac{1}{\alpha}}$.

The simulation is performed on right censoring and left truncation separately, in order to observe the effects for different bias. For right censor simulation, the start point for all the observations are set as 0, and stop point is equal to the survival time t_i for ith observation where $T = (t_1, t_2, \ldots, t_n)$. After that, a survival time histogram is generated. The censor time C is set as first quarter, median, third quarter, and maximum of the survival time, respectively, to get 75%, 50%, 25% and 0% censoring proportions. When the survival time t_i for ith observation is not greater than the censor time (c_i), the stop point is survival time and censor variable δ_i is 1. When survival time t_i for ith observation is greater than censor time (c_i), the stop point is change to censor time (c_i) and censor variable δ_i is 0. The start point and the stop point are dependent variable in the cox regression model. δ is censor variable. And age is predictor or covariate. All these variables are applied to cox model using R EHA package.

For left truncation simulation, the start point U is generated as uniform distribution with $a = 0$ and $b = max(T)$ which indicates an observation start randomly from time 0 to time $max(T)$. The stop point S is $U + T$. The histogram is generated for S. The truncation time L is set as 0, first quarter, median, and third quarter of S, respectively, to get 0%, 25%, 50%, and 75% truncation proportions. When start point u_i for ith observation is less than truncation time l_i, the start point is reset as truncation time l_i. When start point u_i for ith observation is not less than truncation time l_i, the start point does not change (u_i). In left truncation, the censor variable δ for all the observations are equal 1. All these variables are also used to build Cox PH model in R EHA package. The Cox PH model is conducted using "phreg" function in eha package and using weibull distribution to estimate the baseline for both right censoring and left truncation simulation. The "phreg" function performs Cox PH model and also provides a parametric baseline hazards estimation (Broström, 2012). Total predicted failure number is calculated as shown in equation 7 and 9. The actual and forecast failure number are plotted to compare the effects on different levels of bias.

5 Results

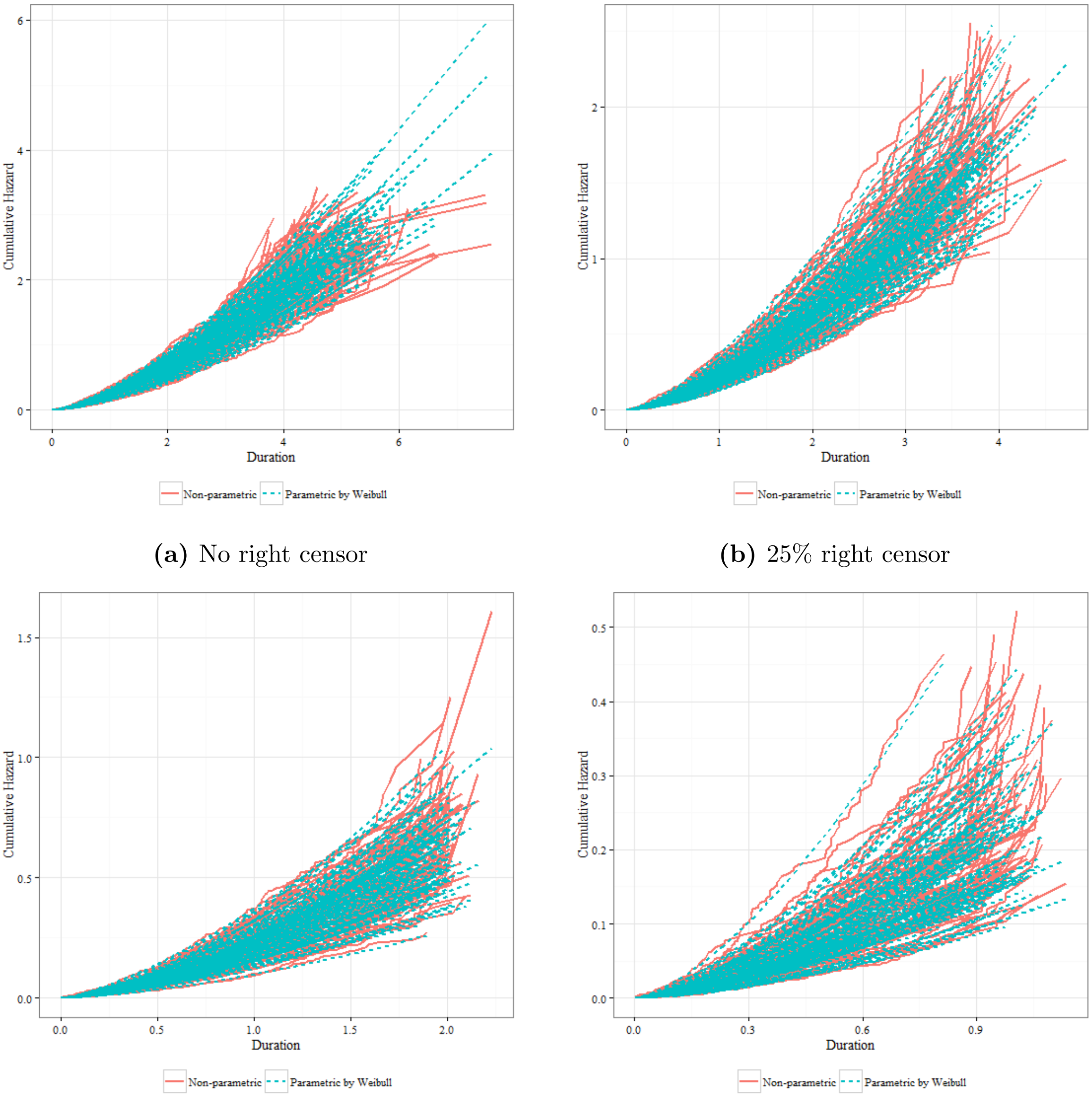
5.1 Right censor and left truncation simulation results

The right censoring simulation result is shown by the average values of Cox PH model coefficient estimate and baseline parameter estimates for Weibull distribution using PHreg from R eha package based on 100 replication with no censoring, 25%, 50%, and 75% censoring proportion models as shown in the left part of Table 1. The events in the second column of the table is the actual total failure events in the dataset without including censoring, which is $events = \sum_{i=0}^{n} \delta_i$ where δ is censor variable, and $n = 100, 200, 500, 1000, 2000,$ and 4000. The simulation results shows censor proportion and the number of events are two influential factors for the coefficient estimation. The model overestimates the coefficients of age, λ, and α, when the data has high proportion of censoring and total number of events is less than 200. For example, when 75% of the data are censored with only 25 events, the estimates for three parameters are 0.028, 4.043, and 1.564, respectively, which are the highest among all the estimates. The estimates of age and $alpha$ tend to approach to the true value, as the event number increasing. The predicted events in the sixth column is the total predicted failure number calculated based on the coefficient estimates and the non-parametric baseline from Cox PH models to calculate the data with all known events (no censoring). The predicted events using censoring models are all lower than the actual total failure number as shown in figure 2, but close to the events number after censoring. For example, the predicted events are 24.84 when using the estimates of age and the baseline from Cox PH model with 75% censoring to calculate the data with 100 events, which is close to 25. The model cannot predict more than 25 events, because the baseline is lack of information for the events after censored time. All the observations start at time 0 in the right censored study, thus only the observation with long life time are censored. As a result, the baseline does not include hazards or survival probability information for the long life time observation, i.e. there is no failure probability for long life time observation. Therefore, a baseline can be highly variable in the extremes leading to poor predictions.

Table 1: Right censoring and left truncation simulation statistics

Censor proportion	Events	Variable Estimates			Predicted Events	Truncation proportion	Events	Variable Estimates			Predicted Events
		Age	λ	α				Age	λ	α	
0%	100	0.026	2.931	1.509	97.42	0%	100	0.027	2.865	1.534	96.60
25%	100	0.027	2.962	1.527	74.31	25%	75	0.027	2.917	1.546	72.86
50%	100	0.028	3.237	1.530	49.66	50%	50	0.027	2.899	1.577	47.32
75%	100	0.028	4.043	1.564	24.84	75%	25	0.029	3.280	1.757	21.75
0%	200	0.026	2.841	1.508	197.23	0%	200	0.025	2.777	1.506	196.13
25%	200	0.026	2.856	1.513	149.34	25%	150	0.025	2.756	1.515	147.97
50%	200	0.026	2.925	1.527	99.68	50%	100	0.025	2.825	1.532	97.44
75%	200	0.026	3.167	1.540	49.86	75%	50	0.026	2.927	1.572	47.17
0%	500	0.025	2.731	1.500	496.77	0%	500	0.025	2.732	1.509	494.78
25%	500	0.025	2.718	1.508	374.31	25%	375	0.025	2.737	1.514	373.42
50%	500	0.025	2.744	1.514	249.65	50%	250	0.026	2.778	1.514	247.70
75%	500	0.025	2.787	1.525	124.89	75%	125	0.026	2.835	1.547	124.15
0%	1000	0.025	2.748	1.509	996.41	0%	1000	0.025	2.710	1.504	993.77
25%	1000	0.025	2.747	1.512	749.23	25%	750	0.025	2.709	1.504	748.94
50%	1000	0.025	2.748	1.514	499.61	50%	500	0.025	2.715	1.506	503.37
75%	1000	0.026	2.844	1.509	249.80	75%	250	0.025	2.694	1.524	250.93
0%	2000	0.025	2.714	1.502	1996.19	0%	2000	0.025	2.740	1.503	1993.65
25%	2000	0.025	2.713	1.503	1499.29	25%	1500	0.025	2.731	1.502	1507.94
50%	2000	0.025	2.742	1.500	999.68	50%	1000	0.025	2.724	1.503	1012.37
75%	2000	0.025	2.733	1.502	499.89	75%	500	0.025	2.718	1.508	512.30
0%	4000	0.025	2.719	1.504	3995.80	0%	4000	0.025	2.720	1.500	3988.90
25%	4000	0.025	2.718	1.505	2998.86	25%	3000	0.025	2.722	1.501	3014.31
50%	4000	0.025	2.724	1.503	1999.52	50%	2000	0.025	2.710	1.502	2032.03
75%	4000	0.025	2.729	1.513	999.86	75%	1000	0.025	2.703	1.503	1028.39

To further test the baseline impacts on the predictions, another simulation study is conducted. The data with 2000 events and no censoring are modeled by Cox PH model using coxreg and phreg from R eha package to generate three baselines: a non-parametric and two parametric (Weibull, and Extreme Value(EV)) baselines as shown in figure 5. Because the data is generated by Weibull distribution, the non-parametric baseline (red solid line) is close to the parametric baseline by Weibull distribution (blue dash line). The parametric



(a) No right censor

(b) 25% right censor

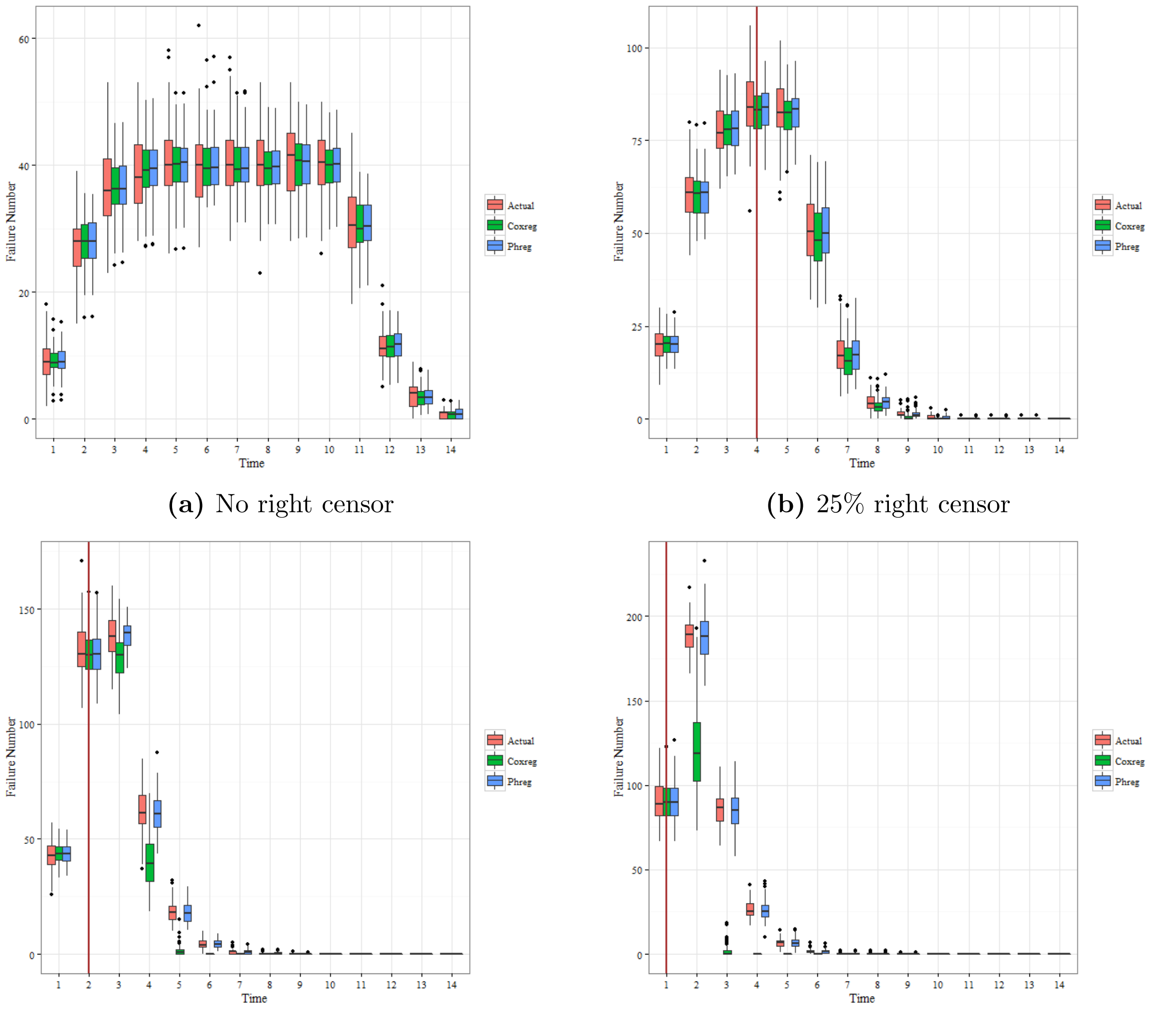
(c) 50% right censor

(d) 75% right censor

Figure 2: Baseline comparison by various censoring

baseline (green dash line) by EV is much lower than the other two baselines as shown in figure 3. The predicted failure number based on the EV baseline (green dash line) are also much lower than the actual failure number across the time. And the prediction (blue dash line) for the non-parametric baseline is close to actual failure number (red solid line) as shown in the figure **??**. This study shows the inaccuracy estimation of the baseline lead to poor prediction. Therefore, accurate estimation of coefficients and the baseline are two key factors impacts the perdition of the events of the Cox PH model.

Table 1 right part shows the results of the left truncation bias simulation statistics for testing Cox PH model function using R eha package. All the values shown are the average



(a) No right censor

(b) 25% right censor

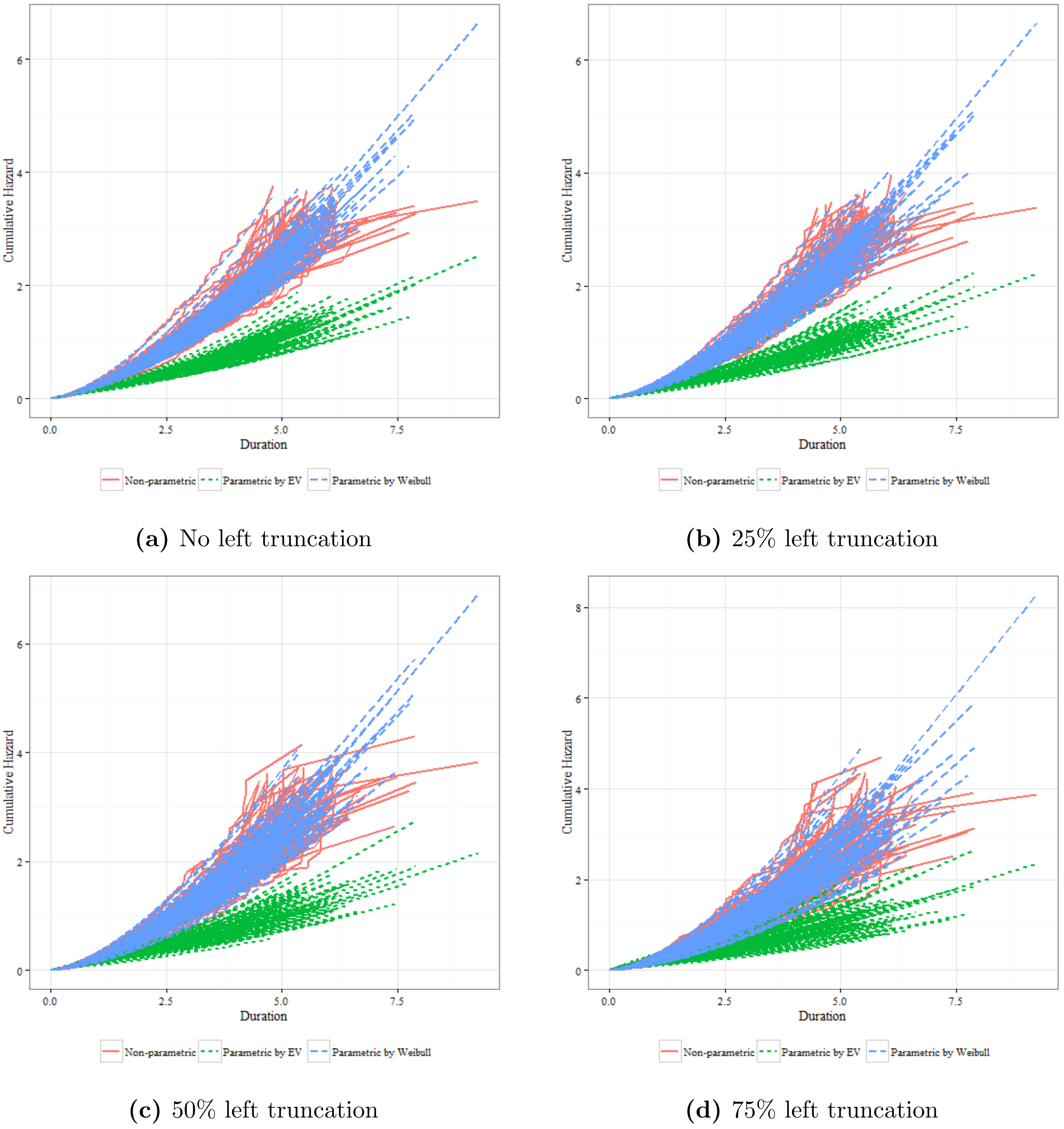
(c) 50% right censor

(d) 75% right censor

Figure 3: Right censor simulation results: actual vs. predicted failure number

values for coefficient estimates for age, baseline parameter estimates for scale and shape by Weibull distribution, and total predicted failure number based on 100 replications. Similar as the right censor simulation result, left truncation proportion and the number of events are two factors for coefficients estimation. As table 1 shown, the coefficients are overestimated when the left truncation proportion is about 75% and the events is less then 1000. The coefficients are all overestimated when the number of events is less than 200. For the predicted events, it underestimates when the number of events is less than 1000, but it over predicts afterwards when the data has left truncation. However, if the number of events is high, it can offset (reduce) the left truncation effects. For example, the estimates for age and are all close to true simulation value, and predicted events are slightly overestimated (about 3% error rate), when events number is 1000 with 75% truncation proportion. The predicted events are smoothed and close to the actual one as shown in the figure 6, which is an example of total 4000 events simulated with various proportion of left truncation. It is overestimates at the beginning of the time line when left truncation proportion is 50% and 75% as shown in figure 6c and 6d, where the dash lines are both higher than the solid lines at the beginning.

Therefore, the simulation test shows that the Cox PH model can accurately estimates the coefficients when events number is more than 1000 even with high proportion of censor and left truncation. The baseline is another key factor to predict the failure number. The prediction will be accurate if the parametric distribution of the baseline is known. Otherwise, a wrong baseline can also deterious the prediction. Compared to parametric baseline, a non-parametric baseline is more robust. However, it still needs high number of events to get a smooth baseline to predict accurately.



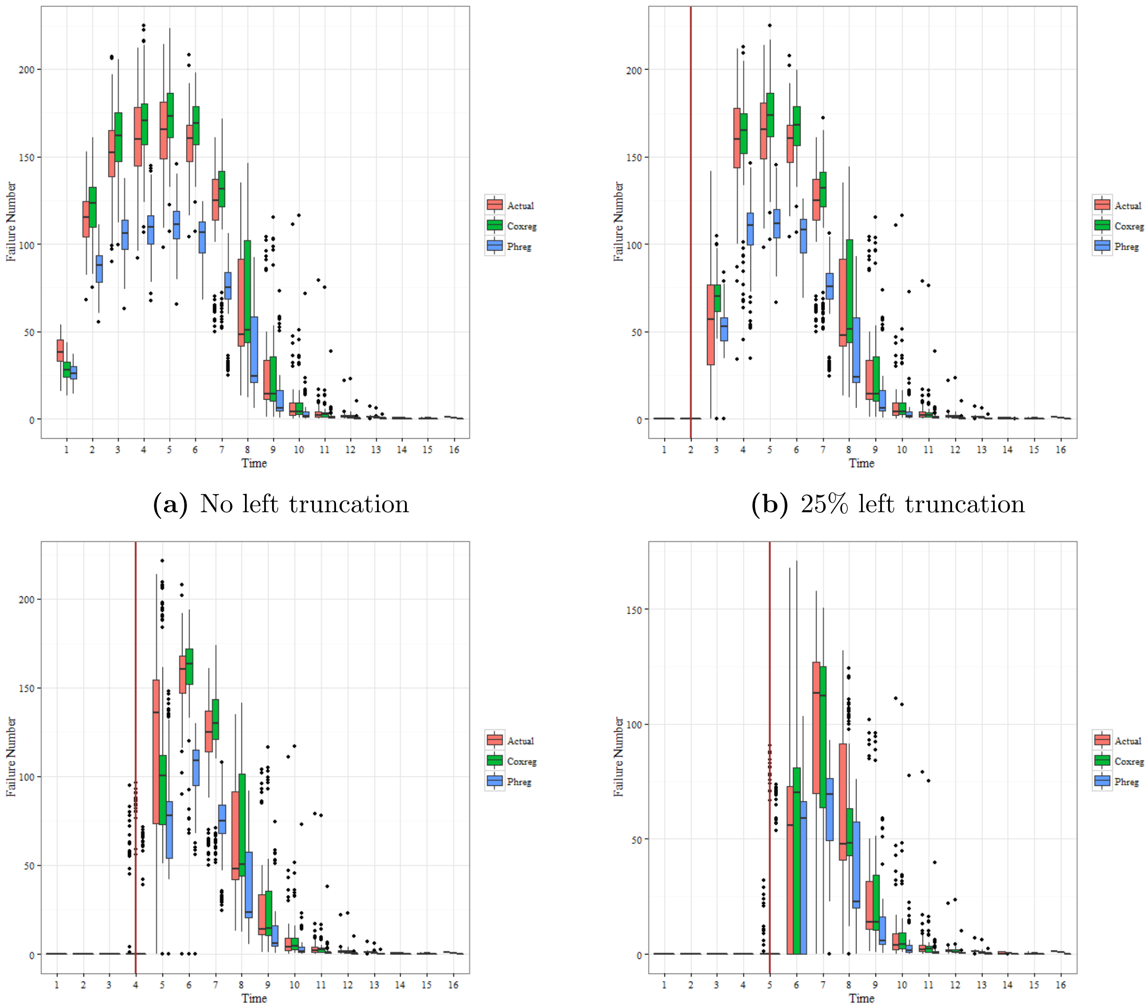
(a) No left truncation

(b) 25% left truncation

(c) 50% left truncation

(d) 75% left truncation

Figure 4: Left truncation simulation results: Baseline Comparison



(a) No left truncation

(b) 25% left truncation

(c) 50% left truncation

(d) 75% left truncation

Figure 5: Left truncation simulation results: actual vs. predicted failure number

5.2 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)

5.3 Retirement model with external variables

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

5.4 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)

5.5 Voluntary quit model with external variables

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

6 Conclusions and Managerial Implications

References

D. G. Allen, P. C. Bryant, and J. M. Vardaman. Retaining talent: Replacing misconceptions with evidence-based strategies. *The Academy of Management Perspectives*, 24(2):48–64, 2010.

P. D. Allison. *Survival analysis using SAS: A practical guide.* Sas Institute, 2010.

G. Broström. eha: Event history analysis. r package version 2.0-7, 2012.

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.

F.-L. Chu. Forecasting tourist arrivals: nonlinear sine wave or arima? *Journal of Travel Research*, 36(3):79–84, 1998.

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.

D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression (3rd Edition).* New York, NY, USA: John Wiley & Sons, 2013.

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.

K. Lemke. Building a predictive model for 30-day inpatient readmission using proc phreg. *NESUG.org*, page 13, 2012.

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. C. Selden and D. P. Moynihan. A model of voluntary turnover in state government. *Review of Public Personnel Administration*, 20(2):63–74, 2000.

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.

M. Tableman and J. S. Kim. *Survival analysis using S: analysis of time-to-event data.* CRC press, 2003.

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.