Quantifying longevity gaps using micro-level lifetime data

SUPPLEMENTARY MATERIAL

Frank van Berkum[1], Katrien Antonio[1,2], and Michel Vellekoop[1]

[1]Faculty of Economics and Business, University of Amsterdam, The Netherlands.
[2]Faculty of Economics and Business, KU Leuven, Belgium.

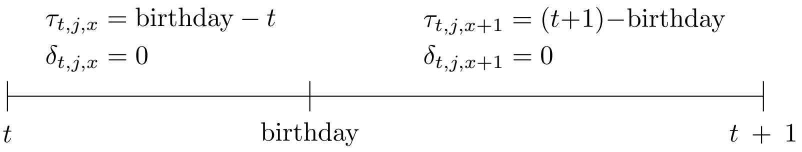
This version: June 2020

1 Taking into account the actual date of birth

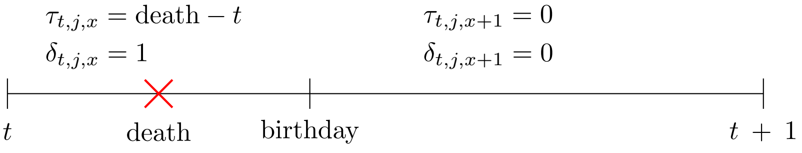
Our proposed framework for statistical modeling of portfolio mortality is explained under the assumption that all participants celebrate their birthday on January 1st. In reality, the dates of birth are spread throughout the year, and in this section we show which formulas of the statistical framework for modeling portfolio mortality are affected, and we illustrate how the date of birth can be taken into account appropriately.

Definition of deaths and exposures. We consider integer years t, and integer and non-integer ages x and $\tilde{x}$ respectively. The distinction between integer and non-integer ages is only needed in this paragraph; in the remainder of this document we only consider integer ages. An individual aged $\tilde{x} = x + \iota$ at the beginning of year t with $\iota \in [0, 1)$, celebrates his birthday at $t + (1 - \iota)$. Define $\tau_{t,j,x}$ as the fraction of the year lived by participant j in calender year t at age x and define the corresponding indicator variable $\delta_{t,j,x}$ which equals 1 if the participant died at age x and 0 otherwise. There are three possible outcomes with respect to survival in year t, and death and exposure observations are defined as follows:

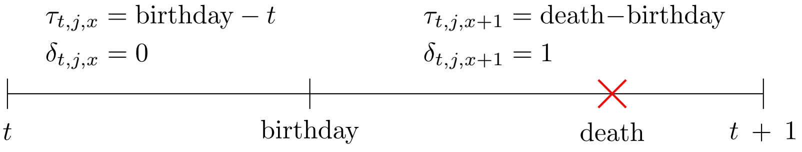
- The participant survives year t:


$\tau_{t,j,x} = \text{birthday} - t$
$\delta_{t,j,x} = 0$
$\tau_{t,j,x+1} = (t{+}1){-}\text{birthday}$
$\delta_{t,j,x+1} = 0$
t
birthday
$t + 1$

- The participant dies during year t, *before* his birthday:



- The participant dies during year t, *at or after* his birthday:


$\tau_{t,j,x} = \text{birthday} - t$
$\delta_{t,j,x} = 0$
$\tau_{t,j,x+1} = \text{death}{-}\text{birthday}$
$\delta_{t,j,x+1} = 1$
t
birthday
death
$t + 1$

Model likelihood. We define the force of mortality for participant j at time t and age x by $\mu_{tjx} = \mu_{tjx}^{\text{pop}} \cdot \eta_{tjx}$. We assume that in calendar year t the baseline population force of mortality $\mu_{tjx}^{\text{pop}} = \mu_{tx}^{\text{pop},g(j)}$ is given for participant j aged x during calendar year t with gender $g(j) \in \{M, F\}$. The factor η_{tjx} represents the ratio between the population force of mortality and the force of mortality of participant j in calendar year t at age x, and this factor must be estimated from the data.

In the likelihood defined below, we take into account that participant j has his birthday at $t+(1-\iota_j)$. The factor η_{tjx} appears for $x = x(j,t)$ and $x = x(j,t)+1$ in the likelihood, but if age is not included as a risk factor, the factors $\eta_{t,j,x(j,t)}$ and $\eta_{t,j,x(j,t)+1}$ are the same. The likelihood for all individual survival observations combined is given by:

$$\mathcal{L}(\eta_{tjx}) = \prod_{t=2006}^{2011} \prod_{j=1}^{J_t} \prod_{x=x(j,t)}^{x(j,t)+1} \exp[-\tau_{tjx}\mu_{tjx}^{\text{pop}}\eta_{tjx}](\mu_{tjx}^{\text{pop}}\eta_{tjx})^{\delta_{tjx}}, \tag{1}$$

where J_t is the number of participants in year t. In total, there are 11,321,861 observations from participants in the different years (summed over t and j), and the likelihood is built up from 22,625,142 contributions from participants in the portfolio (summed over t, j and x).

Cross-validation statistics and robustness analysis. Denote by $\mathcal{F}_{-t}$ all observations in the dataset excluding observations from year t. We define ℓ_{tj} as the likelihood of observed death or survival for participant $j = 1, \ldots, J_t$ in calendar year t given the predictive distribution that follows from $\mathcal{F}_{-t}$. If τ_{tjx} represents the fraction of year t that participant j was alive at age x, ℓ_{tj} is computed as follows:

$$\ell_{tj} = \prod_{x=x(j,t)}^{x(j,t)+1} \exp[-\tau_{tjx}\mu_{tjx}^{\text{pop}}\hat{\eta}_{j,x}^{-t}] \left(\tau_{tjx}\mu_{tjx}^{\text{pop}}\hat{\eta}_{j,x}^{-t}\right)^{\delta_{tjx}}. \tag{2}$$

2 Detailed description of the financial backtest

We describe a financial backtest that is performed for observations in the year 2011. Using the observations from the years 2006 until 2010 as the training sample, we aim to predict the liabilities that are required for those participants that are still alive at the end of 2011. The model that comes closest to the actual required liabilities is said to be the best predicting model.

2.1 Random value of the liabilities

We consider participants $j = 1, \ldots, J_{2011}$ who are alive at the beginning of year 2011 and for which the annual benefit is given by b_j. We assume b_j to remain constant over time and this benefit is paid in case $x(j,t) \geq x_r$ where $x(j,t)$ is the age of participant j at the beginning of year t, and $x_r = 65$ was the retirement age in the Netherlands in the year 2011. The present value of the liabilities for participant j at the end of 2011 in case the participant is still alive is then given by $a_j b_j$, where a_j represents an annuity factor that pays 1 euro if $x(j,t) \geq x_r$. This annuity factor is the same across all models considered; this way we ensure that we compare the different models in a consistent way. See Section 2.3 for details on the calculation of these annuity factors.

The factor $\hat{\eta}_{j,x}^{-2011}$ is estimated using the training sample 2006 to 2010. We model the uncertainty in participant j surviving the year 2011 using a Bernoulli distributed random variable $Y_{2011,j}$, with $P(Y_{2011,j} = 1) = p_{2011,j}$, and the $Y_{2011,j}$ are assumed independent for different j's. Under the assumption of a constant force of mortality μ_{tx} on the interval $[t, t+1) \times [x, x+1)$, the one-year survival probability for participant j with his birthday at $t + (1 - \iota_j)$ is given by:

$$\begin{aligned} p_{2011,j} &= \exp\left[-\int_0^{1-\iota_j} \mu^{\text{pop}}_{2011,j,x(j,2011)} \cdot \hat{\eta}^{-2011}_{j,x(j,2011)} dt - \int_{1-\iota_j}^{1} \mu^{\text{pop}}_{2011,j,x(j,2011)+1} \cdot \hat{\eta}^{-2011}_{j,x(j,2011)+1} dt\right] \\ &= \exp\left[-(1-\iota_j) \cdot \mu^{\text{pop}}_{2011,j,x(j,2011)} \cdot \hat{\eta}^{-2011}_{j,x(j,2011)} - \iota_j \cdot \mu^{\text{pop}}_{2011,j,x(j,2011)+1} \cdot \hat{\eta}^{-2011}_{j,x(j,2011)+1}\right]. \end{aligned}$$

As described in Section 3.3 in the article, the stochastic value of the liabilities Γ on December 31st 2011 is given by

$$\Gamma = \sum_{j=1}^{J_{2011}} \left(Y_{2011,j} \cdot b_j a_j + (1 - Y_{2011,j}) \cdot 0\right). \tag{3}$$

The fund faces a liability with present value $b_j a_j$ at the end of 2011 for those participants who survive, whereas this liability is released for participants who die during 2011. J_{2011} is larger than two million, and based on a simulation study we have verified that the distribution of Γ is close to that of a normally distributed random variable. The skewness of Γ is close to but not equal to zero. We use a parametric approximation to construct prediction intervals for Γ, and we allow for the non-zero skewness by using the skew-normal distribution for this approximation, see for example Vernic (2006). Given the specification of Γ in (3) and using only the point estimate $\hat{\eta}^{-2011}_{j,x}$, the expected value, the variance and the skewness of the random liabilities Γ are given by

$$\mathbb{E}(\Gamma|\hat{\eta}^{-2011}_{j,x}) = \sum_{j=1}^{L_{2011}} p_{2011,j} \cdot b_j a_j \tag{4}$$

$$\text{Var}(\Gamma|\hat{\eta}^{-2011}_{j,x}) = \sum_{j=1}^{L_{2011}} (b_j a_j)^2 \cdot p_{2011,j} \cdot (1 - p_{2011,j}) \tag{5}$$

$$\text{Skew}(\Gamma|\hat{\eta}^{-2011}_{j,x}) = \frac{\sum_{j=1}^{L_{2011}} (b_j a_j)^3 \cdot p_{2011,j} \cdot (1 - p_{2011,j}) \cdot (1 - 2p_{2011,j})}{\left[\text{Var}(\Gamma|\hat{\eta}^{-2011}_{j,x})\right]^{3/2}}. \tag{6}$$

2.2 Population mortality model

We use the AG2014 model to obtain population mortality forecasts, see Koninklijk Actuarieel Genootschap (2014). This model is an application of the Li and Lee (2005) model to a selection of West-European countries under a Poisson assumption for the observed death counts (Brouhns et al. (2002)).

The model is fitted using a group of West-European countries: Austria, Belgium, Denmark, England and Wales, West-Germany, Finland, France, Iceland, Ireland, Luxembourg, the Netherlands, Norway, Sweden and Switzerland. For each country c we downloaded the observed death counts d^c_{tx} and corresponding risk exposures E^c_{tx} from the Human Mortality Database[1], and we defined $d^{\text{EU}}_{tx} = \sum_c d^c_{tx}$ and $E^{\text{EU}}_{tx} = \sum_c E^c_{tx}$ for $t = 1970, \ldots, 2010$ and $x = 0, \ldots, 90$.

The mortality model is specified as

$$D^{\text{EU}}_{tx} \sim \text{Poisson}(E^{\text{EU}}_{tx} \mu^{\text{EU}}_{tx}) \tag{7}$$

$$D^{\text{NL}}_{tx} \sim \text{Poisson}(E^{\text{NL}}_{tx} \mu^{\text{NL}}_{tx}), \tag{8}$$

[1] http://www.mortality.org

with

$$\ln \mu_{tx}^{\mathrm{EU}} = A_x + B_x K_t \tag{9}$$

$$\ln \Delta_{tx}^{\mathrm{NL}} = \alpha_x + \beta_x \kappa_t \tag{10}$$

$$\ln \mu_{tx}^{\mathrm{NL}} = \ln \mu_{tx}^{\mathrm{EU}} + \ln \Delta_{tx}^{\mathrm{NL}}. \tag{11}$$

Here, μ_{tx}^{EU} is the force of mortality for the collection of West-European countries, $\Delta_{tx}^{\mathrm{NL}}$ is the difference in force of mortality between West-Europe and the Netherlands, and μ_{tx}^{NL} is the resulting force of mortality for the Netherlands. The model is applied to both genders separately, so dependence on gender g is not shown.

For K_t the model prescribes a random walk with drift and for κ_t a mean reverting process that reverts back to zero:

$$K_t = K_{t-1} + \theta + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2) \tag{12}$$

$$\kappa_t = a\kappa_{t-1} + \nu_t, \qquad \nu_t \sim \mathrm{N}(0, \sigma_\nu^2), \tag{13}$$

with $|a| < 1$. We further assume the error terms ε_t and ν_t to be correlated for fixed t but uncorrelated over different t, so we can use Seemingly Unrelated Regression techniques to estimate this model. To compute annuities (see the next section) we use the most likely mortality path, i.e. we set error terms ε_t and ν_t equal to zero when forecasting mortality. We calibrate our model to the ages 0 to 90, but mortality estimates for higher ages are needed in order to obtain more accurate (complete) estimates of annuities. We use the closure method of Kannistö (1992) to obtain mortality rates for ages higher than 90, using ages $80, \dots, 90$ in the Kannisto regression.

2.3 Annuity factors

The financial backtest uses the expected present value of direct or deferred annuities a_j for participant j starting at December 31st 2011, as predicted on December 31st 2010. We make the following assumptions when computing the expected present value of the annuities:

1. All participants have integer age at January 1st 2011 (rounded down);

2. Benefits are paid halfway during the year;

3. There are no payments after age 121;

4. The relative difference between mortality in the population and the pension fund is the same for all participants in the pension fund. This difference is represented by the factor η which is computed as

$$\sum_{tjx} \delta_{tjx} / \sum_{tjx} \tau_{tjx} \hat{\mu}_{tx}^{\mathrm{AG},g(j)},$$

   where the summation is taken over all individuals in the training sample $t = \{2006, \dots, 2010\}$, and including both information before and after the birthday of participant j.

Here, $\hat{\mu}_{tx}^{\mathrm{AG},g}$ is obtained by calibrating the AG2014 model on mortality data from 1970-2010.

We use the discount curve published by the Dutch Central Bank for December 31st 2010, and we define z_k as the k-year zero rate for $k \geq 0$ and $z_0 = 0$. We define the mid-year discount factor in year k as

$$DF_{k+\frac{1}{2}} = [(1 + z_k)^k \cdot (1 + z_{k+1})^{k+1}]^{-1/2}, \tag{14}$$

with $k \geq 0$.

In calculating the expected present value of the annuities a_j we only consider the risk factors age x and gender g. The one-year survival probability for participant j in year $2011+k$ is given by

$$p_{j,k} = \exp[-\hat{\mu}^{\text{AG},g(j)}_{2011+k,x(j,2011)+k} \cdot \eta], \quad \text{for } k \geq 0,$$

where the force of mortality $\hat{\mu}^{\text{AG},g}_{tx}$ is obtained from a calibration using mortality data up to and including the year 2010. Given the set of assumptions listed above, we compute an annuity a_j as follows:

$$a_j = \sum_{k \geq 0} DF_{k+\frac{1}{2}} \cdot {}_{k+\frac{1}{2}}\, p_j,$$

with

$${}_{k+\frac{1}{2}}p_j = \begin{cases} (p_{j,0})^{\frac{1}{2}} & \text{for } k = 0, \\ \left(\prod_{l=0}^{k-1} p_{j,l}\right) (p_{j,k})^{\frac{1}{2}} & \text{for } k \geq 1. \end{cases}$$

The expected present value of the annuity a_j is used in (4)–(6).

3 Overdispersion

We estimate (smooth) effects for the risk factors that are available in our dataset using Poisson regression. It is insightful to investigate whether there is overdispersion in our observations. Recall that for each participant in the dataset, the variable δ_{tjx} is an indicator variable that is 1 if participant j died in calendar year t at age x and 0 otherwise, and τ_{tjx} is the fraction of the year lived by participant j in calendar year t at age x $(0 \leq \tau_{tjx} \leq 1)$. We investigate the presence of overdispersion by comparing results from Poisson regression and results from Negative Binomial regression. These regression approaches make use of different mean-variance relationships which may have an impact on our regression results.

- If $X \sim$ Poisson with $\text{E}[X] = \mu$, then $\text{Var}[X] = \mu = \text{E}[X]$;
- If $Y \sim$ Negative Binomial with $\text{E}[Y] = \mu$ and $\theta > 0$, then $\text{Var}[Y] = \mu + \mu^2/\theta > \text{E}[Y]$.

To investigate the mean-variance relationship, we have performed the following analysis:

1. Estimate a regression model as specified in Section 3.1 of the article assuming either a Poisson or a Negative Binomial distribution to predict the participant-specific factors;
2. Sort the estimated participant-specific factors in ascending order;
3. Construct 20 groups which each represent five percent (in successive order) of the distribution of the participant-specific factors;
4. For each group, determine the mean and variance of the observations, corrected for the exposure τ_{tjx}.

We present the outcomes for a model that includes the risk factors `DisPerc`, `Sal`, `IA`, `AFPP`, `Edu` and `PC`; this is the most complex model considered in the article. Results for other specifications were similar. In Figure 1 we have plotted the results from this analysis. The dots represent the empirical mean-variance combinations that have been computed from the dataset after estimation.

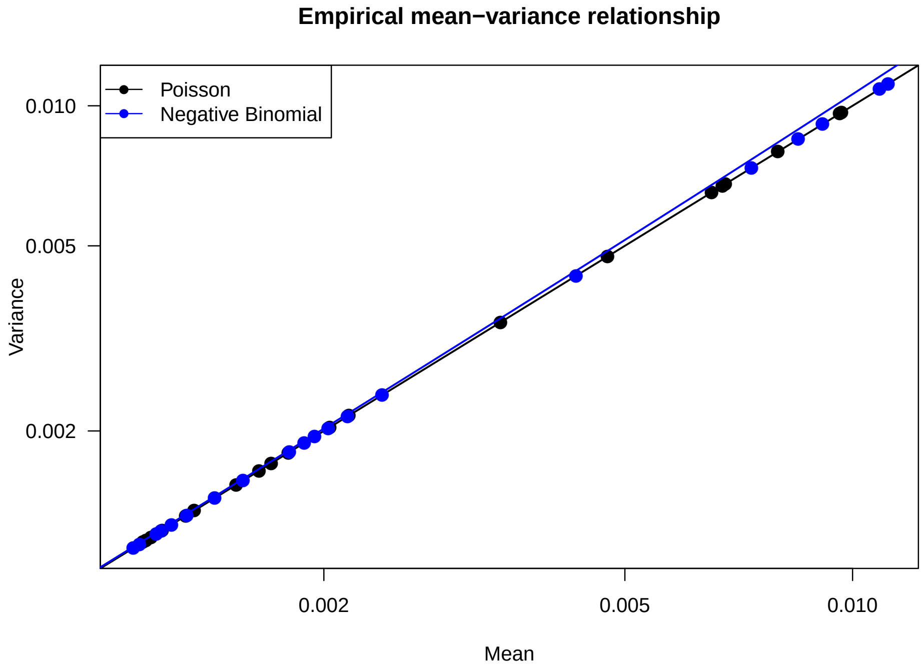


Figure 1: Empirical evidence on the mean-variance relationship in the dataset. The dots represent the mean-variance observations as computed for the 20 groups in the dataset as described in the text. The solid lines represent the implied mean-variance relationships for different distributional assumptions based on regression estimates.

If there is strong overdispersion in our data, we would expect the dots to lie above the black line for the Poisson mean-variance relationship. However, the dots lie close to the black line and do not show a pattern deviating from this black line for higher values of the mean. Further, for the Negative Binomial regression we find a value of $\theta = 0.17$. This implies a very low level of overdispersion: for a mean value of 0.002, the implied variance equals $0.002 + 0.002^2/0.17$ which is 0.00202. The estimated effects for the risk factors that result from Poisson regression and from Negative Binomial regression also turned out to be very similar.

Based on the analysis shown here, we conclude that there is no evidence for overdispersion in our dataset.

References

N. Brouhns, M. Denuit, and J.K. Vermunt. A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3):373 – 393, 2002.

V. Kannistö. *Development of the oldest - old mortality, 1950-1980: evidence from 28 developed countries.* Odense University Press, 1992.

Koninklijk Actuarieel Genootschap. Projection table AG 2014, 2014. Available online at: http://www.ag-ai.nl/view.php?action=view&Pagina_Id=625.

N. Li and R.D. Lee. Coherent mortality forecasts for a group of populations: an extension of the Lee-Carter method. *Demography*, 42(3):575 – 594, 2005.

R. Vernic. Multivariate skew-normal distributions with applications in insurance. *Insurance: Mathematics and Economics*, 38:413–426, 2006.