

DISCUSSION PAPER

2016/45

Risk Classification in Life Insurance:
Extension to Continuous Covariates

M. Denuit

RISK CLASSIFICATION IN LIFE INSURANCE: EXTENSION TO CONTINUOUS COVARIATES

Michel Denuit and Catherine Legrand
Institut de Statistique, Biostatistique et Sciences Actuarielles
UCL, Louvain-la-Neuve, Belgium
`michel.denuit@uclouvain.be` and `catherine.legrand@uclouvain.be`

December 20, 2016

Abstract

This short note supplements the paper by Gschlossl et al. (2011) with an efficient method allowing actuaries to include continuous covariates in their life tables, such as the sum insured for instance. Compared to the classical approach based on grouped data adopted in the majority of actuarial mortality studies, individual observations recorded at the policy level are included in the Poisson regression model. The proposed procedure avoids any preliminary, subjective banding of the range of continuous covariates that may bias the resulting life tables.

Key words and phrases: Life tables, Poisson regression, GLM, GAM.

1 Introduction

Actuaries are aware that a lot of heterogeneity is present in policyholders' survival data. Within insurance portfolios, differences in mortality exist for instance with regard to gender and to social class, as assessed through occupation, income or education. The reader is referred e.g. to Brown and McDaid (2003), Kwon and Jones (2006) or Fong (2015) for examples of risk factors affecting mortality.

The Poisson regression model has been used by actuaries as a method for quantifying the impact of such differences in mortality. Early references include Renshaw (1988, 1991) and Haberman and Renshaw (1990). However, such studies have been conducted on grouped data, where numbers of deaths are recorded in a given risk class, defined by a combination of categorical covariates, from a known exposure-to-risk. Now that individual mortality data are readily available, there is a need for new tools to perform studies at the policy level, to avoid any preliminary, subjective banding of the range of continuous covariates that may bias the resulting life tables.

There are different approaches to graduate death rates, ranging from purely parametric models (like Gompertz, Makeham or Heligman-Pollard laws, for instance) to fully nonparametric ones consisting in smoothing the observed mortality, possibly with the help of a reference life table (relational model). As the parametric approach is subject to the risk of misspecification, actuaries now tend to favor the nonparametric alternatives. There are essentially two kinds of nonparametric approaches for analyzing mortality statistics:

1. either actuaries assume that the force of mortality is piecewise constant and then resort to Poisson regression techniques for grouped data as described in Gschlossl et al. (2011). The response is here the number of deaths observed within a group of policyholders sharing the same characteristics.
2. or the force of mortality is assumed to be smooth (without jumps as under the piecewise constantness hypothesis) and nonparametric techniques based on integral products and Nelson-Aalen estimators are applied, instead of the more conventional Poisson regression approach. We refer the reader e.g. to Guibert and Planchet (2014) for a convincing application of this second approach.

The numerical treatment of the data is more complex under the second approach and this restricts the integration of covariates in the analysis (in general, only categorical covariates are included in the life table so that continuous ones require preliminary banding). In this paper, we show that Poisson regression can also be used to study individual survival times and allows the actuary to deal with large data sets and to include continuous covariates (such as the sum insured, for instance).

Actuarial studies generally involve a large number of individuals. As pointed out by Gschlossl et al. (2011), the Poisson regression model can be used for risk classification in life insurance, without loss of generality. This is particularly important for applications because the Poisson distribution belongs to the exponential dispersion family for which efficient computational tools are widely available. We refer the reader e.g. to Denuit and Lang (2004) as well as Klein et al. (2014) for insurance data analyses in large nonlife portfolios

and to Gschlossl et al. (2011) and Tomas and Planchet (2014) for mortality statistics analyses in life insurance.

Compared to previous actuarial studies adopting the Poisson regression approach, which only used grouped observations and categorical covariates, we show in this paper how to work with individual data, including continuous covariates in the life tables. This allows the actuary to let the force of mortality depend on the sum insured for instance, without preliminary, subjective banding. The remainder of this paper is organized as follows. In Section 2, we describe the mortality model. In Sections 3-4, we demonstrate how to use Poisson regression in the context of actuarial mortality studies to exploit individual survival data. The final Section 5 discusses the proposed approach and concludes the paper.

2 Mortality model

Assume that a life insurance portfolio has been observed during a given period of time (typically, 3 to 5 years). Policyholder i , $i = 1, 2, \dots, n$, has been observed from age a_i to age b_i . Here, a_i may be the policyholder's age at the beginning of the observation period in case of an existing contract, or the age at entry in the portfolio for a new contract, issued during the observation period. The policyholder stopped being observed at age b_i , either because of death at that age, because the observation period terminated or because the policyholder left the portfolio due to policy cancellation or contract arriving at maturity. Notice that actuaries analyze mortality in function of attained age, so that age at death is the variable of interest.

In addition to attained age, we have at our disposal a set of possible risk factors denoted as $\mathbf{z}_i$ for policyholder i . The force of mortality is a function of attained age x and of risk factors in $\mathbf{z}_i$. It is denoted as $\mu(x|\mathbf{z}_i)$ and is assumed to be piecewise constant, i.e.

$$\mu(x + \xi|\mathbf{z}_i) = \mu(x|\mathbf{z}_i) \text{ for all integer } x \text{ and } 0 \leq \xi < 1. \quad (2.1)$$

Thus, we assume that the force of mortality is constant over each year of age, but allowed to vary between ages, in accordance with actuarial practice. Let us nevertheless mention that we could use arbitrarily small age intervals (months, weeks or even days), subject to data availability.

3 Likelihood

Let δ_i be the death indicator for policyholder i , i.e.

$$\delta_i = \begin{cases} 1 & \text{if policyholder } i \text{ dies during the observation period} \\ 0 & \text{otherwise.} \end{cases}$$

The contribution of policyholder i to the likelihood is then given by

$$\ell_i = \exp \left(- \int_{a_i}^{b_i} \mu(\xi|\mathbf{z}_i) d\xi \right) (\mu(b_i|\mathbf{z}_i))^{\delta_i}.$$

Under assumption (2.1), the integrated force of mortality appearing in the exponential function can be further simplified into a sum over each year of age between a_i and b_i , as shown next.

Given a real number ξ , let $\lfloor \xi \rfloor$ denote ξ rounded from below and let $\lceil \xi \rceil$ denote ξ rounded from above. Precisely, $\lfloor \xi \rfloor$ is the largest integer that is smaller than, or equal to ξ and $\lceil \xi \rceil = \lfloor \xi \rfloor + 1$ is the smallest integer that is larger than, or equal to ξ . If $\lfloor a_i \rfloor < \lfloor b_i \rfloor$ then the integral appearing in ℓ_i can be splitted as follows:

$$\begin{aligned} \ell_i = & \exp \left(- \int_{a_i}^{\lceil a_i \rceil} \mu(\xi | \mathbf{z}_i) d\xi \right) \\ & \exp \left(- \sum_{k=\lceil a_i \rceil}^{\lfloor b_i \rfloor - 1} \int_k^{k+1} \mu(\xi | \mathbf{z}_i) d\xi \right) \\ & \exp \left(- \int_{\lfloor b_i \rfloor}^{b_i} \mu(\xi | \mathbf{z}_i) d\xi \right) (\mu(b_i | \mathbf{z}_i))^{\delta_i}. \end{aligned}$$

with the convention that the sum over k is equal to 0 if $\lceil a_i \rceil = \lfloor b_i \rfloor$. Now, the force of mortality appearing in each integral is constant in accordance with our assumption (2.1), so that the contribution of policyholder i to the likelihood can be written as

$$\begin{aligned} \ell_i = & \exp \left(- (\lceil a_i \rceil - a_i) \mu(\lfloor a_i \rfloor | \mathbf{z}_i) \right) \prod_{k=\lceil a_i \rceil}^{\lfloor b_i \rfloor - 1} \exp \left(- \mu(k | \mathbf{z}_i) \right) \\ & \exp \left(- (b_i - \lfloor b_i \rfloor) \mu(\lfloor b_i \rfloor | \mathbf{z}_i) \right) (\mu(\lfloor b_i \rfloor | \mathbf{z}_i))^{\delta_i}, \end{aligned}$$

with the convention that the product over k is equal to 1 if $\lceil a_i \rceil = \lfloor b_i \rfloor$. If $\lfloor a_i \rfloor = \lfloor b_i \rfloor$ then this contribution reduces to

$$\ell_i = \exp \left(- (\lceil a_i \rceil - a_i) \mu(\lfloor a_i \rfloor | \mathbf{z}_i) \right) (\mu(\lfloor a_i \rfloor | \mathbf{z}_i))^{\delta_i}.$$

Assuming independent lifetimes, the likelihood is then obtained by multiplying the individual contributions ℓ_i over all policyholders $i = 1, \dots, n$.

4 Independent Poisson counts

Let us now relate the likelihood obtained in the preceding section to Poisson distributed random variables. To this end, define independent random variables $N_k^{(i)}$, $k = \lfloor a_i \rfloor, \dots, \lfloor b_i \rfloor$, that are assumed to be Poisson distributed with respective means

$$\mathbb{E}[N_k^{(i)}] = \begin{cases} (\lceil a_i \rceil - a_i) \mu(\lfloor a_i \rfloor | \mathbf{z}_i) & \text{for } k = \lfloor a_i \rfloor, \\ \mu(k | \mathbf{z}_i) & \text{for } k = \lceil a_i \rceil, \dots, \lfloor b_i \rfloor - 1, \\ (b_i - \lfloor b_i \rfloor) \mu(\lfloor b_i \rfloor | \mathbf{z}_i) & \text{for } k = \lfloor b_i \rfloor. \end{cases}$$

Hence,

$$P[N_k^{(i)} = 0] = \begin{cases} \exp(-(\lceil a_i \rceil - a_i)\mu(\lfloor a_i \rfloor | \mathbf{z}_i)) & \text{for } k = \lfloor a_i \rfloor, \\ \exp(-\mu(k | \mathbf{z}_i)) & \text{for } k = \lceil a_i \rceil, \dots, \lfloor b_i \rfloor - 1, \end{cases}$$

and for $k = \lfloor b_i \rfloor$,

$$P[N_{\lfloor b_i \rfloor}^{(i)} = \delta_i] = \exp(-(b_i - \lfloor b_i \rfloor)\mu(\lfloor b_i \rfloor | \mathbf{z}_i))((b_i - \lfloor b_i \rfloor)\mu(\lfloor b_i \rfloor | \mathbf{z}_i))^{\delta_i}.$$

This shows that ℓ_i can be rewritten as

$$\ell_i = \left(\prod_{k=\lceil a_i \rceil}^{\lfloor b_i \rfloor - 1} P[N_k^{(i)} = 0] \right) \frac{P[N_{\lfloor b_i \rfloor}^{(i)} = \delta_i]}{(b_i - \lfloor b_i \rfloor)^{\delta_i}}$$

with the convention that the product over k is equal to 1 if $\lfloor a_i \rfloor = \lfloor b_i \rfloor$. Hence, the contribution of each policyholder to the likelihood can be written as the product of Poisson probabilities, up to the factor $(b_i - \lfloor b_i \rfloor)^{-\delta_i}$. Therefore, we are allowed to perform inference using Poisson regression provided we convert the unique observation $(a_i, b_i, \delta_i, \mathbf{z}_i)$ related to policyholder i into a sequence of independent Poisson counts $N_k^{(i)}$, $k = \lfloor a_i \rfloor, \dots, \lfloor b_i \rfloor$, that are all equal to 0, except possibly the last one that equals $\delta_i \in \{0, 1\}$. Formulating the inference problem in terms of Poisson regression is important for practical purposes because tools performing GLM and GAM analyses are widely available and computationally efficient, not to mention that actuaries throughout the world are now used to conduct this kind of regression study.

In practice, the record $(a_i, b_i, \delta_i, \mathbf{z}_i)$ related to policyholder i in the available data basis is replaced with a block of $\lfloor b_i \rfloor - \lfloor a_i \rfloor + 1$ records $(N_k^{(i)}, \delta_i, \mathbf{z}_i)$, $k = \lfloor a_i \rfloor, \dots, \lfloor b_i \rfloor$. The Poisson regression analysis is then conducted on the responses $N_k^{(i)}$, assuming their mutual independence. Typical actuarial mortality studies are performed on data gathered during 3 to 5 years so that the expanded data basis on which Poisson regression is conducted is three to five times bigger compared to the initial one. Even with large portfolios, this is not expected to be a problem as Poisson regression techniques can deal with very large data sets.

Example 4.1. Assume that we only have the sum insured z_i at our disposal. Then, the following models could be considered. The general specification $\ln \mu(x|z) = f(x, z)$ allowing for all interactions between age x and sum insured z is often simplified into an additive decomposition of the form

$$\ln \mu(x|z) = f_1(x) + f_2(z) \tag{4.1}$$

where the functions f_1 and f_2 are left unspecified but assumed to be smooth. These functions can be estimated from the portfolio mortality experience, using local polynomial techniques or spline representations, for instance.

Often, mortality studies are conducted using a reference life table, corresponding to the market where the insurer operates. In such a case, the force of mortality is expressed as

$$\ln \mu(x|z) = f_1(\ln \mu_x^{\text{ref}}) + f_2(z) \tag{4.2}$$

in terms of a set of reference death rates μ_x^{ref} (treated as known constants satisfying (2.1)). Notice that this reference life table can be distorted by the function f_1 to better reflect portfolio experience. Even if there is, *stricto sensu*, no difference between the specifications (4.1)-(4.2), as μ_x^{ref} is itself a function of age, the second one performs generally much better in empirical illustrations as it suffices to distort the curve $x \mapsto \ln \mu_x^{\text{ref}}$ which looks similar to the portfolio experience life table.

Often, the estimated function f_1 in (4.2) appear to be approximately linear so that a linear hazard transform model can be used instead:

$$\ln \mu(x|z) = \beta_0 + \beta_1 \ln \mu_x^{\text{ref}} + f_2(z).$$

We refer the interested reader e.g. Brouhns et al. (2002) for empirical evidence supporting the linearity of f_1 in (4.2).

5 Discussion

Most often, mortality statistics are aggregated over groups of policyholders sharing the same characteristics, such as age, gender, type of product, etc. Death counts originating from given exposures can then be studied using Poisson regression techniques, as explained in Gschlossl et al. (2011). When continuous covariates are included in the analysis, such as the sum insured for instance, this aggregation requires a preliminary, subjective banding. The approach proposed in this paper avoids this problematic step and allows the actuary to analyze individual mortality data using Poisson regression. This requires to augment the data basis by splitting each individual observation $(a_i, b_i, \delta_i, z_i)$ into a set of $\lfloor b_i \rfloor - \lfloor a_i \rfloor + 1$ independent realizations of Poisson counts $N_k^{(i)}$, $k = \lfloor a_i \rfloor, \dots, \lfloor b_i \rfloor$, sharing the same time-invariant explanatory variables, or letting these covariates evolve over time in case they are dynamic (such as attained age, for instance). Given the wide availability of commercial and open-source softwares dealing with Poisson regression, this makes the majority of actuaries in a position to analyze individual mortality statistics by means of familiar tools.

To end with, it is worth mentioning that transition rates in multistate models for life and health insurance can be analyzed in a way similar to mortality rates. The reason is that the multistate model likelihood is proportional to a product of Poisson likelihoods when the transition rates are assumed to be piecewise constant. The techniques proposed in the present note are thus also helpful to graduate transition rates in the presence of covariates.

Acknowledgements

The present work originated from interesting discussions with Jérémy Goris, Marine Habart and Tom Popa from the Retirement and Pension Funds team of the AXA Group, based in Paris, France. The authors wish to thank the AXA team for these fruitful interactions that have inspired the present note to a large extent.

References

- Brouhns, N., Denuit, M., Vermunt, J.K. (2002). A Poisson log-bilinear approach to the construction of projected life tables. *Insurance: Mathematics and Economics* 31, 373-393.
- Brown, R. L., McDaid, J. (2003). Factors affecting retirement mortality. *North American Actuarial Journal* 7, 24-43.
- Denuit, M., Lang, S. (2004). Nonlife ratemaking with Bayesian GAMs. *Insurance: Mathematics and Economics* 35, 627-647.
- Fong, J. H. (2015). Beyond age and sex: Enhancing annuity pricing. *Geneva Risk and Insurance Review* 40, 133-170.
- Gschlossl, S., Schoenmaekers, P., Denuit, M. (2011). Risk classification in life insurance: Methodology and case study. *European Actuarial Journal* 1, 23-41.
- Guibert, Q., Planchet, F. (2014). Construction de lois d'expérience en présence d'évènements concurrents: Application à l'estimation des lois d'incidence d'un contrat dépendance. *Bulletin Français d'Actuariat* 14, 5-28.
- Haberman, S., Renshaw, A.E. (1990). Generalised linear models and excess mortality from peptic ulcers. *Insurance: Mathematics and Economics* 9, 21-32.
- Kwon, H. S., Jones, B. L. (2006). The impact of the determinants of mortality on life insurance and annuities. *Insurance: Mathematics and Economics* 38, 271-288.
- Klein, N., Denuit, M., Lang, S., Kneib, Th. (2014). Nonlife ratemaking and risk management with Bayesian additive models for location, scale and shape. *Insurance: Mathematics and Economics* 55, 225-249.
- Renshaw, A.E. (1988). Modelling excess mortality using GLIM. *Journal of the Institute of Actuaries* 115, 299-315.
- Renshaw, A.E. (1991). Actuarial graduation practice and generalized linear and non-linear models. *Journal of the Institute of Actuaries* 118, 295-312.
- Tomas, J., Planchet, F. (2014). Constructing entity specific projected mortality table: Adjustment to a reference. *European Actuarial Journal* 4, 247-279.