Modelling and quantifying mortality and longevity risk

Module D3 : Heterogeneity in Mortality Models

Michel Vellekoop

Actuarial Summer School

Warsaw, Sept 18-19, 2025

Overview

In this module:

- Heterogeneity due to frailty
- Socio-economic factors
- Some takeaway points.

Heterogeneity due to frailty

Motivating example: Mortality at high ages

- Data limit/unavailable for high ages necessitates extrapolation.
- Debate whether ages at death can be considered to be bounded from above.
- Popular choice to close tables (Kannisto, 1992) makes assumption that for high ages logit-transformed hazard rate $\ln \frac{\mu_{xtg}}{1-\mu_{xtg}}$ becomes linear in the age x
- This implies that for given (t, g), there are $a \in \mathbb{R}$, $b \in \mathbb{R}^+$,

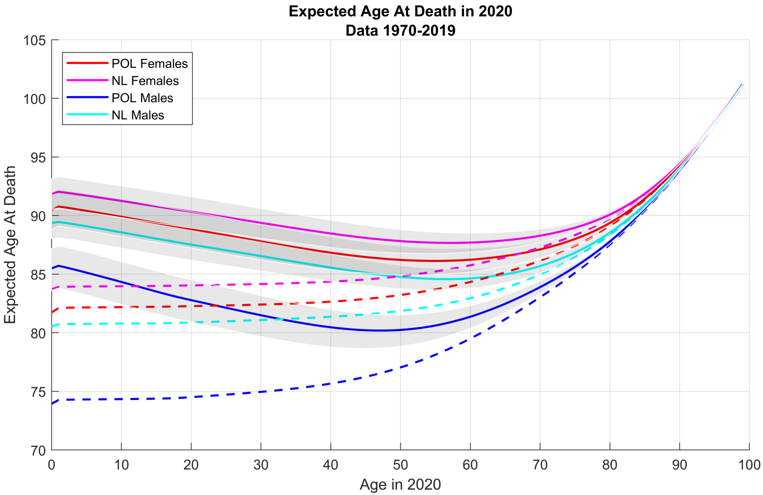
$$\mu_{xtg} \approx \frac{1}{1 + e^{-a-bx}}, \qquad x \gg 0.$$

- Assumption that hazard rate is constant during calendar year then leads to

$$\lim_{x\to\infty} \mu_{xtg} = 1, \quad \Rightarrow \quad \lim_{x\to\infty} q_{xtg} = 1 - e^{-1} \approx 0.632,$$

i.e. a so-called mortality plateau.

Convergence of remaining life expectancy ...



Closing Mortality Tables

- If we define $H(x) = \ln \frac{x}{1-x}$ then regression

$$H(\mu_{xtg}) = a_{tg}x + b_{tg} + \epsilon_{xtg}$$

leads to

$$H(\mu_{xtg}) \;=\; \sum_{k=1}^{n} w_k(x) H(\mu_{y_k tg}),$$

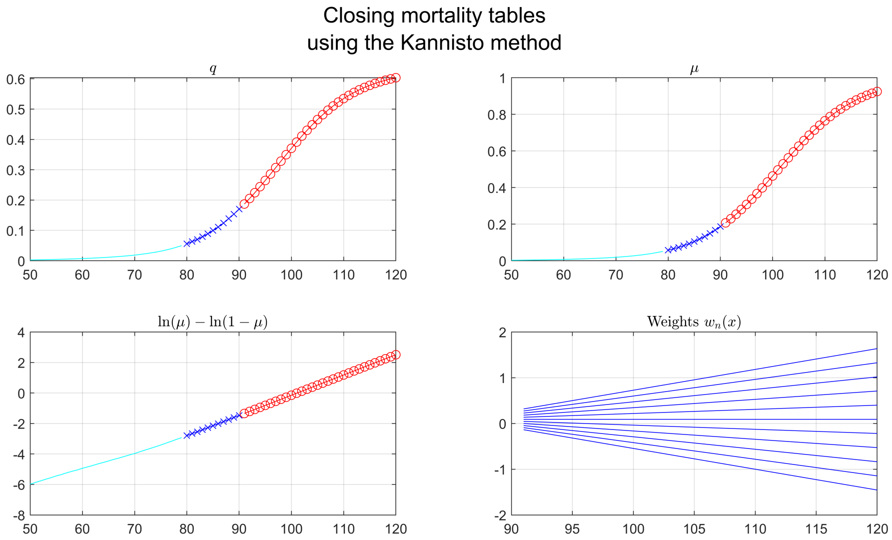
if we use ages $(y_1, \ldots, y_n)$ to extrapolate from, with regression weights

$$w_k(x) \;=\; \frac{1}{n} + \frac{(y_k - \bar{y})(x - \bar{y})}{\sum_{j=1}^{n}(y_j - \bar{y})^2}, \qquad \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j.$$

- Kannisto transformation is thus, for $x > y_n$:

$$\mu_{xtg} \;=\; H^{-1}\left(\sum_{k=1}^{n} w_k(x) H(\mu_{y_k tg})\right).$$

Kannisto method



Closing Mortality Tables

- Alternatively we can write, for some $h_{xg} \in (0, \infty)$,

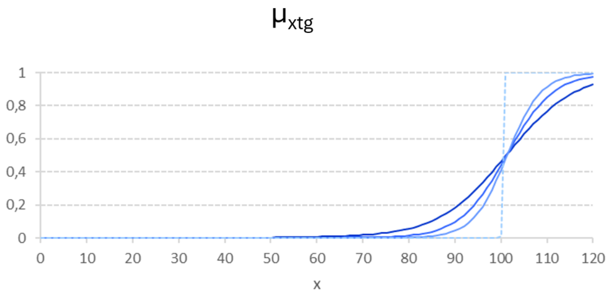
$$\begin{aligned} \frac{1}{\mu_{xtg}} &= 1 + \exp(-\sum_{k=1}^{n} w_k(x) \ln(\mu_{y_k tg})\ ) \\ &= 1 + h_{xg} \exp(-K_t \sum_{k=1}^{n} w_k(x) B_{y_k} - \kappa_t \sum_{k=1}^{n} w_k(x) \beta_{y_k}). \end{aligned}$$

- But then

$$\begin{aligned} \sum_{k=1}^{n} w_k(x) B_{y_k} > 0 \quad &\Rightarrow \quad \lim_{t\to\infty} \mu_{xtg} = 1, \quad \lim_{t\to\infty} q_{xgt} = 1 - e^{-1}, \\ \sum_{k=1}^{n} w_k(x) B_{y_k} < 0 \quad &\Rightarrow \quad \lim_{t\to\infty} \mu_{xtg} = 0, \quad \lim_{t\to\infty} q_{xtg} = 0, \end{aligned}$$

and μ_{xtg} becomes (almost) deterministic for the age x where $\sum_{k=1}^{n} w_k(x) B_{y_k} \approx 0$.

Kannisto method



$$\sum_{k=1}^{n} w_k(x) B_{y_k} > 0 \quad \Rightarrow \quad \lim_{t\to\infty} \mu_{xtg} = 1, \quad \lim_{t\to\infty} q_{xgt} = 1 - e^{-1},$$

$$\sum_{k=1}^{n} w_k(x) B_{y_k} < 0 \quad \Rightarrow \quad \lim_{t\to\infty} \mu_{xtg} = 0, \quad \lim_{t\to\infty} q_{xtg} = 0.$$

Rationale behind Kannisto method

- Logit-linear assumption in Kannisto method can be motivated by observation that
  - Hazard rates seem to become linear on logarithmic scale (i.e. exponential) at high ages (Gompertz, 1825).
  - But heterogeneity in mortality characteristics (frailty) should change composition of survivors: average frailty should decrease (Vaupel, 2014).
- Model for heterogeneity:
  - start from 'average' Gompertz hazard rate per age for whole population, so for fixed (t, g)

$$\ln \mu_x = ax + b,$$

  - but assume time dynamics are due to changing distribution of frailty; for individual i

$$\ln \mu_x^i = \ln \mu_x + \ln Z_i = ax + b + \ln Z_i,$$

with $Z_i \sim \Gamma(\lambda, \kappa)$ iid and $\kappa = \lambda = \sigma^{-2}$ which implies that

$$\mathbb{E}[Z_i] = 1, \quad \mathbb{V}[Z_i] = \sigma^2.$$

High age mortality in AG2022

- If $Z_i \sim \Gamma(\lambda, \kappa)$ iid and $\kappa = \lambda = \sigma^{-2}$ then, for all $s \geq 0$,
$$\mathbb{E}[e^{-sZ_i}] = (1 + \sigma^2 s)^{-1/\sigma^2}.$$
- Probability of survival over period h becomes
$$\bar{s}_x(h) \quad = \quad \mathbb{E}[e^{-\int_0^h \mu^i_{x+s} ds}] = \mathbb{E}[e^{-Z_i \int_0^h \mu_{x+s} ds}] = (1 + \sigma^2 \int_0^h \mu_{x+s} ds)^{-1/\sigma^2}.$$
- So observed Gamma-Gompertz force of mortality over whole population equals
$$\begin{aligned} \bar{\mu}_x(h) &= -\frac{\partial}{\partial h} \ln \bar{s}_x(h) = \frac{\mu_{x+h}}{1 + \sigma^2 \int_0^h \mu_{x+s} ds} \\ &= \frac{e^{a+b(x+h)}}{1 + \sigma^2 e^{a+bx}(e^{bh} - 1)/b} = \frac{e^{bh}}{e^{-a-bx} + c}, \qquad c = \sigma^2(e^{bh} - 1)/b. \end{aligned}$$

Modelling cause-of-death mortality using socio-economic factors

Construction of annual pre-pandemic data

(ZonMw research project Antonio, Kleinow, Simonetti, van Berkum & Vellekoop)

- Merge the microdata to individual spells:
  - time $t \in \mathcal{T} = \{2016, ..., 2019\}$
  - individuals $j \in \mathcal{J}_t = \{1, ..., J_t\}$
  - individual-specific spells $i \in \mathcal{I}_{t,j} = \{1, ..., I_{t,j}\}$ with constant socio-economic factors
- For each (t, j, i) combination we have:
  - exposure-to-risk $E_{t,j,i}$
  - death indicator $\delta_{t,j,i}$ and cause-specific death indicator $\delta^c_{t,j,i}$
  - combination of constant risk factors.

Microdata : sources

- (static) date of birth, gender, migration background
- (dynamic, health care expenses) such as expenses for hospital, pharmacy, nursing, mental, and total
- (dynamic, wealth and income) such as personal income, household income, property value, home ownership
- (dynamic, socio-economic) based on neighbourhood: prosperity, education, job history, urbanity
- (per spell) start and end date of residence spells in the Netherlands
- (per event) cause, date and location of death + (during pandemic) vaccination uptake and COVID-19 tests

125 jaar
Centraal Bureau voor de Statistiek

Construction of annual pre-pandemic data

- Force of mortality is assumed to be constant during spell (t, j, i) and modelled by $\mu_{t,j,i}(\theta)$ for parameter θ we need to calibrate.
- Survival likelihood during spell (t, j, i)

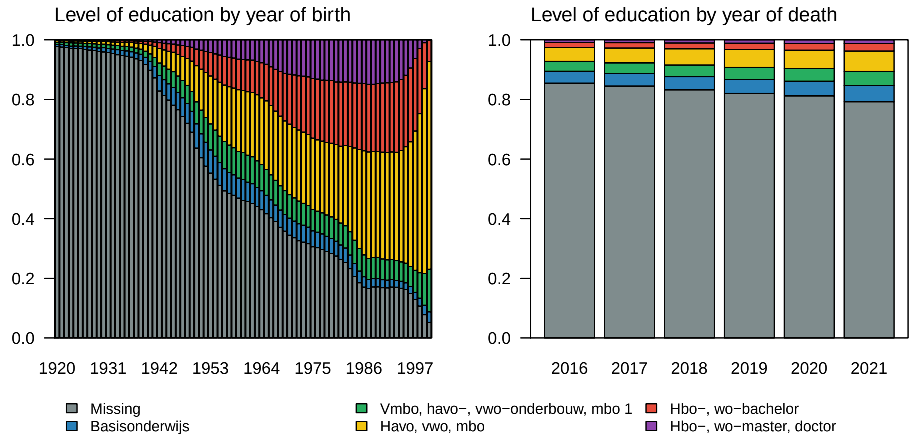
$$
\begin{aligned}
\mathbb{P}(\delta_{t,j,i} = 0 | E_{t,i,j}) &\approx e^{-E_{t,j,i}\mu_{t,j,i}(\theta)} \\
\mathbb{P}(\delta_{t,j,i} = 1 | E_{t,i,j}) &\approx e^{-E_{t,j,i}\mu_{t,j,i}(\theta)} - e^{-(E_{t,j,i}+\Delta E)\mu_{t,j,i}(\theta)} \\
&\approx \quad \mu_{t,j,i}(\theta) e^{-E_{t,j,i}\mu_{t,j,i}(\theta)} \Delta E.
\end{aligned}
$$

- Total likelihood of parameter θ over all observed spells:

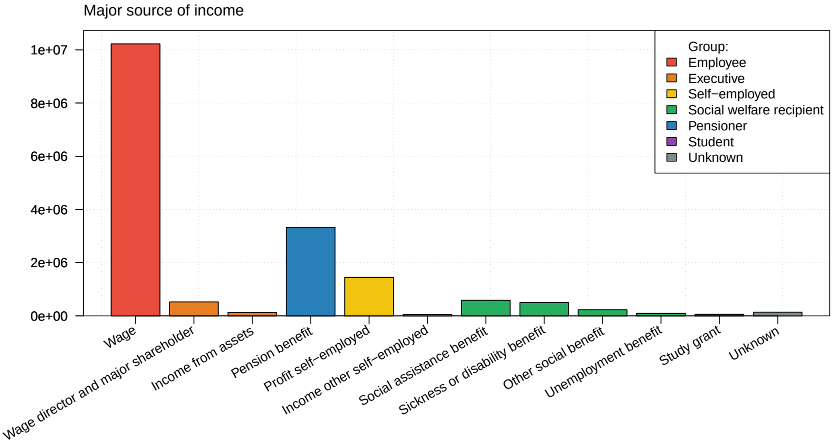
$$
\mathcal{L}(\theta) = \prod_{t \in \mathcal{T}} \prod_{j \in \mathcal{J}_t} \prod_{i \in \mathcal{I}_{t,j}} e^{-E_{t,j,i} \cdot \mu_{t,j,i}(\theta)} \left(\mu_{t,j,i}(\theta) \Delta E \right)^{\delta_{t,j,i}} .
$$

$$
\ln \mathcal{L}(\theta) = \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}_t} \sum_{i \in \mathcal{I}_{t,j}} \left[-E_{t,j,i}\mu_{t,j,i}(\theta) + \delta_{t,j,i} \ln \mu_{t,j,i}(\theta) \right] + c.
$$

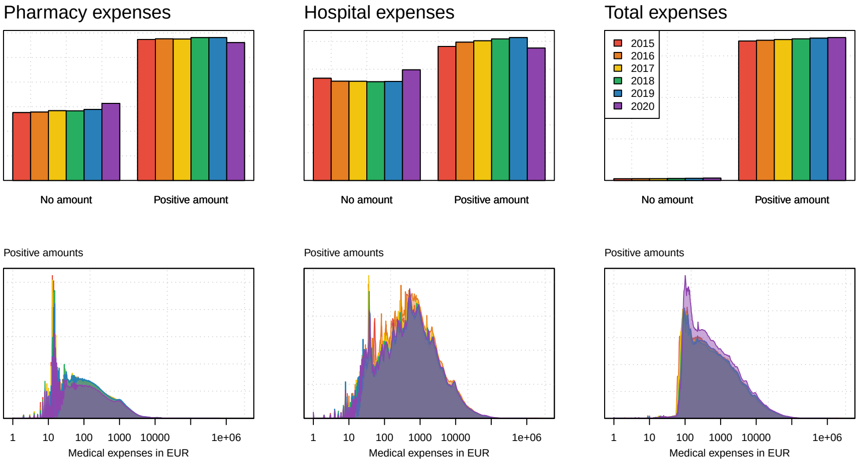
CBS microdata for Education



CBS microdata for Source of income



CBS microdata for Medical Expenses



Pre-pandemic mortality model using socio-economic factors

- Generalized Additive Model for covariates $X^k_{t,j,i}$ $(k = 1, \dots, K)$:

$$\ln \mu_{t,j,i}(\theta) \;=\; \sum_{k=1}^{K} f^k(X^k_{t,j,i}, \theta).$$

- The functions f^k assign a different value for every realization of a categorical covariate k (or a covariate k which has a finite number of possible outcomes), and a smooth function for covariates k with values in a continuum:

$$(k \in \mathcal{K}_d): \quad f^k(x,\theta) \;=\; \sum_n \theta^k_n \mathbf{1}_{x = x^k_n}, \quad (k \in \mathcal{K}_c): \quad f^k(x,\theta) \;=\; \sum_n \theta^k_n f^k_n(x).$$

- We would like the basis functions f^k_n and the optimization process to lead to smooth effects, so we optimize log-likelihood plus a penalty term:

$$\max_{\theta} \Big(\ln \mathcal{L}(\theta) - \sum_{k \in \mathcal{K}_c} \lambda_k \int \Big[\sum_n \theta^k_n (f^k_n)''(u)\Big]^2 du \Big).$$

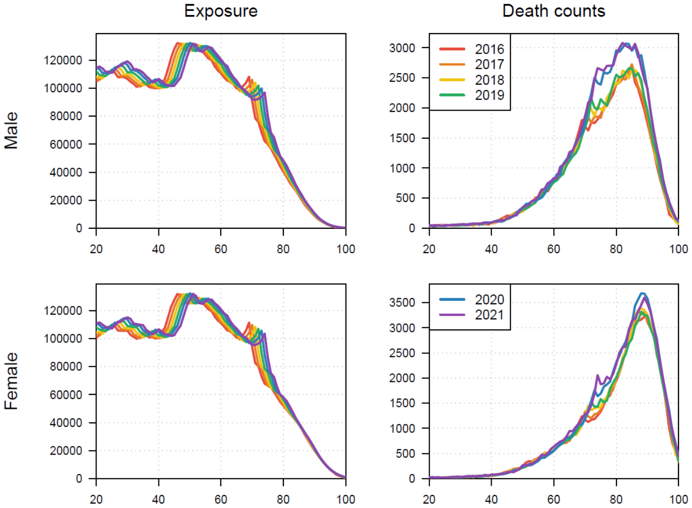
Pre-pandemic mortality model using socio-economic factors

- To assess pandemic excess mortality we want to account for pre-pandemic existing differences in mortality among socio-economic groups, so we specify $\mu_{t,j,i}(\theta)$ as:

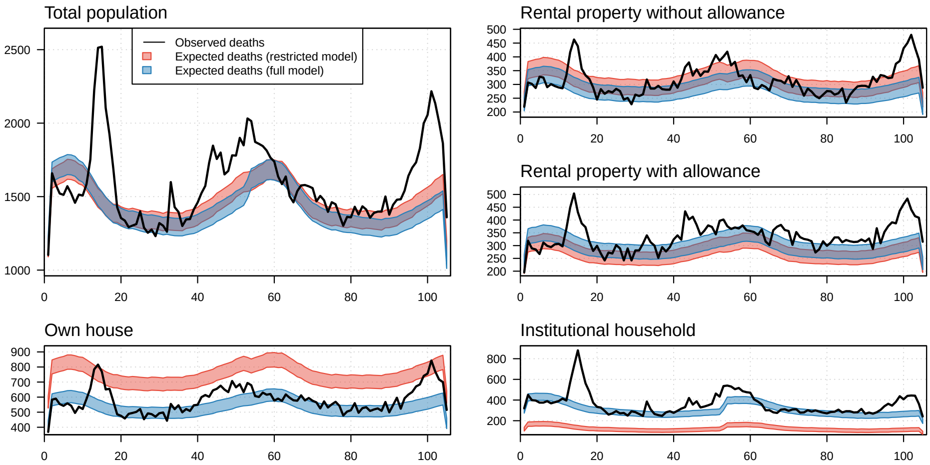
$$
\begin{aligned}
\ln \mu_{t,j,i}(\theta) \;=\; & \sum_n \theta^{\alpha}_{n,g_j} f^{\alpha}_n(x_i) + (\bar{t} - t) \sum_n \theta^{\beta}_{n,g_j} f^{\beta}_n(x_i) \\
+ \;& \mathbf{1}_{\mathrm{ME}_i > 0} \sum_n \theta^{\mathrm{ME}}_n f^{\mathrm{ME}}_n(\ln \mathrm{ME}_i) + \theta^{\mathrm{ME}}_0 \mathbf{1}_{\mathrm{ME}_i = 0} \\
+ \;& \mathbf{1}_{\mathrm{Wealth}_i \text{ known}} \sum_n \theta^{\mathrm{W}}_n f^{\mathrm{W}}_n(\mathrm{Wealth}_i) \\
+ \;& \sum_\ell \theta^{\mathrm{PI}}_\ell \mathbf{1}_{\mathrm{PersIncSrc}_i = \ell} + \sum_\ell \theta^{\mathrm{HO}}_\ell \mathbf{1}_{\mathrm{HomeOwn}_i = \ell} + \sum_\ell \theta^{\mathrm{G}}_\ell \mathbf{1}_{\mathrm{Geo}_i = \ell}.
\end{aligned}
$$

- Based on medical expenses + wealth (property value or income quantiles), source of personal income and home ownership + migration background.

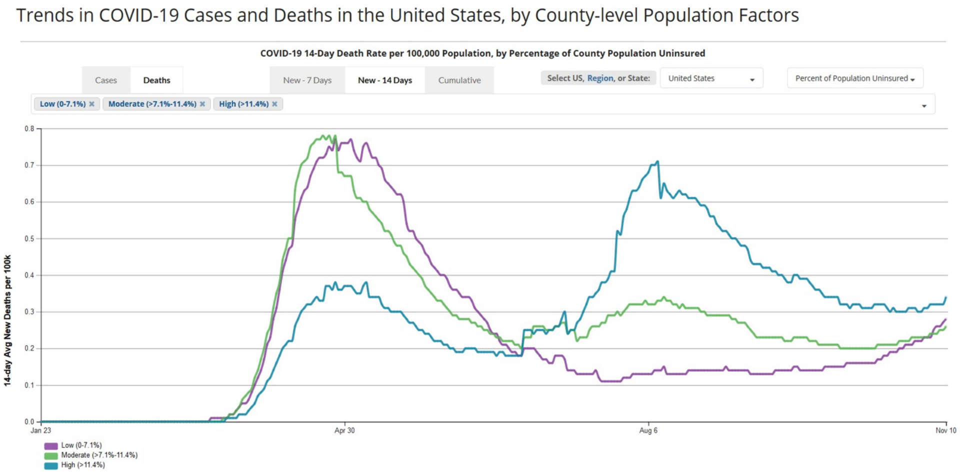
Pre- and Post Pandemic Statistics



Quantifying excess mortality: importance of baseline



Insured vs Uninsured population ...



Source: CDC. Purple: more insurance, Blue: less insurance.

Takeaway points

Concluding

Some takeaway points after two days of model equations ...

- Future survival rates are stochastic, and scenarios are required for proper analysis,
- Post-pandemic model may require newly estimated age effects, and finer data,
- "Uniform" (re-)distribution of longevity risk need not be "fair"
- Longevity derivatives are insurance products, and should be priced accordingly,
- Statistical and machine learning models can help to explain deviations from baseline mortality levels, using proprietary or open source data,
- Portfolios may be very heterogeneous due to socio-economic differences.