Estimating Mortality of Insured Advanced-age Population With Cox Regression Model

By Zhiwei Zhu, Ph.D.; Michael Hoag, FSA; Stéphane Julien, FSA; Sufang Cui, Ph.D. [1]

Risk Management, Transamerica Reinsurance

ABSTRACT

As with other populations, the survival curve of advanced-age populations is influenced by a combination of many risk factors such as age, gender, smoking status, etc. Survival analysis provides an array of statistical modeling tools for estimating the blended impact from multiple factors on the underlying force of mortality. This is a report of a study in which we applied the Cox proportional hazard model, a widely used survival analysis tool, to profiling the mortality of an insured advanced-age population.

Transamerica Reinsurance (TARe) used data from 14 companies in this analysis. This data represents policies issued between 01/01/1997 and 12/31/2000 to policyholders of age 60 or older. It is reasonable to conclude that our data is a good sample of the US insured advanced-age population. The major study results include measures of mortality variations by selected risk factors, estimated mortality experience of the underlying population, and a comparison of mortality experience estimated by direct calculation, the Society of Actuaries' (SOA) 1990-95mortality study, and our extended Cox modeling technique.

[1] All of the authors were employees of Transamerica Reinsurance at the time the paper was written.

1. Introduction

Mortality is the most important risk of the life insurance industry. As the population in the U.S. ages, both the government and the insurance industry recognize the importance of a better understanding of mortality in advanced-age populations. In order to design products that meet the financial needs of customers in different market segments, insurance companies need to understand the mortality of individuals with similar risk profiles. This is most important in creating financially sound social and private insurance programs for the elderly.

Regularly, the SOA compiles mortality experience reports (see [1]) based on policy data voluntarily contributed by insurers. At Transamerica Reinsurance (TARe), mortality has also been studied using similar life insurance policy data. When such detailed data is collected and used for studies, researchers usually face such challenges as:

a) Massive amount of data to analyze and process;
b) Lack of a common standard across various organizations;
c) Insufficient death claims for estimating mortality of certain market segments (e.g., advanced-age segment);
d) Significant data preparation requirements for utilizing advanced statistical modeling methodologies.

In this paper, we focus on presenting one of the solutions--applying Cox regression modeling technique to analyze mortality of the advanced-age insured population--to the last two challenges mentioned in the previous paragraph. By utilizing a standard statistical software package, SAS, not only are we able to estimate mortality experience, but we are also able to quantify the degree

of mortality variations based on policyholders' risk factors such as gender, issue age, smoking status and product type.

The following data elements were also used in our study:

- Issue age
- Duration
- Gender
- Product type
- Smoking status
- Number of Claims
- Issue Date
- Termination Date

In the remaining four sections of this paper, we will present:

1. Data description – providing more details regarding the data sources and definitions;
2. Model and methodology description – presenting how and why the specific statistical model was selected;
3. Result interpretation –quantifying mortality variation by multivariate risk factors, extending the Cox model to generate experience mortality, and comparing these findings to direct experience and to SOA's experience studies.
4. Discussion – commenting on the strengths and weaknesses of this study.

1. Data Description

This study used a total of 66,989 policies that were issued from 1997 to 2000 by 14 US insurance companies to policyholders of age 60 or older. Among these policies, 424 filed for death benefit claims. Though the amount of data to process is overwhelming, the selected policies and claims

were concentrated between issue ages 60 and 70. In other words, very few death observations are available from policies that were issued to people of age over 70, which is one complication in the estimation of advanced-age mortality. (The scarcity of observed deaths in ages above 70 makes it difficult to compute a stable estimate of advanced-age mortality.)

Consider each policy a case study and each claim an event. Our goal, then, is to define the risk of death and to formulate the relationship between this risk and the observable policy characteristics. In terms of survival analysis, this risk is called the *hazard rate,* and the observable characteristics are called *risk factors*.

The proportional hazard model proposed by Cox (see [2]) is an ideal tool for formulating the relationship between event risks and their associated factors. When the underlying data is properly formatted, the estimated parameters, or coefficients, of the model provide intuitive measurements of risk variations for a given factor. Therefore, reformatting the data to achieve more informative results is one of the key steps in our modeling effort. The next two sections present more theoretical and technical explanations on this issue.

Some of the data elements selected for our study have categorical values (e.g. gender, smoking status, etc.) and others are continuous (issue age, etc.). After testing various model configurations, it has been determined that, in order to achieve the best modeling results possible, we need to convert all continuous variables into categorical variables, and reduce the number of value categories for some of the variables. This resulted in the following categories for our preliminary data preparation.

Issue Age Category: 60-65, 66-70, 71-75, 76-80, 81-85, and 86 +

Gender: M: Males

F: Females

Smoking Status: Non-Smoker (NS)

Smoker (SM)

Unknown (UNK)

Product Type: Term

Whole Life

Universal Life

Other

Table I gives a summary of the data.

Table I: Inforce and Claims Summary

Variables	Description	Inforce Policies		Claims	Claims
		Number	Percent	Number	Percent
Issue Age	60-65	46892	70.00%	258	60.85%
	66-70	14219	21.23%	99	23.35%
	71-75	4823	7.20%	57	13.44%
	76-80	853	1.27%	6	1.42%
	81-85	198	0.30%	4	0.94%
	86+	4	0.01%	0	0.00%
Gender	F	17236	25.73%	88	20.75%
	M	49753	74.27%	336	79.25%
Smk Status	NS	53912	80.48%	337	79.48%
	SM	4094	6.11%	43	7.31%
	UNK	8983	13.41%	44	10.38%
Product Type	Term	19079	28.48%	201	47.41%
	Whole Life	4277	6.38%	19	4.48%
	U.L.	3905	5.83%	46	10.85%
	Other	39728	59.31%	158	37.26%
Total		**66989**	**100.00%**	**424**	**100.00%**

2. Model and Methodology Description

The Cox proportional hazards model evaluates risk factors to determine the magnitude and significance of their effects on survival or failure time. Each risk factor is incorporated in the model as a variable, and its value, or factor level, is used to describe the category to which each individual belongs.

Let *T* be the number of months from a policy's issue date to the date of the policyholder's death. The probability of an individual surviving beyond a given time *t* is called a *survival function,* which is given by:

$$S(t) = P(T > t) \tag{2.1}$$

Similarly, given the vector of risk factors Z, the conditional probability that a policyholder will die between in force months t_1 and t_2 can be written as

$$P(t_1 \le T < t_2 \mid T > t_1, Z) = \frac{P(T \ge t_1 \mid Z) - P(T \ge t_2 \mid Z)}{P(T > t_1 \mid Z)} = \frac{S(t_1 \mid Z) - S(t_2 \mid Z)}{S(t_1 \mid Z)}$$

If we denote $Q_{x|Z}$ as the mortality rate for duration X, given a risk factors Z, then:

$$\begin{aligned} Q_{1|Z} &= P(0 \le T < 12 \mid T > 0, Z) \\ Q_{2|Z} &= P(12 \le T < 24 \mid T > 12, Z) \end{aligned} \tag{2.2}$$

Another fundamental quantity in survival analysis is the *hazard function* (known as the force of mortality in demography), which measures the risk at time *t* of an individual who has survived up to time *t*, and is defined as:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t \mid T \ge t)}{\Delta t}$$

The risk that an individual dies in a period *[0, t]*, denoted by $\Lambda(t)$, is the cumulative hazard in this period, and can be expressed as:

$$\Lambda(t) = \int_0^t \lambda(u) du$$

Similarly, we can derive mortality for any given period. For example, the mortality during the second policy year will be:

$$\Lambda(24) - \Lambda(12) = \int_{12}^{24} \lambda(u)du$$

If T is a continuous random variable, the survival function $S(t)$, and the cumulative hazard function $\Lambda(t)$ are related by:

$$S(t) = \exp[-\Lambda(t)].$$

Therefore, given $S(t)$, one can derive $\Lambda(t)$, and vice versa. This property is useful later in our calculation of mortality.

Consider a policy data file with n observations. The information available from the observations can be described by the triples (T_j,δ_j, $\mathbf{Z}_j$), j=1,…, n, where T_j is the time of study for the *j*th policy, δ_j is the death indicator, and $\mathbf{Z}_j = (Z_{j1}, \ldots, Z_{jp})$ is the vector of risk factors for the *j*th policy. If a policy has lapsed, or is in force beyond the study period, it is called a censored case. In our study δ_j has a value of 1 if a death is observed, and 0 otherwise.

Let $\lambda(t \mid Z)$ be the hazard rate at time *t* for an individual with risk factors $\mathbf{Z}$. The basic Cox regression model is as follows:

$$\lambda(t \mid Z) = \lambda_0(t)e^{\beta' Z} = \lambda_0(t)e^{\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \ldots \beta_p Z_p}.$$

where $\lambda_0(t)$ is an arbitrary baseline hazard function, and $\beta' = (\beta_1, \ldots, \beta_p)'$ is the coefficient vector to be estimated. The Cox model is often called a proportional hazard model because, given two individuals with risk-factor values Z_1 and Z_2, the ratio of their hazard rates is:

$$\frac{\lambda(t \mid Z_1)}{\lambda(t \mid Z_2)} = \frac{\lambda_0(t)\exp(\beta^t Z_1)}{\lambda_0(t)\exp(\beta^t Z_2)} = \exp\{\beta^t Z_1 - \beta^t Z_2\}$$

This ratio is constant over time and is independent from *t;* hence, the hazard rates of any two cases are proportional. Their constant hazard ratio measures the relative risk of an individual with risk factor Z_1 comparing to the other individual with risk factor Z_2.

If the data is properly coded, the ratio of hazard rates will be a function of the coefficient vector. Consider a model in which Z consists of only one risk factor, 'Gender'. If we assign a value of 1 for males and 0 for females, then the ratio of hazard rates will be:

$$\frac{\lambda(t \mid Z_1 = 1)}{\lambda(t \mid Z_2 = 0)} = \frac{\lambda_0(t)\exp(\beta\ )}{\lambda_0(t)} = \exp\{\beta\}$$

If β is estimated as 0.26,

$$\frac{\lambda(t \mid Z_1 = 1)}{\lambda(t \mid Z_2 = 0)} = \exp\{\beta\} = e^{0.26} = 1.3$$

which means that the probability of dying for males is 30% higher than the one for females.

In general, if a variable has *K* categories, this variable can always be represented by a set of *(K-1)* indicator variables. For example, in our data, **Smoking Status** has three categories. We can create two (3-1=2) indicator variables, (Smk2, Smk3), to represent the three categories of the variable **Smoking Status** as follows:

(Smk2, Smk3) = (0,0) if Smoking Status = NS

(Smk2, Smk3) = (1,0) if Smoking Status = SM

(Smk2, Smk3) = (0,1) if Smoking Status = UNK

This form of coding is known as indicator coding. When incorporating these indicators into a Cox model, we let each set of indicator variables represent one category of the underlying risk factor. The set represented by zeros is the reference category, the category to which all others will be compared. In the above example, if we replace the risk factor **Smoking Status** with Smk2, and Smk3 in a Cox model, the category Smoking Status = NS, or (Smk2, Smk3)=(0,0) will serve as the reference category. The exponential of the regression coefficients of each of the indicator variables gives the relative risk of the underlying category to that of the reference category. See [3] for a more detailed description of this coding method.

3. Outcome interpretation

By utilizing the SAS statistical computer software, re-coding data as indicators, and applying the Cox model as described in previous sections, our study revealed some valuable and interesting results. The modeling results are presented in the next two subsections.

3.1 Relative Risks

First, we interpret the analysis results directly from the PROC PHREG procedure of SAS, which measures the relative risks.

Figure I: COX Model Output

1. Summary of the Number of Events and Censored Values

Total	Event	Censored	% Censored
66989	424	66565	99.37%

2. Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	123.1884	11	<.0001
Score	140.2791	11	<.0001
Wald	132.6577	11	<.0001

3. Analysis of Maximum Likelihood Estimates

	Parameter	Standard		Hazard

Variable	Estimate	Error	Chi-Square	Pr > ChiSq	Ratio
ISSUE AGE					
60-65					
66-70	0.33275	0.11925	7.7860	0.00523	1.395
71-75	0.99149	0.15228	42.3933	<.0001	2.695
76-80	0.52654	0.42681	1.5219	0.2173	1.693
81-85	1.21429	0.52264	5.3980	0.0202	3.368
86+	-5.18688	214.88307	0.0006	0.9807	0.006
SMOKING STATUS					
NS					
SM	0.46771	0.16781	7.7680	0.0053	1.596
UNK	0.20744	0.16492	1.5821	0.2085	1.231
GENDER					
M					
F	-0.35679	0.12144	8.6323	0.0033	0.700
PRODUCT TYPE					
TERM					
WHOLE LIFE	-1.24342	0.25056	24.6270	<.0001	0.288
U.L.	0.29987	0.17745	2.8556	0.0911	1.350
OTHER	-0.69471	0.11161	38.7449	<.0001	0.499

1. **Summary of the Number of Events and Censored Values**

The study file has a total of 66,989 policies and 424 claims. Except for the policies with claims, the rest (99%) are censored (lapsed, converted, or still in force on the study-ending date).

2. Testing the Global Null Hypothesis: BETA=0 (Model Fit)

The null hypothesis is that none of the risk factors has a statistically significant influence on the hazard rate. Three statistical testing procedures, Likelihood, Score, and Wald, were applied. The results showed that there is a less than 0.01% chance (values in the Pr > ChiSq column) that one will make a faulty rejection of the null hypothesis. So, the data strongly suggests that the hazard rate is dependent on the selected risk factors.

3. Analysis of the Maximum Likelihood Estimates

In the third block of Figure 1, the **'Variable'** column displays the *Zi*'s that are included in the Cox model. In this column, the rows with a variable name and a category value represent the reference categories of the corresponding risk factor, and are not explicitly included in the model as a *Zi.* The **'Parameter Estimate'** column contains the estimated coefficients *bi.* The **'Pr >**

ChiSq' column shows the *p-value* for rejecting the null hypothesis $b_i = 0$; the corresponding risk factor category has the same hazard rate as the reference category. Finally, the **'Hazard Ratio'** column provides the exponential of b_i.

According to the explanations in Section 2, each of these numbers has a straightforward meaning. For example, the numbers in the I.3. block can be translated as follows:

*In this model, the risk factor Issue Age is broken down into 6 categories (levels). The first category, **'60-65'**, is chosen to be the reference category.*

*Statement 1: For I.3. 66-70, the **hazard ratio** = 1.395, which implies that, when all the other conditions are equal, policies issued between ages 66 and 70 are 39.5% (1.395-1) more likely to have death claims than those issued between ages 60 and 65. The fact that **Pr > ChiSq** = 0.0052 implies that there is a 0.52% chance for this statement to be faulty, due to unexpected random effects included in the study data.*

*Statement 2: For I.3. 86+, the **hazard ratio** = 0.006, which indicates that, when all the other conditions are equal, policies issued to policyholders aged 86 or above are 99.4% (0.006-1) less likely to have claims than those issued between 60 and 65. However, there is a 98.07% chance (**Pr > ChiSq** is 0.9807) for this statement to be wrong due to unexpected random effects included in the study data.*

We should consider the Statement 2 a faulty statement and reject it. By revisiting Table I in Section 1, it is apparent that the cause of such a faulty statement is that only 4 policies were issued to policyholders aged 86 or above; furthermore, none of them had had a claim by the end of this study. With so few policies, one can hardly make a reasonable mortality projection. From a modeling aspect, this suggests that the two Issue Age categories 81-85 and 86+ should be combined as a single one.

We can similarly interpret the *p-values* and the **hazard ratios** of the other risk factors from Figure I. When all the other conditions are equal:

- Smokers are 59.6% more likely to have claims than Nonsmokers are (**Hazard Ratio** = 1.596 and the p-value <0.05). However, because their p-values are greater than 0.05, the hazard rates of policyholders with an Unknown Smoking Status have no statistically significant difference from those of Nonsmokes. Female policyholders are 30% less likely to file claims than Male policyholders.
- Whole Life policies are 71.2% less likely to have claims than Term policies. Statistically, Universal Life policies are as likely to have claims as Term policies (p-value > 0.05).

3.2 Estimated Mortality Experience

As discussed in Section 2, the conditional survival functions can be estimated from the Cox model. Therefore, the mortality in a given durationcan be calculated based on formula (2.2).

Table II: Mortality Comparison

Issue Age	Gender	UWClass	Duration	Mortality Rates		
				Actual	Model	SOA 9095
60-70	F	A	1	1.4525	1.9986	1.8480
60-70	F	A	2	3.2262	3.1759	2.9292
60-70	F	A	3	3.5348	3.5089	3.8449
60-70	F	A	4	2.5290	4.6952	4.8072
71-80	F	A	1	4.4484	4.3163	4.9874
71-80	F	A	2	4.8023	6.8540	6.6535
71-80	F	A	3	17.5023	7.5713	9.2040
71-80	F	A	4	9.0158	10.1241	11.9136
81-100	F	A	1	0.0000	8.2352	13.9126
81-100	F	A	2	54.7945	13.0618	20.5672
81-100	F	A	3	82.1918	14.4239	28.3184
81-100	F	A	4	0.0000	19.2647	28.8600
60-70	M	A	1	2.6133	2.5897	2.6486
60-70	M	A	2	4.1549	4.1144	4.2605
60-70	M	A	3	4.1273	4.5456	5.8663
60-70	M	A	4	4.9601	6.0814	7.5835
71-80	M	A	1	6.2779	5.5909	8.8338
71-80	M	A	2	5.6752	8.8747	12.7019
71-80	M	A	3	7.3148	9.8023	16.8462

71-80	M	A	4	16.7209	13.1024	20.9671
81-100	M	A	1	0.0000	10.6608	21.4982
81-100	M	A	2	0.0000	16.8968	34.5062
81-100	M	A	3	0.0000	18.6551	49.4707
81-100	M	A	4	0.0000	24.8979	65.4475

The last three columns of Table II present different ways of obtaining mortality rates. The column labeled **Actual** gives mortality rates derived directly from the actual claims and exposure. The column labeled **Model** consists of mortality rates estimated by the Cox model. The **SOA 9095** column displays the expected mortality rates based on the SOA 9095 mortality experience tables, which, as industry averages, can be used as benchmarks for assessing TARe's own mortality.

As expected, the fluctuation of a product mortality experience depends heavily on the size and the life cycle of that product. Therefore, the experience in any short time period may not accurately represent the product's lifetime performance. For example, in comparing the **Actual** mortality to the **SOA9095** mortality in Tables II, one can see that female policyholders between Issue Ages 81 and 100 included in TARe's study experienced significantly higher mortality than the industry average. This result holds for the policies included in this study period. This is mainly due to the fact that a very small number of policies were issued to this age group in the study sample, and any claim will have a more significant effect on the mortality experience. On the other hand, the **Model** mortality derived by the Cox model suggests that TARe will have better mortality experience than the industry average, assuming the same distribution of policies within that age group.

When no claims are recorded, the calculation of actual mortality rates provides no information for estimating future mortality (See Males, Issue Ages 81-100, in Table II). However, the Cox

model will still project estimation based on experience of other age groups. As more data becomes available, the estimation becomes more accurate.

4. **Discussion**

When conducting a study that relies on large amount of data gathered from multiple sources, the accuracy of the results is assured and/or constrained by:

A. The merit of the study design, and the analytical techniques used.

B. The integrity of data processing and management.

C. The quality of the source data.

A. As explained earlier, the Cox model is a proven and widely used analytical tool for analyzing event-occurrence data. It has many advantages for analyzing life insurance data, such as:

   - Achieving the most efficient use of incomplete/censored data;
   - Formulating the relationship between risk (hazard rate) and risk factors;
   - Quantifying risk variations by risk factors;
   - Generating smoothed mortality/survival predictions.

The limitation of the Cox model is that only mortality tables by count can be generated through the estimation process. In practice, insureds will have policies with different face amounts, so their claims will weight differently on the overall mortality.

B. Pooling millions of records into one data system, with millions more flooding in every quarter, presents a great challenge to an IT department and its analysts. To turn this massive amount of data into useful information that can be used in taking business decisions is truly a joint venture of data processing and data analysis. The success of this venture relies on dedicated administrative support, strong IT operation, and knowledgeable business guidance.

For this very reason, conducting high quality mortality studies requires coordinating multi-disciplinary professionals in processing, managing, and analyzing data, as well as adapting highly sophisticated analytical tools, such as SAS, and SPSS.

C. The fact that each organization has its own preferences in collecting, managing, and transmitting data, necessitates its unification before any analysis can be performed. For example, it takes tremendous efforts and resources to group the hundreds of underwriting classes and tens of thousands of plan codes defined by various companies into a small enough number of categories in order for analytical tools, such as SAS, and modeling techniques, such as Cox, to yield meaningful outcomes. The data and mappings used for purposes of this study have been reviewed for reasonableness; however, results may vary based upon company methodology used for mapping data into the specified categories. In assessing business risks, strengthening industry-wide collaboration on data collection and standardization is as important as mortality studies themselves.

References:

[1] Society of Actuaries Reinsurance Council, Reinsurance Advanced Age Mortality Study, 1998

[2] D.R. Cox, Regression models and life-tables (with discussion). J. Royal Stat. Soc. B, 34:187-220, 1972

[3] Terry M Therneau, Patricai M. Grambsch, Modeling Survival Data: Extending the Cox Model. Springer-Verlag, 2000.