SCSUG 2017 – Paper TK57

The Cox Hazard Model for Claims Data: a Bayesian Non-Parametric Approach

Samuel Berestizhevsky, InProfix Inc, Boca Raton, FL
Tanya Kolosova, InProfix Inc, Boca Raton, FL

ABSTRACT

General insurance protects individuals and organizations from financial losses due to damages or legal liabilities. It allows policyholders to exchange the risk of a large loss for the certainty of smaller periodic payments of premiums. Insurers allocate the premiums dollars into investment and claims payments. As it is for an insurer to manage its investment portfolio, it is equally important to manage its claims portfolio. Claim management is the analytics of insurance costs. It requires applying statistical techniques in the analysis and interpretation of the claims data. In the data-driven industry of general insurance, claim management provides useful insights for insurers to make better business decisions.

The central piece of claim management is claims modeling. Two strategies are commonly used by insurers to analyze claims: the two-part approach that decomposes claims cost into frequency and severity components, and the pure premium approach that uses the Tweedie distribution. In this article we evaluate additional approach to claims analysis – time-to-event modeling.

In this article, we provide a general framework to look into the process of modeling of claims using Cox hazard model. This model is a standard tool in survival analysis for studying the dependence of a hazard rate on covariates and time. Although the Cox hazard model is very popular in statistics, in practice data to be analyzed often fails to hold assumptions underlying this model. This article also is a case study intended to indicate a possible application of Cox hazard model to workers' compensation insurance, particularly occurrence of claims (disregarding claims size).

INTRODUCTION

The term "survival data" has been used in a wide meaning for data involving time to a certain event. This event may be the appearance of a tumor, the development of some disease, cessation of smoking, etc. Applications of the statistical methods for survival data analysis have been extended beyond the biomedical field and used in areas of reliability engineering (lifetime of electronic devices, components or systems), criminology (felons' time to parole), sociology (duration of first marriage), etc. Depending on the area of application, different terms are used: survival analysis – in biological science, reliability analysis – in engineering, duration analysis – in social science. Further, in the article, we will use term "time-to-event analysis" that is more suitable for insurance claims analysis.

A central quantity in survival analysis (time-to-event analysis) is the hazard function or the survival (time-to-event) function. The most common approach to model covariate effects on survival (time-to-event) is the Cox hazard model developed and introduced by Cox (1972), which takes into account the effect of censored observations. Although the Cox hazard model is very popular in statistics, in practice data to be analyzed often fails to hold assumptions. There are several important assumptions which need be assessed before the model results can be safely applied. First, the proportional hazards assumption means that hazard functions are proportional over time. Second, the explanatory variable acts directly on the baseline hazard function and not on the failure time, and remains constant over time. For example, when a cause of claims interacts with time, the proportional hazard assumption fails. Or, when the hazard ratio changes over time and survival curves are not "parallel", the proportional hazard assumption is violated. We present application of Bayesian approach to survival analysis (time-to-event analysis) that allows dealing with violations of assumptions of Cox hazard model.

This article is a case study intended to indicate possible applications to workers' compensation insurance, particularly occurrence of claims. We studied workers' compensation claims during the 2 years period from November 01, 2014 till October 31, 2016. Claims data was provided by a leading insurance company. The risk of occurrence of claims was studied for selected industries and locations (USA states).

MODEL

THE COX MODEL FOR CLAIMS EVENTS ANALYSIS

Time-to-event (or survival) function $\boldsymbol{S(t)}$ describes the proportion of policies "surviving" without a claim to or beyond a given time (in days):

$$\boldsymbol{S(t)\ =\ P(T>t)}$$

where:

$\boldsymbol{T}$ – time to claim in a randomly selected policy

$\boldsymbol{t}$ – a specific point in time

Hazard function $\boldsymbol{h(t)}$ describes instantaneous failure rate at time $\boldsymbol{t}$**:**

$$\boldsymbol{h(t)\ =\lim_{\Delta t\to 0}\frac{P(t\le T<t+\Delta t|T\ge t)}{\Delta t}}$$

Cox (1972) proposed a model which doesn't require assumption that events times follow certain probability distribution. As a consequence, Cox model is considerably robust.

Cox hazard model can be written as:

$$\boldsymbol{h_i(t)=h_0(t)exp\sum_{j=1}^{k}\beta_j x_{ij}}$$

where:

$\boldsymbol{h_i(t)}$ – the hazard function for subject $\boldsymbol{i}$ at time $\boldsymbol{t}$

$\boldsymbol{x_1,\ \dots,x_k}$ – the covariates

$\boldsymbol{h_0(t)}$ – the baseline hazard function, that is the hazard function for the subject whose covariates $\boldsymbol{x_1,\ \dots,x_k}$ all have values of 0

$\boldsymbol{\beta_1,\dots,\beta_k}$ – the coefficients of Cox model.

Cox hazard model is also called Proportional Hazard Model because the hazard for any subject is a fixed proportion (hazard ratio) to the hazard for any other subject:

$$\boldsymbol{HR=\ h_i(t)/h_p(t)=(h_0(t)exp\sum_{j=1}^{k}\beta_j x_{ij})/(h_0(t)exp\sum_{j=1}^{k}\beta_j x_{pj})}$$

Baseline hazard $\boldsymbol{h_0(t)}$ cancels out, and $\boldsymbol{HR}$ is constant with respect to time:

$$\boldsymbol{HR=\ exp\sum_{j=1}^{k}\beta_j(x_{ij}-x_{pj})}$$

Estimated survival (time-to-event) probability at time $\boldsymbol{t}$ can be calculated using estimated baseline hazard function $\boldsymbol{h_0(t)}$ and estimated $\boldsymbol{\beta}$ coefficients:

$$\boldsymbol{S_i(t)=exp\left\{-h_0(t)exp\sum_{j=1}^{k}\beta_j x_{ij}\right\}}$$

where:

$\boldsymbol{S_i(t)}$ – the time-to-event function for subject $\boldsymbol{i}$ at time $\boldsymbol{t}$

$\boldsymbol{x_1,\ \dots,x_k}$ – the covariates

$\boldsymbol{h_0(t)}$ – the baseline hazard function, that is the hazard function for the subject whose covariates $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}$ all have values of 0

$\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_k}$ – the coefficients of Cox model.

APPLICATION OF THE COX MODEL FOR CLAIMS EVENTS ANALYSIS

We identified 3 main goals of time-to-event analysis for workers' compensation claims:

1. Estimate time-to-event function $\boldsymbol{S(t)}$
2. Estimate effects of industry covariate
3. Compare time-to-event functions for different industries:

$$\boldsymbol{H_0: S_i(t) = S_p(t)\ vs.\ H_1: S_i(t) \neq S_p(t)}$$

In order to build an appropriate model, we had to address the nature of claims process. In contrast with biomedical applications where an event of interest, for example, is death and thus can happen only once, in workers' compensation insurance claims happen multiple times, because for each policy there are possible multiple claims. There are many different approaches that one could use to model repeated events in a time-to-event analysis. The choice depends on the data to be analyzed and the research question to be answered. We considered each claim as a single event, and built models that didn't account for claims dependence within the same policy.

Below is a short review of different models.

The counting process model

In the counting process model, each event is assumed to be independent, and a subject contributes to the risk set for an event as long as the subject is under observation at the time the event occurs. The data for each subject with multiple events is described as data for multiple subjects where each has delayed entry and is followed until the next event. This model ignores the order of the events, leaving each subject to be at risk for any event as long as it is still under observation at the time of the event. This model doesn't fit our application needs because the entry time is considered as a time of the previous event, and time-to-event is calculated as a time between consecutive events.

The conditional model A

This conditional model assumes that it is not possible to be at risk for a subsequent event without having experienced the previous event (i.e. you cannot be at risk for event 2 without having experienced event 1). In this model, the time interval of a subsequent event starts at the end of the time interval for the previous event. This model doesn't fit our application needs because it introduces dependence between consecutive claims.

The conditional model B

This model only differs from the previous model in the way the time intervals are structured. In this model each time interval starts at zero and ends at the length of time until the next event. This model doesn't fit our application needs because it introduces dependence between claims within the same policy.

The marginal model

In the marginal model each event is considered as a separate process. The time for each event starts at the beginning of follow up time for each subject. Furthermore, each subject is considered to be at risk for all events, regardless of how many events each subject actually experienced. Thus, the marginal model considers each event separately and models all the available data for each event. This model fits our application needs and was used for the analysis.

DATA

Our case study is based on claims data from one of leading insurance companies. For each policy, the following data was used:

1. The start and the end date of the policy
2. Industry in which policy was issued
3. Date of claim occurrence
4. Date of claim reported
5. State where claim was reported

In this study, we focused our analysis on claims that led to payments.

DATA TRANSFORMATION

We analyzed workers compensation claims data for the 2 years period. Each claim was associated with an industry of the policy and with a state (location) of the claim. To prepare this data for the marginal model, each claim event was considered as a separate process. The time to each event was calculated starting from the beginning of follow up time or from the beginning of the policy, whichever happened later. If there were no claim events for a policy during the observation period, the policy was censored at the end of the observation or at the end of the policy, whichever happened earlier. To note, a subject is said to be censored, if it is lost to follow up, or dropped out of the study, or if a claim event didn't happen during the observation period.

An example is presented in Figure 1:

- Policy A started before January; there were 2 claims that happened in May and June; policy ended in August.
- Policy B started in March; there was one claim in August; policy was cancelled in October.
- Policy C started in April; there were no claims in the observed period of time.

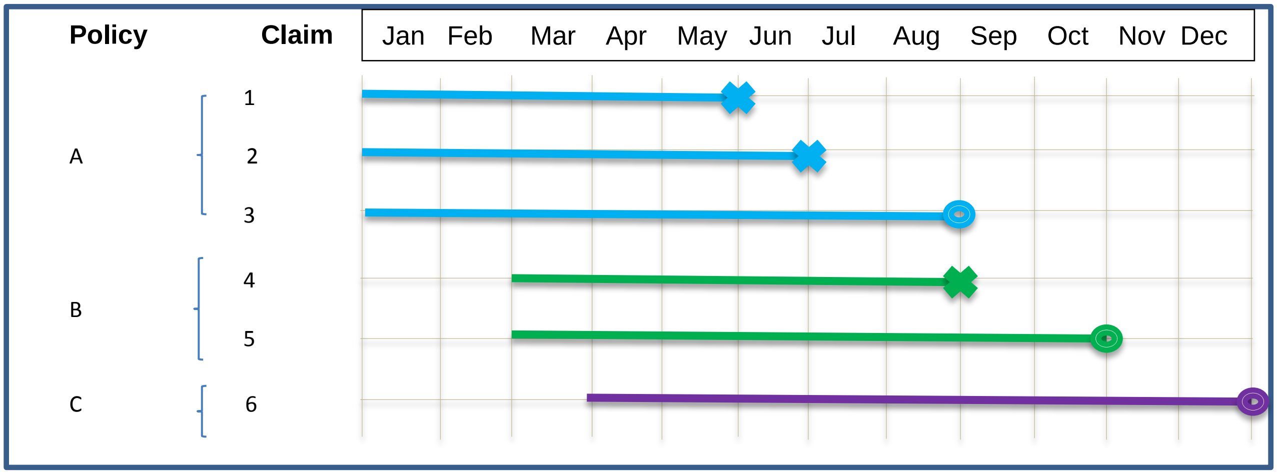

Figure 1. Example of Claims Data

For this example, data is shown in Table 1.

We separated US states into groups with statistically similar frequencies of claims in a month. This produced 4 groups of claims by states (41 states in total):

1st group: CA

2nd group: FL, GA, IL, NJ, NY, PA, TX

3rd group: AL, AZ, CO, IA, IN, LA, MA, MD, MI, MO, NC, NV, OK, OR, SC, TN, VA, WI

4th group: AR, CT, DC, DE, ID, KS, KY, MN, MS, NE, NH, NM, SD, VT, WV

Policy	Claim	Time-to-event	Event	Censor
A	1	5	1	1
A	2	6	2	1
A	3	8	3	0
B	4	6	1	1
B	5	8	2	0
C	6	9	1	0

Table 1. Example of Claims Data Set

In this article we present case study of analysis and modeling performed on claims data of 3rd group of US states.

In our analysis we assumed that each claim event independent, and a policy within the same industry did not contribute to the risk set for an event as long as the policy is under observation at the time of the claim. The data for each policy with multiple claim events was described as data for multiple claims, where each claim has an entry time at the beginning of the policy or beginning of the observation period – whichever is later.

Thus $\boldsymbol{HR}$ model looks like:

$$\boldsymbol{HR = \ exp\sum_{j=1}^{k}\beta_j(x_{ij} - x_{pj})}$$

where:

$\boldsymbol{HR}$ – the hazard ratio of hazard function for industry $\boldsymbol{i}$ related to the baseline hazard function for the industry $\boldsymbol{p}$.

In most insurance risk papers, the authors take the proportional assumption for granted and make no attempts to check that it has not been violated in their data. However, it is a strong assumption indeed. In order to verify proportional hazard assumption, we used Kaplan-Meier empirical product-limit survival estimates, as well as ‘log-negative-log’ Kaplan-Meier estimated survival functions. Kaplan-Meier empirical plots allow to evaluate the assumption visually: if the proportional hazard assumption holds, the curves of the Kaplan-Meier empirical product-limit survival estimates should not intersect, and the curves of the ‘log-negative-log’ plot should be parallel, with distance between them constant over time. The plots on

Figure 2 and Figure 3 show that the proportional hazard model assumption doesn’t hold.

We used industry as a categorical covariate assuming that time-to-event functions vary by industries. It would be wrong to assume that there was no time-dependent impact on the baseline hazard function for different values of this covariate variable. For example, hazard changes for Agriculture depending on seasons, or for Transportation – depending on weather, or for Hospitality – depending on school breaks schedule.

All these conditions are latently depend on time, which means that the impact of industry categorical variable does not remain constant over time. In order to account for season dependency we introduced time-dependent covariate for winter season and used extended Cox model:

$$\boldsymbol{h_i(t) = h_0(t)exp\left(\sum_{j=1}^{k}\beta_j x_{ij} + \sum_{n=1}^{m}\gamma_n x_{in} g_n(t)\right)}$$

where:

$\boldsymbol{h_i(t)}$ – the hazard function for subject $\boldsymbol{i}$ at time $\boldsymbol{t}$

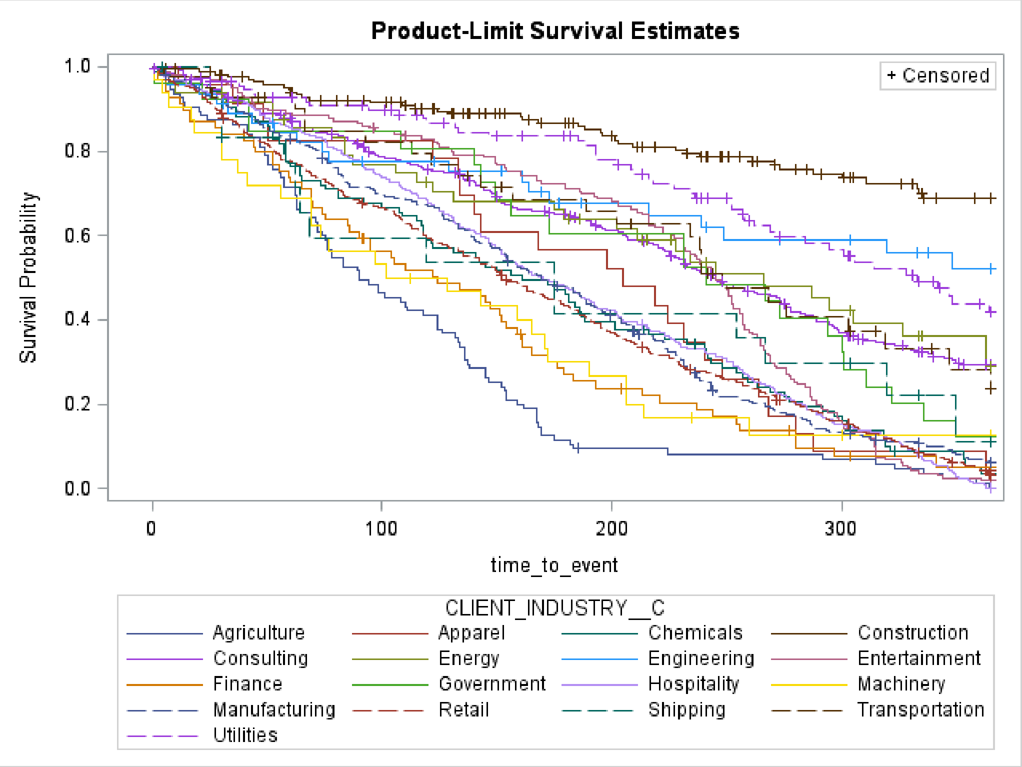

Figure 2. Kaplan-Meier Empirical Product-Limit Survival Estimates

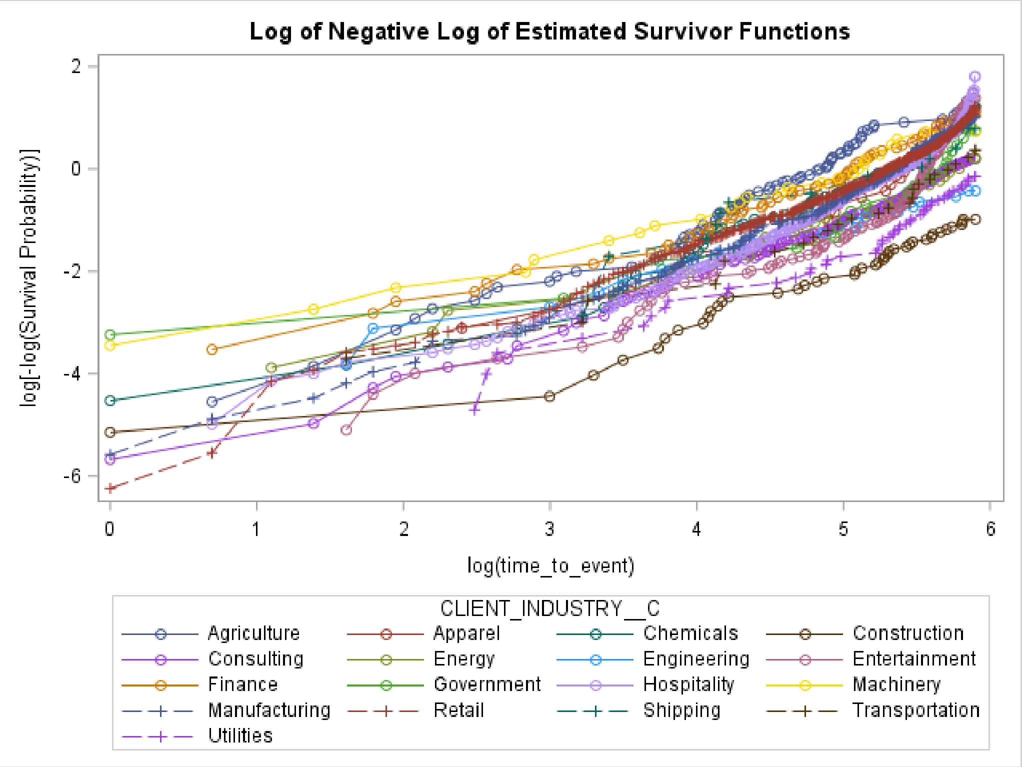

Figure 3. ‘Log-Negative-Log’ Kaplan-Meier Estimated Survival Functions

$\boldsymbol{x_1}, \ \dots, \boldsymbol{x_k}$ – the covariates

$\boldsymbol{h_0(t)}$ – the baseline hazard function, that is the hazard function for the subject whose covariates $\boldsymbol{x_1}, \ \dots, \boldsymbol{x_k}$ all have values of 0

$\boldsymbol{g_n(t)}$ – the function of time (time itself, log time, etc.)

$\boldsymbol{\beta_1}, \dots, \boldsymbol{\beta_k}$ – the coefficients of Cox model.

Applying this approach, our model looks like:

$$\boldsymbol{h_i(t) = h_0(t) exp\left(\sum_{j=1}^{k} \beta_j x_j + \gamma \times season \times ln(t)\right)}$$

where:

$\boldsymbol{h_i(t)}$ – the hazard function for industry $\boldsymbol{i}$ at time $\boldsymbol{t}$

$\boldsymbol{h_0(t)}$ – the baseline hazard function, in our case - the hazard function for one selected industry

$$\boldsymbol{x_j} = \begin{cases} \mathbf{1}, if \ i = j \\ \mathbf{0}, if \ i \neq j \end{cases}$$

$$\boldsymbol{season} = \begin{cases} \mathbf{1}, if \ an \ event \ (claim) \ happened \ during \ winter \ season \ (months \ 11, 12, 1, 2, 3) \\ \mathbf{0}, if \ an \ event \ (claim) \ didn't \ happen \ during \ winter \ season \ \ (months \ 4-10) \end{cases}$$

Calculation of survival functions when we have time-varying covariates is a little bit more complicated, because we need to specify a path or trajectory for each variable (Rodriguez G. 2007). For example, if a policy started on 1st of April, survival function should be calculated using hazard corresponding to $\boldsymbol{season} = \mathbf{0}$ for time-to-event $\boldsymbol{t} < \ \mathbf{214}$ days (from 1st of April till the 1st of November), while for time-to-event $\boldsymbol{t} \geq \ \mathbf{214}$ – using hazard corresponding to $\boldsymbol{n} = \mathbf{1}$. For another example, if a policy started on 1st of August, survival function should be calculated using hazard corresponding to $\boldsymbol{season} = \mathbf{0}$ for time-to-event $\boldsymbol{t} < \ \mathbf{92}$ days (from 1st of August till 1st of November), and $\boldsymbol{t} \geq \ \mathbf{243}$ days (from 1st of August till 1st of April), while for time-to-event $\mathbf{92} \leq \boldsymbol{t} < \ \mathbf{243}$ – using hazard corresponding to $\boldsymbol{season} = \mathbf{1}$.

Unfortunately, the simplicity of calculation of $\boldsymbol{S_i(t)}$ is lost: we can no longer simply raise the baseline survival function to a power.

Yet additional challenge in our data was reliability of claims date. There were 2 dates available – the date of the event caused the claim, and the date when the claim was reported. Wide variability of time intervals between these 2 dates created additional challenge in application of Cox hazard model, as time-to-event became essentially random variable. We used dates of claims occurrence.

Assumptions violation of the Cox hazard model should be taken into account. If possible, appropriate modification of the model should be used to enable more precise interpretation (Hosmer, Lemeshow, 1999), however we had vast amount of unobserved data. Another possibility was to use Bayesian non-parametric approach.

BAYESIAN APPROACH

Bayesian approach is based on a solid theoretical framework. The validity and application of the Bayesian approach do not rely on the proportional hazards assumption of the Cox model, thus, generalizing the method to other time-to-event models and incorporating a variety of techniques in Bayesian inference and diagnostics are straightforward. In addition, inference doesn't rely on large sample approximation theory and can be used for small samples. Information from prior research studies, if available, can be readily incorporated into the analysis as prior probabilities. Although choosing prior distribution is difficult, the non-informative uniform prior probability is proved to lead to proper posterior probability (Gelfand, Mallick, 1994). Instead of using partial Maximum Likelihood Estimation in Cox Hazard model, Bayesian method uses Markov Chain Monte Carlo method to generate posterior distribution by the Gibbs sampler: sample from a specified prior probability distribution so that the Markov chain converges to the desired proper posterior distribution. Only disadvantage of this method is that this process is computation intensive.

DEPLOYMENT WITH SAS® PROC PHREG

The estimation of the Cox hazard model using Bayesian approach was implemented using the PHREG procedure:

```
proc phreg data=claims_data_group_3 ;
  by accident_state ;
  class client_industry ;
  model (zero_time time_to_event)*censor(0) = client_industry season_event;
  season_event = season*log(time_to_event);
  bayes seed = 1 outpost = post;
run ;
```

The data set claims_data_group_3 contained data ready for analysis, where:

- season = 1 means that claim happened in the period from November to March, otherwise season = 0
- censor =1 means presence of claim, and censor = 0 means no claims during the observed period

A fragment of claims_data_group_3 data set presented in Table 2.

accident_state	policy	client_industry	zero_time	time_to_event	season	sensor
…	…	…	…	…	…	…
MD	10	Construction	0	119	1	0
MD	20	Construction	0	162	0	0
CO	10	Consulting	0	220	0	1
AL	10	Electronics	0	263	0	0
AL	20	Electronics	0	365	1	0
VA	10	Entertainment	0	237	1	1
WI	10	Finance	0	95	0	1
MA	10	Hospitality	0	108	0	1
IN	10	Machinery	0	7	0	1
…	…	…	…	…	…	…

Table 2. Fragment from claims_data_group_3 Data Set

We obtained separate analyses on observations in US states using the statement BY accident_state.

The variable client_industry was defined as a covariate in the analysis by the statement CLASS client_industry, and by effect variable in the MODEL statement.

The time-dependent covariate season_event is defined in the MODEL statement. The separate statement defines how to calculate season_event.

The BAYES statement requests a Bayesian analysis of the model by using Gibbs sampling.

In the PROC PHREG above we specified a seed value as a constant to reproduce identical Markov chains for the same input data. However, during our analysis we didn’t specify seed value at all, and used this option as SEED= . It allowed us to use a random seed (derived from the time of day) and then compare multiple results to evaluate robustness of the model. We didn’t specify prior distribution, thus applying uniform non-informative prior.

The result of estimation of Cox hazard models using Bayesian method is estimation of $\boldsymbol{\beta}$ coefficients.

PROC PHREG does not produce baseline survival function when time-dependent covariate is defined. To calculate the baseline survival function, we used the following work around:

```
data ds ;
  set claims_data_group_3  ;
  season_event = season*log(time_to_event);
run ;

data industry;
  client_industry = "Utilities" ;
  season_event = 0;
run ;

proc phreg data=ds ;
  by accident_state ;
  class client_industry ;
  model (zero_time time_to_event)*censor(0) = client_industry season_event;
  bayes seed=1 ;
  baseline out=baseline survival=s covariates=industry;
run ;
```

INTERPRETATION OF RESULTS

Utilities industry was used as a baseline for hazard, meaning that hazard for all other industries were estimated relatively to Utilities industry. Below are results for several US states along with explanations.

State of Colorado

For Colorado, data represents claims for 8 industries. Estimations of $\boldsymbol{\beta}$ coefficients of Cox model for each industry except Utilities, and for season_event covariate are presented in Table 3. As Utilities industry was used as baseline for hazard, $\boldsymbol{\beta}$ coefficient for Utilities equal 0.

Parameter	Mean estimate of $\boldsymbol{\beta}$	Standard Deviation	95% HPD Interval	
Consulting	1.0589	0.9106	-0.6557	2.9269
Energy	1.3495	0.8135	-0.1058	3.1137
Entertainment	2.7912	0.9851	0.8430	4.7817
Finance	1.2907	0.9998	-0.7032	3.2725
Hospitality	1.9079	0.7654	0.4868	3.4654
Manufacturing	2.0940	0.8167	0.4742	3.7191
Retail	2.0322	0.7840	0.5722	3.6240
Season_event	0.2476	0.0553	0.1384	0.3559

Table 3. Mean estimate of $\boldsymbol{\beta}$ for Colorado

Calculation of survival functions when we have time-varying covariates is not straightforward, because we need to specify exactly when a specific policy started, and when related to the start date of the policy the winter season occurred. For purposes to compare survival functions of different industries, we built time-to-event functions for each industry and season=0. According to the time-to-event function for Consulting industry, for example, there is 84% chances that there will be no claims before 100th day of a policy, and there is 16% chances that there will be no claims at all for one year policy (see Figure 4).

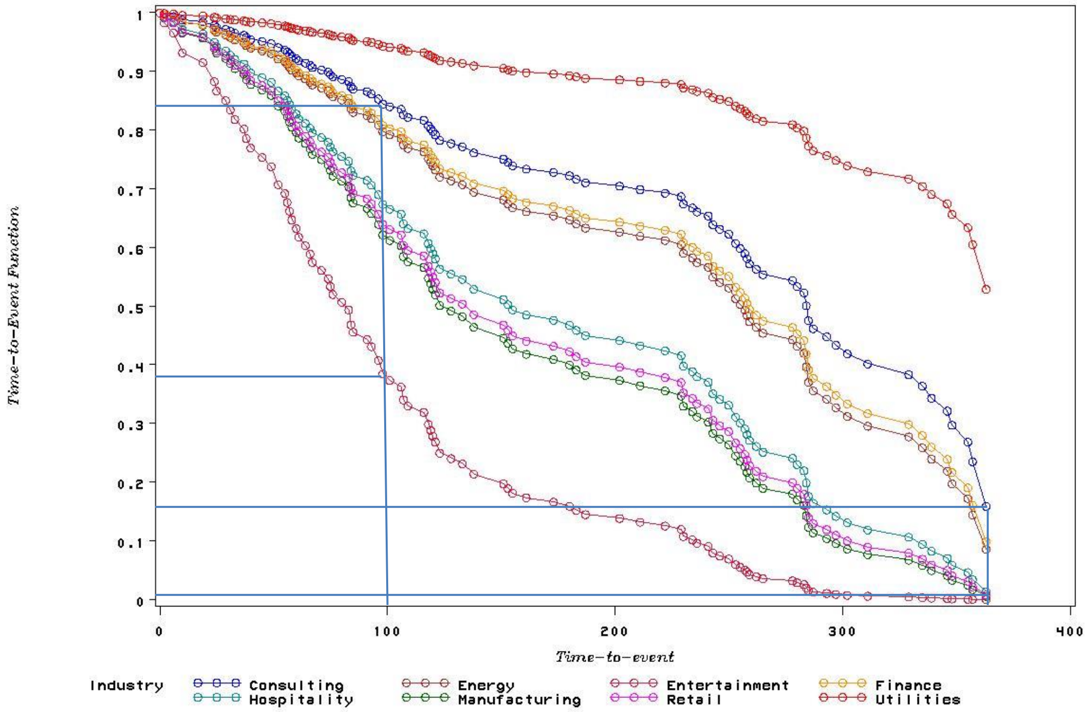

Figure 4. Time-to-Event Function for Claims per Industry in Colorado

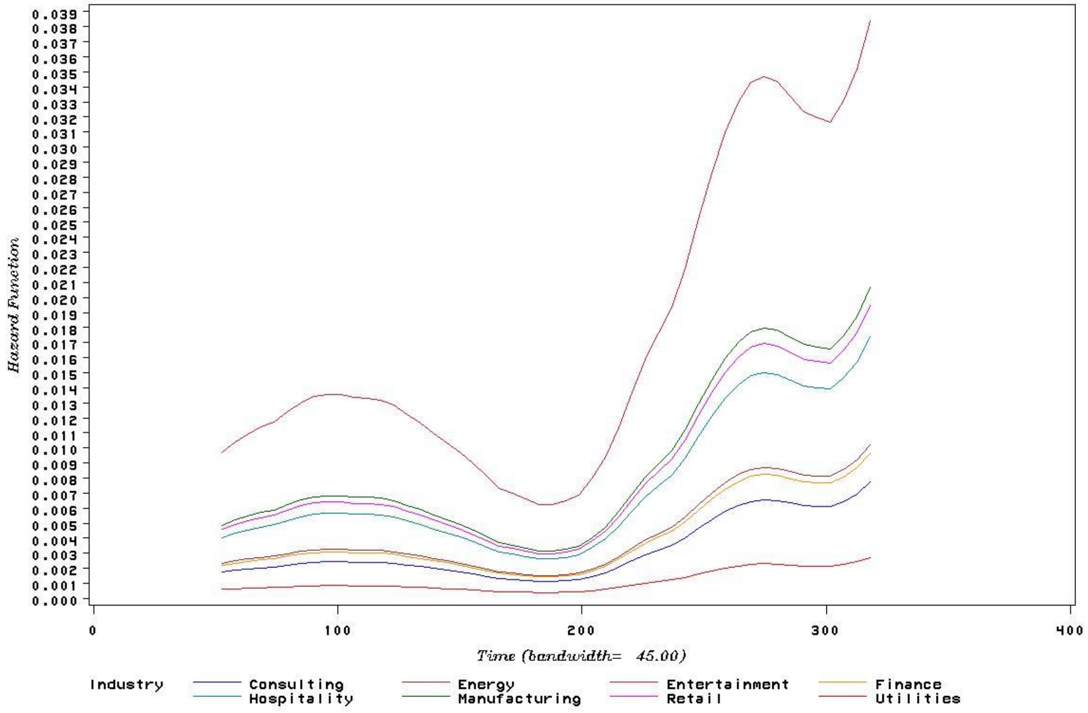

Figure 5. Hazard Function for Claims per Industry in Colorado

The time-to-event functions allow to estimate and to compare chances of claims among industries. For example, for Entertainment industry there is 37% chances that there will be no claims before 100th day of a policy, and 0% chances that there will be no claims at all for a one year policy. In other words, Consulting industry in state of Colorado presents 16% higher chances to have no claims during a one year policy than Entertainment.

Also, we can observe that Hospitality, Manufacturing and Retail have very similar risks of claims.

Hazard function presented on Figure 5 shows that the hazard of claims reaches first high hazard around 3 months from the beginning of the policy, and then starting from the second half of the policy year continuously increases, achieving highest risk at the end of the policy term.

Figure 5 was produced with the SMOOTH macro (Allison, 2012).

Time-dependent covariate is significant with $\boldsymbol{\beta}$=0.2476. This means that hazard ratio during winter season in Colorado is 28% higher, controlling for the other covariates ($\boldsymbol{exp}(\mathbf{0.2476}) - \mathbf{1} \approx \mathbf{0.28}$). An estimation of time-to-event function for a specific policy should take into consideration when the policy started – and thus, when during this policy chances of claims will increase due to the winter season.

State of Massachusetts

For Massachusetts, data represents claims for 9 industries. Estimations of $\boldsymbol{\beta}$ coefficients of Cox model for each industry except Utilities, and for season_event covariate are presented in the Table 4. As Utilities industry was used as baseline for hazard, $\boldsymbol{\beta}$ coefficient for Utilities equal 0.

The time-to-event function presented on Figure 6 demonstrates that Consulting industry has 85% chances that there will be no claims before 100th day of a policy, and only 12% chances that there will be no claims at all for a one year policy.

The time-to-event functions also allow seeing that Manufacturing and Retail industries in Massachusetts have similar risks of claims. These industries have 77% chances that there will be no claims before 100th day of a policy, and 3% chances that there will be no claims at all for a one year policy for each of these industries.

Another group of industries with similar risk of claims is Chemicals, Apparel and Entertainment industries. There are 74% chances that there will be no claims before 100th day of a policy, and 1.7% chances that there will be no claims at all for a one year policy for each industry.

Parameter	Mean estimate of $\boldsymbol{\beta}$	Standard Deviation	95% HPD Interval	
Apparel	3.3516	1.3602	0.8272	6.2793
Chemicals	3.3323	1.3057	0.8944	6.0357
Consulting	2.7005	1.2588	0.4473	5.3216
Entertainment	3.3618	1.2291	1.2855	6.0712
Hospitality	3.6521	1.2287	1.5277	6.2932
Machinery	4.3370	1.3759	1.7317	7.1790
Manufacturing	3.2123	1.2362	0.9801	5.7963
Retail	3.2018	1.2407	1.0116	5.8300
Season_event	0.2520	0.0479	0.1628	0.3511

Table 4. Mean estimate of $\boldsymbol{\beta}$ for Massachusetts

Hazard function presented on Figure 7 shows that the hazard of claims grows through the term of policies and reaches its highest risk at the end of the term.

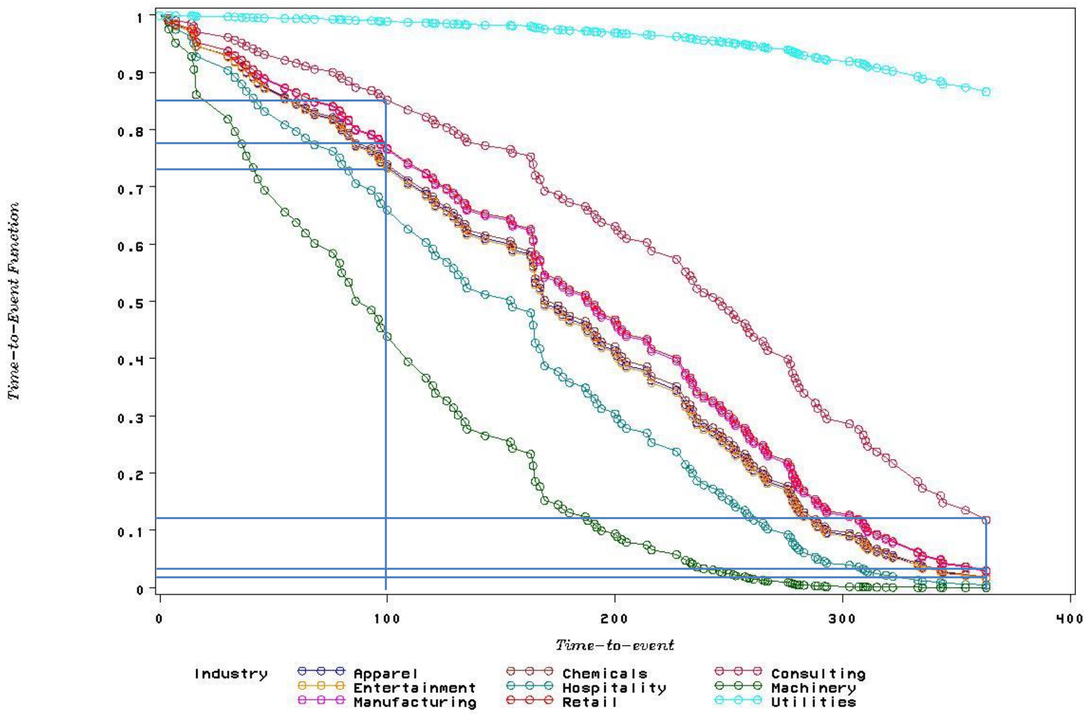

Figure 6. Time-to-Event Function for Claims per Industry in Massachusetts

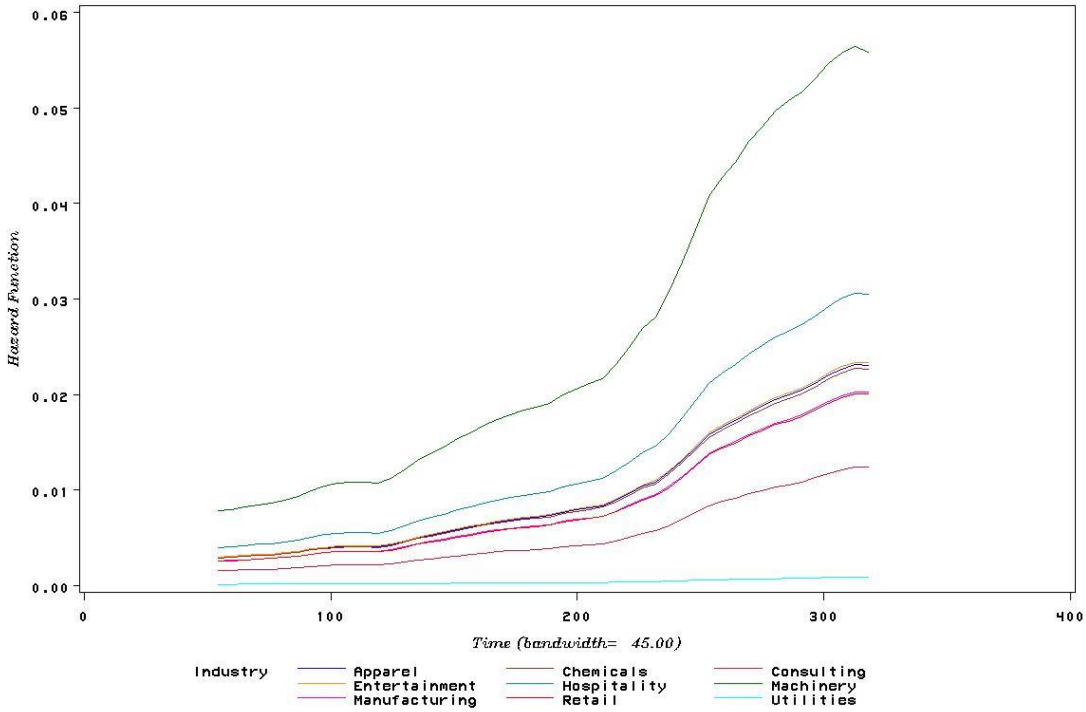

Figure 7. Hazard Function for Claims per Industry in Massachusetts

Time-dependent covariate is significant with $\boldsymbol{\beta}$=0.2520. This means that hazard ratio during winter season in Massachusetts is 29% higher, controlling for the other covariates ($\boldsymbol{exp}(\mathbf{0.2520}) - \mathbf{1} \approx \mathbf{0.29}$).

State of Maryland

For Maryland, data represents claims for 6 industries. Estimations of $\boldsymbol{\beta}$ coefficients of Cox model for each industry except Utilities, and for season_event covariate are presented in the Table 5.

Parameter	Mean estimate of $\boldsymbol{\beta}$	Standard Deviation	95% HPD Interval	
Agriculture	2.4279	0.4945	1.4263	3.3414
Consulting	-1.6205	0.8543	-3.3536	-0.0350
Entertainment	0.3380	0.5154	-0.6588	1.3387
Hospitality	0.7939	0.4190	-0.0275	1.6244
Retail	0.4811	0.3786	-0.2558	1.2183
season_event	0.2669	0.0699	0.1298	0.4010

Table 5. Mean estimate of $\boldsymbol{\beta}$ for Maryland

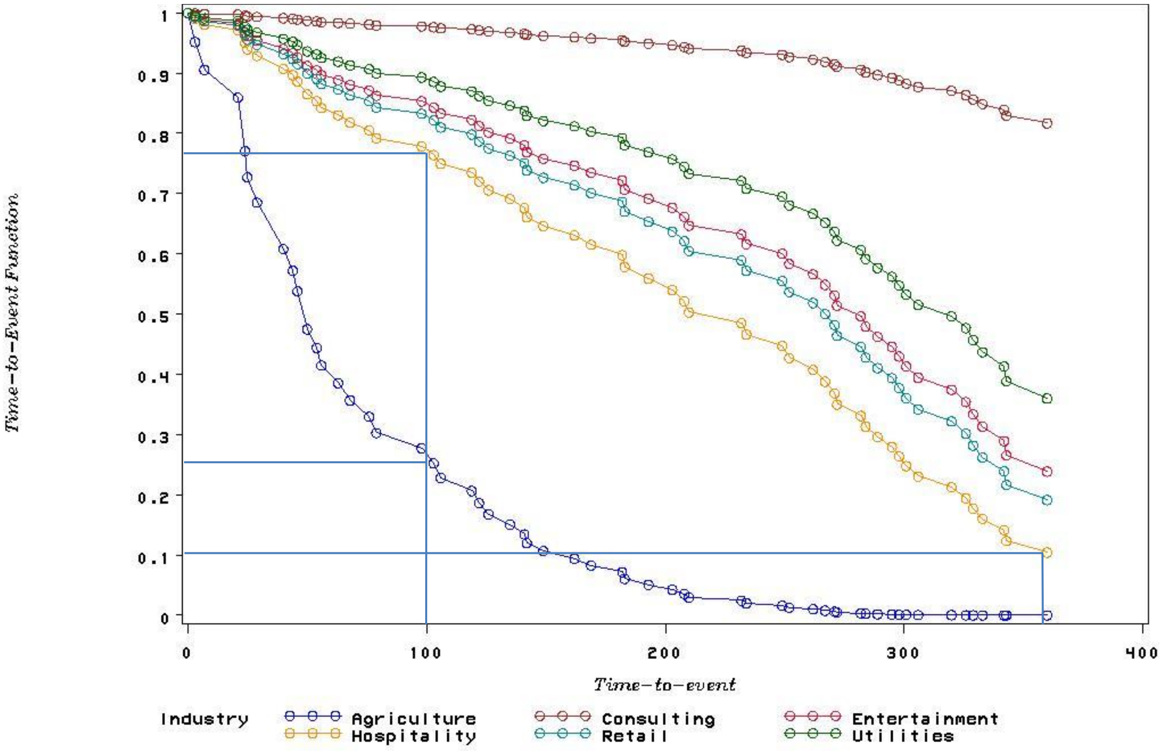

Figure 8. Time-to-Event Function for Claims per Industry in Maryland

The time-to-event function on Figure 8 demonstrates that in state of Maryland, Agriculture industry has only 25% chances that there will be no claims before 100th day of a policy, and practically 0% chances that there will be no claims at all for a one year policy.

For comparison, Hospitality has 76% chance that there will be no claims before 100th day of a policy, and 10% chances that there will be no claims at all for a one year policy.

Hazard function presented on Figure 9 shows that the hazard of claims grows through the term of policies and reaches its highest risk at the end of the term.

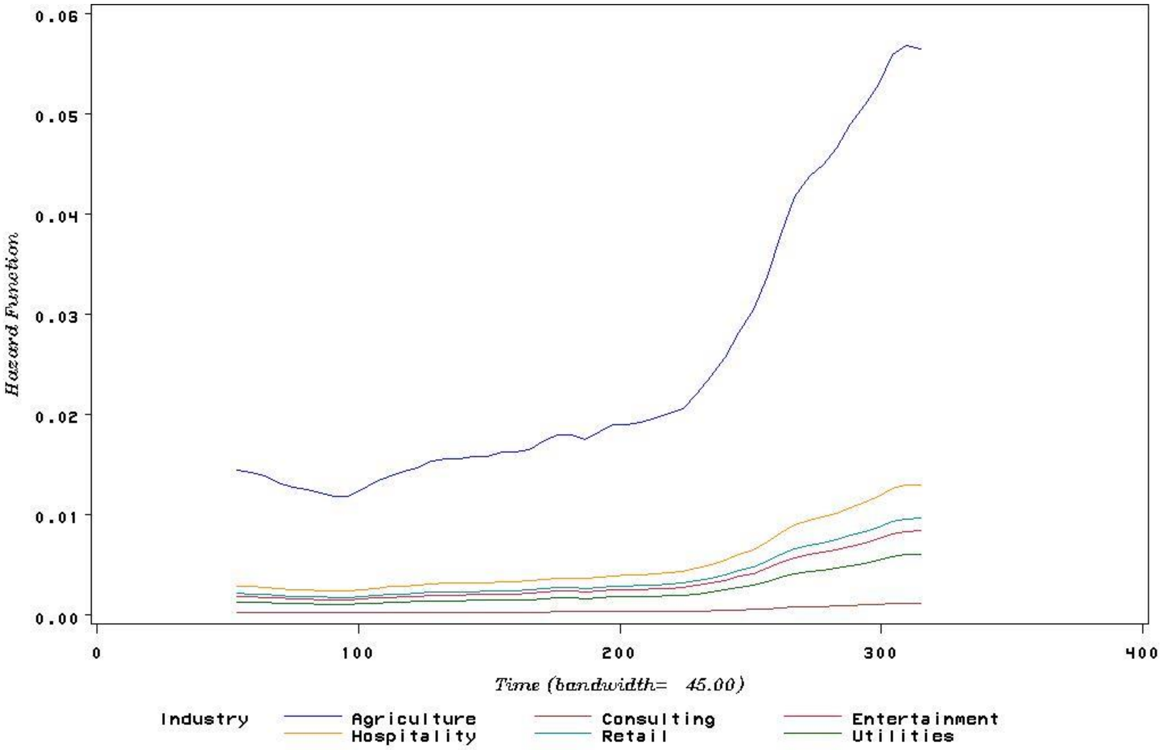

Figure 9. Hazard Function for Claims per Industry in Maryland

Time-dependent covariate is significant with $\boldsymbol{\beta}$=0.2669. This means that hazard ratio during winter season in Maryland is 31% higher, controlling for the other covariates ($\boldsymbol{exp(0.2669) - 1 \approx 0.31}$).

State of Oklahoma

For Oklahoma, data represents claims for 8 industries. Estimations of $\boldsymbol{\beta}$ coefficients of Cox model for each industry except Utilities, and for season_event covariate are presented in the Table 5.

Parameter	Mean estimate of $\boldsymbol{\beta}$	Standard Deviation	95% HPD Interval	
Agriculture	2.7484	0.8288	1.2316	4.4317
Consulting	1.5334	0.9108	-0.2242	3.2846
Energy	1.1123	0.8142	-0.3626	2.7749
Finance	2.4426	0.7848	0.9997	4.0137
Manufacturing	0.8842	0.8501	-0.7545	2.5391
Retail	1.4189	0.9627	-0.4599	3.3081
Transportation	1.3947	0.7950	-0.0437	3.0287
season_event	0.2608	0.0628	0.1384	0.3814

Table 6. Mean estimate of $\boldsymbol{\beta}$ for Oklahoma

According to the time-to-event function for Oklahoma presented on Figure 10, Agriculture industry has only 40% chances that there will be no claims before 100th day of a policy, and practically 0% chances that there will be no claims at all for a one year policy.

For comparison, Consulting has 76% chance that there will be no claims before 100th day of a policy, and 20% chances that there will be no claims at all for a one year policy.

Hazard function presented on Figure 7 shows that the hazard of claims increases around 60 days of the policy term, and then grows through the term of policies and reaches its highest risk approximately 30 days before the end of the term.

Time-dependent covariate is significant with $\boldsymbol{\beta}$=0.2608. This means that hazard ratio during winter season in Oklahoma is 30% higher, controlling for the other covariates ($\boldsymbol{exp(0.2608) - 1 \approx 0.30}$).

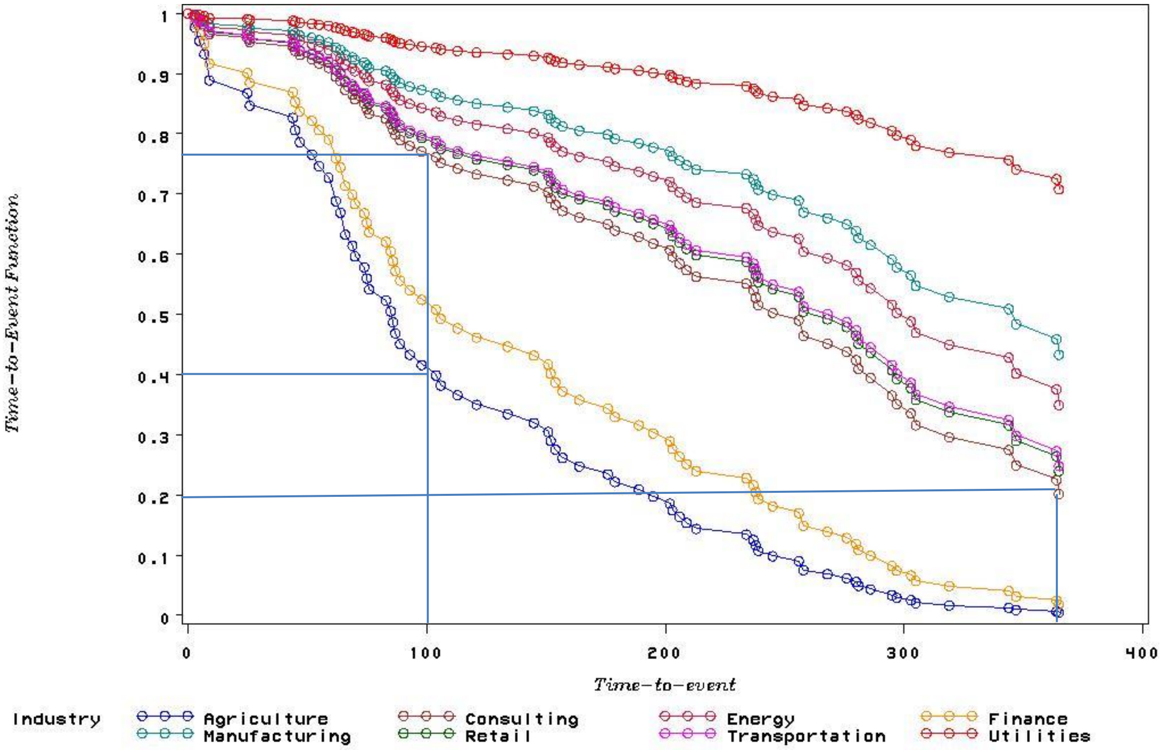

Figure 10. Time-to-Event Function for Claims per Industry in Oklahoma

CONCLUSION

An ultimate goal of insurance risk management is to create a profitable portfolio and to fit right price to right risk. This complex problem consists of multiple parts, including estimation of risk, estimation of price, monitoring of market changes, and more. In our article, we discussed one part of this complex problem – estimation of risk of workers' compensation claims for different industries and states. Our approach to estimate hazard function using Bayesian approach allowed estimating risk of claims per industry and state, as well as ranking industries by risk within states. As a next step to build profitable portfolio, the severity of claims should be included in the analysis, which eventually will allow re-evaluating premiums and insurance products to increase profitability of portfolios.

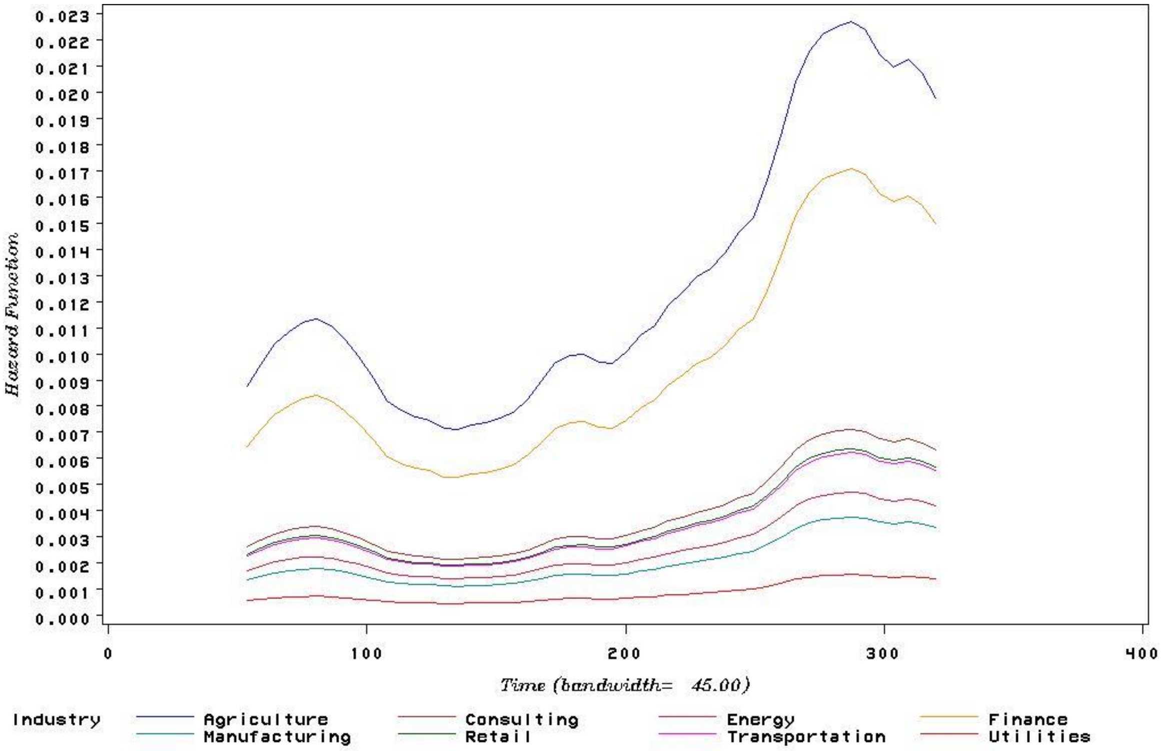

Figure 11. Hazard Function for Claims per Industry in Oklahoma

REFERENCES

Allison, P.D. 2012. *Survival Analysis Using SAS.* SAS Publication.

Arjas, E. 1988. “A Graphical Method for Assessing Goodness of Fit in Cox’s Proportional Hazards Model.” *American Statistical Association*, 83:204-212.

Breslow, N.E. 1974. “Covariance Analysis of Censored Survival Data.” *Biometrics*, 30:89-99.

Cox, D.R. 1972. “Regression Models and Life-Tables (with discussion).” *Journal of the Royal Statistical Society – Series B*, 34:187-220.

Gelfand, A.E., and Mallick, B.K. 1994. “Bayesian analysis of semiparametric proportional hazards models.” Technical Report No. 479, Department of Statistics, Stanford University.

Gill, R. and Schumacher, M. 1987. “A simple test of the proportional hazards assumption.” *Biometrika*, 74, 2:289-300.

Hosmer, D.W., and Lemeshow, S. 1999. *Regression Modeling of Time To Event Data*. New York: John Wiley & Sons, Inc.

Ibrahim JG, Chen MH, Sinha D. 2005. Bayesian survival analysis. Wiley Online Library. Available at onlinelibrary.wiley.com/

Kalbfleisch J.D. “Non-parametric Bayesian analysis of survival time data.” *Journal of the Royal Statistical Society. Series B (Methodological),* 1978:214–221.

Kaplan, E.L. and Meier, P. 1958. “Nonparametric estimation from incomplete observations.” *Journal of the American Statistical Association*, 53: 457-481.

Lee, E.T. 1992. *Statistical Methods for Survival Data Analysis,* 2nd Ed. Oklahoma City, John Wiley & Sons, Inc.

Rodriguez, G. 2007 “Lecture Notes on Generalized Linear Models.” Available at http://data.princeton.edu/wws509/notes/

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Samuel Berestizhevsky
InProfix Inc
samuelb@inprofix.com
inprofix.com

Tanya Kolosova
InProfix Inc
tanyak@inprofix.com
inprofix.com