Modelling pensioner longevity

A variety of options are open to actuaries who want to model pensioner mortality, including Generalised Linear Models (GLMs) and survival models. This article compares and contrasts these two classes of model, and explains the circumstances when survival models are preferable to GLMs

BY S.J.RICHARDS

WITH A BACKDROP OF SUBDUED INFLATION, lower long-term interest rates and generally lower investment returns, the importance of pensioner longevity has increased greatly in recent years. In the United Kingdom, the growth of money-purchase private pensions since 1988 has meant that the volume of funds seeking an annuity now exceeds £7 billion per annum (Source: ABI). However, by far the biggest impact is in defined-benefit ('final-salary') pension schemes provided by employers. These schemes carry perhaps around £1 trillion of liabilities to current and future pensioners in the UK, so changes in life expectancy have a large macro-economic impact on the private sector.

A pension (or an annuity) is a liability to pay a set amount to the pensioner each year until death. In this sense, annuities from life-insurance companies and pensions from occupational schemes are identical liabilities. How long these pensioners live is clearly a crucial factor in the ultimate cost of providing pensions. There are two elements to assessing life expectancy: first, and most important, the current level of mortality; and second, the possible future path of mortality rates. This article is concerned with the first element, namely measuring the current level of mortality. It explains the statistical methods currently in use amongst UK life insurers, and their connection to well-known actuarial mortality laws. The article spells out some of the limitations of GLMs, and describes the survival-analysis techniques the life industry is switching to.

Table 1: Impact of selected risk factors on pension reserves			
Factor	**Step change**	**Reserve**	**Change**
Base case	–	16.97	–
Gender	Female-male	15.02	-11.5%
Lifestyle	Top-bottom	13.31	-11.4%
Duration	Short-long	11.88	-10.8%
Pension size	Large- small	11.18	-5.9%
Region	South-north	10.52	-5.9%
Overall	–	–	-38.0%
Source: Richards and Jones (2004), p39. Annuities payable annually in advance valued at net 2.5% interest p.a.			

Table 2: Death counts by age band and gender		
Age band	**Females**	**Males**
60-40	3	9
65-69	3	14
70-74	19	22
75-79	26	33
80-84	19	23
85-89	20	11
90-94	19	7

Financial significance of longevity differentials

Richards and Jones (2004) presented results on the financial impact of various risk factors for longevity (see Table 1). This table shows the change in annuity reserve which results from a single stepwise change in a risk factor. The base case is a high-income, high-social-status female living in the South of England, and each step away from this base case shows the change in the size of the reserve. For example, the impact of changing the gender in the base case in Table 1 has been to reduce the annuity reserve by 11.5%. Bearing in mind that an insurer's profit at the time of writing is perhaps around 4-5% of premium, a change in any one of the listed risk factors will have a material impact on the profitability of annuity business. Equally, for an occupational pension scheme it is very important to know the socio-economic mortality differentials (here 'Lifestyle') to properly reserve for pension costs.

Models of Group Mortality

One of the simplest approaches to modelling mortality is to look at groups. First consider the situation with a group of n_x identical lives all aged exactly x. The number of deaths observed can be treated as a random variable, D_x, which is distributed binomially with probability of death q_x:

$$D_x \sim \text{Binomial}\,(n_x, q_x) \qquad (1)$$

Another common alternative for population data is where we do not have n_x but instead have the central exposed-to-risk of death aged $x+\frac{1}{2}$, denoted $E^c_{x+½}$. In this case, the number of deaths aged x last birthday is treated as a random variable with Poission distribution:

Table 3: Death counts by age band, gender and membership category				
Age band	**Females Group A**	**Females Group B**	**Males Group A**	**Males Group B**
60-64	0	3	3	6
65-69	2	1	11	3
70-74	13	6	19	3
75-79	12	14	27	6
80-84	12	7	14	9
85-89	10	10	5	6
90-94	11	8	7	0

$$D_x \sim \text{Poisson}\left(E^c_{x+\frac{1}{2}}\mu_{x+\frac{1}{2}}\right) \qquad (2)$$

where $\mu_{x+½}$ is the force of mortality at age $x+\frac{1}{2}$. Statisticians call μ_x the *hazard* and E^c_x the waiting time.

If the data set is large enough, these sorts of model can work well. Indeed, for population data sets, this sort of model is often the only option as population data is usually only available in grouped form. Such models also work well for organisations such as the CMI in the U.K., which collects data from many life offices and publishes grouped data for assured lives going back to 1947.

For the insured data sets for a particular portfolio, however, this approach is less satisfactory for a number of reasons. As the data volumes are usually very much smaller, it is often necessary to group into age bands instead of individual ages, thus creating a so-called contingency table. A common approach is to use five-year age bands, which obviously involves loss of information and means mortality is being modelled at the mid-point of a range of mortality rates.

The consequence of grouping is that it limits the number of risk factors which can be investigated. This is not a problem for population-based data, which is usually only available split by gender, but it is a serious limitation for insured data sets where one is very interested in a variety of other risk factors. In a contingency table, each cell represents a specific combination of risks: in order to perform tests of model fit, a minimum expected number of deaths in each cell is required (usually five in each cell). For many portfolios, this restricts the number of risk factors which can be investigated using group mortality. This is shown in Table 2 for a medium-sized pension scheme, where we see how the 228 deaths in the pension scheme were distributed by age band and gender. We can see that even with just two risk factors – age and gender – the contingency table has some difficulties. There are fewer than five deaths for females in age bands 60-64 and 65-69, and the expected values are therefore likely to be on the borderline of acceptability for applying any tests of model fit, such as a χ^2 test. However, for this pension scheme there are also two categories of membership on which we have data, groups A and B. Table 3 shows that a contingency-table approach simply will not be valid, even though we only wish to investigate three risk factors. The only way this could be made to work with the current data would be to use fewer cells with wider age bands, such as collapsing the first two age ranges into 60-69 and the last two into 85-94. This would obviously throw away even

Table 4: Some actuarial mortality laws and their equivalence as a generalised linear model

Mortality law	Form for q_x	Link name	Link function
Gompertz (1825)	$\exp(\alpha+\beta x)$	logarithm	$\log q_x$
Perks (1932, simplified)	$\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$	logit	$\log\left(\frac{q_x}{1-q_x}\right)$
Extreme value	$1-exp(-exp(\alpha+\beta x))$	complementary log-log	$\log(-\log(1-q_x))$
Reversed extreme value	$\exp\left(\frac{-1}{\exp(\alpha+\beta x)}\right)$	log-log	$\log(-\log q_x)$
Probit-Gompertz	$\Phi(\alpha+\beta x)$	probit	$\Phi^{-1}(q_x)$

more information on age. The alternative is to add more years of experience, which might simply not be available.

One solution to the problem of minimum expected cell deaths and information loss is to model mortality at the level of each individual, not of the group. This approach is not available to users of population data sets, as only grouped death counts and exposures are generally available. However, this approach is well-suited to portfolios of pensioners and annuitants: not only is the exact age and gender known for each pensioner, but also the exact date of death and a number of other relevant individual variables as well, including postcode, product type or membership category. Since the requirement for a minimum number of expected deaths vanishes in an individual model, this enables us to model an unlimited number of risk factors simultaneously for any size of scheme. There are two approaches we can adopt for modelling mortality at the level of the individual: to estimate q_x in a Bernoulli model (usually as a GLM), or to estimate μ_x in a survival model.

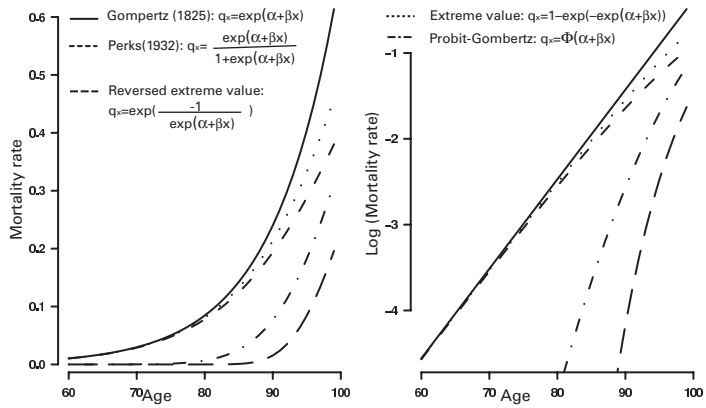
Models of individual mortality: GLMs

Actuaries are most familiar with the one-year mortality rate, denoted ${}_1q_x$ or just simply q_x. This is the probability of a life aged exactly x dying before reaching age x+1. GLMs for individual mortality are typically based around the Bernoulli model using q_x (the Bernoulli model is simply the Binomial model with n=1). Richards and Jones (2004) presented the results of such a generalised linear model for pensioner mortality in a large portfolio of annuitants. The model chosen was logistic regression, which was the best of a variety of choices of GLM. The basic idea behind logistic regression can be expressed as:

$$\log\left(\frac{q_x}{1-q_x}\right)=\alpha+\beta x \qquad 3$$

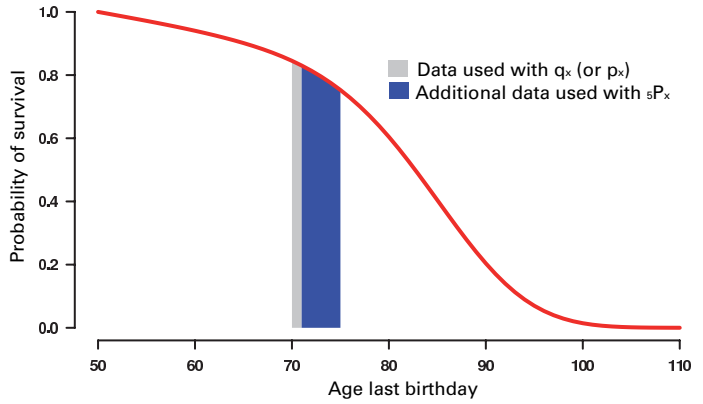
where q_x is the probability that a life now aged x will die before reaching age x+1. The expression on the right is known in GLM terms as the *linear predictor*, as it is a linear combination of age and a constant. The function on the left is called the *log-odds ratio*, or *logit*. In GLM terms, any function which relates the mortality measure to the linear predictor like this is called the link function. Equation 3 can be re-arranged to give a formula for q_x:

Figure 1. Five mortality laws for qx, shown on original scale (left panel) and on log scale (right panel)



The key point about the log-scale plot is that only the Gompertz mortality law has a constant proportionate increase in mortality with age, i.e. a straight line on the log scale. The other laws all demonstrate a slowing down in the increase of the mortality rate with age, the so-called late-life mortality deceleration. Source: Richards (2006).

Figure 2. Survival function ${}_{x-50}p_{50}$ according to PMA00



The lighter grey area shows the data used when modelling q_x, while the darker blue area shows the extra data used when modelling the t-year survival probability, ${}_tp_x$. Source: Longevitas Ltd, using CMIB data.

$$q_x=\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \qquad 4$$

which is a simplified version of the actuarial 'law' proposed by Perks (1932). Richards (2007) describes a number of mortality laws which can all be expressed in GLM terms (see Table 4). In this table, $\Phi()$ is the distribution function of the $N(0, 1)$ distribution, and $\Phi^{-1}()$ is the inverse. With the possible exception of the reversed extreme-value law, any one of these models might be chosen to represent pensioner mortality, as illustrated in Figure 1.

Table 5: Width of 95% confidence intervals relative to parameter estimates for a sample five-year data set				
Model type	**Intercept ($\alpha_{baseline}$)**	**Age ($\beta_{baseline}$)**	**Gender (α_{male})**	**Age: Gender (β_{male})**
(a) GLM	0.153	0.206	1.073	1.388
(b) Survival model	0.038	0.053	0.203	0.239
(b) as % of (a)	25%	26%	19%	17%

Models of individual mortality: survival models

An alternative measure to q_x is the *instantaneous hazard rate*, known to actuaries as the force of mortality, μ_x, and defined as:

5
$$\mu_{x+t} = \lim_{dt \to 0+} \frac{\Pr\ (T_x \in [t, t+dt))}{dt}$$

where $T_x \geq 0$ is a random variable representing the future lifetime of a life now aged x. For small dt, Equation 5 is effectively saying:

6
$$\mu_x \approx \frac{{}_{dt}q_x}{dt}$$

where ${}_{dt}q_x$ is the probability of dying in the small interval of time dt. Actuaries working with longevity risk are naturally more focused on survival probabilities, rather than mortality per se. The t-year survival probability, ${}_tp_x$, is given by:

7
$${}_t p_x = e^{-H(t)}$$

where $H(t)=\int_0^t \mu_{x+s} ds$ is called the *integrated hazard function*. Equation 7 therefore gives us a simple relationship between the rate of mortality, q_x, and the force of mortality, μ_x:

8
$$q_x = 1 - \exp\left(-\int_0^1 \mu_{x+s} ds\right)$$

As with q_x, a functional form for the force of mortality can be used for pensioner data. Below is the Gompertz (1825) model:

9
$$\mu_x = \exp(\alpha + \beta x)$$

which can be written in a log-linear form:

10
$$\log \mu_x = \alpha + \beta x$$

The right-hand side of Equation 10 is identical to the right-hand side of Equation 3, so here we can explain how the individual models of mortality – both GLMs and survival models – differ from the group models. Simply put, each life i has its own specific parameters which describe its own combination of risks, that is α_i and β_i instead of a group α and β. Thus:

11
$$\alpha_i = \sum_{j=1}^{m} z_{ij}\alpha_j$$
$$\beta_i = \sum_{j=1}^{m} z_{ij}\beta_j$$

where there are m components (factors) to the overall risk, each α_j and β_j is a parameter estimating a particular risk component (α_j) and its interaction with age (β_j), and z_{ij} is a binary indicator variable taking the value 1 when life i belongs to the group with risk factor j and the value 0 otherwise. For example, in a model with risk factors for both gender and smoker status:

12
$$\alpha_i = \alpha_{baseline} + z_{i,\,male}\alpha_{male} + z_{i,\,smoker}\alpha_{smoker}$$
$$\beta_i = \beta_{baseline} + z_{i,\,male}\beta_{male} + z_{i,\,smoker}\beta_{smoker}$$

Note that the model is structured as measuring differences from a baseline profile. In the model specified in Equation 12, the baseline is a female non-smoker, while the model parameters measure male mortality as a departure from the female baseline, and smoker mortality is measured as a departure from the non-smoker baseline. The $z_{i,\,male}$ are zero-one indicator variables for whether life i is male, and the $z_{i,\,smoker}$ are similar zero-one indicators for whether a life is a smoker.

By way of illustration, Table 1 shows the results of an individual-level GLM with six risk factors: age, gender, lifestyle, duration since retirement, region and pension size. There is no limit to the number of different risk factors which can be investigated when using individual-level models of mortality, whether this is done as a GLM or a survival model. This is the crucial advantage of models for individuals instead of groups.

Advantages of survival models

Thus far we have seen that models for individual-level mortality are better-suited to insured datasets than models of group mortality. We have also seen that such individual-level models can be either GLMs (for q_x) or survival models (for μ_x). In this section we explain why survival models are usually preferable.

A full survey of models for actuarial use is given by Macdonald (1996a, 1996b, 1996c), who also gives wider justification for using survival models for μ_x in preference to modelling q_x. However, we will focus on pensioner mortality in this section. One common feature of pension schemes and annuity portfolios is the nature of the data. People enter the portfolio more or less constantly, and the exact date of entry is known, as is the exact date of death when it occurs. Exact dates of birth are usually also known, as are gender, pension size and a number of other potentially useful pieces of information. Thus, for each pensioner their survival from start to finish is usually known exactly. Intuitively, we want to use all this data, not just part of it.

Another common feature is that we are not so interested in pensioners' mortality as much as their longevity. A more natural focal point is therefore not the rate of mortality, q_x, or even the force of mortality, μ_x, but the survival function, ${}_tp_x$. We saw in Equation 7 that there was a simple and direct relationship between μ_x and ${}_tp_x$ for any real value $t>0$.

Figure 2 shows the information contributing to a model of q_x: a single age's exposure and the deaths (if any) which are observed at that age.

Regardless of just how much observed data you might have for an individual, only a single year's exposure can actually be used in logistic regression (or any of the other individual-level GLMs). This is because if ${}_1q_x$ can be transformed into a linear function (as in Equation 3), then ${}_2q_x$ and other multi-year mortality probabilities cannot. The alternative suggestion – simply putting each individual into the data set several times for q_x, q_{x+1},... etc, will violate the crucial independence assumption on which the model is based and produce unreliable results.

Figure 2 also shows the information contributing to a model of ${}_tp_x$, which clearly benefits from both greater exposed-to-risk and a larger number of deaths. Thus, if we can switch from modelling mortality rates to modelling survival, we can use all available data and thus boost the overall power of the model. The result is not only efficient use of all available data, but also better estimation and smaller standard errors.

The power of the individual survival model manifests itself in a number of ways relative to the individual GLM, as shown in Table 5:

(i) The standard errors for a survival model are usually a fraction of their GLM equivalent, thus yielding more-accurate parameter estimation. Table 5 shows the results of a GLM and survival model applied to the same data set, showing the smaller relative confidence intervals (and therefore greater accuracy) of the survival model.

(ii) This greater accuracy enables the user to find mortality differentials (and thus risk factors for pricing and reserving) which would otherwise be hidden from a GLM.

(iii) The 'critical mass' of data required for a survival model is measured in terms of life-years of exposure overall, not lives in a single year. While a few insurers will have enough data to work with GLMs, survival analysis puts the same power within reach of many more than just the very largest insurers.

(iv) The greater stability of estimates from using data spanning different calendar years, thus minimising the impact of period effects.

Conclusions

For insured data sets, modelling mortality at the level of the individual is preferable to modelling the mortality of groups. Both GLMs and survival models are capable of individual-level modelling, but GLMs do not make efficient use of all the data usually available and require larger data sets to achieve the same results. In contrast, survival models for ${}_tp_x$ and μ_x make full use of every piece of information, and permit more stable estimates than the equivalent GLM.

Stephen Richards is the developer of the Longevitas system for modelling demographic risks. He can be contacted at stephen@richardsconsulting.co.uk or +44(0)131 315 4470

References

Gompertz, B. (1825).
The nature of the function expressive of the law of human mortality,
Philosophical Transactions of the Royal Society.

Macdonald, A. S. (1996a).
An actuarial survey of statistical models for decrement and transition data, I: multiplestate, Poisson and Binomial models,
British Actuarial Journal, 2 (1) 129–155.

Macdonald, A. S. (1996b).
An actuarial survey of statistical models for decrement and transition data, II: competing risks, non-parametric and regression models,
British Actuarial Journal, 2(2) 429–448.

Macdonald, A. S. (1996c).
An actuarial survey of statistical models for decrement and transition data, III: counting process models,
British Actuarial Journal, 2(3) 703–726.

Perks, W. (1932).
On some experiments in the graduation of mortality statistics,
Journal of the Institute of Actuaries, Vol. 63, pp.12–40.

R Development Core Team (2007).
R: A language and environment for statistical computing,
R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org.

Richards, S. J. and Jones, G. L. (2004).
Financial aspects of longevity risk,
Staple Inn Actuarial Society (SIAS).

Richards, S. J. (2007).
Detecting Year-of-Birth Mortality Patterns with Limited Data,
Journal of the Royal Statistical Society, Series A (to appear).