Logistic Regression for Insured Mortality Experience Studies

Zhiwei Zhu,[1] Zhi Li[2]

Presented at the Living to 100 Symposium

Orlando, Fla.

January 8–10, 2014

[1] Zhiwei Zhu, Ph.D., is vice president of risk modeling and analytics at SCOR Global Life Americas, zzhu@scor.com.
[2] Zhi Li, Ph.D., ASA, CERA, is associate director of financial risk modeling at SCOR Global Life Americas, zli@scor.com.

ABSTRACT

Insured population mortality estimation is essential to (re)insurers' developing liability expectations and maintaining required solvency capital. In practice, insured mortality measurement needs to deal with a broad range of data and analytical challenges. In this paper, we introduce a logistic regression-based modeling approach for analyzing the U.S. insured mortality experience, including at advanced ages where less credible experience data are available. As a validation, we create a version of industry basic experience tables based on the model-estimated mortality and compare them to standard industry experience tables produced by the Society of Actuaries (SOA). The conclusion is that properly designed logistic modeling processes can more efficiently utilize available data to deliver solutions for multiple needs, including: a) testing mortality drivers' statistical significances in explaining mortality variations; b) estimating normalized mortality slopes and mortality differentials such as how mortality increases by duration or varies between underwriting classes while product and attained-age distributions are controlled; and c) addressing analytical challenges such as extrapolating for ultimate mortality, smoothing between select and ultimate estimations, and constructing multidimensional basic experience tables.

---

1. INTRODUCTION

The following three aspects are equally important in life insurance industry mortality studies.

a) Mortality trend: how mortality improves or deteriorates over 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

Aspects a and b are related but not duplicative. Although the life insurance industry collects an enormous amount of data, the collections usually do not consistently cover sufficient time periods for credible trend analyses. Insured mortality trends are often approximated based on general population mortality-trend studies. This dependency on general population studies for insured mortality-trend understanding is likely to last at least until the life insurance industry establishes an adequate insured-experience data repository, similar to the Human Mortality Database for the general population, to support its own comprehensive and true-experience data-driven studies.

As to understanding insured mortality slopes and mortality differentials, there are insured-specific challenges unlikely to be addressed with the general population data: dynamic insured segmentation, high mortality disparity among the segments and the need for multidimensional normalization. First, compared to the general population, the U.S. insured population is highly unstable. Insurers constantly initiate risk selection efforts through improving underwriting, adjusting pricing strategies, expanding markets and developing new products, which not only

attracts various levels of risks but also causes current insureds' antiselection reactions such as policy lapsation and conversion. These risk selection and antiselection activities form and reshape numerous insured cohorts. In addition to age, insured cohorts may be defined by smokers vs. nonsmokers, preferred class vs. standard class, permanent policyholders vs. term policyholders, groups vs. individuals, or a mix of these characteristics that usually are neither relevant to nor captured in any general population data. Second, by design, mortality and mortality-trend patterns differ significantly among these insured segments. Companies often compete on properly pricing by segment. Third, the insured segments often have relatively small sizes, short histories and multidimensional characteristics (e.g., male preferred smokers). Analyzing the true values of each of the characteristics in differentiating mortality and trend requires controlling the distributions by the others (normalization). As the demand for risk management and competition for market share increases, the demand for more thorough understanding of insured mortality increases.

In this paper, we apply a logistic regression-based modeling approach to analyze the U.S. insured mortality experience based on a large amount of data collected by a major consulting firm and a global reinsurer. The adaption of the multiple-variable modeling approach (Advanced Analytics), the availability of a large amount of policy data (Big Data) and the use of modern computing technology provide many advantages over the conventional insured mortality study methods, including

- Empirical data-driven models with multiple explanatory variables (drivers)
- Projection of ultimate and advanced-age mortality by combining past experience and model extrapolation
- Smooth bridging of select and ultimate mortality with the models' link functions
- Derivation of normalized mortality slopes and differentials between policy segments with model coefficients, which is difficult to do with conventional methods
- Verification of reliability of the overall study with model fit statistics and not solely relying on the number of claims available for credibility verification
- Construction of multidimensional industry experience tables by using the models as predictive models, overcoming some of the weaknesses of pivot tables

In summary, a logistic regression modeling approach allows use of less but more relevant data to address multiple challenges in quantifying insured mortality. Examples and interpretations of findings from our modeling process are represented in section 3.

The rest of the paper is organized as follows: Section 2 summarizes the data used for this study; section 3 describes logistic models and how to model mortality slopes and differentials; section 4 reviews the issue of "death censorship by policy lapsation" and a logistic regression-based solution; and section 5 discusses the limitations and possible enhancements in using logistic regression models for industry experience studies.

2. THE DATA SOURCES

The Human Mortality Database (HMD) is our source for U.S. general population mortality experience. At the time of our study, the database covered 1933 to 2010. The data is mainly used for comparison purpose, not for insured mortality approximation.

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 more than 60 insurers with exposure from 2000 to 2009. The included policies were issued as early as 1912. A total of 174 million policy exposure years and 1.6 million death claims are available for study.

Actuaries and analysts traditionally use as much experience data as they can get to perform insured mortality analyses, whether the purpose is to evaluate in-force blocks (policies issued in the past) or to price new business. This simple strategy can be a double-edged sword. While increased volume of data may benefit the credibility of analysis, it may also diminish the accuracy if less relevant data are included. Studies (e.g., Vaupel 2014) found that in some developed countries, life expectancy expands about 2.5 years per decade, which implies that using policies issued 50 years ago to estimate future policyholders' mortality can result in significant bias. Instead of trusting or subjectively adjusting biased analysis finding, we explored using less but more relevant data to model and to extrapolate estimations. It achieved satisfactory model performance.

For illustration purpose, we elected to apply the following data-selection criteria for estimating mortality that is more relevant to pricing future fully underwritten policies. Other filters may be applied for different purposes.

- Policies issued since 1950
- Face amount $\geqslant$ \$50,000

The following table summarizes the total and the filtered study data.

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	1,926	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**	3,525	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

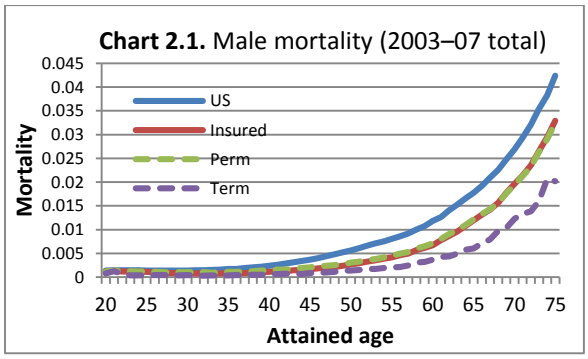
Notes: The q in the table is defined as the number of deaths divided by exposure. In this paper, mortality, mortality rate, death probability and death rate all refer to the same q, unless specified otherwise.

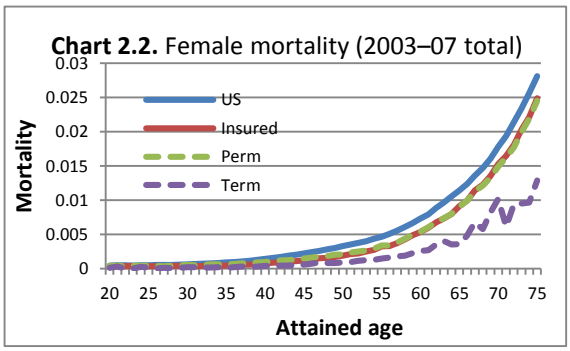
The probability of death q and the odds of death $q / (1 - q)$ are approximately equal for nearly all age groups because $1 - q \approx 1$. This implies that many of the odds ratio-based interpretations of the logistic q model can be reasonably interpreted in the terms of probability ratios or mortality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

The following two charts compare the five-year 2003–07 total mortality 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.
- These mortality rates are derived without normalization by any distributions such as duration, issue year and underwriting class. According to the charts, for any given age group, permanent policyholders have about 50 percent higher mortality than term

policyholders. Later, the differences after normalizing by nine study variables will be quantified with our models.





3. MODELING INSURED MORTALITY WITH LOGISTIC q MODELS

Researchers have long been using statistical models to study general population mortality. Since the introduction of Gompertz Law of Mortality (1825), the effort of modeling the human mortality trajectory by age has only accelerated. Thatcher (1999) provided an excellent description and comparison of four mortality-by-age models. With some simplifications in reducing the number of parameters and 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 (1994) $\mu = \frac{\alpha * \exp(\beta * x)}{1 + \alpha * \exp(\beta * x)}$

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 he also confirmed with various population data (Thatcher, Kannisto and Vaupel 1998; Thatcher 1999) that the Kannisto model fits and approximates old-age mortality the best.

We modify the Kannisto model in two ways for insured mortality study: use mortality rate q instead of force of mortality μ as the dependent variable and include not only age but also many other insured explanatory variables. Our logistic mortality model has a general form of

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

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

where

- q is probability of death in an exposure year, given a policyholder survived to the beginning of the year
- x_i are explanatory variables (e.g., age, sex, duration, product)
- α is the 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 (see appendix B).

To distinguish from the logistic force-of-mortality model or *logistic μ model* (1.4) by Kannisto (1994) and Thatcher (1999), let us call our model (3.1) *logistic q model*. According to Thatcher's illustration, a simplified Heligman and Pollard model (1.3) with only one explanatory variable, age, is a special form of our logistic q model.

A logistic q model has many advantages for insured experience studies.

- It models mortality q that is directly used in business operation and risk management.
- It can be flexibly configured 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 and experience table construction that are not simply to do with conventional experience study methods (Harrell 2001).
- It can be developed with widely available commercial software systems 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: Four eras defined by issue year to reflect key underwriting evolutions such as smoker and preferred ratings
- Face category: $50,000–$99,000; $100,000–$499,000; $500,000+ (inflation adjusted)

These variables are selected because they have the 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 policy issue age and policy duration instead of attained age and calendar year to represent age and time. The chosen pair has better reflection of insured characteristics and is the key dimension of insured mortality tables.

Recall that our logistic *q* model has the general form of

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

The right hand side of the model has three components: 1) the intercept α, 2) the main effect component that is a weighted sum of individual explanatory variables, and 3) the interaction component that is a weighted sum of products of two or more explanatory variables.

When interaction terms are omitted, models (3.1) or (3.1a) become

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

and

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

In this model, the model coefficients α and β_i can simply be transformed as estimations for mortality level, slopes and 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 the 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 Society of Actuaries' (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 smoker, male nonsmoker, female smoker and female nonsmoker. 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 without being constrained by the other three groups.

We use SAS software for our data preparation and model development. In the upcoming sections, we highlight the interpretations and usefulness of the three types of outputs from the SAS modeling process.

3.1 Analysis of effects for the mortality driver significance test

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

Table 3.1. Analysis of effects

Pr > ChiSq (p-value)	Degree of Freedom	Female		Male	
		Nonsmoker	Smoker	Nonsmoker	Smoker
Duration[1]	1	<.0001	<.0001	<.0001	<.0001
Issue age[1]	1	<.0001	<.0001	<.0001	<.0001
Study year[2]	1	0.1714	0.4597	0.1719	0.0017
Face band[1]	2	0.0051	0.0040	<.0001	<.0001
Product[3]	1	0.0157	0.9533	0.1363	<.0001
Issue year[1]	2	<.0001	0.0003	<.0001	<.0001
Class[1]	2	<.0001	<.0001	<.0001	<.0001

Several items should be noted.

1. As expected, insured mortality varies significantly statistically 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.

2. Study year, or exposure year, is included as a placeholder for mortality improvement in the 10-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 gender and smoker 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.

3. 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.

4. At 95 percent 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.

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, and 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 the ultimate mortality estimate.

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 percent confidence intervals. For the three continuous variables, the odds ratios estimate average mortality increases 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 percent confidence intervals provide a means to verify the credibility of the corresponding slope or differential estimate.

Table 3.2. Odds ratio estimates

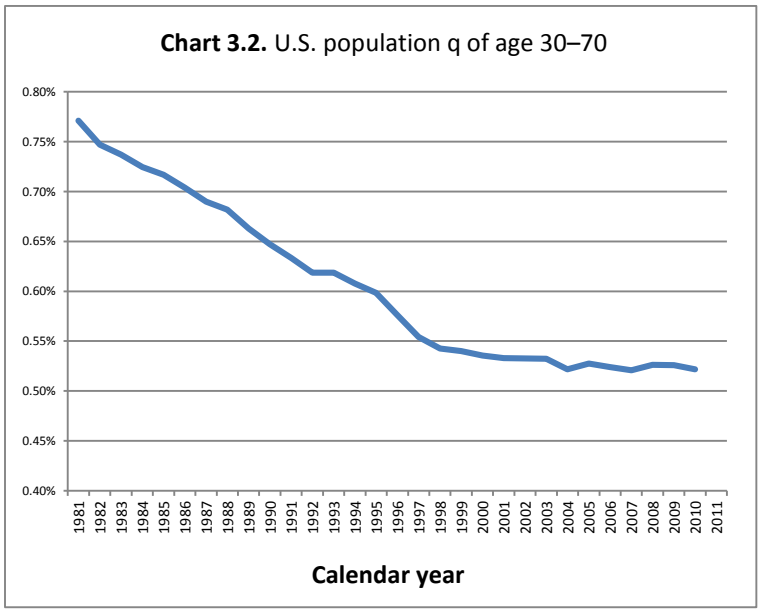
	Male nonsmoker			Male smoker			Female nonsmoker			Female smoker		
Effect	Point estimate	95% Wald confidence limits		Point estimate	95% Wald confidence limits		Point estimate	95% Wald confidence limits		Point estimate	95% Wald confidence limits	
Duration	1.141[1]	1.139	1.143	1.118	1.114	1.122	1.157	1.153	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. nonmed	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

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 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 the male nonsmokers 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 percent per duration and about 10 percent per issue age (odds ratio = 1.14 and 1.10, respectively). The 10 percent per issue age increase is known to be also true for the general population (Thatcher 1999).

2. If everything else were equal, there is a statistically insignificant 0.2 percent annual mortality improvement (odds ratio = 0.998, the 95 percent confidence interval including 1). This finding may seem to be inconsistent with the common thought of higher mortality improvement. There are three possible explanations. First, due to the short time period and inconsistent data contributions from insurers, the study data may have not captured the true insured mortality improvement. Second, in the past decade or so, U.S. population mortality improvement has been leveling off as shown in chart 3.2 (data are from the Human Mortality Database; the age range reflects the most commonly insured ages). This may also be true of 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 for 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 many more mortality variations.



3. If everything else were equal, compared to large policies with face amount at least $500,000, the polices sized between $50,000–$99,000 and $100,000–$499,000 would have 28 percent and 12 percent 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 percent 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 percent 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. 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. The logistic mortality model provides an effective means to perform normalization.

6. If everything else were equal, the mortality of the preferred class would be about 27 percent lower than that of the residual standard class while mortality of the aggregate class (one class plus unknown) is about 4 percent higher (odds ratio = 0.73 and 1.041).

As mentioned before, normalized mortality information is essential in identifying underlying causes and avoiding miscounting the mortality differentiation values when setting pricing factors. Findings of this analysis can also be useful in validating industry tables split from an aggregated table, like the American Council of Life Insurers' 2001 Commissioner's Standard Ordinary (CSO) preferred class structure tables.

3.3 Model fit for overall study reliability measurement

Compared to health or property and casualty insurance claims, mortality claims occur at a much lower frequency and with a much more stable pattern. Relatively scarce claim counts 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 (ROC) curve. Table 3.3 displays the c-statistics for the four models.

Table 3.3. Model fit

Association of predicted probabilities and observed responses	Female		Male	
	Nonsmoker	smoker	Nonsmoker	Smoker
c	0.682	0.753	0.679	0.747

Several items should be noted.

- Vinsonhaler et al. (2001) analyzed private pension plan experience data with a 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 nonsmoker 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 percent

higher than for nonsmokers. This 10 percent gain in death predictability is likely from knowing smoking status.

4. IMPACT OF DEATH CENSORSHIP BY POLICY LAPSATION

Along with the advantages mentioned before, the adaption of a statistical model for insured mortality study brings a new issue that the conventional descriptive 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 five generated claims and the other five died after termination of coverage, the death rate of the group would be 10 percent but claim rate would be only 5 percent. Insured mortality, or claim rate, is conditioned on policy in-force and only reflects claim risk. It is not equivalent to mortality measurement for a general population. 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 a 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 is to modify logistic q model (3.1) and (3.2) to discount possible future policy lapse from contributing claim.

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 (q_c) and in-force (q_i) so that

$$q_l + q_c + q_i = 100\%.$$

With the same explanatory variables x_i as used in (3.2), we can use a multinomial logistic model to model the three probabilities as follows (see Hosmer, Lemeshow and Sturdivant 2013, chapter 8, for more descriptions).

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

Let us call this model a logistic q_c model to emphasize claim-rate estimation. 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 models (3.2) and (4.1), we have

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

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

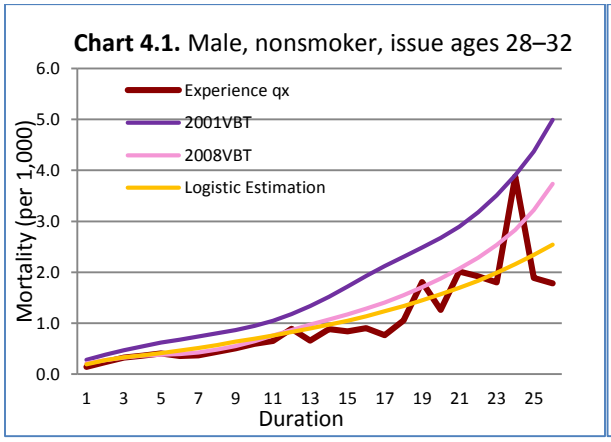
For projection purposes, α_l and α_c are usually related to initial lapse and claim levels; β_{l1} and β_{c1} are related to lapse and claim slopes. A highly simplified interpretation of (4.3) is that, depending on if the death rate asymptotically increases faster than, slower than or equal to the lapse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

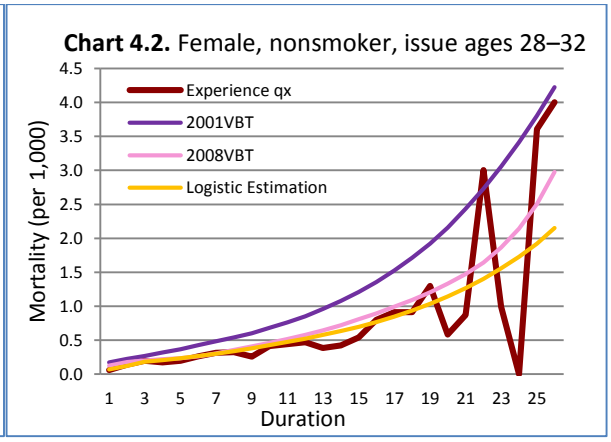
It is understood that insured lapse rates are driven by many long- and short-term factors and do not necessarily have a relationship as regular with duration as claim rate or death rate. The lapse component of model (4.1) may not have as good a fit to insured lapse experience. However, it is reasonable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

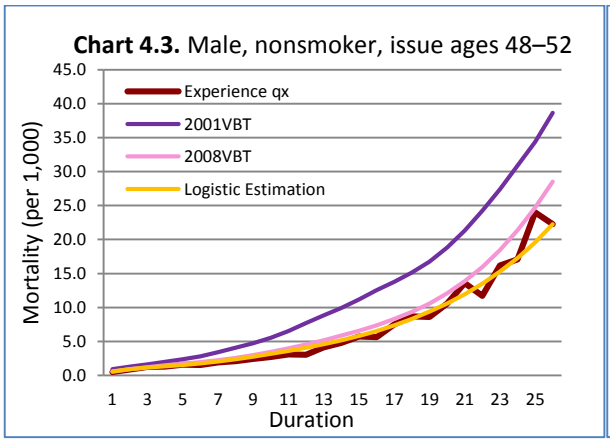
Model (4.1) is especially useful for estimating ultimate mortality. Due to a data-usage agreement issue, we do not have access to the policy-lapse detail for this study and are unable to demonstrate a real output for model (4.1). A follow-up study is planned.

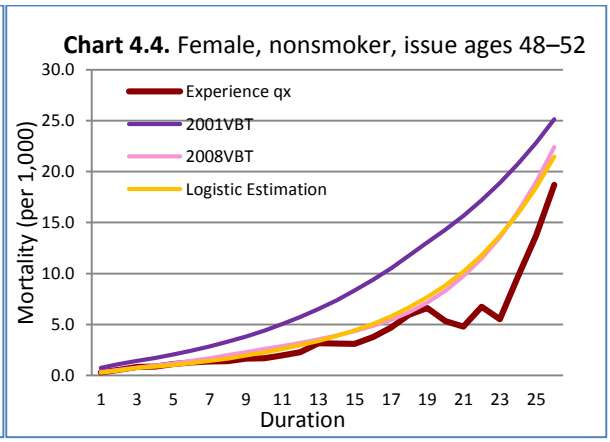
As an alternative, we applied some industry expert opinions on insured ultimate mortality to create a simplified version of model (4.1), used the model to produce model-estimated industry experience tables, and compared the tables with SOA's 2001 and 2008 Valuation Basic Tables (VBTs). The result is very positive (see the following six charts). This supports a point we made earlier: Using less but more relevant data may achieve equal or better results than using more but less relevant data.

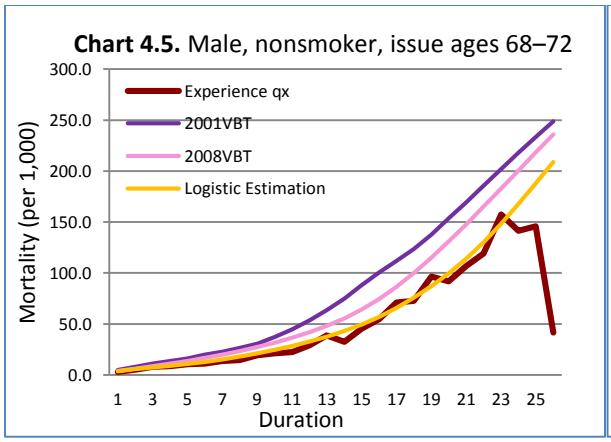
Because this alternative involves various subjective assumptions, it is not presented in detail here. However, we are open to inquiries and interested in discussion.

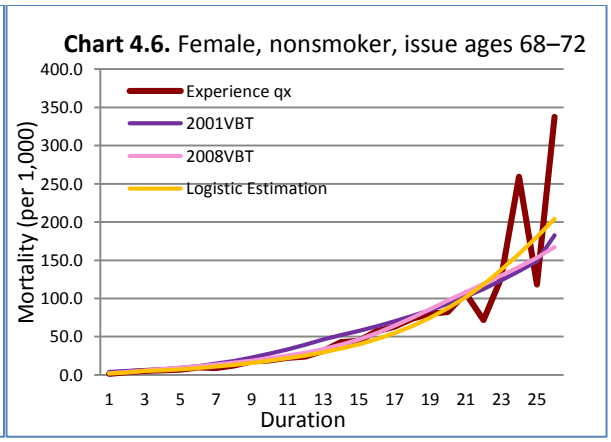












5. CONSTRAINTS AND POSSIBLE ENHANCEMENTS

Among others, three types of biases can occur in an 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 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 has plenty of data biases such as missing data and inconsistent data coding between companies, which 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 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 a 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), a logistic function 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 occur 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 the level period effect for a specific portfolio will need more than a logistic mortality model. However, at an industry aggregated level and for constructing insured mortality tables, our study shows logistic models deliver reasonable results.

4. The current lack of a consistently collected long-term insured experience database is limiting the optimization of any industry experience studies, 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-processing technology and analytical methodology advance, it is our hope 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 much potential for insured mortality experience studies, including

- Testing for statistical strength of mortality drivers in explaining mortality variations with effect analysis
- Generating normalized mortality metrics such as slopes and differentials with odds ratio analysis
- Extrapolating for advanced age or ultimate mortality with modeled estimation
- Bridging or smoothing between select and ultimate mortality with model link function
- Quantifying overall study reliability with model fit statistics
- Helping construct multidimensional experience tables by using the model as a predictive model
- Being 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

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.

Harrell, Frank E. Jr. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer-Verlag.

Heligman, Larry, and John H. Pollard. 1980. "The Age Pattern of Mortality." *Journal of the Institute of Actuaries* 107 (01): 49–80

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, N.J.: John Wiley & Sons Inc.

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

Kannisto, Vaino.1994. Development of Oldest-Old Mortality, 1950-1990: Evidence from 28 Developed Countries. Odense Monographs on Population Aging 1. Odense University Press. Odense, Denmark.

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

Olshansky, S. Jay. 1998. "On the Biodemography of Aging: A Review Essay." *Population and Development Review* 24 (2): 381–93.

Society of Actuaries. 2001 Valuation Basic Table (VBT) Report & Tables. http://www.soa.org/Research/Experience-Study/Ind-Life/Valuation/2001-vbt-report-tables.aspx.

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. Roger, Vaino Kannisto, and James W. Vaupel. 1998. "The Force of Mortality at Ages 80 to 120." Odense Monographs on Population Aging 5, Odense University Press, Odense, Denmark.

Thatcher, A. Roger. 1999. "The Long-Term Pattern of Adult Mortality and the Highest Attained Age." *Journal of the Royal Statistical Society: Series A* 162 (1): 5–43.

Vaupel, James. 2014. "The Advancing Frontier of Human Survival." General Session 1 presentation, Living to 100 Symposium, Jan. 8, Orlando.

Vinsonhaler, Charles, Nalini Ravishanker, Jeyaraj Vadiveloo, and Guy Rasoanaivo. 2001. "Multivariate analysis of Pension Plan Mortality Data." North American Actuarial Journal, Vol 5, Issue 2, 126-135.

Weibull, Waloddi A. 1951. "A Statistical Distribution Function of Wide Applicability." *Journal of Applied Mechanics* 18 (September): 293–97.

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),

(A.2) $$\ln(\frac{q_{female}}{1-q_{female}}) - \ln(\frac{q_{male}}{1-q_{male}}) = \beta_2$$

or

(A.2a) $$e^{\beta_2} = (\frac{q_{female}}{1-q_{female}}) / (\frac{q_{male}}{1-q_{male}})$$

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

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 for comparing mortality of a segment relative to the overall means. See Hosmer, Lemeshow and Sturdivant (2013, chapter 3) for more discussion.

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 in-force, X denotes the vector of explanatory variable X = { x_1, x_2,…,x_k}, and β = { β_1 ,…, β_k } are the coefficients. Then,

$$q = \text{Prob}(y = 1 | X)$$

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). Consult McCullagh and Nelder (1989) for discussions of the methods commonly used by statistical modeling computer software.