Paper 2445-2018

The Cox Hazard Model for Claims Data

Tanya Kolosova, InProfix Inc.; Samuel Berestizhevsky, InProfix Inc.

ABSTRACT

Claim management requires applying statistical techniques in the analysis and interpretation of the claims data. The central piece of claim management is claims modeling and prediction. 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 an additional approach: time-to-event modeling. We provide a general framework to look into the process of modeling and prediction of claims using Cox hazard model. The Cox hazard 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 the Cox model. We use a Bayesian approach to survival analysis to deal with violations of assumptions of the Cox hazard model.

This article is a case study intended to indicate a possible application of the Cox hazard model to workers' compensation insurance, particularly occurrence of claims, while dealing with violation of assumptions of this model.

Keywords: claims modeling, claims prediction, insurance analytics, risk assessment, Cox model assumptions validation, time-to-event analysis, Cox hazard model, Bayesian approach, SAS.

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), insurance (workers compensation claims), 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, and time-to-event analysis – in insurance. Further, in the article, we use terms that are more often used in insurance.

A central quantity in survival (time-to-event) analysis is the hazard 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). 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 remains constant over time. Although the Cox hazard model is very popular in statistics, in practice data to be analyzed often fails to hold assumptions. For example, when a cause of claims interacts with time, the proportional hazard assumption fails. Or, when the hazard ratio changes over time, the proportional hazard assumption is violated. We present application of Bayesian approach to survival (time-to-event) analysis that allows dealing with violations of assumptions of Cox hazard model, thus assuring that model results can be trusted.

This article is a case study intended to indicate possible applications to workers' compensation insurance, particularly occurrence of claims. We study workers' compensation claims for the period of 2 years from November 01, 2014 till October 31, 2016. Claims data was provided by a worker compensation insurer that writes approximately $900 million of direct premium annually on a countrywide basis. The risk of occurrence of claims is studied, modeled and predicted for different industries within several USA states.

DATA

The present case study is based on the following policy and claims data:

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 focus our analysis on claims that led to payments.

THE COX MODEL FOR CLAIMS EVENTS ANALYSIS

Survival (or time-to-event) 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:

T – survival time of a randomly selected policy

t – a specific point in time

Hazard function $\mathrm{h(t)}$ describes instantaneous claims rate at time t:

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

In other words, hazard function $\boldsymbol{h(t)}$ at a time $\boldsymbol{t}$ specifies an instantaneous rate at which a claim happens, given that it haven't happened up to time $\boldsymbol{t}$. Hazard function is usually more informative about underlying mechanism of claims than survival function.

Cox (1972) proposed a model which doesn't require assumption that times of events 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 $\boldsymbol{t}$

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

Cox hazard model is also called Proportional Hazard Model if the hazard for any subject is a fixed hazard ratio (HR) relative to 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:

$$\boldsymbol{S_i(t) = S_0(t)^{exp \sum_{j=1}^{k} \beta_j x_{ij}}}$$

$$\boldsymbol{S_0(t) = \int_0^t h_0(u) du}$$

where:

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

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

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

$\boldsymbol{S_0(t)}$ – the baseline survival function, that is the survival function for the subject whose covariates $\boldsymbol{x_i, \dots, x_k}$ all have values of 0

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

APPLICATION OF THE COX MODEL FOR CLAIMS ANALYSIS

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

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

In order to build an appropriate model, we have 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 models that one can use to model repeated events in a time-to-event analysis. The choice depends on the data to be analyzed and the research questions to be answered.

A possible approach is to treat each claim as a distinct observation, but in this case we have to consider dependence of multiple claims that belong to the same policy. The dependence might arise from unobserved heterogeneity. Using some simple ad-hoc ways to detect dependence (Allison, 2012), we come to conclusion that the dependence among time-to-event intervals of claims that belong to the same policy is so small that it has negligible effect on the estimates of the model. Thus, we consider each claim as a single event, and can build models that do not 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 does not 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 I

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. a subject 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 dependency between consecutive claims.

The conditional model II

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 because it introduces dependency 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 the specific event. This model fits our application needs and is used for the analysis.

DATA TRANSFORMATION

We analyze workers compensation claims data for the 2 years period, so called observation period, from November 01, 2014 till October 31, 2016. Each claim is associated with an industry to which employer belongs, and with a state where the accident happened. For example, an employer that belongs to Entertainment industry with headquarters in NY state may have company offices in different other states, where accidents happen. To prepare this data for the marginal model, each claim event is considered as a separate process. The time to each event is calculated starting from the beginning of the observation period or from the beginning of the policy, whichever happens later. If there are no claim events for a policy during the observation period, the policy is censored at the end of the observation or at the end of the policy, whichever happens earlier. To note, a subject is said to be censored (Censor = 0) if a policy expired or canceled, or if a claim event didn't happen during the observation period.

Example of data prepared for marginal model is presented in the Figure 1:

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

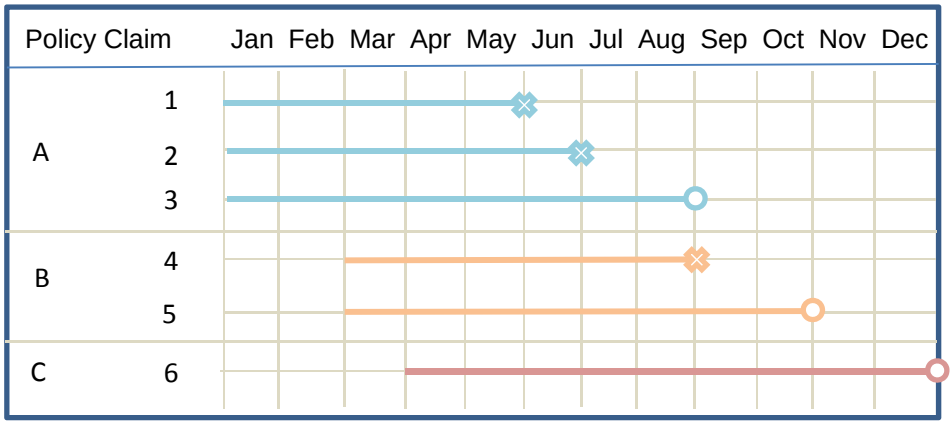

Figure 1. Claims Data Presentation

For this example, data is presented in the following way (Table 1):

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. Claims Data

In this article we present a case study of analysis and modeling performed on claims data for Illinois. Illinois data contained claims for Construction, Consulting, Entertainment, Finance, Hospitality, Manufacturing, Retail, and Utilities industries.

In our analysis we assume that each claim event is independent within the policy and the industry. For example, if two claims are covered by the same policy, we consider these claims independent. As well, if two claims are covered by different policies, we assume that the claims are independent and that the policies have no effect on risk. The data for each policy with multiple claim events is described as multiple claims, where each claim has an entry time at the beginning of the policy or beginning of the observation period – whichever is later.

COX MODEL ASSUMPTIONS VALIDATION

In most insurance risk papers, the authors take the proportional hazard 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. Note that, when used inappropriately, statistical models may give rise to misleading conclusions. Therefore, it's highly important to check underlying assumptions.

Perhaps the easiest and most commonly used graphical method for checking proportional hazard is so called 'log-negative-log' plot. For this method, one should plot $\boldsymbol{ln}(-\boldsymbol{ln}(\boldsymbol{S_i}(\boldsymbol{t}))$ vs. $\boldsymbol{ln}(\boldsymbol{t})$ and look for

parallelism – the constant distance between curves over time. This can be done only for categorical covariates. If the curves show non-parallel pattern, then the assumption of proportional hazard is violated, and, as a result, analytical estimation of $\boldsymbol{\beta}$ coefficients is incorrect.

For claims in 8 industries in Illinois, log-negative-log plot is presented on Figure 2. This plot shows that the proportional hazard model assumption does not hold: the lines of log-negative-log plot not parallel and intersect.

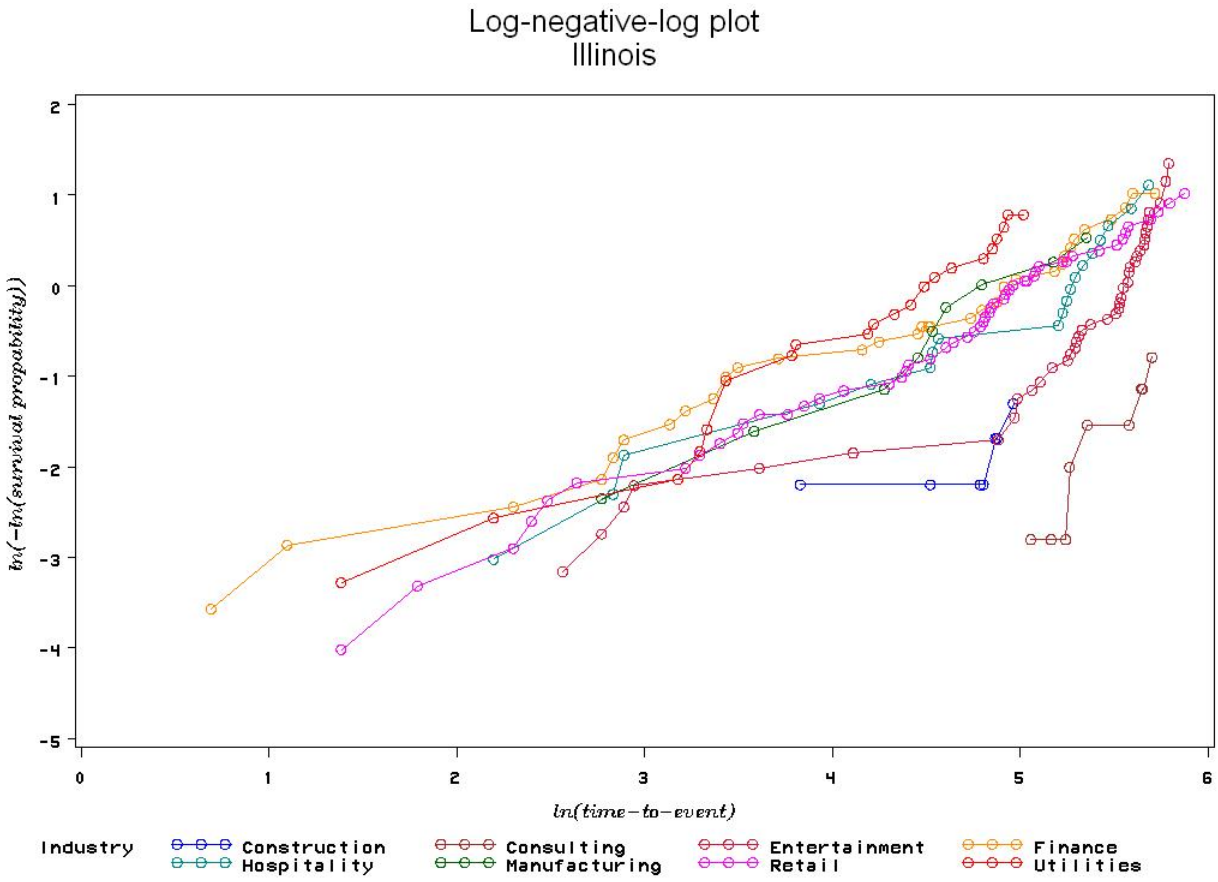

Figure 2. Log-Negative-Log Plot for Illinois

We use industry as a categorical covariate, assuming that time-to-event (survival) functions vary by industries. It is wrong to assume that there is no 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, thus violating assumptions of Cox model. In order to account for season dependency we introduce time-dependent covariate for winter season, and use 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 $\boldsymbol{t}$

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

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

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

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

Applying this approach to the case of Illinois, our model looks like:

$$\boldsymbol{h_i(t) = h_0(t) exp\left(\sum_{j=1}^{7} \beta_j x_j + \gamma \times season \times ln(t)\right)}$$

where:

$\boldsymbol{h_i(t)}$ – the hazard function for industry $\boldsymbol{i = 1, \ldots, 7}$ at time $\boldsymbol{t}$, where industries are: Construction, Consulting, Entertainment, Finance, Hospitality, Manufacturing, and Retail

$\boldsymbol{h_0(t)}$ – the baseline hazard function, in our case - the hazard function for one selected industry, by default the last alphabetically ordered industry – Utilities.

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

$$\boldsymbol{season} = \begin{cases} \boldsymbol{1}, if\ the\ claim\ happened\ during\ winter\ season\ (months\ 11, 12, 1, 2, 3) \\ \boldsymbol{0}, if\ the\ claim\ happened\ during\ non-winter\ season\ \ (months\ 4-10) \end{cases}$$

As there is no reason to prefer any specific industry for a baseline, we choose the last alphabetically ordered industry – Utilities. Selection of Utilities industry as a baseline for hazard means that hazards for all other industries are estimated relatively to that of Utilities industry.

Calculation of time-to-event (survival) functions when we have time-varying covariates becomes 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 $\boldsymbol{season = 0}$ for time-to-event $\boldsymbol{t \leq 214}$ days (from 1st of April till the 1st of November), while for time-to-event $\boldsymbol{t > 214}$ – using hazard corresponding to $\boldsymbol{n = 1}$. For another example, if a policy started on 1st of August, survival function should be calculated using hazard corresponding to $\boldsymbol{season = 0}$ for time-to-event $\boldsymbol{t \leq 92}$ days (from 1st of August till 1st of November), and $\boldsymbol{t > 244}$ days (from 1st of April till 31st of July), while for time-to-event $\boldsymbol{92 < t \leq 244}$ – using hazard corresponding to $\boldsymbol{season = 1}$.

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. For our model we develop an appropriate formula for calculation of $\boldsymbol{S_i(t)}$:

$$\boldsymbol{S_i(t) = \left(\frac{S_0(t|t < t_1) S_0(t|t > t_2)}{S_0(t|t_1 \leq t \leq t_2)}\right)^{exp\left(\sum_{j=1}^{k} \beta_j x_{ij}\right)} * exp\left(-exp\left(\sum_{j=1}^{k} \beta_j x_{ij}\right) * \int_{t_1}^{t|t_1 \leq t \leq t_2} h_0(u)\, u^{\gamma} du\right)}$$

where:

$\boldsymbol{t_1}$ – the start day of the winter season relatively to the beginning of the policy

$\boldsymbol{t_2}$ – the end day of the winter season relatively to the beginning of the policy

Yet additional challenge in our data is reliability of dates related to claims. There are 2 dates available – the date of the accident caused the claim, and the date when the claim was reported. Wide variability of time intervals between these 2 dates creates additional challenge in application of Cox Hazard model, as

time-to-event becomes essentially random variable. To address these problems, as well as assumptions violations, we use Bayesian non-parametric approach to estimate coefficients of the extended Cox hazard model.

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. In addition, 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. However, a known disadvantage of this method is that it is computation intensive.

DEPLOYMENT WITH SAS® SOFTWARE

To estimate coefficients of Cox hazard model, we use SAS® software, specifically PHREG procedure, which performs analysis of survival data. The estimation of the Cox hazard model using Bayesian approach by SAS PROC PHREG is implemented in the following way:

```
proc phreg data= CLAIMS_DATA_IL;
   class CLIENT_INDUSTRY;
   model  TIME_TO_EVENT*CENSOR(0) = CLIENT_INDUSTRY SEASON_EVENT;
   SEASON_EVENT = SEASON*log(TIME_TO_EVENT);
   bayes seed = 1 outpost = POST;
run;
```

CLAIMS_DATA_IL is a SAS data set that contains data for state of Illinois like industries, time intervals from the beginning of policies to date of claims, etc. Sample of rows from CLAIMS_DATA_IL is presented in

Table 2.

CLIENT_INDUSTRY	TIME_TO_EVENT	CENSOR	SEASON
…	…	…	…
Construction	119	0	0
Construction	162	0	0
Consulting	220	1	0
Retail	365	0	0
Retail	237	1	0
Transportation	95	1	0
Transportation	108	1	1
Utilities	7	1	0
…	…	…	…

Table 2. Selected Rows from CLAIMS_DATA_IL Data Set

CLIENT_INDUSTRY column contains names of industries to which claims are related. TIME_TO_EVENT column contains number of days to an event calculated starting from the beginning of the observation period or from the beginning of the policy, whichever happens later. CENSOR column indicates if the event is a claim (CENSOR=1), or if the event is end of a policy (CENSOR=0). SEASON column indicates if the event happened during winter season (SEASON=1) or not (SEASON=0).

There are 2 covariates in the model: CLIENT_INDUSTRY and SEASON_EVENT. CLIENT_INDUSTRY is a categorical variable, so it is defined as the covariate in CLASS statement and in MODEL statement. SEASON_EVENT is the time-dependent covariate that represents the following component of the model: $\boldsymbol{season} \times \boldsymbol{ln}(\boldsymbol{t})$. SEASON_EVENT is defined in MODEL statement and in the expression that follows the MODEL statement:

```
class CLIENT_INDUSTRY;
model TIME_TO_EVENT*CENSOR(0) = CLIENT_INDUSTRY SEASON_EVENT;
SEASON_EVENT = SEASON*log(TIME_TO_EVENT);
```

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

In the BAYES statement we specify a seed value as a constant to reproduce identical Markov chains for the same input data. We didn't specify prior distribution, thus applying uniform non-informative prior.

The described PHREG procedure produces estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ coefficients.

However, PROC PHREG does not produce baseline survival function $\boldsymbol{S_0(t)}$ when time-dependent covariate is defined. To calculate the baseline survival function, we use the following work around (Thomas, Reyes, 2014):

```
data DS;
   set CLAIMS_DATA_IL;
   SEASON_EVENT = SEASON*log(TIME_TO_EVENT);
run;

data INDUSTRY;
   CLIENT_INDUSTRY = "Utilities";
   SEASON_EVENT = 0;
run;

proc phreg data=DS;
   class CLIENT_INDUSTRY;
   model TIME_TO_EVENT *censor(0) = CLIENT_INDUSTRY SEASON_EVENT;
   bayes seed=1;
   baseline out = BASELINE survival = S covariates = INDUSTRY;
run;
```

This step produces baseline survival function $\boldsymbol{S_0(t)}$.

INTERPRETATION OF RESULTS

Estimations of $\boldsymbol{\beta}$ coefficients of the Cox model for each industry except Utilities are presented in Table 3. Because Utilities industry is used as a baseline for hazard, $\boldsymbol{\beta}$ coefficient for Utilities is equal 0 and is not presented in the table. Table 3 also contains $\boldsymbol{\gamma}$ coefficient for SEASON_EVENT covariate.

Industry	Mean estimate of β
Construction	-2.482
Consulting	-2.240
Entertainment	-0.483
Finance	-0.214

Industry	Mean estimate of β
Hospitality	-0.127
Manufacturing	-0.428
Retail	-0.503
SEASON_EVENT	0.281

Table 3. Estimations of the Model Coefficients

For the purposes to compare risk of claims for different industries, we build survival functions for each industry, and $\boldsymbol{season} = \mathbf{0}$ (Figure 3). According to the survival function for Utilities industry, for example, there is 58% chances that there will be no claims before 100th day of a policy, and there is 1% chances that there will be no claims at all for one year policy.

The survival functions allow to estimate and to compare risk of claims among industries. For example, for Entertainment industry there is 72% chances that there will be no claims before 100th day of a policy, and 6.5% chances that there will be no claims at all for a one year policy. In other words, Entertainment industry in Illinois presents 5.5% higher chances than Utilities industry to have no claims during a one year policy. Also, we can observe that Entertainment, Manufacturing, and Retail have very similar risks of claims in Illinois. In addition, there is strong evidence that Construction and Consulting industries have significantly lower risk than other industries.

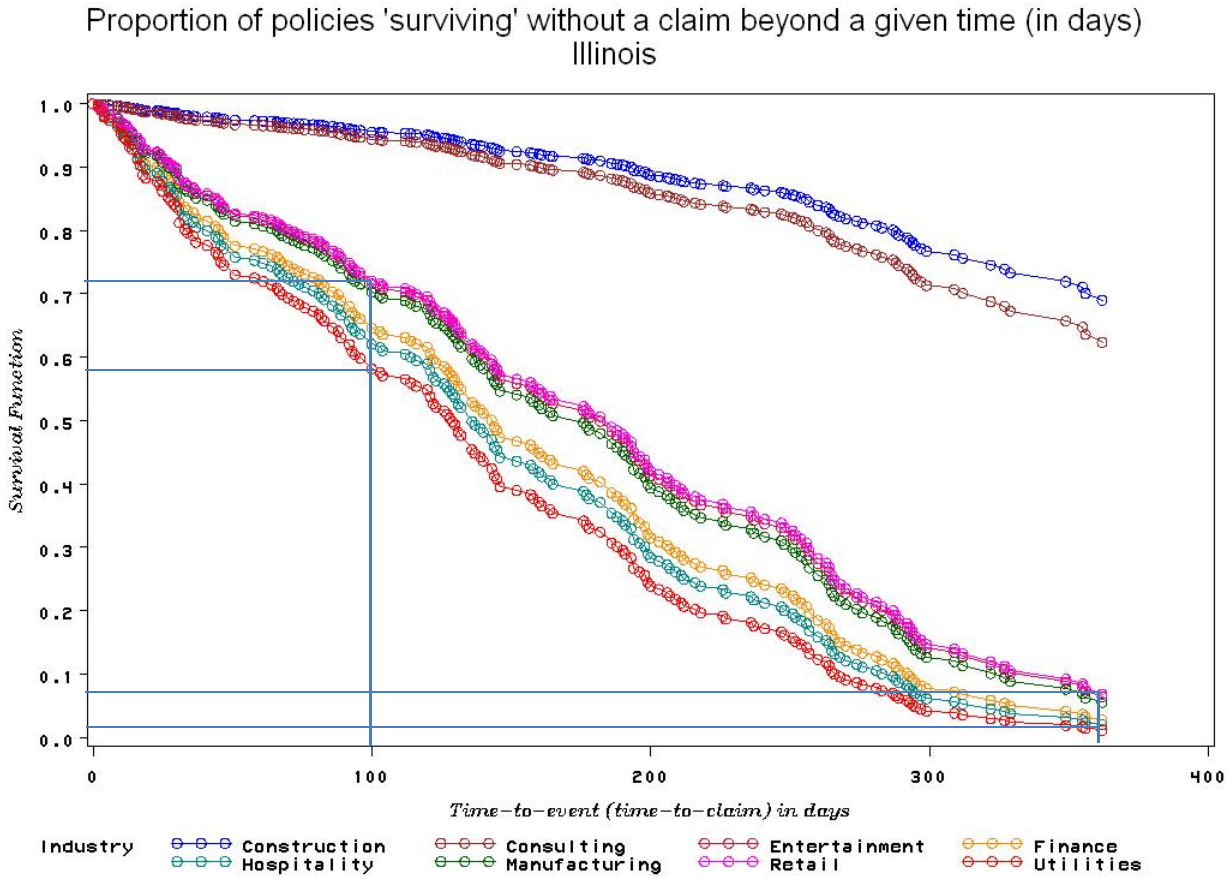

Figure 3. Survival Functions for Industries in Illinois

Hazard function presented on Figure 4 shows that the instantaneous claims rate continuously increases, achieving highest claims rate around 280th day of policy, and then slightly decreasing. We can also observe that Construction and Consulting industries have somewhat constant and relatively low claims rate through the duration of a policy. Hazard function on Figure 4 (as well as on Figure 6 and Figure 8) was produced with the SMOOTH SAS macro program (Allison, 2012).

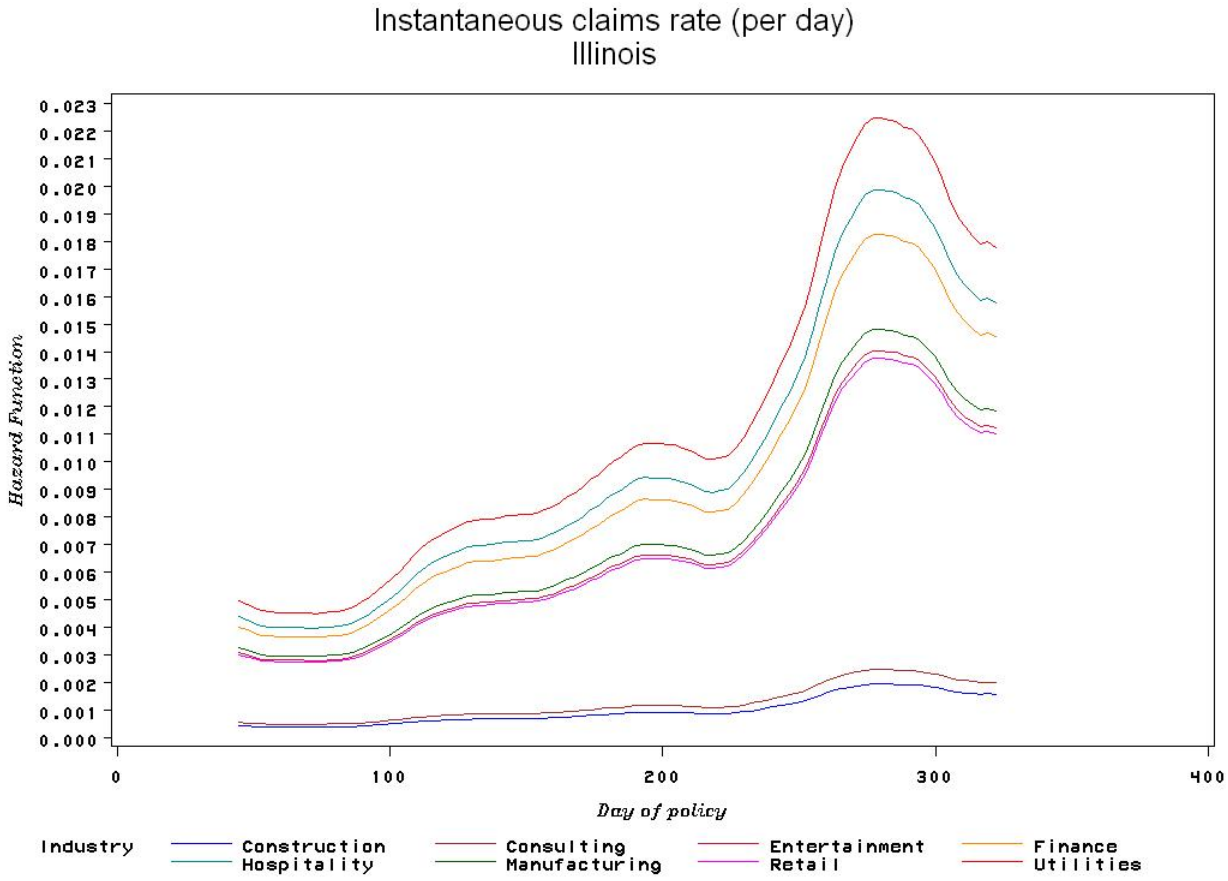

Figure 4. Hazard Functions for Industries in Illinois

Time-dependent covariate SEASON_EVENT is significant with $\boldsymbol{\gamma = 0.281}$. This means that hazard ratio during winter season in Illinois is 32% higher, controlling for the other covariates:

$$\boldsymbol{exp(0.281) - 1 \approx 0.32 = 32\%}$$

An estimation of survival (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 increase due to the winter season.

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, relatively to the start date of the policy, the winter season occurred. A proprietary computer program is developed by the authors to calculate $\boldsymbol{S_i(t)}$ for each industry with the time-dependent covariate.

Below we compare 2 examples mentioned earlier: the case when the policy started on 1st of April, and the case when the policy started on 1st of August.

If a policy started on 1st of April, then during time $\boldsymbol{t \leq 214}$ days (from 1st of April till the 1st of November) $\boldsymbol{season = 0}$. Then, for the duration of time $\boldsymbol{t > 214}$ days till the end of the policy, $\boldsymbol{season = 1}$. Thus, survival function is calculated using hazard corresponding to $\boldsymbol{season = 0}$ for time-to-event $\boldsymbol{t \leq 214}$ days,

and for time-to-event $t > 214$ – using hazard corresponding to $son = 1$. Survival function for this case is presented on Figure 5**.**

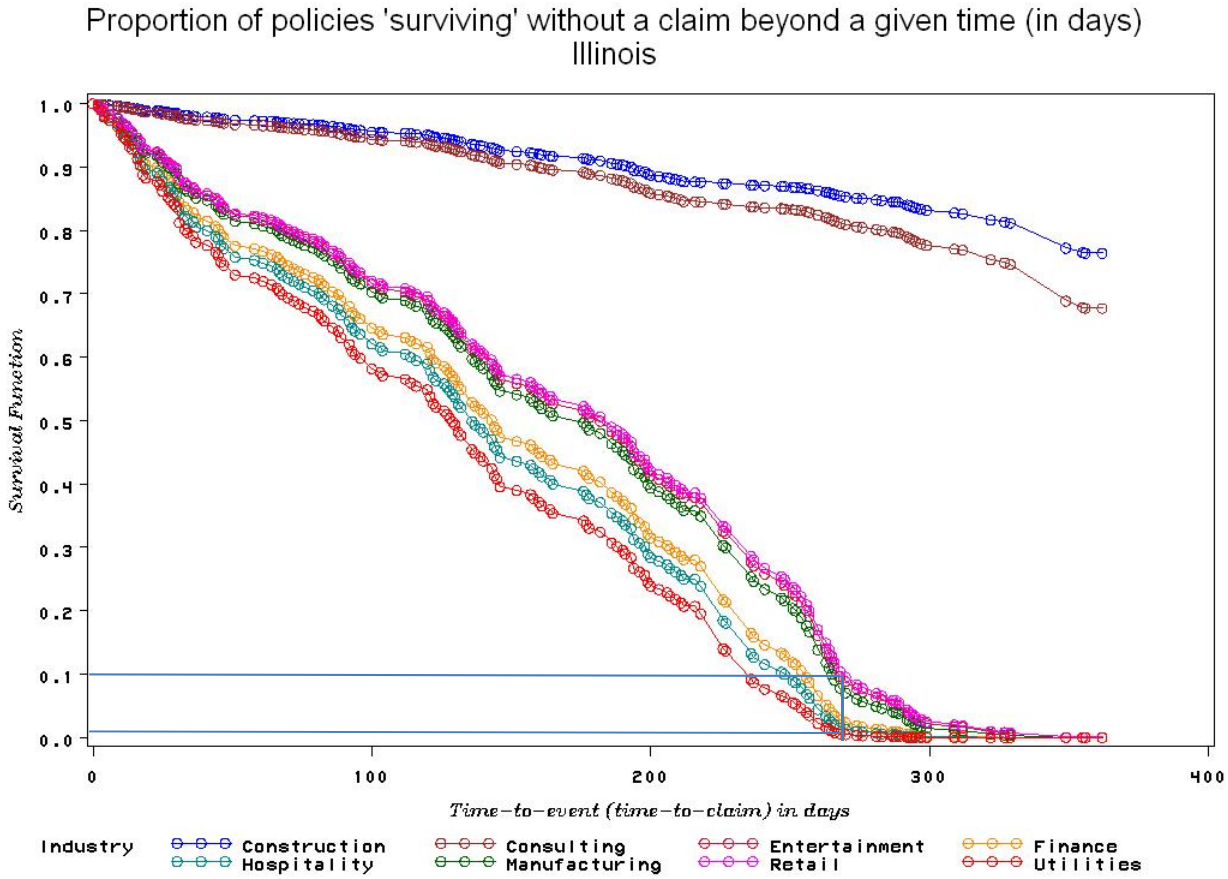

Figure 5. Survival Functions of Industries in Illinois for Policies Starting on April 1

In comparison with Figure 3 where winter season was not taken into consideration, we can see that the proportion of survival drops staring from 214th day of the policy.

Both Entertainment and Utilities industries have 0% chances that there will be no claims at all for a one year policy when we take the winter season into consideration.

In fact, for Utilities industry there is 0% chance that there will be no claims even before 270th day of a policy. However, the chances that Entertainment will "survive" without claims by 270th day are about 9%.

Hazard function presented on Figure 6 shows that the instantaneous hazard of claims sharply increases after $t > 214$, achieving highest claims rate around 280th day of a policy term.

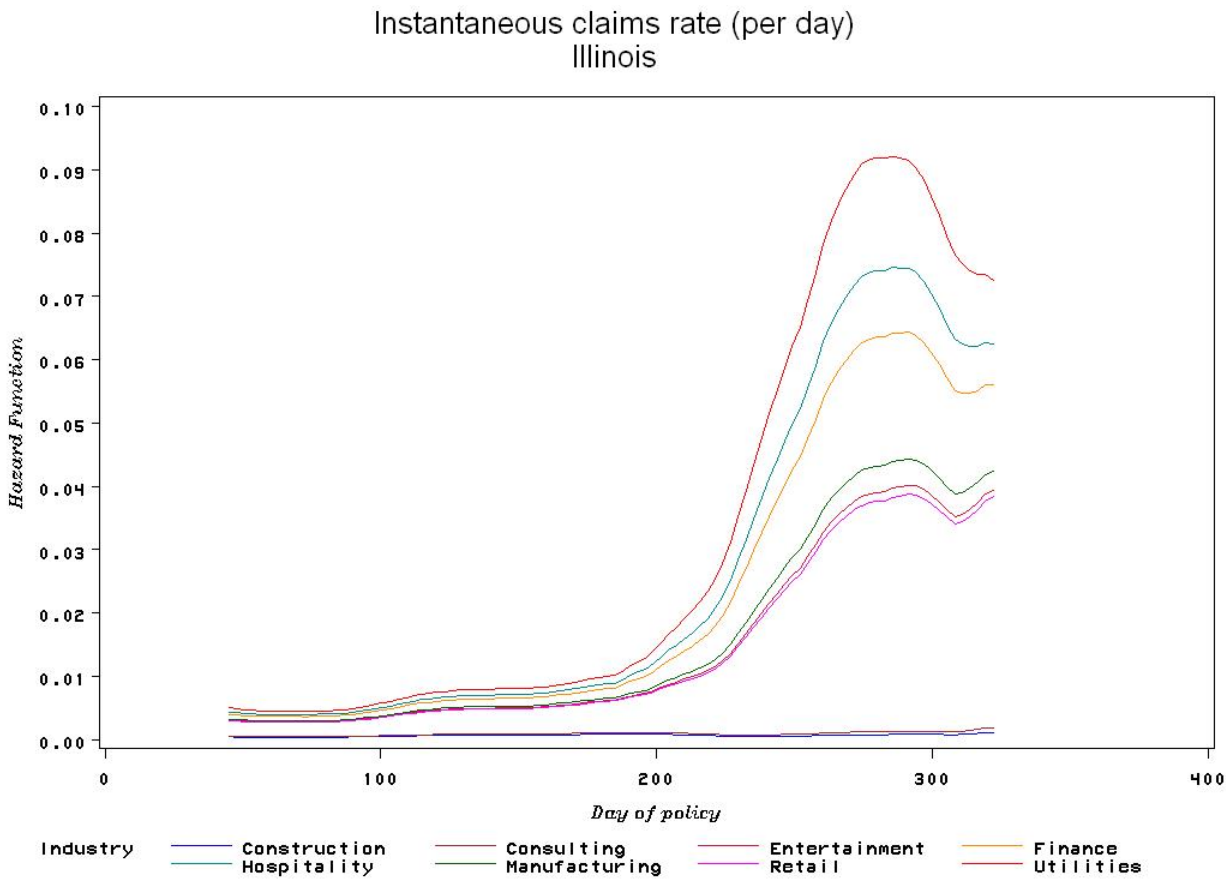

Figure 6. Hazard Functions for Industries in Illinois for policies starting on April 1

For the second example, if a policy started on 1st of August, then during time $\boldsymbol{t \leq 92}$ days (from 1st of August till the 1st of November) and $\boldsymbol{t > 244}$ days (from 1st of April till 31st of July), $\boldsymbol{season = 0}$. Then, for the duration of time $\boldsymbol{92 < t \leq 244}$ days of the policy, $\boldsymbol{season = 1}$. Thus, survival function is calculated using hazard corresponding to $\boldsymbol{season = 0}$ for time-to-event $\boldsymbol{t \leq 92}$ and $\boldsymbol{t > 244}$ days, and for time-to-event $\boldsymbol{92 < t \leq 244}$ – using hazard corresponding to $\boldsymbol{season = 1}$**.** Survival function for this case is presented on Figure 7**.**

In comparison with Figure 3 where winter season was not taken into consideration, we can see that the proportion of survival drops before 100th day of the policy.

For Utilities industry there is 44% chances that there will be no claims before 100th day of a policy accounting for winter season vs. 65% without accounting for winter season. After that the chances are dropping, and by 210th day of the policy there are 0% chances that there will be no claims in Utilities industry, accounting for winter season.

For Entertainment industry there are 66% chances that there will be no claims by the 100th day of a policy when we take winter season in consideration – vs. 72% otherwise. Also, there is about 1% chances that there will be no claims at all for a one year policy — in comparison with 6.5% chances when we don't take the winter season into consideration.

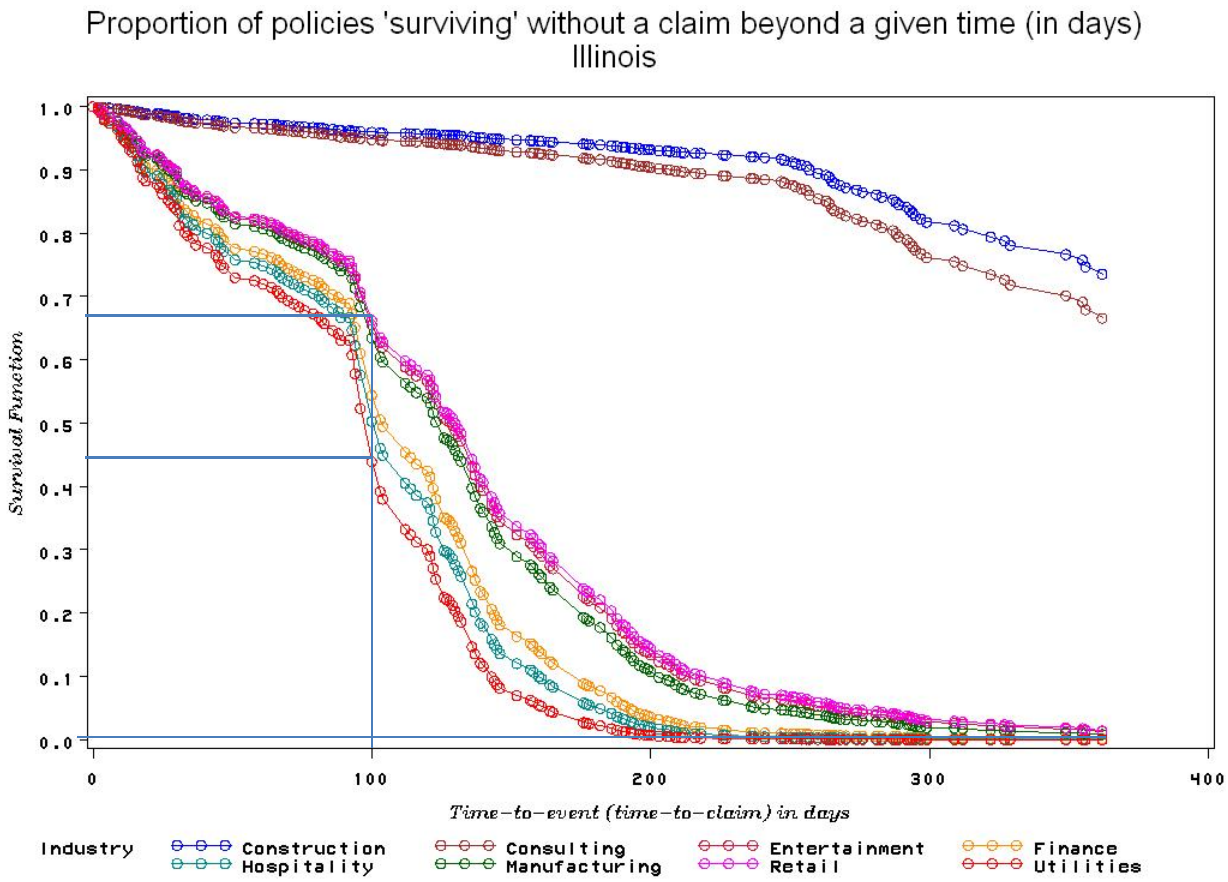

Figure 7. Survival Functions for Industries in Illinois for Policies Starting on August 1

The information revealed by the presented models can be used for purposes of underwriting and pricing, for development of new insurance products, as well as for marketing. For example, insurers can estimate risk of claims more accurately depending not only on industry, but also on the time period when the policy is started. Insurers can better manage anticipation of losses related to claims. In addition, insurers can develop new workers compensation products for the duration shorter than 1 year. In this case the insurance during periods of lower risk will have lower premium and therefore higher acceptance rate by customers. Referring to the example when a policy starts on 1st of April, 6-month policy would have significantly lower risk and will justify lower premiums. Marketing of such new products will attract companies seeking workers compensation.

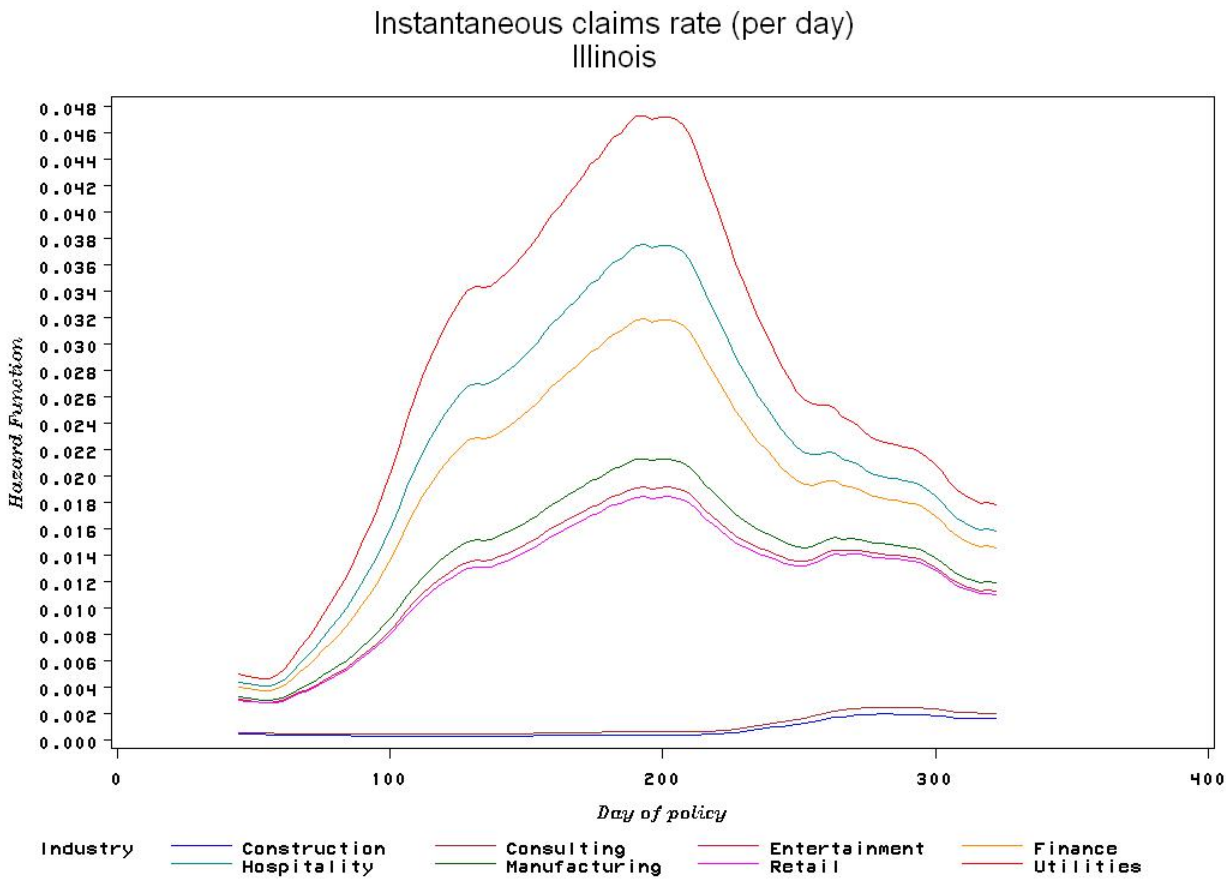

Figure 8. Hazard Functions for Industries in Illinois for Policies Staring on August 1

CONCLUSION

An ultimate goal of insurance risk assessment is to create a profitable portfolio and to fit right price to right risk. This complex problem comprises from 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 with season-dependent factor. Our method to estimate hazard function using Bayesian approach allows estimating risk of claims per industry and state, ranking industries by risk within states, as well as estimate risk depending on time-varying covariates like season. 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.

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

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

Hosmer, D.W., 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.

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

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

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

Thomas, L., Reyes, E.M. 2014. “Tutorial: Survival Estimation for Cox Regression Models with Time-Varying Coefficients Using SAS and R.” *Journal of Statistical Software,* 2014:61.

CONTACT INFORMATION

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

Tanya Kolosova
InProfix Inc.
425-241-6846
tanyak@inprofix.com

Samuel Berestizhevsky
InProfix Inc.
862-215-4611
samuelb@inprofix.com

www.inprofix.com