Logistic Regression for Insured Mortality Experience Studies

SCOR *in*FORM - December 2013

Global Life

Authors
Zhiwei ZHU
Zhi LI
SCOR Global Life Americas

Publisher
Gilles Meyer

Abstract

Findings of industry mortality experience studies are used by (re)insurers and regulators as the basis for developing liability expectations, reserve guidelines, and solvency capital requirements. In this paper, we introduce a logistic regression based modeling approach for analyzing the US insured mortality experience, including at advanced ages where less credible experience data are available. As a validation for applications, we create industry experience tables based on the model-estimated mortality and compare them to standard industry experience tables produced by the Society of Actuaries (SOA).

Our conclusion is that a properly designed logistic modeling approach can enhance industry experience studies in: a) testing mortality drivers' statistical significance in explaining mortality variations; b) estimating normalized mortality slopes and differentials such as how mortality varies by duration or between underwriting classes while product and age distributions are controlled; and c) addressing analytical challenges such as extrapolating for ultimate mortality, smoothing between select and ultimate estimations, and constructing multi-dimensional experience tables.

1. Introduction

There are three equally important aspects in industry mortality studies:

a) Mortality trend: how mortality improved or will improve by time
b) Mortality slope: how mortality increases by age or duration
c) Mortality differential: how mortality, mortality trend, and/or mortality slope vary between insured segments such as males vs. females or preferred class vs. residual standard class

a) and b) are related but not duplicative. c) needs to be studied at more granular levels. Data availability is one of the key determining factors of what and how studies can be done.

Due to insufficient insured experience data collection by time, insured mortality trends are often approximated based on general population's trend studies that take advantage of the long term general population experience data collection. For example, Lee-Carter (1992) introduced a twin-model method for studying general population mortality: one model for fitting the population's past experiences and the other for extrapolating future expectations. Eilers and Marx (1996) proposed the use of p-spline regression for fitting empirical data and smoothing the fit. Currie, Durban, and Eilers (2004) applied the p-spline concept to projecting the U.K. insured lives experience. Both Lee-Carter and Currie's studies formed the foundation of the insured mortality stochastic estimation tool developed by the U.K.'s Continuous Mortality Investigation Bureau (e.g., CMI 2005). Similarly, Hardy, Li, and Tan (2006) employed both Lee-Carte models and Currie's p-spline to fit and to project Canadian general population mortality and then calibrated the learning to estimate Canadian insured mortality improvement. Their study provided supporting evidence for changes in the Standards of Practice for the valuation of insurance and annuity business in Canada (Canadian Institute of Actuaries 2010).

Also due to scarce death claims at advanced ages and underwriting wear-off in high policy durations (mortality becomes more similar between the insured and uninsured), significant amount of advanced age mortality research is shared by and applied to both insured and general populations. Since the introduction of Gompertz Law of Mortality (1825), the effort of modeling human mortality trajectory by age has only accelerated. Thatcher (1999) provided an excellent description and comparison of four mortality models by age. With some simplifications in reducing number of parameters and unifications of using force of mortality as the dependent variable, the four models are:

(1.1) Gompertz(1825): $\mu \approx \alpha * \exp(\beta * x)$

(1.2) Weibull(1951): $\mu = \alpha * x^{\beta}$

(1.3) Heligman and Pollard(1980): $\mu \approx \alpha - \frac{1}{2}\beta + \beta * x$

(1.4) Kannisto(1992): $\mu = \frac{\alpha * \exp(\beta * x)}{1 + \alpha * \exp(\beta * x)}$

Of the four models, only the Kannisto model assumes that force of mortality has a finite asymptote. Thatcher's conclusion was: "when these four models are fitted to actual (general population) data, they are all relatively close to the data at ages where most of the deaths are concentrated, and hence relatively close to each other." It is not surprising that he also confirmed with various population data (Thatcher et al., 1998, Thatcher 1999) that the Kannisto model fits and approximates old age mortality the best.

Insured mortality differential studies have to stand on their own ground because general population data are simply not suitable for analyzing the disparity of insured mortality. In the US, insurers' risk selection activities (e.g., underwriting, pricing, marketing, product development) and insureds' anti-selection behaviors (e.g., policy choice, lapsesation, conversion, etc.) formed numerous "insured cohorts", or segments, within the insured population. These segments can be identified by variables, such as underwriting class, product type, policy size, etc., that are not captured in general population data. Mortality, trend, and slope do differ considerably among these segments. Companies thrive or fail by targeting some or all of these insured segments based on their knowledge and specialties.

One way to analyze selection impact and mortality disparity in conventional insured experience studies is to separately analyze so called 'select mortality' and 'ultimate mortality'. Select mortality occurs in earlier policy durations when the industry's risk selection activities are most effective. Ultimate mortality reflects later duration experience or expectation when the selection effectiveness has worn off and the mortality of insured and uninsured become closer. Oftentimes, an additional 'graduation' act is also taken to bridge the gap between the estimated select and ultimate mortality, depending on how the two are studied.

Table 2.1: Summary of the Insured Data

		Total Data				Selected Data			
Sex	Attained Age	Claim Count	Exposed Count	q	q/(1-q)	Claim Count	Exposed Count	q	q/(1-q)
Female	00-22	1,371	5,919,604	0.00023	0.00023	286	1,758,271	0.00016	0.00016
	23-27	1,096	3,194,034	0.00034	0.00034	291	1,425,257	0.00020	0.00020
	28-32	1926	5,493,708	0.00035	0.00035	598	3,133,315	0.00019	0.00019
	33-37	3,442	8,419,013	0.00041	0.00041	1,240	5,186,770	0.00024	0.00024
	38-42	6,636	10,403,257	0.00064	0.00064	2,467	6,266,131	0.00039	0.00039
	43-47	11,571	11,203,952	0.00103	0.00103	3,888	6,323,438	0.00061	0.00062
	48-52	17,935	10,672,817	0.00168	0.00168	5,206	5,405,578	0.00096	0.00096
	53-57	24,972	9,073,003	0.00275	0.00276	5,947	3,975,759	0.00150	0.00150
	58-62	32,389	6,817,009	0.00475	0.00477	5,541	2,408,077	0.00230	0.00231
	63-67	39,066	4,673,083	0.00836	0.00843	4,668	1,204,946	0.00387	0.00389
	68-72	50,894	3,551,700	0.01433	0.01454	4,099	642,219	0.00638	0.00642
	73-77	74,868	3,116,261	0.02402	0.02462	4,552	413,902	0.01100	0.01112
	78-high	299,642	4,887,952	0.06130	0.06531	14,515	482,748	0.03007	0.03100
Male	00-22	3525	6,303,991	0.00056	0.00056	768	1,827,197	0.00042	0.00042
	23-27	3,105	3,304,725	0.00094	0.00094	767	1,428,309	0.00054	0.00054
	28-32	4,175	5,964,346	0.00070	0.00070	1,354	3,463,059	0.00039	0.00039
	33-37	7,204	10,166,729	0.00071	0.00071	2,712	6,587,593	0.00041	0.00041
	38-42	13,114	13,884,778	0.00094	0.00095	5,144	8,933,857	0.00058	0.00058
	43-47	22,948	16,060,846	0.00143	0.00143	8,353	9,912,149	0.00084	0.00084
	48-52	36,977	16,212,737	0.00228	0.00229	12,003	9,305,648	0.00129	0.00129
	53-57	54,632	14,819,343	0.00369	0.00370	14,814	7,668,238	0.00193	0.00194
	58-62	73,629	12,072,373	0.00610	0.00614	16,423	5,396,172	0.00304	0.00305
	63-67	89,983	8,450,035	0.01065	0.01076	14,849	3,033,903	0.00489	0.00492
	68-72	109,391	5,993,105	0.01825	0.01859	12,866	1,584,710	0.00812	0.00819
	73-77	146,537	4,640,211	0.03158	0.03261	11,648	840,363	0.01386	0.01406
	78-high	490,630	6,539,738	0.07502	0.08111	19,897	586,764	0.03391	0.03510

2.1 The q in the table is defined as the number of deaths divided by exposure. In this paper, mortality, mortality rate, death probability, and death rate all refer to the same q, unless specified otherwise.

2.2 The probability of death q and the odds of death q / (1 – q) are approximately equal for nearly all age groups because 1 – q ≈ 1. This implies that many of the odds ratio based interpretations of the logistic q model can be reasonably interpreted in terms of probability ratios or mortality differentials. (Appendix A)

There are more challenges in industry mortality experience studies:

- Though the amount of insured experience data can be large, the data usually have uneven claim credibility and collection consistency
- The ‘separately studying’ approach can quickly run into the data credibility ‘ceiling’, especially for mortality differentials by multiple variables
- It is difficult to control or normalize by multiple explanatory variables

In this paper, we modify the Kannisto mode (1.4) for modeling insured mortality q rather than force of mortality μ and expand the model to include multiple explanatory variables rather than just age. The adaption of a multiple variable modeling approach, the availability of large amount of policyholders’ data, and the use of modern computing technology enable us to

- Train the model with model fit
- Project ultimate and advanced age mortality with model extrapolation
- Bridge between select and ultimate mortality with logistic link functions
- Derive normalized mortality slopes and differentials between policy segments with model coefficients
- Verify reliability of the study with model fit statistics
- Construct multi-dimensional industry experience tables by using the model as a predictive model

Combining our mortality experience modeling method and the mortality trend study approaches such as the ones mentioned earlier can provide more complete solutions to serve the industry’s mortality projection needs.

The rest of the paper is organized as follows: Section 2 summarizes the data sources for our study; Section 3 describes our logistic mortality models, their advantages, and modeled mortality slopes and differentials for an insured sub-population; Section 4 reviews the issue of “death censorship by policy lapsation” a logistic regression based solution; Section 5 discusses the limitations and possible enhancements of using logistic regression models for industry experience studies.

2. The Data Sources

General population data: The Human Mortality Database (HMD) is our source for US general population mortality experience. At the time of our study, the database covers a period from 1933 to 2010.

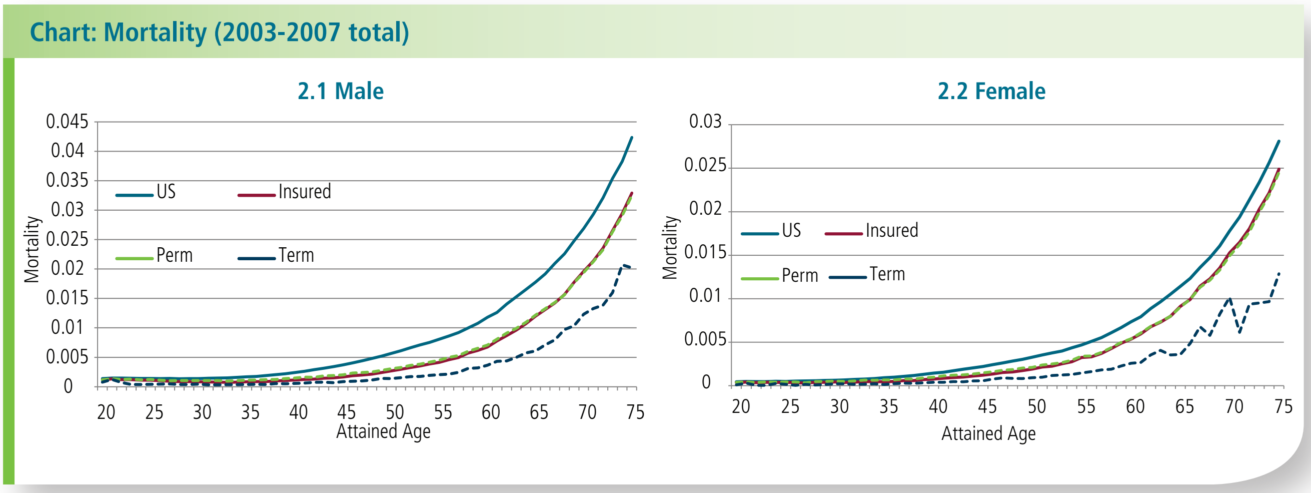
Insured population data: Unlike the general population, there is no centralized experience data repository for the US insured population. Insured experience data usually come from insurers’ ad hoc contributions and cover relatively short exposure periods. The insured experience data used in this study were collected by a major consulting company and a global reinsurer. The data file consists of experiences from over 60 insurers with exposure from 2000 to 2009. A total of 174 million policy exposure years and 1.6 million death claims are available for study.

We do not expect to represent insured mortality disparity well with just one model or one study. In this paper we select and study a relatively homogeneous subset of the insured population with:

Policies issued since 1950
Face amount >= $50,000

The filtered subset has more relevance to recent and near future experience.

Chart: Mortality (2003-2007 total)



The following table summarizes the total and the selected data.

The following two charts compare the five-year 2003-2007 total mortality rates of the general and the insured experiences based on the data we have. Again,

- The general population data source is the Human Mortality Database
- The insured population data source is our total study data
- The insured data are also split into two exclusive subgroups: permanent and term product subgroups
- These mortality rates are derived without normalizing any distributions such as by duration, issue year, and underwriting class. They will be later compared to normalized estimations from our models

3. Model Insured Mortality With Logistic q Models

Our logistic mortality model has a general form of

$$(3.1) \quad q = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_{1,2}x_1 * x_2+\beta_{1,3}x_1 * x_3+...)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_{1,2}x_1 * x_2+\beta_{1,3}x_1 * x_3+...)}} \quad \text{or}$$

$$(3.1a) \quad \ln(\frac{q}{1-q}) = \alpha + \sum_i \beta_i * x_i + \sum_{i,j} \beta_{i,j} * x_i * x_j + ...$$

where

q: is probability of death in an exposure year, given a policy holder survived to the beginning of the year.

x_i: are explanatory variables (e.g., age, sex, duration, product).

α: is intercept, to be estimated with experience data and maximum likelihood method.

β_i: are coefficients of the explanatory variables, to be estimated with experience data and maximum likelihood method (Appendix B).

To distinguish from the logistic force of mortality model or logistic μ model (1.4) studied by Kannisto (1992) and Thatcher (1999), let us call our model (3.1) logistic q model, as they really are. According to Thatcher's illustration, a simplified Heligman and Pollard model (1.3) model with only one explanatory variable 'age' is a special form of our logistic q model.

We chose logistic q model for our study for several reason.

- It models mortality q that are directly used in business operation and risk management
- It is flexible to configure for estimating mortality levels, slopes, and differentials that are key metrics used in business practices (see Attachment A)
- It performs many other analytical functions such as normalization, hypothesis test, risk scoring, experience table construction that are difficult to do with conventional experience study methods (Harrell 2001)
- It can be built with widely available commercial software system such as SAS, SPSS, and R

In addition to the dependent variable q, nine observable explanatory variables are selected as potential independent variables for our model development:

Gender:	male and female
Duration:	as continuous variable
Issue age (last birth):	1 through 99 as continuous variable
Smoker status:	smoker, nonsmoker, unknown
Product:	permanent, term
Underwriting class:	preferred, residual standard, aggregate (one class)
Exposure year:	2000 through 2009 as continuous variable
Underwriting era:	4 eras defined by issue year to reflect key underwriting evolutions such as smoker and preferred ratings
Face Category:	\$50-\$99k, \$100-499k, \$500k+ (inflation adjusted)

These variables are selected for study because they have least missing values and are the most frequently used for pricing decisions, underwriting adjustments, and marketing strategies.

Unlike in general population mortality studies, we chose issue age and policy duration instead of attained age and calendar year to represent age and time, because the chosen pair better reflect insured characteristics and have direct links to pricing tables.

Recall that our logistic q model has the general form of

$$(3.1a) \quad \ln(\frac{q}{1-q}) = \alpha + \sum_i \beta_i * x_i + \sum_{i,j} \beta_{i,j} * x_i * x_j + ...$$

The right hand side of the model has three components: the intercept α, the main effect component that is the weighted sum of individual explanatory variables, and the interaction component that is the sum of products of two or more explanatory variables.

When interaction terms are omitted from model (3.1) or (3.1a),

$$q = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots)}} \qquad (3.2)$$

$$\ln(\frac{q}{1-q}) = \alpha + \sum_i \beta_i * x_i \qquad (3.2a)$$

The model coefficients α and β_i can deliver estimations for mortality level, slope, ratios, depending on how the corresponding variable is coded. Appendix A provides more details on this topic.

Adding the interaction component to a model has the potential to improve model fit. It also adds complexity to interpreting the model coefficients. From our tests, we found that adding interaction term improves our model fit slightly. For simple interpretation, in this paper we only present sample model (3.2) without interactions.

For a better matched comparison with SOA studies, we split the selected study data into four subsets and fit each subset with its own model (3.2). The four subsets are male smokers, male non-smokers, female smokers, and female non-smokers. This separate model design allows each model's coefficients to be estimated independently from the other three models, which means that each of the four policy groups can have its own mortality level, slopes, and differential factors completely different from the other three groups.

SAS software is used for our data preparation and model development. In the following, we highlight learning from the interpretations of three SAS modeling process outputs.

3.1 "Analysis of Effects" for mortality driver significance test:

Of the nine explanatory variables, gender and smoker-status are used to split the study data and seven are left to be included in the models. The following Table 3.1 summarizes the p-values of the significance tests of the seven explanatory variables on each of the four data sets.

3.2 "Odds Ratio Estimate" for mortality slopes and differentials

Of the nine studied explanatory variables, three (issue age, duration, and study year) are treated as continuous for three reasons: 1) to estimate smoothed relationships between q and these variables, 2) to allow the coefficients β of these variables to be transformed as mortality 'slopes', 3) to enable model based mortality extrapolation for older ages and later durations where sparse or no experience data are available. The modeled extrapolation can be used as ultimate mortality estimate.

Table 3.1: Analysis of Effects

Pr > ChiSq (p-value)	DF	Female		Male	
		Non-Smoker	Smoker	Non-Smoker	Smoker
Duration	1	<.0001	<.0001	<.0001	<.0001
Issue Age	1	<.0001	<.0001	<.0001	<.0001
Study Year	1	0.1714	0.4597	0.1719	0.0017
Face Band	2	0.0051	0.004	<.0001	<.0001
Product	1	0.0157	0.9533	0.1363	<.0001
Issue Year	2	<.0001	0.0003	<.0001	<.0001
Class	2	<.0001	<.0001	<.0001	<.0001

As expected, insured mortality varies statistically significantly by duration, issue age, underwriting class, underwriting era (Issue Year), and face band for all four subgroups. This confirms that these variables are among the most reliable mortality predictors.

Study Year, or exposure year, is included as a placeholder for mortality improvement in the ten-year period covered by the study data. The corresponding p-values from the four models imply that, after factoring out what have been explained by the other eight explanatory variables (including sex and smoking status), mortality variation explained by exposure year (or improvement) is statistically significant at $\alpha = 0.05$ only for male smokers. This may imply that more male smokers ceased smoking and resulted in more mortality improvement during the studied period.

Mortality differentiation by product (between permanent and term policyholders) is only statistically significant for female nonsmokers and male smokers, after controlling the other eight explanatory variables.

At 95% confidence level, all seven tested variables have statistical significance in explaining mortality variation in at least one of the four policy groups. We decide to include them in all four logistic q models. Vinsonhaler et al (2001) analyzed private pension plan experience data with similar logistic q models and only found one significant explanatory variable. Since mortality and longevity are the two sides of the same death related 'risk coin', our finding may suggest that more potential longevity risk drivers are yet to be confirmed.

The values of the other six explanatory variables are categorized based on data credibility and recoded as binary variables as described in Appendix A. Therefore, mortality differentials are obtained for these variables.

Table 3.2 below contains the odds ratio estimations (point Estimate columns) and their 95% confidence intervals. For the three continuous variables, the odds ratios estimate average mortality increase per unit increase in the corresponding variables. For the categorized variables, odds ratios represent the mortality ratios as defined in the "Effect" column. The 95% confidence intervals provide a means to verify the credibility of the corresponding slope or differential estimate.

As described in Appendix A, odds(death)= q/(1-q) ≈ q because q is usually very small. Therefore, odds ratios can be viewed as mortality ratios in this table. Also explained in Appendix A is that logistic q model coefficients are estimated assuming the values of all other the explanatory variables are the same (normalized). Therefore, they approximate normalized mortality differentials that may or may not appear to be consistent with results obtained from actual mortality studies. Let's take male non-smokers model as an example and interpret some of the odds ratios:

1. Duration and age slopes: If everything else were equal, on average mortality increases about 14% per duration and about 10% per issue age (Odds ratio=1.14 and 1.10 respectively). The 10% per issue age increase is known to be also true to the general population (Thatcher 1999).
2. If everything else were equal, there is a statistically insignificant 0.2% annual mortality improvement (odds ratio=0.998, the 95% confidence interval including 1). This finding may seem to be inconsistent with the common thought of higher mortality improvement. There are three possible explanations for this. First, due to the short time period and inconsistent data contributions from insurers,

Table 3.2: Odds Ratio Estimates

	Male Non Smoker			Male Smoker			Female Non Smoker			Female Smoker		
Effect	Point Estimate	95% Wald Confidence Limits		Point Estimate	95% Wald Confidence Limits		Point Estimate	95% Wald Confidence Limits		Point Estimate	95% Wald Confidence Limits	
Duration	1.141[1]	1.139	1.143	1.118	1.114	1.122	1.157	1.153	1.160	1.133	1.126	1.139
Issue Age	1.101[1]	1.100	1.102	1.093	1.092	1.094	1.105	1.104	1.105	1.098	1.096	1.099
Study Year	0.998[2]	0.995	1.001	1.009	1.003	1.014	0.997	0.992	1.001	1.004	0.994	1.013
Face 100k-499k vs 500k+	1.115[3]	1.096	1.135	1.203	1.143	1.265	1.000	0.971	1.030	0.926	0.855	1.002
Face 50k-99k vs 500k+	1.284[3]	1.258	1.311	1.407	1.335	1.484	1.037	1.003	1.071	0.988	0.911	1.072
UnderW Med vs Non-med	0.920[4]	0.902	0.939	1.018	0.986	1.050	0.950	0.921	0.981	1.044	0.992	1.099
Product Perm vs Term	1.013[5]	0.996	1.030	0.923	0.890	0.958	1.033	1.006	1.060	0.998	0.939	1.061
Class One-Class vs Standard	1.042[6]	1.027	1.057	0.930	0.893	0.967	1.038	1.014	1.062	0.938	0.881	0.999
Class Preferred vs Standard	0.730[6]	0.719	0.741	0.748	0.717	0.781	0.740	0.722	0.758	0.767	0.715	0.823

the study data may have not captured the true insured mortality improvement. Second, in the past decade or so, US population mortality improvement has been leveling off as shown in the chart (data are from the Human Mortality Database; the age range reflects the most commonly insured ages). This may also be true to the insured population. Third, insurance underwriting has specifically targeted high death rate causes, such as cardiovascular diseases and smoking, and excluded or discouraged these risks being insured, which may have resulted in less benefits of insureds from the advancement in medicine, treatment, and public education. Fourth, unlike a univariate analysis that attributes all the mortality variation to the single study variable, a large portion of the insured mortality improvement over the studied years has been attributed by the model to other variables such as (the introduction of) preferred classes, term products, and flattened age or duration slopes that do explain much more mortality variations.

3. If everything else were equal, compared to large policies with face amount at least $500k, the polices sized between $50-99k and $100-499k would have 28% and 12% higher mortality, respectively (odds ratio=1.28 and 1.115).
4. If everything else were equal, mortality of policies that had medical exams at issue is about 8% lower than that of those without (odds ratio=0.92, significant).
5. If everything else were equal, permanent policy mortality would be about 1.3% higher than that of term policies (odds ratio=1.013, insignificant). This may appear inconsistent with what is shown in Charts 2.1 and 2.2 in Section 2. Keep in mind that the descriptive measures in Charts 2.1 and 2.2 are obtained without controlling any other variables. Most of the differences displayed in Charts 2.1 and 2.2 may be caused by unmatched duration, issue year, and underwriting class distributions. Logistic mortality model provides an effective means to perform normalization.
6. If everything else were equal, the mortality of preferred class would be about 27% lower than that of the residual standard class while mortality of the aggregate class (one class plus unknown) is about 4% higher (odds ratio=0.73 and 1.041).

As mentioned before, normalized mortality information is essential in identifying underlying causes and avoiding miscounting of the mortality differentiation values when setting pricing factors. Findings of this analysis can also be useful in validating industry tables that are split from an aggregated table, like the 2001 CSO preferred class structure tables.

3.3 "Model Fit" for overall reliability test

Compared to health or property & casualty insurance claims, mortality claims occur at a much lower frequency and with a much more stable pattern. Relatively scarce claim count and more consistent claim patterns led us to use all available data for model building, without setting aside data for over-fit verification.

One commonly used model fit measuring statistic is c-statistic, or area under the Receiver Operating Characteristic. Table 3.3 below displays the overall c-statistics for the four models.

Table 3.3: Model Fit

Association of Predicted Probabilities and Observed Responses	Female		Male	
	Non-Smoker	Smoker	Non-Smoker	Smoker
	0.682	0.753	0.679	0.747

Vinsonhaler et al (2001) analyzed private pension plan experience data with similar but simpler logistic q model (only one explanatory variable). Their model had c-statistics in the range of 0.51 - 0.59 for most of the age groups. Though we are not measuring c-statistic by age group, the comparison still gives a sense that our four models have reasonably high c-statistics and fit the corresponding data sets well.

An interesting observation is that the two non-smoker models have lower c-statistics than the two smoker models. If c-statistic is used as a predictability measure, the predictability by the same set of explanatory variables for smokers is about 10% higher than for non-smokers. This 10% gain in death predictability is likely from knowing smoking status.

4. Impact of Death Censorship By Policy Lapsation

The adaption of a statistical model for insured mortality study brings a new issue that the conventional methods do not need to deal with: "death censorship by lapsation".
Think of a group of 100 current policyholders. If 10 died in the next 12 months but only 5 generated claims and the other 5 died after termination of their coverage, the death rate of the group would 10% but claim rate would be only 5%. Insured mortality, or claim rate, is conditioned on policy inforce and only reflects claim risk. It is not equivalent to general population mortality. When models like those in (1.1) – (1.4) or our logistic q model are used for estimating insured mortality, or claim rate, they do not recognize or discount policy lapse and tend to overestimate claim rate. This overestimation may not be a material issue for mortality differential study because differential is usually measured in aggregate and by ratios. If overestimation occurs to both the numerator and the denominator by a same factor, the ratio will cancel out the overestimation and remains relatively accurate. However, when the model is used for individual policy or policy group mortality extrapolation, such as in experience table development, the overestimation can be significant. One solution to address the issue is to discount possible future policy lapse from contributing claim by the model default.

Censoring Based Adjustment: Let's reserve q for death rate and assume that each insured policy has three observable statuses (and corresponding probabilities) at the end of an exposure year: lapse (q_l), claim (qc), and Inforce (qi) so that

$q_l + q_c + q_i = 100\%$

With the same explanatory variables xi as used in (3.2), we can use multinomial logistic model to model the three probabilities as follows (see Chapter 8 of Hosmer et al. 2013 for more descriptions):

$$(4.1)\quad \begin{cases} q_c = \dfrac{e^{(\alpha_c + \beta_{c1}x_1 + \beta_{c2}x_2 + \ldots)}}{1 + e^{(\alpha_l + \beta_{l1}x_1 + \beta_{l2}x_2 + \ldots)} + e^{(\alpha_c + \beta_{c1}x_1 + \beta_{c2}x_2 + \ldots)}} \\[2ex] q_l = \dfrac{e^{(\alpha_l + \beta_{l1}x_1 + \beta_{l2}x_2 + \ldots)}}{1 + e^{(\alpha_l + \beta_{l1}x_1 + \beta_{l2}x_2 + \ldots)} + e^{(\alpha_c + \beta_{c1}x_1 + \beta_{c2}x_2 + \ldots)}} \\[2ex] q_i = \dfrac{1}{1 + e^{(\alpha_l + \beta_{l1}x_1 + \beta_{l2}x_2 + \ldots)} + e^{(\alpha_c + \beta_{c1}x_1 + \beta_{c2}x_2 + \ldots)}} \end{cases}$$

Let us call this model logistic q_c model. The added lapse component q_l in (4.1) plays a role of estimating the to-be-lapsed portion of exposures and excluding them from contributing deaths to claim rate q_c estimation.

Asymptotically, by comparing Model (3.2) and Model (4.1), we have

$$(4.2)\quad \lim_{duration\to\infty} q = \lim_{duration\to\infty}(q_c + q_l) = 1$$

which implies that Model (4.2) asymptotically splits the total death rate into a claimed portion and a lapsed portion. As to the asymptote of the claimed portion,

(4.3)

$$\lim_{duration\to\infty} q_c = \lim_{duration\to\infty} \frac{1}{1+e^{(\alpha_l-\alpha_c)+(\beta_{l1}-\beta_{c1})*duration}} = \begin{cases} 1 & \beta_{l1} < \beta_{c1} \\ \dfrac{1}{1+e^{(\alpha_{l1}-\alpha_{c1})}} & \beta_{l1} = \beta_{c1} \\ 0 & \beta_{l1} > \beta_{c1} \end{cases}$$

For projection purpose, α_l and α_c are usually related to initial lapse and clam levels; β_{l1} and β_{c1} are related to lapse and claim slopes. A highly simplified interpretation of (4.3) is that, depending on if the death rate asymptotically increases faster than, slower than, or equal to the lapse rate of a portfolio, the portfolio's claim rate will approaching 100%, 0%, or something in between.

It is understood that insured lapse rates are driven by many long and short term factors and do not necessarily have as a regular relationship with duration as claim rate or death rate has. The lapse component of model (4.1) may not have an as good fit to insured lapse experience. However, it is reasonable to view the lapse component of model (4.1) as an empirical data driven adjustment for the unknown portion of the non-claim generating exposures. No matter which of the three asymptotes in (4.3) occurs, the overall effect of (4.1) on qc is to flatten the modeled qc slope by duration and to result in lower modeled qc than modeled q by Model (3.2), especially for advanced ages or later durations. Model (4.1) allows qc not to approach 100%, which is not achievable with model (1.1) – (1.4) mentioned in Section 1.

Due to a data usage agreement issue, we do not have access to the lapse information for this study and unable to demonstrate a real example of model (4.1) that is especially useful for estimating ultimate mortality. A follow up study is planned.

As an alternative, we did apply some industry expert opinions on insured ultimate mortality to create a simplified version of the model (4.1), used the model to produce model-estimated industry experience tables, and compared the tables with SOA's 2001 and 2008 VBTs (Valuation Basic Table). The result is very positive. Because this alternative involves various subjective assumptions, it is not presented in detail here.

5. Constraints and Possible Enhancements

Among others, three types of biases can occur in a insured mortality experience study: parameter bias, sampling bias, and data bias. A parameter bias is a systemic bias that reflects technical limitations of a study method (e.g., using a linear model to fit U shaped experience). A sampling bias happens when a substitute dataset is used to represent a target population but the substitute does not have the same characteristics of the target (e.g., using a small sample to represent a large population, or using past experience to approximate future outcomes). A data bias is the discrepancy between data and actuality (e.g., misreported ages of deaths or unrecorded lapse).

Some of logistic models' parameter bias (e.g., a logistic q model overestimates claim rate q_c) and sampling bias (e.g., uncontrolled company contributions causing inconsistent representation of the industry) have been discussed in the previous sections. As in any other large database, insured experience data have plenty of data biases such as missing data and inconsistent data coding among companies that may compromise the quality of logistic modeling or other experience studies. The following are a few more constraints of using logistic regression for insured experience studies.

1. Logistic q or q_c models may not fit infant and pre-marriage attained age experience well (parameter bias). Mortality is usually high in these ages due to causes such as accidents and suicides. As the excess causes level off with age, mortality regresses back to a more normal pattern that fits better with logistic q function. The main strengths of logistic models are in aggregated mortality slope/differential estimation and model extrapolation. To improve fit, a possible solution could be to further customize logistic q or q_c model with some spline or localized regression methods to fit the ages that have less regular mortality patterns.
2. When scarce experience data are available such as at very old issue ages or later durations (data bias), logistic regression will be the primary driver for estimating modeled q or q_c. For more accurate estimations, calibrations with expert knowledge are usually necessary.
3. Shock lapse and shock mortality that occurs at the end of the level premium period or during rare events like pandemics cannot be fit or reflected well by a continuous function based model (parameter bias). At a more granular level, modeling issues such as quantifying the end of level period effect for a specific portfolio will need more than logistic mortality model. However, at an industry aggregated level and for constructing insured mortality tables, our study shows that logistic models deliver reasonable results.
4. The current lack of a consistently collected long term insured experience data are limiting the optimization of any modeling efforts including logistic mortality models (sampling and data biases). For example, not all companies and not all product information are consistently or proportionally presented in an ad hoc industry experience data collection. Special cares are necessary in interpreting model outputs that implicitly assume the consistency. As data process technology and analytical methodology advance, it is our hope that the industry will establish a mechanism to consistently collect comprehensive experience data for in-depth experience studies.

In summary, logistic regression models have many strengths and potentials for insured mortality experience studies:

- Test for statistically significant mortality drivers in explaining mortality variations with "Effect Analysis"
- Generate normalized mortality metrics such as slopes and differentials with "Odds Ratio Analysis"
- Extrapolate for advanced age or ultimate mortality with "Modeled Estimation"
- Quantify overall study reliability with "Model Fit Statistics"
- Help construct multi-dimensional experience tables through using the model as a "Predictive Model"
- Be implementable with widely available software systems.

Acknowledgement

We would like to thank David Wylde, George Hrischenko, Mike Failor, Jun Han, Daria Ossipova-Kachakhidze, Doris Azarcon, and Pat Bresina for their insightful inputs and helpful suggestions.

References

Canadian Institute of Actuaries, 2010. **Mortality Improvement Research Pape**r, Document 210065.

Continuous Mortality Investigation Bureau, 2005. **Projecting Future Mortality: Towards a proposal for a stochastic methodology.** CMI Working Paper no. 15. London: Institute of Actuaries and Faculty of Actuaries.

Continuous Mortality Investigation Bureau, 2007. Working Paper 20: **Stochastic projection methodologies: Further progress and p-spline model features, example results an implications.** London: Institute of Actuaries and Faculty of Actuaries.

Continuous Mortality Investigation Bureau, 2007. Working Paper 25: **Stochastic projection methodologies: Lee-Carter model features, example results and implications.** London: Institute of Actuaries and Faculty of Actuaries.

Currie, I. D., Durban, M., and Eilers, P. H. C. 2004. **Smoothing and Forecasting Mortality Rates.** *Statistical Modeling* 4: 279–298.

Eilers, P.H.C. and Marx, B.D. 1996 **Flexible Smoothing with B-splines and Penalties.** Statistical Science, 11(2):89-121.

Gompertz, Benjamin. 1825. **On the Nature of the Function Expressive of the Law of Human Mortality and on a New Mode of Determining Life Contingencies.** Royal Society of London, Philosophical Transactions, Series A 115: 513–85.

Hardy, M., Li, S., and Tan, K. 2006. **Report on Mortality Improvement Scales for Canadian Insured Lives.** http://www.soa.org/Research/Experience-Study/Ind-Life/Mortality/cia-mortality-rpt.aspx

Harrell, F. Jr. 2001. **Regression Modeling Strategies.** Springer-Verlag New York, Inc.

Heligman,L. and Pollard,J.H. (1980) **The age pattern of mortality.** J. Inst. Act., 107, 49-80

Hosmer, D., Lemeshow, S. and Sturdivant R. 2013, **Applied Logistic Regression, 3rd edition,** John Wiley & Sons, inc.

Human Mortality Database. University of California, Berkeley. http://www.mortality.org

Kannisto,V. (1992) **Workshop Old Age Mortality**, Odense, June

Lee, R., and L. Carter. 1992. **Modeling and Forecasting U.S. Mortality.** *Journal of the American Statistical Association* 87: 659–671.

Li, Johnny Siu-Hang, Hardy, R. Mary, and Tan, KenSeng. 2008. **Threshold Life Tables and Their Applications.** NAAJ 2008, Vol 12, 99-115.

Li, Johnny Siu-Hang, Hardy, R. Mary, and Tan, KenSeng. 2010. **Developing Mortality Improvement Formulas: The Canadian Insured Lives Case Study.** NAAJ 2010, Vol 14, 381-399.

McCullagh, P. and Nelder, J.A. 1989. **Generalized Linear Models, 2nd edition.** Chapman & Hall, London.

Olshansky, S. J. 1998, **On the Biodemography of Aging: A review Essay.** Population and Development Review 24, 381-393.

Society of Actuaries. **2008 Valuation Basic Table (VBT) Report & Tables**. http://www.soa.org/Research/Experience-Study/Ind-Life/Valuation/2008-vbt-report-tables.aspx

Thatcher, A. R., Kannisto, V., and Vaupel, J. W. 1998. **The Force of Mortality at Ages 80 to 120.** Odense: Odense University Press.

Thatcher, A. R. 1999. **The Long-Term Pattern of Adult Mortality and the Highest Attained Age.** Journal of the Royal Statistical Society Series A 162: 5–43.

Weigull,W.A. (1951) **A statistical distribution function of wide applicability.** J. Appl. Mech., 18, 293-297

Appendix A: Logistic q model coefficient interpretation

Consider a logistic q model

(A.1) $\ln(\frac{q}{1-q}) = \alpha + \beta_1 * age + \beta_2 * sex$

with two explanatory variables: $x_1 = age$ as continuous and $x_2 = sex$ as a binary variable having male and female two value categories. For the categorical variable sex, there could be many different ways to code the variable for analysis. The most commonly used coding scheme is Reference coding: Code one category as 1 and the other as 0 and call the category 0 the reference category (e.g. 1 for female and 0 for male and male is the reference category). Reference coding is useful when the primary goal of a study is to compare mortality between two segments of policies.

Under this coding scheme, we can calculate the difference of log of odds between females and males for the same age (controlling age),

$$\ln(\frac{q_{female}}{1-q_{female}}) - \ln(\frac{q_{male}}{1-q_{male}}) = \beta_2$$

or

(A.2)

$$e^{\beta_2} = (\frac{q_{female}}{1-q_{female}})/(\frac{q_{male}}{1-q_{male}})$$

which is the odds ratio of death between females and males. For the continuous variable age, if we take the difference of log of odds between any age x and x+1 for the same sex (controlling sex), we can derive:

(A.3) $e^{\beta_1} = (\frac{q_{age=x+1}}{1-q_{age=x+1}})/(\frac{q_{age=x}}{1-q_{age=x}})$

This is the odds ratio of death when age increases by 1 unit. If we set age=0 and sex=0 (or male) and consider this as the overall reference group, we have

(A.4) $e^{\alpha} = \frac{q_{male,age=0}}{1-q_{male,age=0}}$

In summary, (A.2), (A.3), and (A.4) illustrate how the coefficients of a logistic q model can be interpreted as odds ratios under the reference coding:

- The exponential of a binary variable's coefficient represents the odds ratio of the non-reference category vs. the reference category.
- The exponential of a continuous variable's coefficient represents the odds ratio when the variable value increases by 1 unit.
- The exponential of the intercept represent the odds of the overall reference subset that have value 0 for all the explanatory variables. In this case the males of age 0.
- Through variable transformation and recoding, we may choose any category as the reference.

In a more general situation, if a categorical variable has k categories of values and k>2, we can replace it with a set of k-1 binary variables and retain the reference coding advantages. For example, if in model (A.1) sex has three values: female, male, and unknown, we can replace sex with 3-1=2 binary variables y_1 and y_2. And the three sex categories can be represented by the paired (y_1, y_2) as:

	y_1	y_2
female	1	0
male	0	1
unknown	0	0

This means that y_1 serves as female indicator, y_2 as male indicator, and the pair of (0,0) as the reference. Model (A.1) is reformatted as:

(A.2) $\ln(\frac{q}{1-q}) = \alpha + \beta_1 * age + \beta_2 * y_1 + \beta_3 * y_2$

This model has only continuous and binary explanatory variables. Its coefficients can be interpreted as summarized before.

There are also other useful coding schemes for categorical variables, under which the model coefficients can be interpreted differently. For example, the 'deviation from means coding' codes the binary variables with values of 1 and -1 instead of 1 and 0. With this coding, the reference category is always the total controlled mean and the coefficient of a binary variable estimates the odds ratio between the represented variable category and the overall mean. This coding scheme is very useful in comparing the mortality of a segment relative to the overall means. See Chapter 3 of Hosmer et al (2013) for more discussions.

Appendix B: Logistic q model coefficient estimation

Consider logistic q model,

(B.1) $$q = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots)}}$$

Let y be the death indicator, with value 1 for death and 0 for inforce, X denote the vector of explanatory variable X ={ x_1, x_2 ,…,x_k}, and β ={ β_1,…, β_k} are the coefficients. Then,

q=Prob(y=1|X).

is a function of β when a sample value of X is given. Suppose we have a sample of n independent observation pairs (y_i, X_i), *i*=1, …, n. Since the likelihood of one observed y_i given X_i is

$$q_i^{y_i}(1-q_i)^{1-y_i}$$

the joint likelihood of all n observations is the product of these likelihoods:

(B.2) $$l(\beta)=\prod_{i=1}^{n} q_i^{y_i}(1-q_i)^{1-y_i}$$

To solve for the β that maximize the likelihood function (B.2), it is equivalent and easier to solve for β that maximizes the log likelihood

(B.3) $$L(\beta)=\ln(l(\beta))=\sum\{y_i\ln(q_i)+(1-y_i)\ln(1-q_i)\}$$

Unfortunately, the maximum likelihood estimate of β cannot be written explicitly. A Newton-Raphson method is usually used to solve iteratively for the value of β that maximize (B.3). One may consult McCullagh and Nelder(1989) for discussions of the methods commonly used by statistical modeling computer software.

SCOR Global Life
1, avenue du Général de Gaulle
92074 Paris La Défense Cedex
France
www.scor.com

SCOR Global Life