A Modern Conceptual Framework for the Analysis of Factors in Retirement Decisions

Xiaojuan Zhu

December 18, 2015

Abstract

This article focuses on analytical methodologies useful for analyzing and forecasting retirement decision making in large organizations. We discuss a variety of biases that can occur in retirement databases along with modelling strategies and factors that may play a roll in retirement decision making with employees. The models presented are applied to the analysis of prior early retirement incentives and the prediction of future retirement behavior as a function of demographic, behavioral and external economic factors. Simulation is used to demonstrate the potential impact of sampling biases on predictions. We find that a key factor in retirement among defined benefit employees is achieving full vesting but that those that do not retire immediately ...

Key words: Cox model, proportional hazards, defined benefit pension, early retirement incentive, left truncation, censoring

1 Introduction

Employee turnover is a topic that has drawn the attention of management researchers and practitioners for decades because it is both costly and disruptive to the functioning of most organizations (Staw, 1980; Mueller and Price, 1989; Kacmar et al., 2006). Both private firms and governments spend billions of dollars every year managing the issue according to Leonard (2001). In addition to identifying factors that lead to employee satisfaction and productivity, the ability to identify both causes and timing of attrition are key goals of human resource analytics systems (IBM, 2013). For many mature firms with large workforces, an important piece of this puzzle is the development predictive model of retirement. While commercial tools may exist in this space, very little discussion of applied predictive models has appeared in the academic literature. From an operational perspective, the ability to accurately predict retirement across a range of organizations and job types is a highly beneficial to both the front line management of these organizations as well as financial, human resources, and actuarial concerns of the company and its supporting partners. The ability to forecast retirement and other employee attrition becomes even more valuable in specialized industries and government agencies with long hiring lead times. From a research perspective, a predictive retirement model is a platform that can allow investigators to evaluate both external economic and demographic factors as well as internal policies that can influence retirement decisions (any reference for this? Tim's Survival thing?)

The current study focuses on behavior of individuals between the years 2000-2012 employed at a single firm that provided employees with a defined benefit retirement plan. The study has three objectives: 1) Develop a probabilistic model of the employee lifetime as a function as a function of basic demographic, employment, and external factors. 2) Evaluate the aggregate predictive accuracy of the model in a particular time frame such as 1 or 2 years as a tool to facilitate planning. 3) To determine the impact of internal and external economic variables on retirement. 4) Quantify the impact of a early retirement policy on retirement behavior. Because of the sampling approach used to collect the data, an integral part of the study was to ensure that the modelling strategy was robust to biases introduced by left truncation and right censoring.

In order to analyze the data, we used the Cox proportional hazards model. The strength of the Cox model is its semiparametric form that incorporates a non-parameteric baseline estimate that avoids the impact of truncation ... Also discuss counting process formulation, time varying covariates.

1.1 Motivation of the Study

In many industries with older worker populations, retirement is a major source of HR disruption causing delays and other problems in processing the flow of work. Replacing retired workers can also be a major expense both in HR staff time and recruiting costs. Effective predictive models can also help managers identify divisions, departments, or potentially individuals in an organization that are likely to leave giving the organization more lead time for planning and recruitment of replacements. For example, in large organizations that utilize skilled workers in both white collar and blue collar jobs, accurate forecasts of predictive models of openings in the next 6 to 12 months can be extremely beneficial in maintaining continuity of operations.

While aggregate forecast models of attrition and, more specifically, retirement exist (Zhu et al., 2015), such models have limited ability to provide estimates at division or job category levels. Such models also do not take into account demographic structure of the population such as age, years of service, pension type, and potentially numerous other factors that can influence the probability of retirement. By modeling the distribution of time until retirement at the individual level we can include much more relevant information such as age, retirement plan type, job classification, organizational division, years of service, pension benefit details, individual survey responses if they exist, as well as external social and economic trends that are geographically relevant. This model could potentially provide both accurate predictions as well as giving managers and researchers feedback on how different factors and incentives may influence retirement and other HR decisions such as early retirement incentives.

2 Literature Review

2.1

Wang and Shultz (2010) summarizes key theoretical and empirical research between 1986 and 2010 as well as identifying inconsistent findings. This work draws from the literature in a variety of social science fields. Most of the work cited in that review focuses on the process of retirement and factors driving the retirement decision from the individual perspective. Most relevant to the current work is research in retirement decision making and human resource management.

The current work bridges the gap between individual retirement decision making and the factors involved and the importance of work force management from the perspective of human resources. Here, the focus is on the prediction of retirement for strategic planning in large organizations such as government agencies, large corporations,large academic institutions in order to determine changes in workforce size and plan for eventual loss of critical skills. Particularly applicable to corporations that have a majority of working with defined benefit retirement plans. Can also be used by human resources, benefits managers, and actuaries to determine how much funding is left.(needs editing)

2.2 Methods for retirement forecasting

Employee attrition is a general term and captures the loss of employees to a wide range of causes such as retirement, death, quitting, termination, and potentially promotion or reassignment. Each of these modes of attrition has different foundational causes and may be more or less prevalent during different points in ones career. Forecasting or prediction of turnover may be accomplished at both the aggregate level or may be broken down by organizational factors or by the mode of loss. For example, using a similar data source to that considered here, Zhu et al. (2015) use a time series approach to predict future aggregate attrition based on losses in previous years. The weakness of such an approach is that it does not use known characteristics of the employee population such as age, skill set, performance evaluation, salary, years of service, or numerous other factors to attempt to predict attrition. Obviously, one would expect that the use of these factors as well as further factors identified as critical in the retirement decision making process Wang and Shultz (2010) would be beneficial in predicting future patterns, most obviously in areas such as retirement.

Given access to internal human resource data, regression models for lifetime data offer the potential to make use of this valuable information when making predictions. Such models have been widely applied in academic settings such as engineering(reliability) (Lawless, 2011) , social sciences(event studies) (Allison, 2010; Long and Freese, 2006), and medicine and epidemiology (Kalbfleisch and Prentice, 2011; Moeschberger and Klein, 2003). In the organizational and business settings both academic and professional researchers have began to use these methods to solve practical problems in industry. While actuarial scientists have been using these methods since their inception to create models in risk and insurance (Brockett et al., 2008), more recently, researchers in finance have explored the use of these models to model lifetimes of banks (Lane et al., 1986), as well as time until default of financial instruments such as fixed income securities (LeClere, 2005). However, until now, little has been done in the area of human resources. While Berger and Chen (1993) considers statistical modelling of retirement of tenured faculty within a university setting, the emphasis is on applying the Bayesian statistical approach to modelling retirement outcome. More recently, major analytics consulting firms such as PWC and IBM have begun to offer human resource analytics software and services (IBM, 2013; PWC, 2015). Within this area, some consultants have proposed basic survival models for employee churn (Briggs, 2014) but few details are available and complexities of real world situations tend to be avoided.

2.3 Survival analysis application

As discussed above survival analysis is widely used to analyze lifetime data in many areas, particularly in the health care and engineering areas. These method have been applied to thousands of epidemiological studies retrospective biomedical studies, and clinical trials over the past 40 years. provide a detailed review of the methodology. For example, Claus et al. (1991) investigated the familial risk of breast cancer in a large population-based, case-control study using recurrent life time analysis and found that the risks of breast cancer are a function of women's age. Moeschberger and Klein (2003) provide a thorough book length overview of the methods and include specific case studies focused on medical applications. Survival analysis is also widely used in reliability area. For example, Carrión et al. (2010) estimates the time to failure of the pipes in water supply network dataset under left-truncation and right-censoring by using the extend Nelson estimator (Pan and Chappell, 1998). Book length treatments of these methods in engineering applications include Lawless (2011); Meeker and Escobar (2014).

Although less frequent than the other applications, the method is also heavily utilized in The method is also , Lu (2002) applied survival analysis techniques to predict customer churn by using data from a telecommunications company. Their study provided a tool for telecommunication companies to make retention plan to reduce the customer churn. Also, Braun and Schweidel (2011) used a hierarchical competing risks analysis to model when and why customers terminate their service by using the data from a provider of land-based telecommunication services.

3 Data Preparation

The dataset analyzed was provided by a large multipurpose research organization in the U.S. and consists 4316 active and 3782 former full-time employees. This population of employees is followed across a 12 year window from November 2000 to December 2012. Records of employees that retired or left before November 2000 or that began employment after December 2012 truncated from the dataset. In addition, for 4316 current employees there is no termination date. The sampling approach taken, capturing only employees active in a fixed window creates two forms of bias in the sample that must be accounted for, right censoring and left truncation. A subject is right censored if their endpoint, retirement in this case, is unknown at the time of the study since they are still actively employed. Right censored observations do provide information and should not be dropped but must be analyzed differently than complete observations. Left truncation results from a failure to include cases that fail before the beginning of the study window. This results in a biased sample since only those cases that survive long enough will be represented in a sample. Both of these potential biases will be considered during the discussion of models below.

A number of static employee attributes are provided including payroll type(hourly, weekly, monthly), date of hire, start date, termination date, age at time of hiring , years of service at time of hiring (YCSH), gender, occupation code, and division (organization level). The company credit service date is the date that the organization starts to credit the employee retirement plan. Years of service (YCS) is the total years credit for employees' pension plan. In this study, all of the employees analyzed were hired before January 2012 are eligible for a defined benefit retirement plan. Beginning in January of 2012 the organization began to offer new employees defined contributions plans but such plans were not available during the period of the current study.

An employee is eligible for full pension benefits after achieving either 65 years of age, 85 points of retirement credit. An employee's points are the sum of their age and YCS. YCS can be nonzero at hire if an employee has credit from earlier employment in the same organization. The occupational code is a standardized code used to describe the job category in the organization for reporting purposes. In this study, occupational codes are highly correlated with payroll category: managers, engineers, administrative, and scientists are monthly payroll, general administrative employees and technicians have weekly payroll, and other categories are paid on an hourly basis. The organization or division level code is used to distinguish the departments. In this study, the definition of each division may not static over the entire period of observation. Over the course of time, divisions can be renamed, reduced, or dismissed in reorganizations. Furthermore, no transfer of employees between divisions is recorded. Therefore, for purposes of prediction, the division variable indicates the organization level that an employee is associated with at the time of the final observation.

The variables identified from the turnover dataset and used to build the models are payroll, gender, division, cocs (Job category), age at hired (Ageh), and year of service at hired:

- Payroll (PR): hourly, weekly, or monthly payroll,
- Gender: male, female
- Division (ORG): ten divisions in the organization.
- Occupational Code: Crafts(C), Engineers (E), General Administrative (G), Laborers (L), General Managers (M), Administrative (P), Operators (O), Scientists (S), Technicians (T))
- Age at hire: age at most recent time that an employee is hired or their age in November, 2000.
- Age at credit: age at most recent time that an employee is credited for their pension.
- Years of service at hire (YCSH): the years of service which accounts for pension credit at the most recent hiring of the employee.

Beyond employee level information, we also investigate the potential impact of external exogenous economic factors on retirement decision making and employee turnover. A number of financial indices were selected to capture the various economic factors that may play a role in decisions to retire or leave a position. Because factors such as job market, housing market, and financial markets potentially influence high impact financial decisions such as these, indices such as the seasonally adjusted unemployment rate published by the U.S. Bureau of Labor Statistics (U.S Bureau of Labor Statistics, 2015), the U.S housing price index, the U.S. and southeastern monthly purchase-only index (Federal Housing Finance Agency, 2015), the S&P 500 stock index and Dow Jones Indices were considered. *More specific investment indices including S&P 500, dividend, earnings, consumer index, long interest rate, real price, real dividend, real earnings, P/E 10 ratio were also explored(S&P Dow Jones Indices, 2015).* The Wilshire 5000 total market full cap index published by Wilshire Associates was also considered as market index in the forecasting model (Wilshire Associates, 2015). All twelve of these indices were operationalized for testing using their using a twelve month lag of their one year averages. This approach ensures that the variable can be useful in forecasting since the one year lag will be known at the time of forecasting. The economic indices are originally reported at the daily or monthly level. The yearly average is computed as the average value of the index over the previous twelve months.

4 Model Development and Evaluation

This study aims to develop accurate predictive models of retirement and quitting behavior. The model will be used to address several key questions:

1. How accurately can retirement(quitting) be predicted?
2. What factors indicate an individual is more likely to retire(quit)?
3. Which external economic factors are most predictive of retirement or quitting?
4. What is the magnitude of impact of an Early Retirement Incentive Programs (ERIP)?
5. How do the tenure and age impact retirement?
6. How many employees will retire(quit) next year by occupational category and division?

Survival or lifetime data analysis is methodology that is used to study the distribution of time that it takes for a subject from a population to experience an event such as mechanical failure, death or recovery. Survival regression models relate aspects of the lifetime distribution such as the hazard function to a linear function of explanatory variables. Statistical survival models are often separated into two categories: parametric survival models and semi-parametric proportional hazards (PH) models or Cox models. In this study, the Cox PH model is employed to build predictive models of retirement and quitting, to estimate a employees' baseline hazard of retirement or quitting, and to identify significant factors that may impact turnover. The parametric models are not appropriate for this study for several reasons. First, it is unlikely that hazards for events like retirement would match common parametric distributions such as weibull or log-normal since the risk should stay close to zero until the usual range of retirement at which time it spikes and then drops quickly again. Furthermore, as mentioned in the introduction, the sampling scheme used in this data involved several biases, which can most easily be adjusted for using the Cox PH model. A third advantage of the cox model is the ability to incorporate time dependent covariates in the model. In the case of the current model these are required to incorporate the impact of a 2008 early retirement intervention promotion (ERIP) in the organization, as well as examining the effects of two retirement key variables and outside economic indicators. A special version of the model known as the competing risks analysis is applied for modeling a population which can experience two types of events, in this case employee retirement and voluntary quitting. In addition to the model fitting, a simulation study is performed to examine the impact of data bias on the forecasting capability of Cox proportional hazard model.

4.1 Missing data biases: right censoring and left truncation

Right censoring and left truncation are both commonly observed forms of missing data in survival analysis data sets. In the current, the study window is from November 2000 to December 2012 as shown in the figure 1. Anyone that was an active employee during this period has their complete record included in the dataset whether or not their tenure began or ended outside of the study window. Conversely, those employees whose tenure ended before the study window or for whom the start date occurred after the study window ended are not included in the study.

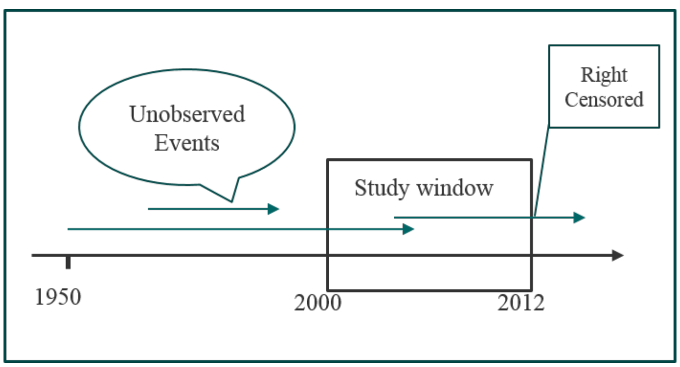


Figure 1: Right censor and left truncation

Let T_i be the time at which the i^th individual experiences the event of interest and let C_i denote the final time at which the individual is observed. An observation is called *right censored* if $T_i > C_i$, indicating that actual event time for the individual is not recorded but is only known to be greater than C_i. Thus, employees that are currently active at the end of the observation window are right censored. These right censored observations do contain information, although incomplete, about a subjects lifetime and require special treatment in order to draw proper inference.

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

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 another interesting artifact of the window sampling scheme. Let T again denote the time that the event of interest occurs and let X denotes the time an individual enters the study. Only the individuals with $T \geq X$ are observed in the study window. Those individuals with $T \leq X$ are referred to as left truncated because they cannot be included in this study based on the sampling window as shown in the figure 1. Left truncation exaggerates the number of longer life individuals leading to a biased sample. To see this consider Figure 1. The longest arrow represents a life span for an employee hired in 1950 and retiring in 2006. While this employee and any in his cohort that are still active are in the sample, others that began in 1950 but retired in 1998, for example, are not. Hence, because of the sampling window approach the longer someone continues working, the more likely they are to appear in our dataset. We therefore see an overabundance of longer lived individuals in our data. The presence of left truncation and the associated bias in the data must be taken into account to achieve accurate estimation of survival analysis (Carrión et al., 2010).

4.2 Cox PH regression model

The Cox proportional hazards (PH) regression model is the most widely used method for modelling lifetime data. The model was introduced in a seminal paper by Cox (1972), one of the most cited papers in history. The Cox PH model is the canonical example of the semi-parametric family of models, specifying a parametric form for the effect of effect of the covariates on an unspecified baseline hazard rate which is estimated non-parametrically.

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 $x_i = (x_i1, x_i2, \ldots, x_ik)$ are characteristics of individual i, $h_0(t)$ is the baseline hazard, and β is a vector of regression coefficients. The model provides an estimator of the hazard at time t for an individual with a given set of explanatory variables denoted by x_i. In the standard Cox model, the linear combination $\sum_{i=1}^{k} \beta_i x_i$, is not a function of time t, and is called time-independent. If $x_i(t)$ is a function of time the model is called the extended Cox model and is discussed in Section 4.3. A key assumption for the model is the proportional hazards assumption. Proportional hazards assumes that explanatory factors have a strictly multiplicative impact on the hazard function so that different groups maintain a constant hazard ratio at all times. However, Cox regression can be extended to handle non proportional hazards using time-dependent variables or stratification; see **?**.

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 $\alpha(t) = \alpha$ i.e. constant, then the model reduces to the exponential distribution. As noted earlier, the general Cox PH model puts no restrictions on $\alpha(t)$. The partial likelihood method is used to estimate the model parameters. The partial likelihood is estimated b baseline (Allison, 1995).

4.3 Time dependent variable and counting process

Some explanatory variable values change over the course of the study. The extended Cox PH regression is a modification of the model that incorporates both unchanging time-independent variables as well as variables that change with time or time-dependent variables,

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

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. In order to fit this model, a modification of the partial likelihood is required and we often present the data set in a format called the *counting process* format in order to facilitate this calculation.

The current study considers four internal time dependent variables that are functions of the individuals characteristics. *Policy* is a a variable that indicates a specific time during the study window, part of the 2008 calendar year, when the organization offered an early retirement incentive program(ERIP); see (Clark, 2002). It is time varying in the sense that it occurs at a different age for each individual in the study population and is set at 0 during the period when no program exists and 1 during the period when the program does exist. The variable *P85*, an indicator that an employee has amassed 85 service points, the sum of their years of service and age, and qualifies for full retirement benefits, and *A65*, an indicator that an employee qualifies for retirement by exceeding the age 65 threshold, are two additional time varying variables that capture important changes in individuals hazard level throughout the study.

The counting process format allows software packages to handle time dependent variables, by creating multiple intervals for each employee. Each interval is defined so that the time varying variables are constant within the interval. For example, for an individual that remained active until the end of the study and that achieved 85 points in December 2003, exceeded age 65 in December 2007, and received a retirement incentive in the 2008 calendar year, our study would include 5 records for the individual, November 2000 - November 2003, December 2003 - November 2007, December 2007, Jan 2008-December 2008, and Jan 2009 - December 2012, which is the end of the study.

Economic indicators represent another form of time dependent variables. These are referred to as external variables because they depend upon factors external to the employee. In models considering economic variables, monthly observations of economic indicators are aggregated at the yearly level leading to a larger number of time periods in the counting process format. The economic variables are included at a one year lag since it is assumed that retirement and quitting decisions are made a significant time in advance of the actual event and therefore depend on older values of these indicators. Importantly, using lagged quantities allows the model to be used for forecasting for a 12 month period since these numbers will be known at the time of forecast.

The start and end points of each interval are given in the definitions below.

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

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 variables across. If the proportional hazards assumption holds within these subgroups then this model allows us to get valid common estimates of variable effects using all of the observations. Equation 6 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 variables ac. Note that the strata variables cannot be the variables in the Cox PH model.

The proportional hazard assumption can be tested using Schoenfeld residuals which works even if the model includes time-dependent variables; see Allison (2010); Collett (2015). An alternative is to test the interaction between time-dependent and time-independent variables in the Cox PH model. The assumption is valid if the interaction is not statistically significant ($P > 0.05$). Including a stratified variable, 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). When the proportional hazards assumption fails there are several alternative approaches to capturing observed hazard patterns. One approach for handling nonproportional hazards is stratification. A stratified model contains a separate baseline hazard for each subgroup defined by the analyst while allowing shared effects of other factors. If the proportional hazards assumption holds within these subgroups then this model allows us to get valid common estimates of variable effects using all of the observations. Equation 6 below represents the hazard function for strata z;

$$h(t, x, z) = h_0^z(t)e^{(\sum_{i=1}^{k} \beta_i x_i)} \tag{6}$$

where z represents one level of the grouping variable Z, $h^z\sigma_0(t)$ is a baseline hazard for this individual subgroup and β_i are constant effects of variables across all strata. Note that a variable cannot be used as both a stratification variable and a constant effect in the Cox PH model.

4.5 Testing the PH Assumption

Three common approaches areavailable for testing the validity of the proportional hazard assumption: The first approach is to investigate the Schoenfeld residuals. A second alternative is to test the interaction between time-dependent and time-independent variables in the Cox PH model. The PH assumption is valid if the interaction is not statistically significant ($P > 0.05$). Finally, including separate baseline hazards for each strata defined by the analyst can also capture variation in changes of the hazard rate. See Allison (2010); Collett (2015) for more details on these tests.

4.6 Competing risks

One of the many nuances observed within this data set is the fact that currently employees can leave employment in several mutually exclusive ways. These include leaving due to quitting voluntarily, being laid off, dismissed for cause, transferred, retired, or being unable to continue due to disability or death. 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). A competing risks model is a common approach when studying a single mode of leaving such as retirement if subjects at risk may also exit through an alternative mode such as quitting. In the current study, when considering retirement as the event of interest and voluntary quitting as a competing risk, all observations are initially included in the study and outcomes that end in a quitting event are treated as censored which allows their observed work period to be used informatively.

4.7 Variable Selection and Model Choice

In the current study two equivalent time measurements were considered as response variables for modelling, *age* in years and *years of service*. Because of the inclusion of time varying explanatory variables, and the need to estimate the baseline hazard for purposes of forecasting, the data must be formulated as a counting process , see Section 4.3. After some consideration, age was considered the better option for analysis because the more condensed distribution of values allows more accurate estimates of the baseline.

The model selection initiated by considering gender, dept, age, and other time independent variables as well as policy, p85, and A65. We removed non-significant variables using the criteria that p-values should be less than .05 and starting with the largest p-value first, i.e. backwards selection. This continued until only statistically significant variables remained. We also tested stratifying the baseline using occupational code, which did not improve the model. Finally we tested external time varying covariates, which capture economic factors, one at a time and noted the impact on model performance.[MAY NEED A BETTER SUMMARY OF MODELS COMPARED]

4.8 Model evaluation and comparison

In order to evaluate the models considered in this study, the data were first split into two sets, a training set containing all of the observations from years 2000-2010 and a testing set containing events on the same individuals that occurred in calendar years 2011 and 2012. The testing (holdout) sample was included to get an accurate measure of how well the model would forecast beyond the observed data. Because the testing data set is not included in the model fitting process, this out of sample evaluation provides a better estimate of predictive accuracy, see Kuhn and Johnson (2013) for further discussion of this approach;

All of the fitted models considered in this study are first evaluated by four statistical criteria: Akaikes Information Criterion (AIC), Schwartzs Bayesian Criterion (SBC), mean absolute percentage error (MAPE) and likelihood based goodness of fit G^2. Ideally, the optimal model should minimize the values of AIC, SBC, MAPE, and G^2 when fit to the training data. In this study, the model performance on holdout dataset is considered more important than that on the training dataset. AIC and SBC both assess model fit by balancing a larger likelihood value with a penalty that increases with the number of variables included. The inclusion of the penalty term diminishes the potential for over-fitting (Allison, 2010; Hosmer et al., 2013). These measures are generated automatically by the model fitting process.

In order to asses predictive measures such as MAPE and G^2 we first predict the probability of the event of interest, e.g. retirement, for each active individual during each calendar year of the training or testing data set. For each employee, we compute the conditional probability of the event occurring between time t_j and t_{j-1}, given that the employee is active at time t_{j-1}. It is calculated using the baseline hazard and coefficients from Cox PH models as shown in Equation 7 below

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

where, T_k is survival time of the k^{th} individual , t_j is a specific time value, $S_k(t) = S_0(t_j)^{(\sum_{i=1}^{p} \beta_i x_i)}$ is the survival function for the kth individual, $S_0(t)$ is the baseline function generated by Cox PH model, x_i are the individual explanatory variables, and β_i are the respective regression coefficients for the variables.

MAPE is common measure for computing the accuracy of predictions from a forecast model and is often used to compare models since it measures relative performance (Chu, 1998). MAPE is calculated as the average percent deviation of a forecast from the actual observation,

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

Our implementation of MAPE for retirement predictions used the yearly actual and predicted numbers of retirements number as y_t and $\hat{y}_t$. The predicted yearly retirement number, $\hat{y}_t$, is the expected retirement count for a particular year and is computed as the sum of conditional probabilities given by Equation 7 over currently active employees. This follows from the fact that the probability of each employee retiring can be viewed as an independent Bernoulli random variable.

Although less common in forecasting, G^2 is another useful criteria for evaluating model prediction for dichotomous events (**?**). The calculation takes the form,

$$G^2 = 2 \sum_t [y_t \log(\frac{\bar{p}_t}{\hat{p}_t}) + (n_t - y_t) \log(\frac{1 - \bar{p}_t}{1 - \hat{p}_t})] \tag{9}$$

where, y_t is the number of employees retired in year t, n_t is the workforce number in year t, $\bar{p}_t = y_t/n_t$ is the observed proportion of events, and $\hat{p}_t = \hat{y}_t/n_t$ is the model predicted number of events. Small values of G^2 indicate close agreement of observed and predicted numbers of events.

5 Simulation Studies of Proportional Hazards Models

The goal of the proposed data analysis is to create a predictive model for retirement and other types of turnover based on a database of employees. Section **??**, pointed out two sources of bias present in the data, left truncation and right censoring. Because the Cox PH model is estimated using a partial likelihood that depends only upon the cases at risk at the specific failure times, estimates of regression coefficients should remain unbiased and efficient in the presence of left truncation and right censoring (**?**). What is less clear is the impact of the bias on model predictions. This stems from the fact that model predictions depend upon both the baseline, estimated using non-parametric methods, and the parametric estimates of the regression coefficients. In addition, a third ever-present challenge, model selection, may also impact predictions to a significant degree.

In order to better understand the effect of various levels of truncation and censoring on the predictions we perform 3 simulation studies based on Weibull simulated data with varying amounts of bias.

The basic setup in all three simulations is the same. Data sets of sizes $n = 100, 200, 500, 1000, 2000$, and 4000 were generated from a Weibull regression model which was a function of 1 explanatory variable, which we refer to as *Age*. In the simulation, *Age* is uniformly distributed from 22 to 70 years, a range that is chosen to mimic the actual distribution of worker ages observed in our sample.

The baseline hazard for Weibull distribution with shape α and scale λ is $h_0(t) = \alpha(\lambda)^\alpha t^{\alpha-1}$. Extending this to a hazard from a proportional hazards regression model for *Age*, we simply multiply by the exponentiated linear predictor shifting the baseline up or down,

$$\begin{aligned} h(t|Age) &= h_0(t)exp(\beta Age) \\ &= \alpha(\lambda(exp(\beta Age)^{\frac{1}{\alpha}})^\alpha t^{\alpha-1} \\ &= \alpha(\tilde{\lambda})^\alpha t^{\alpha-1} \end{aligned} \tag{10}$$

where, $\tilde{\lambda} = \lambda(exp(\beta \times Age)^{\frac{1}{\alpha}}$.

The survival times T_i are randomly generated from the Weibull distribution with shape parameter $\alpha = 1.5$ and $\tilde{\lambda} = exp(1.5 + 0.025 \times Age)^{\frac{1}{\alpha}}$. It follows that $\lambda = e^1$.

The simulations are performed using the `coxreg` and `phreg` functions from the `R`-package `eha` (**?**) for model fitting. Function `coxreg` performs a standard Cox PH regression using the partial likelihood to fit the model. The `phreg` function performs a parametric proportional hazards regression using both Weibull, Extreme value (EV) baselines.

5.1 Simulation 1: Right Censoring

The first study focuses on understanding the impact of right censoring, a significant effect in the turnover data analyzed later due to the many employees that remain active for the entire observation window. For each of the sample sizes above, survival times T_i are simulated from the Weibull distribution as described earlier. Four censoring times C_j are defined as the first, second, third, and fourth(maximum) quartiles of the simulated sample of lifetimes and refer to 75%, 50%, 25% and 0% censoring proportions respectively. If the survival time T_i of the i_{th} observation is below the censoring time (C_i), then the lifetime T_i is observed and the censoring indicator $\delta_i = 1$. When the survival time T_i for i_{th} observation is greater than the censoring time (C_i), then the censoring time C_i is observed and censoring indicator $\delta_i = 0$.

The results of 100 simulations at each combination of sample size and censoring proportion are shown on the left side of Table 1. Column 1 gives the censoring proportion, column 2 the observed number of events before censoring, and columns 3 and 4 are the average β_{Age} estimates over 100 simulations for both `coxreg` and `phreg`. The values in columns 5 & 6 are the average estimates of λ and α from the parametric fit of `phreg`. The simulation results show censoring 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 dataset has a high proportion of censoring. 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. As the event number increases, the estimates approach the actual value. For example, the estimation of age, λ, and α are close to 0.025, 2.7, and 1.5, respectively, as the number of events exceeds 500.

[Discuss Predictions]The predicted number of events are shown [What does this say about censoring effect on predictions?] in the sixth column is the total predicted failure number calculated by applying the coefficient estimates and the non-parametric baseline from Cox PH models into the dataset without considering censoring. The predicted events using censoring models are all lower than the actual total failure number, but close to the number of events after censoring.

5.2 Simulation 2: Right Censoring with Staggard Entry Times

In order to more accurately capture the nuances and complexities of the data sampling scheme and understand the impact of right censoring we modify the above simulation by staggering the entry times. Starting with the Weibull simulated failure times, we add offset factor S that follows a uniform distribution from 0 to 10 and represents variation in the starting times of the employees within the study window. The event time is equal to the summation of start point and survival time: $S + T$. The censoring time is a single fixed value that ensures a fixed proportion (25%, 50%, and 75%) of censored observations. The observation and censoring indicator are then determined as in the first simulation with the survival time for an individual being $min(C, T_i + S_i) - S_i$. Because some observations start after the cutoff point (censor time) the sample sizes would vary for different censoring proportions. To ensure constant sample size, 4000 observations are initially simulated and 400 whose start point occur before the censoring point are randomly selected.

The results of the simulation are shown in Table 2. As discussed above, both a correctly parametric proportional hazards regression model with a Weibull baseline `phreg` and a semi-parametric Cox regression `coxreg` are fit in order to evaluate differences in efficiency. The estimation of age and α are all close to the true values (0.025 and 1.5) when averaged over 100 simulations. Based on 400 events, the estimates of λ increase from 2.694 with no censoring to over 2.8 when censoring is above 50%. In general, the results indicate that right censoring has little impact on coefficient estimation and that the semi-parametric estimates show similar efficiency to fully parametric estimates. However, right censoring does impact estimation of the baseline function for the Cox PH model as shown in the 6th and 7th columns of the Table and Figure 3. Panel (a) shows no censoring and the parametric baseline matches the non-parametric although as time increases and the number at risk decreases we see that non-parametric estimate deviates more strongly from the parametric fit. Panel (b) shows 25% censoring for 100 replications. We see that, as the range of the data decreases, the duration of the non-parametric baseline estimate is more restricted, while the parametric fit can be extrapolated with the usual caveats. In panel (d), the 75% censoring level, we see increased variability even with the diminished range due to the more restrictive censoring time. Note that these results are indicative of monotonically increasing hazards and may differ if we consider other baseline distributions

[Parametric Bootstrap distribution-look at research, also Simonoff paper

Figure **??**, panels (a-d) illustrate the difference between parametric and non-parametric estimates of the baseline cumulative hazard function, $H_0(t) = -log(\hat{S}(t))$ at different time points over 100 simulated data sets. Panel (a) indicates that without censoring, overall survival estimates including both baseline and covariates are very similar across parametric and semi-parametric approaches. As lifetime, t increases, we see increasing deviations between the two estimates. This due both to the cumulative nature of the estimates and to the increasing variability of hazard estimation of the non-parametric baseline, the Breslow estimator, (**??**) as the number of observations at risk diminishes over time.

When censoring is introduced in panel (b) we see that the range of the x axis is restricted. Although this varies slightly across simulations, it is clear that the non-parametric estimator cannot estimate survival probabilities beyond the maximum observed lifetime due to the lack of a parametric model for the baseline. In the current simulation, this was dictated by the censoring time, which was set to ensure a fixed proportion of censored cases. In real studies, such as the one introduced in Section 6, the maximum observed lifetime will be dictated by the population and the sampling window. Panels (c) and (d) reiterate both of these factors. As the censoring level increases to 50% and finally 75%, the range of non parametric baseline estimate is further restricted but the accuracy improves due to the increasing concentration of observations before the censoring time. Because all simulations had 400 observations, 75% censoring ensures 100 events before the censoring point, which is close to 1 providing a very accurate estimate of the baseline and less dispersion than the uncensored data. Parametric models do not suffer from limitations on the baseline estimate, a seeing advantage, but instead require careful model selection steps which introduce other challenges.

5.3 Left Truncation Simulation

Finally, we perform a simulation to evaluate the impact of left truncation bias on parameter estimation and prediction. Again, we begin by generating a simulated sample of employment times $T_1, \ldots, T_n$. For each observation, i, a uniform random variable $U_i \sim U(0, max(T))$ is then generated and represents the simulated employee starting time. The turnover time is then $S_i = U_i + T_i$. Figure 2 provides a histogram of the simulated population of S_i. Four levels of truncation are introduced by shifting the beginning of the sampling window, L

from 0 across the quartiles of S . When the starting for ith observation, U_i, is less than truncation time l_i, the observation starting point is reset to l_i which is the first point that the employee is observed in the sampling window. If $U_i > l_i$, then the employee is first observed at U_i, which remains the starting point. If $S_i > C$, where C represents the end of the observation window, then the employee is still active at the end of the sampling window and their turnover time is censored. In order to isolate the impact of truncation bias, no censored values were generated in this study.

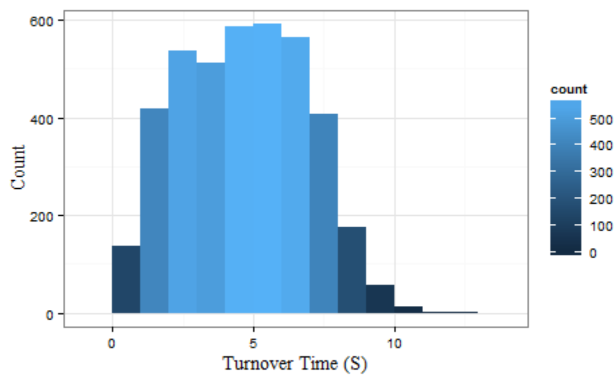


Figure 2: Histogram of Simulated Turnover Time

Sample size and censoring proportions and results follow the same protocol given in Section 5.1. The results of the simulation are shown on the right side of Table 1. As before, values shown represent average parameter estimates over 100 replications for the Cox model and a Weibull PH. The general pattern suggests that as truncation proportion increases for a fixed number of events, the parameter estimates become slightly biased. In the Cox PH case when the sample size is 100, the coefficient for age increases from .027 to .029 as the truncation increases from 0% to 75%. The scale parameter for the parametric model, `phreg`, also increases with truncation percentage. For samples of size 100, with $\lambda = e^1 \approx 2.718$ the average estimate increases from 2.865 with no censoring to 3.280 with 75% truncation. Estimates for $\alpha = 1.5$ increase from 1.534 to 1.757. Such effects disappear as the number of events increase.

Figure 5 compares the non-parametric estimates of the baseline from the Cox PH model with two parametric proportional hazards fits estimated using `eha` (?). The first parametric fit is properly specified and assumes a Weibull distribution as in the prior simulations. The second fit assumes that data follow a type-I extreme value distribution[check manual on this]. Unlike the simulation of Section 5.2 the duration of non-parametric baseline estimates are not limited by the four left truncation proportions. Both Weibull and Cox PH fits overlap with the variability of the Cox baselines showing increased variability with amount of truncation. EV based fits all significantly underestimate the cumulative hazard indicating a potential risk of parametric modelling.

Figure 6 focuses on the number of events predicted by the models. Here 2000 failure time events are generated using the protocol of this section. Both parametric and nonparametric models are fit to the data and procedure is repeated 100 times. The red vertical line is the average left truncation time across the simulations for each truncation percentage. Using each model, the number expected number of failures is then predicted for each 1 unit time interval and compared to the observed number of failures. Boxplots indicate the range of observed values across the 100 replications. The non-parametric Cox PH model tends to slightly overestimate the predicted number of events in each interval suggesting that the baseline estimation is not affected by the left truncation[WHY OVERESTIMATING??]. Matching parametric estimates are generated using the extreme value (EV) distribution instead of the true Weibull model. The underestimation of the baselines by the EV fit shown in Figure 5 lead to drastic under-predictions of the number of events.

[Summarize results of simulation now.] As a result, it is hard to accurately predict the employee turnover for a company just formed recently, or when the employee population characteristic changed, due to high proportion of censor and lack of events. Although this study has more than 50% right censor, it still has more than 3000 events with long duration (around 50 years length).

6 Results

The construction of a predictive model for retirement involves consideration of several factors. Among the variables available for analysis we must choose the set that offers the maximum predictive power, i.e. a model that includes variables that provide the best possible predictions on out of sample testing data sets and not simply on data used to train the model. We must also evaluate potential strengths and weaknesses of the baseline hazard estimate and its impact on prediction accuracy. Finally, it is useful to also understand the impact of various predictors on retirement age.

6.1 Descriptive Analysis

As described in Section 3, the current data provides five demographic and career history variables: payroll(hourly, weekly, or monthly payroll), gender(M,F), division(ten divisions), occupational codes(crafts, engineers, general administrative, laborers, managers, administrative, operators, scientists, technicians), age at hire, and years of service at hire. Table 3 provides marginal counts of the number of workers in the sample within each category. From that we see that among occupation codes there are four large categories (C,E,M,& P) with over one thousand employees observed throughout the data sample, four medium sized groups with 500-700 employees (G,L,R & T), and one small group, S, with 208 employees. Payroll data shows that the largest group is paid monthly, followed by hourly, and weekly. In terms of gender, approximately 72% of employees are male. Finally, divisions, while not fixed over the life of the employee, are distributed similarly to occupational codes with four larger groups and a number of smaller groups.

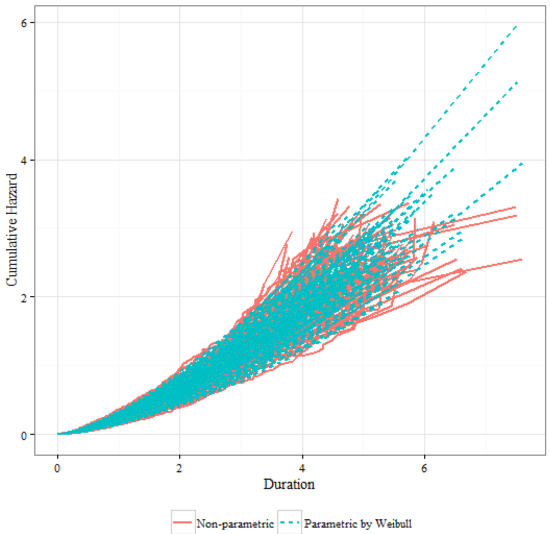
Beyond these demographic and career factors, our models also include behavioral variables derived from policy requirements for retirement and early retirement incentives that occur throughout the observation period. Histograms of retirement age and accumulated pension points are shown in Figure 7. The histogram of for age at retirement is right skewed and shows an anomalous spike at age equals 62 which is the mode, because age 62 is the earliest age for a person to receive social security retirement benefit. The average age is 59.72

Table 1: Right censoring and left truncation simulation statistics

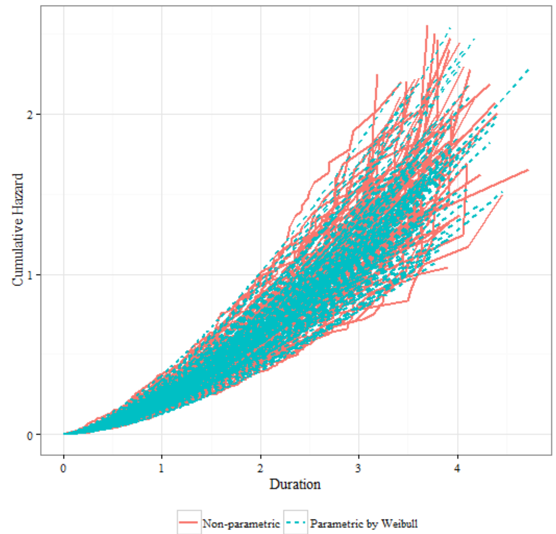
Right Censor						Left Truncation					
Right Censor Proportion	Events	β_{coxreg}	β_{phreg}	λ_{ph}	α_{ph}	Left Truncation Propotion	Events	β_{coxreg}	β_{phreg}	λ_{ph}	α_{ph}
0%	100	0.026	0.027	2.931	1.509	0%	100	0.027	0.027	2.865	1.534
25%	100	0.027	0.027	2.962	1.527	25%	75	0.027	0.027	2.917	1.546
50%	100	0.028	0.028	3.237	1.530	50%	50	0.027	0.027	2.899	1.577
75%	100	0.028	0.028	4.043	1.564	75%	25	0.029	0.029	3.280	1.757
0%	200	0.026	0.026	2.841	1.508	0%	200	0.025	0.025	2.777	1.506
25%	200	0.026	0.026	2.856	1.513	25%	150	0.025	0.025	2.756	1.515
50%	200	0.026	0.026	2.925	1.527	50%	100	0.025	0.025	2.825	1.532
75%	200	0.026	0.026	3.167	1.540	75%	50	0.026	0.026	2.927	1.572
0%	500	0.025	0.025	2.731	1.500	0%	500	0.025	0.025	2.732	1.509
25%	500	0.025	0.025	2.718	1.508	25%	375	0.025	0.025	2.737	1.514
50%	500	0.025	0.025	2.744	1.514	50%	250	0.026	0.026	2.778	1.514
75%	500	0.025	0.025	2.787	1.525	75%	125	0.026	0.026	2.835	1.547
0%	1000	0.025	0.025	2.748	1.509	0%	1000	0.025	0.025	2.710	1.504
25%	1000	0.025	0.025	2.747	1.512	25%	750	0.025	0.025	2.709	1.504
50%	1000	0.025	0.025	2.748	1.514	50%	500	0.025	0.025	2.715	1.506
75%	1000	0.026	0.026	2.844	1.509	75%	250	0.025	0.025	2.694	1.524
0%	2000	0.025	0.025	2.714	1.502	0%	2000	0.025	0.025	2.740	1.503
25%	2000	0.025	0.025	2.713	1.503	25%	1500	0.025	0.025	2.731	1.502
50%	2000	0.025	0.025	2.742	1.500	50%	1000	0.025	0.025	2.724	1.503
75%	2000	0.025	0.025	2.733	1.502	75%	500	0.025	0.025	2.718	1.508
0%	4000	0.025	0.025	2.719	1.504	0%	3999	0.025	0.025	2.720	1.500
25%	4000	0.025	0.025	2.718	1.505	25%	3000	0.025	0.025	2.722	1.501
50%	4000	0.025	0.025	2.724	1.503	50%	2000	0.025	0.025	2.710	1.502
75%	4000	0.025	0.025	2.729	1.513	75%	1000	0.025	0.025	2.703	1.503

Table 2: Right censor simulation results by various start time

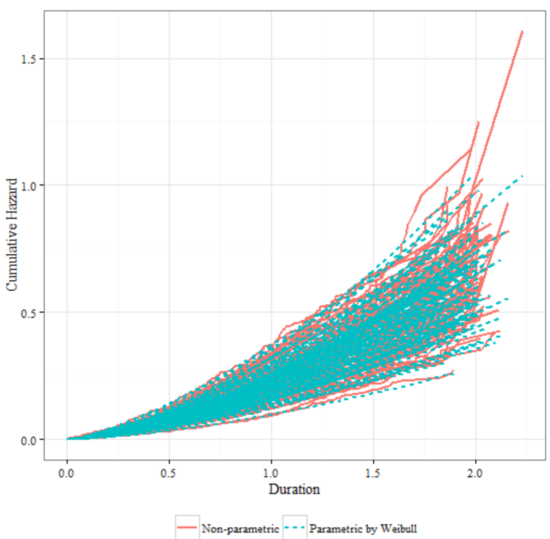
Censor proportion	Events	Variable Estimaties			Predicted Events	
		Age	λ	α	"coxreg"	"phreg"
0%	400	0.025	2.694	1.508	398.52	400.44
25%	400	0.026	2.802	1.518	394.24	401.72
50%	400	0.026	2.828	1.514	340.73	398.51
75%	400	0.025	2.821	1.518	215.92	400.80



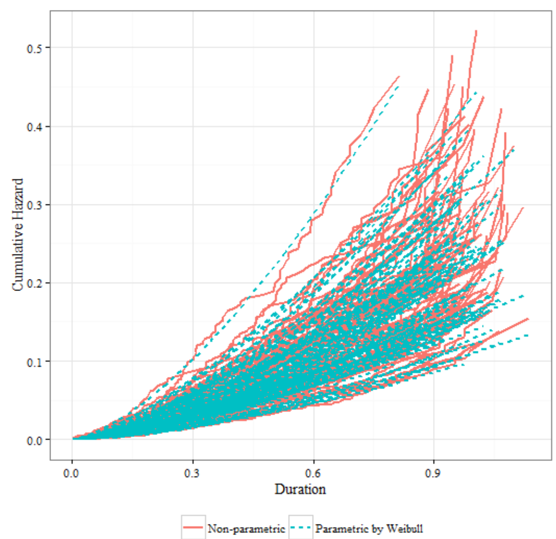
(a) No right censor



(b) 25% right censor



(c) 50% right censor



(d) 75% right censor

Figure 3: Baseline comparison by various censoring

demonstrating that many individuals retire before 62 and most before age 70. In terms of points accrued at time of retirement, recall that points are the sum of years of service plus current age, we see an irregular distribution with the vast majority retiring with point totals in the range of 85 to 100 and relatively few taking a reduced pension are retiring with diminished benefits with points below 85. Again we see that 85 points is the mode indicating that retiring immediately after become fully vested in the pension is a popular choice.

Table 3: Descriptive Statistics

Variable	**Count**	**N %**	**Variable**	**Count**	**N %**
Cocscode			**Gender**		
C	1295	16.0%	F	2296	28.4%
E	1361	16.8%	M	5802	71.6%
G	574	7.1%	**Division**		
L	613	7.6%	divison1	1542	19.0%
M	1178	14.5%	divison2	751	9.3%
P	1621	20.0%	divison3	1042	12.9%
R	595	7.3%	divison4	369	4.6%
S	208	2.6%	divison5	398	4.9%
T	652	8.1%	divison6	1199	14.8%
Missing	1	0.0%	divison7	302	3.8%
Payroll			divison8	823	10.2%
Hourly	2503	30.9%	divison9	404	5.0%
Monthly	4369	54.0%	divison10	1268	15.7%
Weekly	1226	15.1%			

Table 4: Discriptive Statistics 2

	Count	Mean	Median	Mode	Minimum	Maximum	Std. Deviation
Age at Retire	1757	59.72	60.00	62.00	49.00	84.00	4.56
Years of Service at Retire	1757	29.72	30.90	30.06	0.05	55.68	7.74
Point at Retire	1757	89.44	88.67	85.47	51.05	136.66	9.15

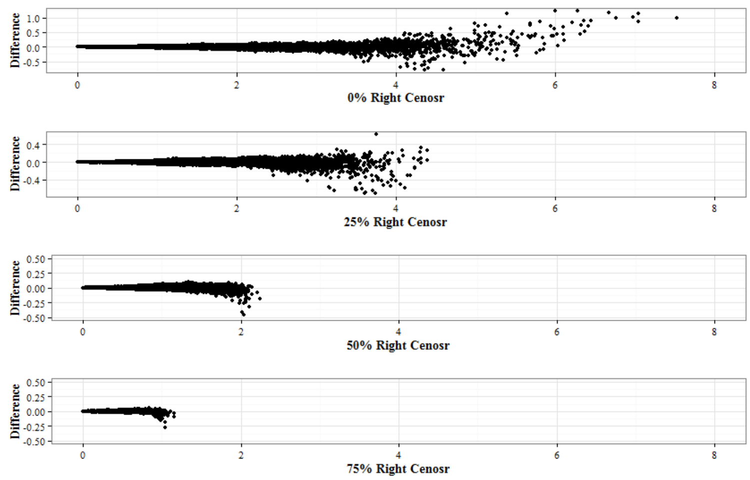


Figure 4: Differences in baseline estimates: parametric - non-parametric for 4 levels of censoring

6.2 Retirement model without external variables

Extensive model selection using a variety of metrics including Log likelihood, AIC, BIC,and out of sample predictive scoring (MAPE and G^2) was used to identify key predictive factors in the model as shown in table 5.

Based on this analysis the optimal modelling variables, excluding aggregate economic factors, chosen for the prediction of age at retirement include *division*, *years of service at hire*, and *age at hire*. In addition, based on our understanding of the covenants and parameters of the retirement program we tested a number of additional variables and found several that increase the predictive power of the model. These include *policy*(an indicator for the 2008 ERI program), *P85*(An indicator that the individual has accrued 85 points and can retire with full benefits), *A65*P85* (An interaction term that moderates the impact of the *P85* effect after the individual has exceeded 65 years of age), and *Policy*P85* (An interaction term that moderates the impact of the *P85* effect while the ERI is in place).

Table 6 describes the fit parameters and hazard ratios. As noted above, we did not find that gender, occupational code and payroll category were not significant predictors in the presence of the other variables. This indicates that employees' gender, job types, and payroll status are not associated with choice of retirement age conditional on the other variables in the model.

P85 is an indicator that a person is eligible for maximum retirement benefits and naturally this has a strong impact on the probability that a person will retire. From a quantitative point of view the hazard ratio is $e^{1.44} = 4.22$. So the hazard of retirement becomes 4.22 times more likely after the person exceeds 85 points. While not surprising, this quantification is important in predicting individual and aggregate retirement time and reflects the modal spike observed in the histogram in Figure 7. The survival function after 85 points is achieved can be su (Julia, why is the hazard ratio not published in the data??? because it has interaction term)

An alternative eligibility criteria for retirement occurs when individuals exceed an age

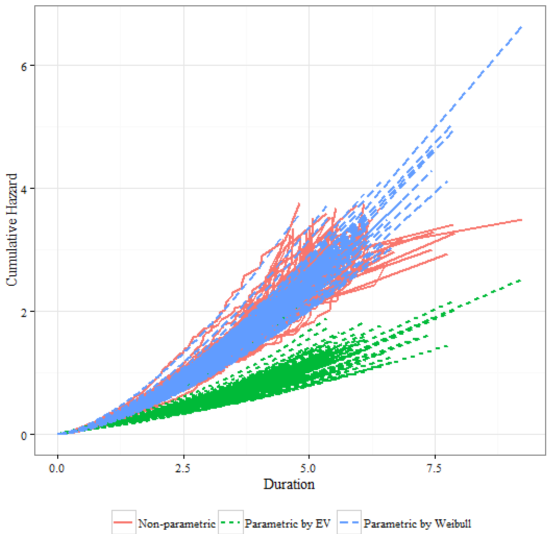
Table 5: Models statistics

Model	LR	AIC0	AIC	SBC0	SBC	Pred. MAPE	Holdout MAPE	Pred. G^2	Holdout G^2
Division Gender Payroll Cocs YCSH Ageh	1194.3	20425.6	19271.3	20425.6	19377.3	39.44	56.78	381.77	85.19
Division Cocs YCSH Ageh	1193.77	20425.6	19269.8	20425.6	19370.5	39.45	56.84	381.92	85.29
Division Policy YCSH Ageh	1451.62	20425.6	18998	20425.6	19061.6	25.91	15.51	128.75	2.64
Division Policy YCSH Ageh Cocs	1469.92	20425.6	18995.7	20425.6	19101.7	25.91	15.24	129.47	2.56
Division Policy YCSH Ageh and Strata on Cocscode	1426.34	14602.1	13199.8	14602.1	13263.4	43.45	87.92	186.89	50.00
Division Policy YCSH Ageh P85	1826.95	20425.6	18624.6	20425.6	18693.6	25.59	19.04	109.17	3.40
Division Policy YCSH Ageh P85 P85*A65	1873.69	20425.6	18579.9	20425.6	18654.1	25.38	7.97	111.55	0.79
Division Policy YCSH Ageh P85 P85*A65 P85*POLICY	1881.02	20425.6	18574.6	20425.6	18654.1	25.42	4.20	112.27	0.81
Logistic regression	4103.91	13618.4	9556.53	13627.3	9752.1	28.20	31.73	232.84	117.38
Time series	N/A	N/A	N/A	N/A	N/A	11.17	34.38	32.10	8.54

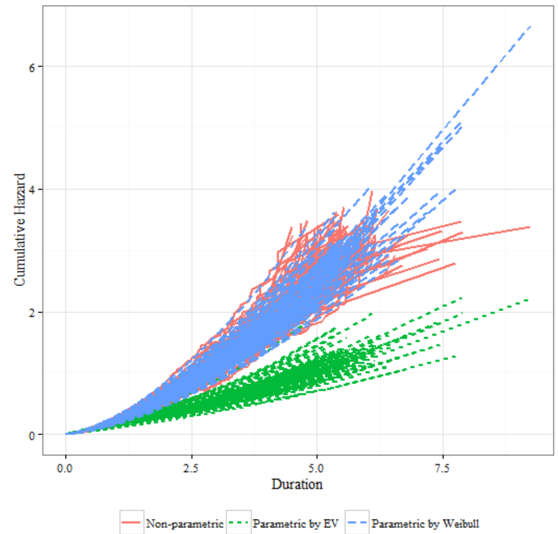
Table 6: Parameter estimates for models

		Period Model		Yearly model	
Parameter	Label	Parameter (Standard Error)	Hazard Ratio	Parameter (Standard Error)	Hazard Ratio
division	dir2	-0.965(0.179)***[1]	0.381	-1.025(0.177)***	0.359
division	dir3	-0.241(0.112)*	0.786	-0.246(0.111)*	0.782
division	dir4	0.078(0.195)	1.081	-0.039(0.195)	0.962
division	dir5	-0.131(0.190)	0.877	-0.246(0.190)	0.782
division	dir6	2.136(0.095)***	8.463	2.252(0.095)***	9.511
division	dir7	2.435(0.129)***	11.418	2.437(0.129)***	11.437
division	dir8	0.864(0.106)***	2.373	0.816(0.106)***	2.261
division	dir9	-3.023(0.581)***	0.049	-2.774(0.504)***	0.062
division	dir10	0.793(0.093)***	2.211	0.709(0.093)***	2.031
YCSH		0.019(0.004)***	1.019	0.043(0.004)***	1.044
Policy	1	0.859(0.169)***	.	0.942(0.109)***	.
Ageh		-0.172(0.013)***	0.842	-0.187(0.013)***	0.829
P85	1	1.435(0.091)***	.	0.682(0.072)***	.
A65*P85	1	-1.610(0.206)***	.	-0.662(0.177)***	.
Policy*P85	1	0.469(0.179)**	.	0.600(0.130)***	.

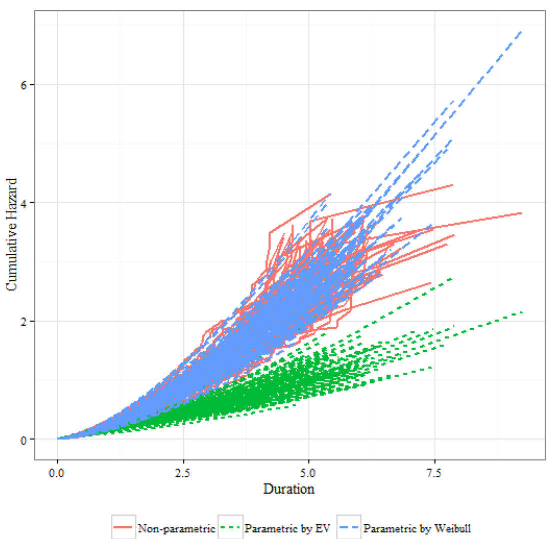
[1] * denotes $P < 0.05$, ** denotes $P < 0.01$, and *** denotes $P < 0.001$.



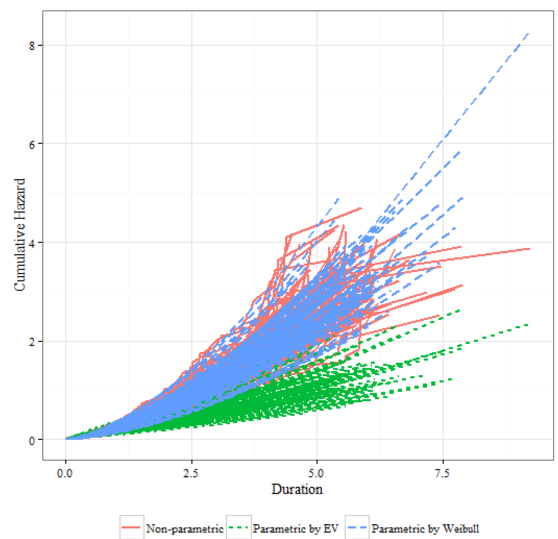
(a) No left truncation



(b) 25% left truncation



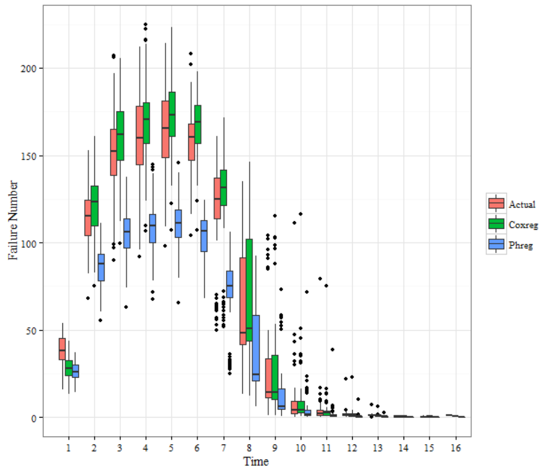
(c) 50% left truncation



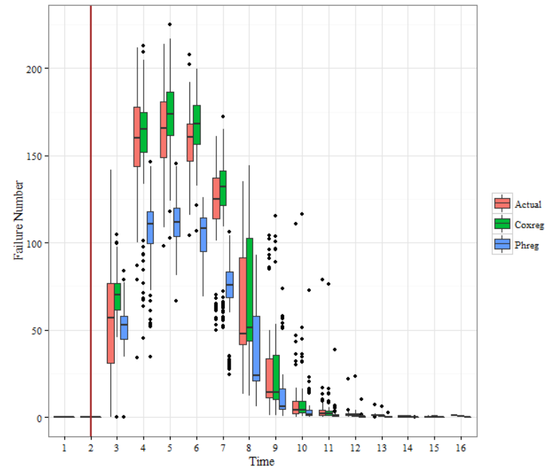
(d) 75% left truncation

Figure 5: Left truncation simulation results: Baseline Comparison

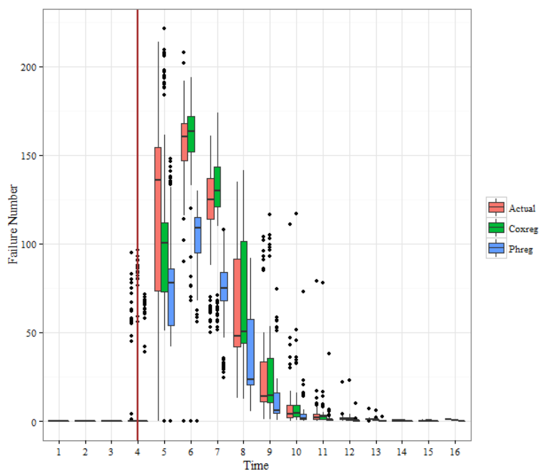
of 65 years and so we would anticipate the hazard increasing at this point in an employees career. Because the response variable in our model is age, we cannot estimate the effect of this within the proportional hazards setting because the impact is included in the baseline hazard which should increase after this point; see figure 9 (explain). However, by including an interaction between the indicators of Age greater than 65 and points greater than 85, *A65*P85*, we can estimate how the impact of reaching 85 points diminishes when a person exceeds regular retirement age. In this case, the interaction term is estimated at -1.61 indicating a diminishing effect on the *P85* criteria to $e^{1.44-1.61} = 0.84$. This suggests that people that exceed both criteria actually have a reduced hazard of retiring over those have



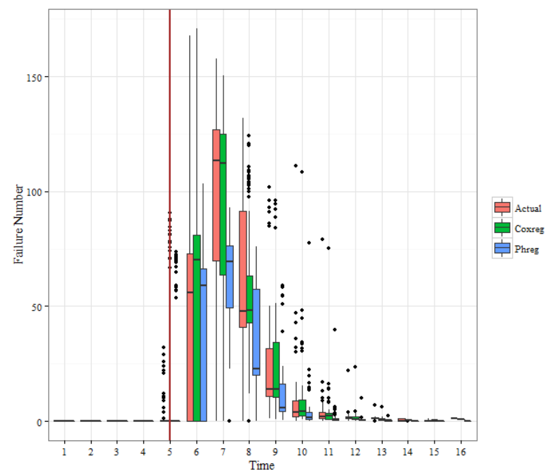
(a) No left truncation



(b) 25% left truncation

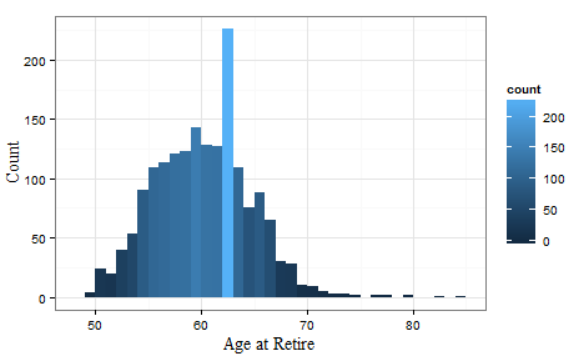


(c) 50% left truncation

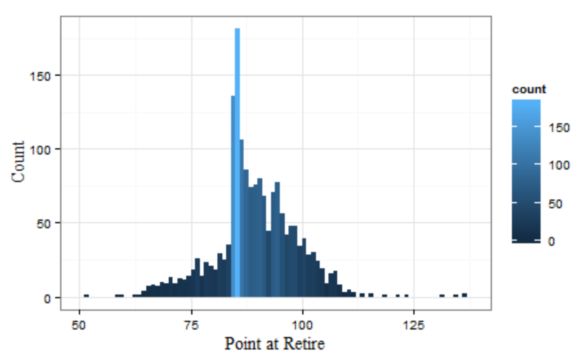


(d) 75% left truncation

Figure 6: Left truncation simulation results: actual vs. predicted failure number



(a)

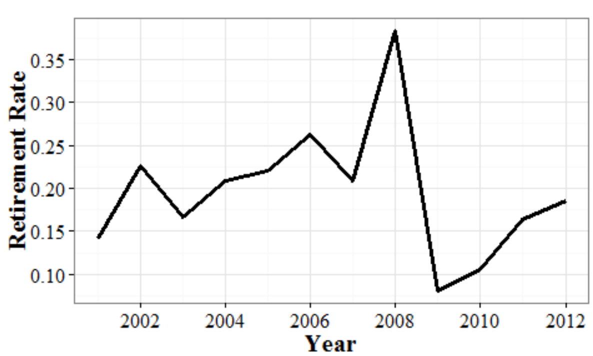


(b)

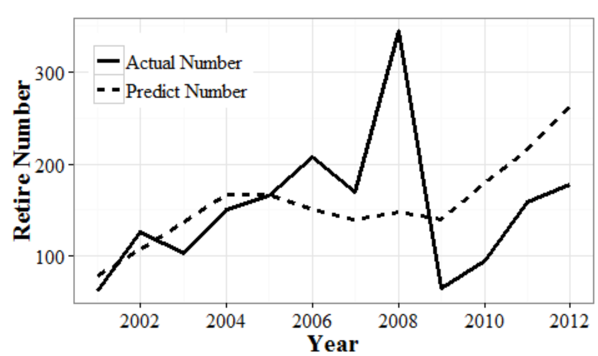
Figure 7: Histogram of Age and point at retire

only met the Age 65 criteria. In other words, the fact that the individual remains on the job after hitting either criteria indicates that the other criteria has less impact (or that they are intending to work longer?). (Discuss that the baseline may increase so that the overall probability is actually higher.)

According to our model, retirement can also be influenced by an employee's age at hire and their years of service at the time of hiring, *YCSH*. The coefficient estimate for age is -0.17. As the reference age is 45.49, this means that the hazard ratio for retirement of an employee that started working at age 46.49 is $exp(-.017) = .84$ indicating a 16% drop in hazard for each additional year later that an employee started. The employee's survival probability at any time, t, can be computed as $S(t)^{1.19} = (S(t)^{e^{0.17}})$ when age at hire is one year below 45.49, where $S(t)$ is the baseline survival probability for a reference employee of average age at the time of hiring. Moving in the other direction, employee's survival probability is $S(t)^{0.84} = (S(t)^{e^{-0.17}})$ for a one year increase beyond 45.49 in the employee's starting age. Together, this implies that at any given retirement age, the employee who starts earlier than 45.49 years old is more likely to retire because they have more years of service and closer to vesting full benefits (85 points) than an equivalent employee who starts working at an older age. Similarly, the employee's years of service at hire show a positive estimate (0.019) with a hazard ratio (1.019) indicating that each year of service at hire beyond the baseline of 2.75 is associated with an approximately 2% increase in the hazard of retirement. This leads to a survival probability $S(t)^{0.98}$ for a one year decrease in the reference years of service at time of hiring. On the other hand, the survival probability is $S(t)^{1.019}$ for an employee with one year of service above average at the time of hire. Together, both age at hire and YCSH effects reflect the intuitive fact that, all else being equal, an employee who has more years of service and therefore is closer to full vesting is more likely to retire. What is non-intuitive about this finding is that while one might suspect that the effects should be of similar magnitude, we actually see that the effect of one year difference in age seems to have about 8 times the impact that one year of previous service does on the hazard of retirement.



(a) Retirement rate



(b) Prediction w/o variable policy

Figure 8: Retirement rate and Prediction Plot

In the fiscal year 2008, the employer in this study created a temporary early retirement buyout option. The response window for this option was 3 months although the specific details beyond this are unknown. In order to deal with the increased level of retirement during this period we including a time dependent indicator variable. The coefficient for this indicator was 0.86 leading to a hazard ratio of $e^{.86} = 2.21$ which indicates that, on average, an individuals hazard of retirement increased by almost 2.2 times during this period. If more information were known about the requirements or targets of this policy, a more case specific estimate may be possible. However, this would not effect the overall aggregate retirement estimates. It is important to include this one-time effect in order to improve the estimates of other factors(work on this last part - why do we need to include).

As a second step, we further test the policy effect on the employees who are eligible for getting a pension. The test results show that the policy had a significant effect for an employee who is eligible for getting the full pension benefit rather than the employee who is only eligible for getting partial retirement as the interaction term of between policy and indicator variables that employees achieve points 75 or points 65 are not statistically significant. The hazard ratio for the policy effect on those employees increases substantially to 15.85 from 2.21, which is more than seven times of the basic policy effect, after the model adding a interaction term of policy and the indicator that a person exceeds 85 points.

The *division* variable was a significant predictor. For analysis, the baseline level was chosen arbitrarily as division 1 so that it's hazard rate is determined by the baseline. Relative to this baseline, divisions 6 and 7 have very high hazard ratios, 8.363 and 11.405 respectively, which indicates, other factors being equal, that the employees in division 6 and division 7 are much more likely to retire at any age than those from division 1. Conversely, division 9 has a hazard ratio of exp(-3.023) = .049 indicating that individuals within this group have 1/20 the hazard of group 1. This may indicate that the division is new and contains younger employees(how to deal with this?? Time varying variable?). In general, differences in retirement rates could be caused by differences in age demographics, leadership, departmental and job function, or departmental leadership.

The baseline survival function and log hazard function are shown in figure 9. The survival probability is 1 before age 49 as shown in figure 9a, which indicates that no employees retire before this age. The survival probability starts to slowly decease from age 50 to age 62. By age 62 the survival probability has decreased to close to 0.75, which indicates that 75% of employees retire at an age greater than 62 years. The slope of survival function decreases sharply at this point indicating the increased retirement rates for workers between age 62 and 65. After 65 the probability drops off even further as most of the remaining population retires by age 68 or 69. Accompanying the survival function is the log of the cumulative hazard ratio. Again, the steep rise in the cumulative hazard between age 62 and 65 indicates the increased retirement activity during this period. After this the cumulative hazard levels off indicating a drop in the hazard rate at these future points.

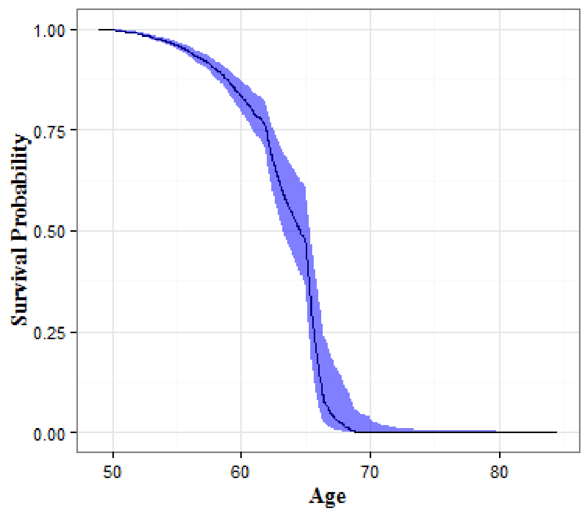
Summarize external

Summarize the predictive capabilities at the individual level. Strengths and weaknesses.

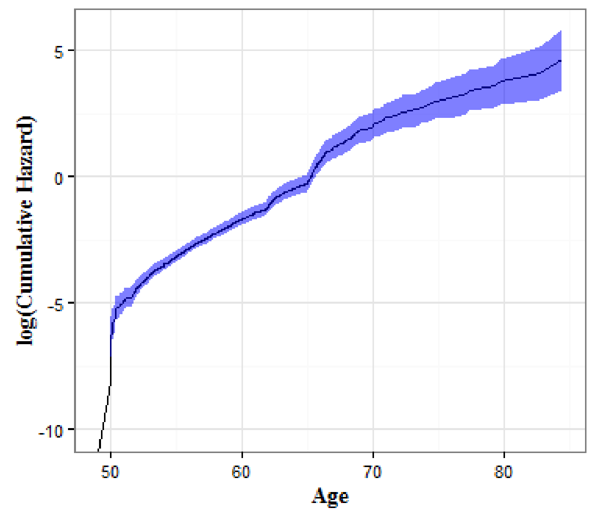
Summarize the predictive capabilities for aggregates. Strengths and weaknesses.

Julia: We need to describe the prediction process results.

The prediction for the employee retirement is shown in the figure 10, table 7 and 8. The model predictions capture the fluctuations in actual retirement, and also capture the peak of the special year 2008 when an early retirement incentive(policy) was introduced as shown in Figure 10. The out of sample predictions for holdout years (2011 and 2012) are very close to the actual number, indicating that the model performs well on both the training



(a) Survival function



(b) Cumulative hazard

Figure 9: Baselines with 95% confident intervals

and holdout samples. Besides predicting the overall retirement, the model can also provide predictions by category. The table 7 and 8 shows the prediction by occupation code and division by year, which is the summation of the retirement probabilities of individuals by job classification and division, separately. As the tables shown, the predicted values are also close to the actual values in these two categories.

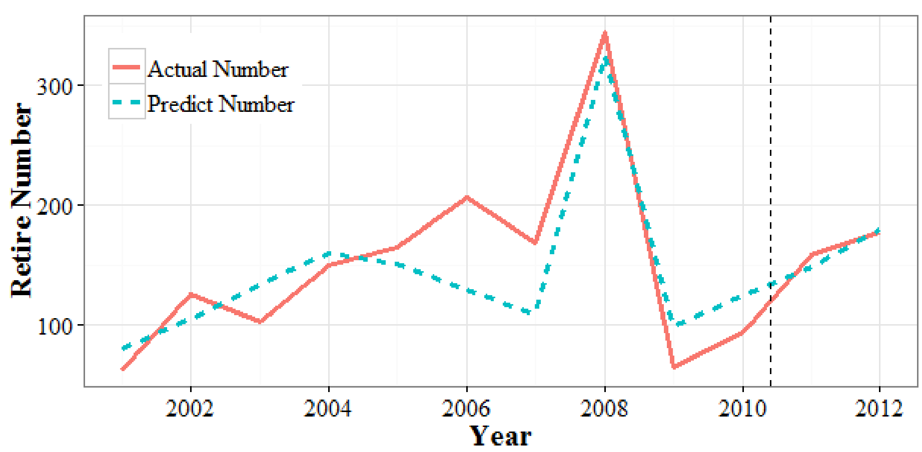


Figure 10: Retirement Forecasting

6.3 Retirement model with external variables

Because the previous model cannot account for the time varying features of the economic indicators, we use a new counting process model with yearly interval based on calendar

Table 7: Predictions by job classification

Year	Crafts	Engineers	General Admin.	Laborers	Managers	Prof. Admin.	Operators	Scientists	Technicians	Total
2001	23[1](16)[2]	12 (10)	4 (1)	6 (5)	11 (14)	11 (9)	8 (2)	2 (1)	4 (4)	81 (62)
2002	31 (29)	15 (16)	5 (15)	7 (11)	15 (26)	13 (18)	11 (6)	2 (1)	6 (4)	105 (126)
2003	37 (26)	18 (13)	6 (5)	9 (6)	17 (21)	18 (13)	16 (15)	4 (2)	8 (2)	133 (103)
2004	43 (32)	23 (23)	8 (9)	11 (7)	21 (30)	23 (25)	15 (15)	4 (3)	10 (6)	158 (150)
2005	40 (39)	24 (17)	8 (13)	9 (7)	20 (27)	24 (31)	12 (15)	3 (4)	11 (12)	151 (165)
2006	31 (58)	20 (29)	6 (10)	8 (9)	19 (32)	25 (37)	9 (13)	2 (4)	9 (15)	129 (207)
2007	19 (44)	14 (25)	7 (9)	6 (9)	18 (26)	27 (40)	8 (6)	3 (4)	7 (6)	109 (169)
2008	55 (71)	30 (33)	19 (20)	22 (12)	64 (63)	79 (84)	27 (32)	6 (7)	21 (23)	323 (345)
2009	16 (14)	9 (6)	7 (3)	7 (7)	20 (8)	25 (10)	8 (11)	1 (1)	5 (5)	98 (65)
2010	18 (19)	11 (17)	9 (1)	7 (8)	28 (23)	34 (16)	8 (4)	1 (3)	7 (3)	123 (94)
2011	22 (36)	13 (25)	11 (8)	9 (9)	34 (27)	40 (34)	9 (5)	2 (1)	8 (13)	148 (158)
2012	24 (29)	16 (23)	14 (11)	12 (10)	41 (46)	49 (36)	12 (4)	3 (2)	9 (16)	180 (177)

[1] the number before the parentheses is predicted retirement number.
[2] the number inside the parentheses is actual retirement number.

Table 8: Prediction by division

Year	Division1	Division2	Division3	Division4	Division5	Division6	Division7	Division8	Division9	Division10
2001	3[1](0)[2]	0 (0)	1 (0)	0 (0)	0 (0)	57 (43)	11 (2)	5 (10)	0 (0)	5 (7)
2002	4 (0)	0 (0)	2 (0)	0 (0)	0 (0)	72 (54)	14 (2)	6 (38)	0 (0)	9 (32)
2003	6 (0)	1 (0)	2 (0)	1 (0)	0 (0)	89 (44)	19 (7)	6 (18)	0 (0)	9 (34)
2004	9 (0)	1 (0)	4 (0)	1 (0)	1 (0)	101 (96)	26 (29)	7 (13)	0 (0)	10 (12)
2005	13 (0)	1 (0)	5 (0)	1 (0)	1 (0)	87 (114)	20 (26)	8 (18)	0 (0)	14 (7)
2006	18 (34)	2 (0)	8 (0)	2 (5)	2 (3)	58 (105)	12 (32)	9 (12)	0 (0)	17 (16)
2007	23 (59)	3 (0)	12 (5)	3 (7)	3 (9)	26 (53)	3 (10)	12 (6)	0 (0)	24 (20)
2008	87 (97)	14 (23)	52 (79)	11 (11)	12 (13)	16 (16)	0 (0)	45 (24)	1 (0)	85 (82)
2009	26 (17)	5 (4)	15 (21)	4 (3)	5 (3)	0 (0)	0 (0)	16 (4)	0 (0)	27 (13)
2010	32 (25)	7 (10)	18 (20)	6 (4)	6 (4)	0 (0)	0 (0)	23 (6)	1 (4)	32 (21)
2011	38 (51)	9 (15)	23 (25)	7 (8)	7 (9)	0 (0)	0 (0)	28 (12)	1 (15)	35 (23)
2012	42 (44)	13 (16)	30 (33)	9 (3)	10 (7)	0 (0)	0 (0)	32 (21)	1 (15)	42 (38)

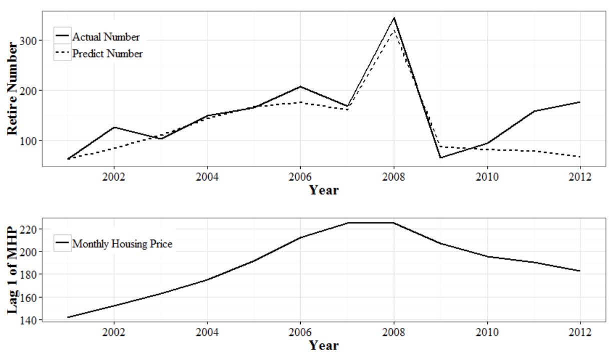
[1] the number before the parentheses is predicted retirement number.
[2] the number inside the parentheses is actual retirement number.

year to test their effects on retirement. This model has the similar parameter estimation as the selected model as shown in the right part of the table 6. We found three indicators are statistically significant and also improve the model forecasting due to lower MAPE and G^2 than the values of selected model without economic indicator, which are S&P500, Real Earnings and Wilshire 5000 as shown in table 9. Although Real price is also statistically significant, its coefficient estimates is 0.0004 leading hazard ratio is 1, which indicate it does not impact the employee's retirement behaviors. As shown in table 9, Real earnings is the most important factor among all the indicators as it has lowest G^2. The test results show that it has strong impact on the retire behaviors. As shown in figure **??**, the fluctuation of retirement plot is corresponding with 1 year lag of the trend of Real earnings. Unadjusted Monthly Housing Price (MHP) is another influential index with the lowest MAPE value. It has significantly impacts on the retire number as it increases as shown in figure 11a. However, the retire number does not decrease coinciding with the decreasing of MHP. Two market indicators(S&P500 and Wilshire 5000) are both significantly impact the employee retirement behavior with >1 hazard ratio, which indicating the employee are more likely to retire when the economics is under good condition.

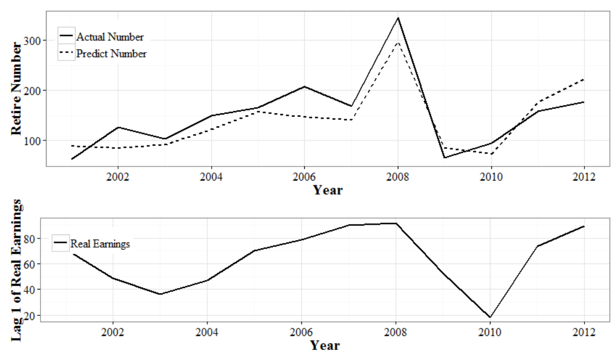
Table 9: Economic index test statistics

Economic Inidcator	χ^2	P-value	Hazard ratio	MAPE	G^2
Without Econmic indicator[1]				21.31	147.43
MHP NSA	129.614	$< .001$	1.020	17.84	220.65
MHP SA	129.516	$< .001$	1.020	17.86	221.66
Southeast MHP NSA	68.055	$< .001$	1.030	23.37	178.38
Southeast MHP SA	67.871	$< .001$	1.030	23.40	179.87
S&P500	13.319	$< .001$	1.001	20.79	129.83
Dividend	1.045	0.307	1.015	22.63	150.03
Earnings	84.895	$< .001$	1.016	21.40	105.06
Consumer Price Index	5.404	0.020	1.013	21.36	133.57
Real Price	5.522	0.019	1.000	21.22	138.72
Real Dividend	1.925	0.165	1.022	23.39	154.84
Real Earnings	80.358	$< .001$	1.013	20.33	95.02
Long Interest Rate	1.539	0.215	1.082	22.03	149.01
Unemployment Rate	32.212	$< .001$	0.849	25.08	179.49
P E10	0.041	0.839	0.998	21.31	147.97
Wilshire5000	22.392	$< .001$	1.028	20.83	121.58

[1] it is the selected model without economic indicator.



(a) Monthly Housing Price



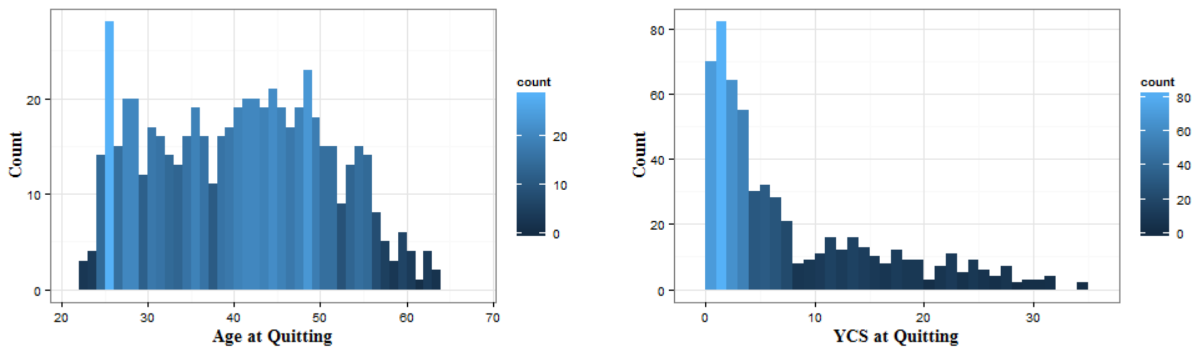
(b) Real Earnings

Figure 11: Economic indicators and retirement prediciting plot

6.3.1 Voluntary quit model without external variables

I. dependent variable are YCSH, because age is not able to predict well. Why plot YCSH by vq and age by vq to discribe why. one employee quit with 67 years old and less than 85 points. all quit before 65 years. II. shorten the length of risk set. how to shorten the risk set. iii. model comparison. (survival model, time seris model, and logsitic regression model)

Voluntary quitting model is built based on the same variables in the retirement model except *P85* and *A65*. The difficulty for building the quitting model is that we only have around 600 out of 8000 employees quitting during the 10-year study window. This causes the high proportion of censor data. Also, the model cannot generate a smooth baseline to achieve a good forecasting model (10) when age is the dependent variable, because employees quit at a wide range of ages (20 to 64) as shown in figure 12a. To better forecast quitting, we use years of service (YCS) as dependent variable, because usually an employee quits at first 10 years of service as shown in figure 12b. Since an employee will not quit if their are eligible for pension, which are 65 years of service and 85 points. We remove those employees from the risk set when they meet either one of requirements.



(a) Age at Quitting **(b)** YCS at Quitting

Figure 12: Histogram of Age and YCS at quitting

Table 10: Voluntary Quitting Models statistics

Model	DF	LR	AIC0	AIC	SBC0	SBC	Pred. MAPE	Holdout MAPE	Pred. G^2	Holdout G^2
Age (Dependent)	19	1226.81	7148.92	5960.11	7148.92	6041.44	83.71	91.35	1213.96	197.21
YCS (Dependent)	19	1012.39	7822.87	6848.48	7822.87	6929.70	23.90	21.05	65.63	2.48
YCS (Dependent) reduced riskset	19	838.42	7711.61	6911.19	7711.61	6992.52	15.16	18.77	26.25	2.08
Logistic regression	24	1870.74	6535.36	4712.62	6544.40	4938.62	15.98	17.55	23.03	3.21
Time series		1	NA	NA	NA	NA	26.41	61.32	30.88	18.33

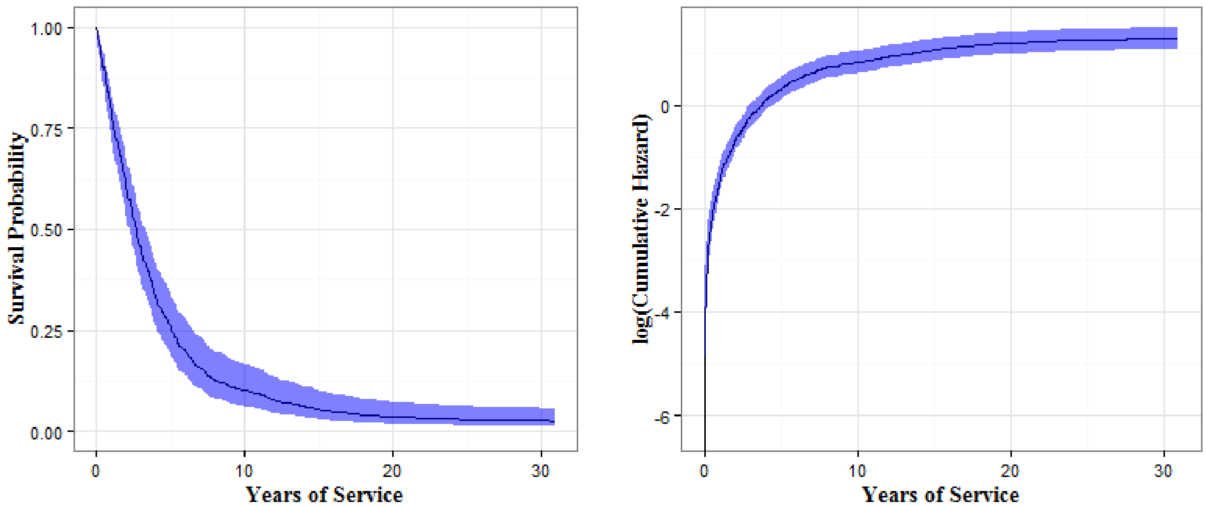
Among all the models shown in table 10, a survival model is selected with *YCS* as dependent variable and reduced risk set, and with *division*, *COCS*, *Age at credited*, and *policy* as explanatory variables. Logistic regression model also performs well as the survival model, as two models have similar MAPE and G_2 values. One reason for the close performance is that the logistic regression is the discrete model of Cox PH model.

According to our model, voluntary quitting is influenced by an employee's age at their years of service started. The coefficient estimate for age is -0.025. As the reference age is 35.44, this means that the hazard ratio for quitting of an employee that started working at age 35.44 is $exp(-.025) = .975$ indicating a 2.5% drop in hazard for each additional year later that an employee started. The employee's survival probability at any time, t, can be computed as $S(t)^{1.025} = (S(t)^{e^{0.025}})$ when age is one year below 35.44, where $S(t)$ is the baseline survival probability for a reference employee of average age at the time of starting their pension. Moving in the other direction, employee's survival probability is $S(t)^{0.975} = (S(t)^{e^{-0.025}})$ for a one year increase beyond 35.44 in the employee's starting age. Together, this implies that at any given voluntary quitting age, the employee who starts

Table 11: Parameter estimates for voluntary quitting models

		Period Model		**Yearly model**	
Parameter	Label	Parameter (Standard Error)	Hazard Ratio	Parameter (Standard Error)	Hazard Ratio
division	dir1	-3.496 (0.264)***	0.03	-3.066 (0.263)***	0.047
division	dir2	-3.100 (0.204)***	0.045	-2.652 (0.198)***	0.071
division	dir3	-3.430 (0.258)***	0.032	-3.015 (0.253)***	0.049
division	dir4	-2.929 (0.288)***	0.053	-2.533 (0.288)***	0.079
division	dir5	-3.106 (0.314)***	0.045	-2.739 (0.314)***	0.065
division	dir7	0.011 (0.144)	1.011	0.028 (0.145)	1.029
division	dir8	-1.117 (0.159)***	0.327	-0.806 (0.157)***	0.447
division	dir9	-4.811 (0.714)***	0.008	-3.985 (0.586)***	0.019
division	dir10	-1.65 (0.138)***	0.192	-1.325 (0.136)***	0.266
COCS	C	-1.165 (0.275)***	0.312	-1.163 (0.274)***	0.313
COCS	G	-0.864 (0.2)***	0.421	-0.794 (0.198)***	0.452
COCS	L	-0.656 (0.229)**	0.519	-0.711 (0.228)**	0.491
COCS	M	-0.488 (0.151)**	0.614	-0.541 (0.15)***	0.582
COCS	P	-0.541 (0.137)***	0.582	-0.530 (0.135)***	0.588
COCS	R	-0.981 (0.318)**	0.375	-0.950 (0.308)**	0.387
COCS	S	-0.682 (0.276)*	0.506	-0.691 (0.275)*	0.501
COCS	T	-0.531 (0.201)**	0.588	-0.608 (0.203)**	0.544
Age at Credited		-0.025 (0.005)***	0.975	-0.025 (0.005)***	0.975
Policy	1	1.111 (0.137)***	3.036	0.851 (0.135)***	2.343

[1] * denotes $P < 0.05$, ** denotes $P < 0.01$, and *** denotes $P < 0.001$.



(a) Survival function **(b)** Log of Cumulative hazard

Figure 13: Voluntary quitting model Baselines with 95% confident intervals

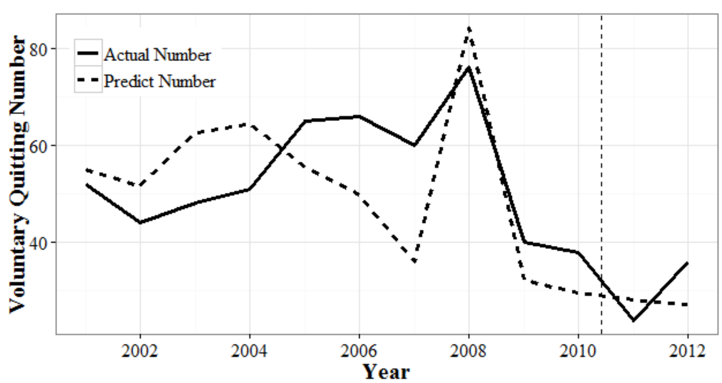


Figure 14: Voluntary Quitting Forecasting

earlier than 35.44 years old is more likely to quit than an equivalent employee who starts working at an older age.

The temporary early retirement buyout option also impact on employee's quitting behavior, as *policy* also significant in the model. The coefficient for this indicator was 1.111 leading to a hazard ratio of $e^{\cdot 1.111} = 3.04$ which indicates that, on average, an individuals hazard of quitting increased by almost 3 times during this period. This indicates that employees are more likely to quit during this specific period.

The *division* variable was a significant predictor. For analysis, the baseline level was chosen arbitrarily as division 6 so that it's hazard rate is determined by the baseline. Relative to this baseline, division 7 has a similar hazard ratio of quitting with division 6, indicating the employees have the similar quitting behavior with division 6. Conversely, the other divisions all have negative coefficients with hazard ratios less than 1 indicating that individuals within these groups have lower the hazard of group 6.

Table 12: Economic index test statistics for voluntary quitting

Economic Inidcator[1]	χ^2	**P-value**	**Hazard ratio**	**MAPE**	G^2
Without Econmic indicator				15.71	28.33
MHP NSA	37.93	<.001	1.012	13.62	18.44
MHP SA	37.75	<.001	1.012	13.64	18.50
Southeast MHP NSA	39.80	<.001	1.020	13.74	18.67
Southeast MHP SA	39.63	<.001	1.020	13.75	18.68
SP500	0.02	0.879	1.000	15.46	27.79
Dividend	31.21	<.001	1.077	13.01	18.01
Earnings	8.83	0.003	1.009	15.81	23.88
Consumer Price Index	35.84	<.001	1.024	24.26	42.95
Real Price	5.34	0.021	1.000	17.94	34.52
Real Dividend	26.66	<.001	1.089	12.40	16.37
Real Earnings	3.71	0.054	1.005	14.16	23.34
Long Interest Rate	12.04	0.001	0.789	19.10	38.73
Unemployment Rate	2.99	0.084	1.068	17.04	34.96
P E10	16.22	<.001	0.968	17.95	35.19
Wilshire5000	10.75	0.001	1.031	15.84	21.31

[1] it is the selected model without economic indicator.

The *COCS* variable was another significant predictor. For analysis, the baseline level was engineer for this variable so that it's hazard rate is determined by the baseline. Relative to this baseline, all the other divisions have negative coefficients with hazard ratios less than 1, especially for Crafts group with -1.165 of coefficient estimation and 0.312 ($exp(-0.165)$) of hazard ratio, indicating that individuals within these groups have lower the hazard of engineer group. In other words, engineers are more likely to quit than the other job categories, because it may be easier for engineer to find a better job in the job market.

In general, differences in quitting rates could be caused by differences in age demographics, leadership, departmental and job function, or departmental leadership.

The baseline survival function and log hazard function are shown in figure 13. The survival probability starts to steeply decease from 0 to 10 years of service. By years of service at 10 the survival probability has decreased to close to 0.25, which indicates that 75% of employees quitting at first 10 years of service. The slope of survival function decreases flatly to 0 from 10 t0 30 years of service. Accompanying the survival function is the log of the cumulative hazard ratio. Again, the steep rise in the cumulative hazard between years of service at 0 to 10 indicates the increased quitting activity during this period. After this the cumulative hazard levels off indicating a drop in the hazard rate at these future points.

6.3.2 Voluntary quit model with external variables

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

7 Conclusions and Managerial Implications

Conclusions and Managerial Implications

1) We were construct an individual level retirement model for current and past workers that accurately predicted yearly retirement levels.

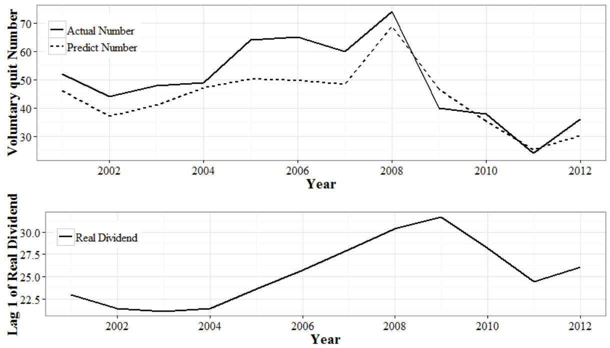


Figure 15: Voluntary quitting with Real Dividend

2) A semi-parametric cox proportional hazards model was used to model age at retirement. The model investigated demographic, organizational, and macroeconomic factors that might impact the retirement decision.

3) Factors that had the strongest influence included years of service, whether or not an ERP was in effect, whether or not the individual exceeded the number of points required for full retirement benefits. The most valuable external factor impacting retirement decisions is the Real Earnings as defined by the bureau of labor statistics.

4) Discuss impacts captured by baseline.

5) Factors that did not seem to significantly impact retirement decisions included gender, payroll classification, job classification code, and a large number of additional macroeconomic indicators.

6) The model was tested and was able to accurately predict number of retirements up to two years into the future. It may be accurate beyond that as well but sufficient data was not available to test this.

7) Why might real earnings be important.

8) Moderating effect of age. Workers that stayed beyond the term required for full benefits were seen to have much reduced probabilities of retiring as indicated by a moderating interaction indicator.

Managerial Implications....

References

P. D. Allison. Survival analysis using the sas system: A practical guide. cary, north carolina: Sas institute, 1995.

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

J. O. Berger and M.-H. Chen. Predicting retirement patterns: Prediction for a multinomial distribution with constrained parameter space. *The Statistician*, pages 427–443, 1993.

M. Braun and D. A. Schweidel. Modeling customer lifetimes with multiple causes of churn. *Marketing Science*, 30(5):881–902, 2011.

T. Briggs. Survival analysis for predicting employee turnover. http://www.slideshare.net/twbriggs/survival-analysis-for-predicting-employee-turnover, November 2014. Accessed on 10/29/2015.

P. L. Brockett, L. L. Golden, M. Guillen, J. P. Nielsen, J. Parner, and A. M. Perez-Marin. Survival analysis of a household portfolio of insurance policies: how much time do you have to stop total customer defection? *Journal of Risk and Insurance*, 75(3):713–737, 2008.

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.

R. L. Clark. Retirement: Early retirement incentives. http://www.encyclopedia.com, 2002. Accessed: 2015-12-02.

E. Claus, N. Risch, and W. Thompson. Genetic analysis of breast cancer in the cancer and steroid hormone study. *American journal of human genetics*, 48(2):232, 1991.

D. Collett. *Modelling survival data in medical research*. CRC press, 2015.

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.

IBM. Hr analytics: Invest in your most valuable asset with hr analytics. http://www-01.ibm.com/software/analytics/solutions/operational-analytics/hr-analytics/, December 2013. Accessed on 10/29/2015.

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.

J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

D. G. Kleinbaum. Survival analysis, a self-learning text. *Biometrical Journal*, 40(1):107–108, 1998.

M. Kuhn and K. Johnson. *Applied predictive modeling*. Springer, 2013.

W. R. Lane, S. W. Looney, and J. W. Wansley. An application of the cox proportional hazards model to bank failure. *Journal of Banking & Finance*, 10(4):511–531, 1986.

J. F. Lawless. *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons, 2011.

M. J. LeClere. Preface modeling time to event: Applications of survival analysis in accounting, economics and finance. *Review of Accounting and Finance*, 4(4):5–12, 2005.

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.

J. S. Long and J. Freese. *Regression models for categorical dependent variables using Stata*. Stata press, 2006.

J. Lu. Predicting customer churn in the telecommunications industry: An application of survival analysis modeling using sas. *SAS User Group International (SUGI27) Online Proceedings*, pages 114–27, 2002.

W. Q. Meeker and L. A. Escobar. *Statistical methods for reliability data*. John Wiley & Sons, 2014.

M. L. Moeschberger and J. Klein. *Survival analysis: Techniques for censored and truncated data: Statistics for biology and health*. Springer, 2003.

C. W. Mueller and J. L. Price. Some consequences of turnover: A work unit analysis. *Human Relations*, 42(5):389–402, 1989.

W. Pan and R. Chappell. A nonparametric estimator of survival functions for arbitrarily truncated and censored data. *Lifetime data analysis*, 4(2):187–202, 1998.

PWC. Dont just collect data. generate insight. `http://www.pwc.com/us/en/hr-saratoga.html`, 2015. Accessed: 2015-11-15.

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.

M. Wang and K. S. Shultz. Employee retirement: A review and recommendations for future investigation. *Journal of Management*, 36(1):172–206, 2010.

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

X. Zhu, W. Seaver, R. Sawhney, S. Ji, B. . Holt, G. Upreti, and G. B. Sanil. Employee turnover forecasting for human resource management based on time series analysis. *Journal of Applied Statistics*, (under review), 2015.