Quantifying longevity gaps using micro-level lifetime data

Frank van Berkum, Katrien Antonio, Michel Vellekoop

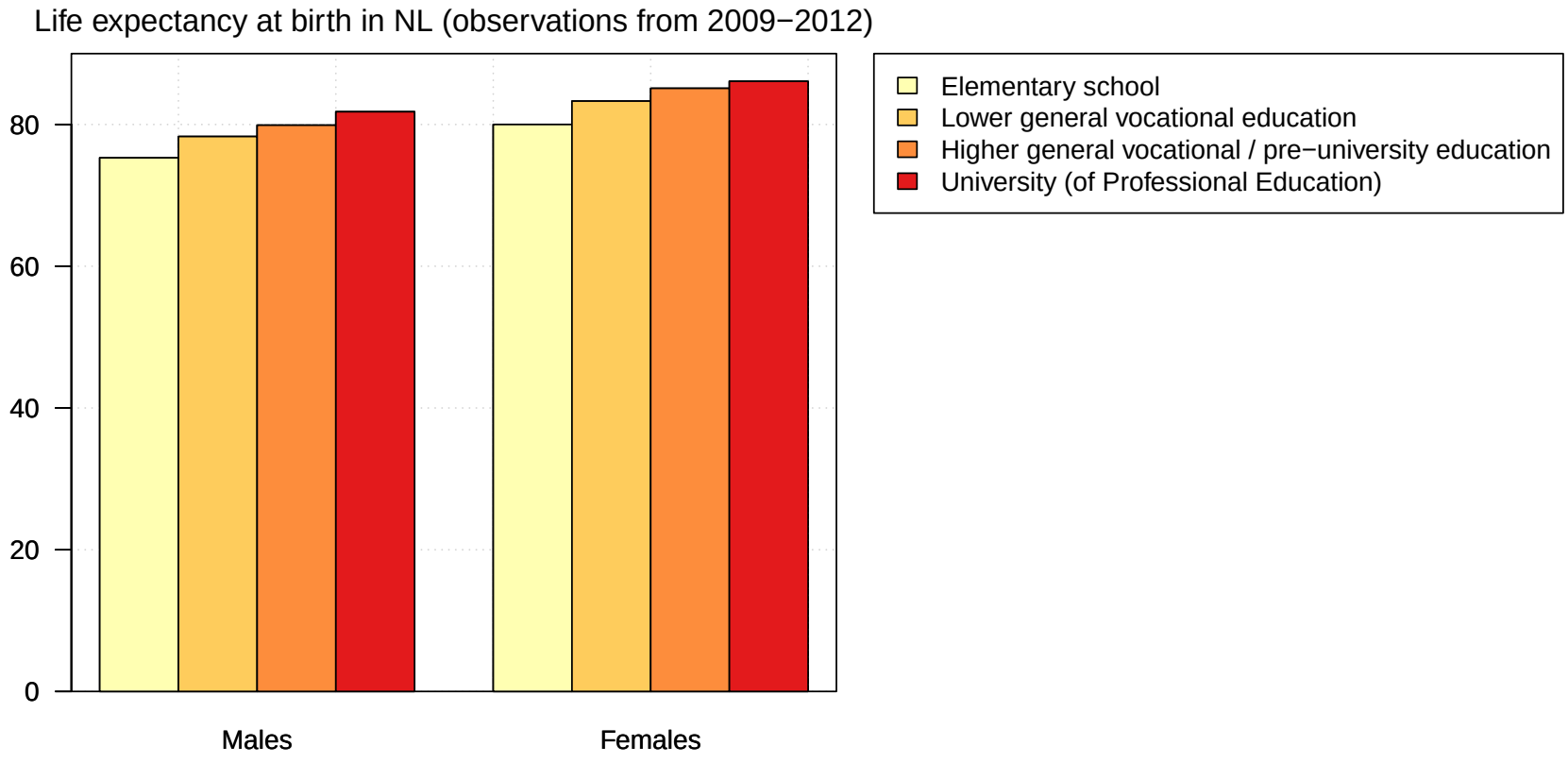
University of Amsterdam & Netspar

Longevity14, 21 September 2018

UNIVERSITY OF AMSTERDAM

Outline

The population of a country is heterogeneous



Source: RIVM publication:
http://www.eengezondernederland.nl/Heden_en_verleden/Levensverwachting

Population vs portfolio-specific mortality

- Life insurers and pension funds value their liabilities using prospective mortality rates to account for future mortality developments
- These mortality rates should be specific to the underlying portfolio, and two general approaches are used to obtain portfolio-specific mortality rates:

  1. Assuming population mortality is known, explain the difference between population and portfolio using risk characteristics:

     Pro: Can be done for relatively small portfolios;
     Con: Population mortality is projected separately.

  2. Use a multiple population approach to model a stochastic 'spread' between population and portfolio:

     Pro: Allows for separate mortality trend in portfolio;
     Con: Time series modeling for the portfolio is required.

- Our dataset has only few historical years, and we shall therefore focus on the first approach.

Literature review

Taking salary information into account

Plat [2009] considers observed portfolio factors that are defined as:

$$P_{t,x} = \frac{q^A_{t,x}}{q^{\text{pop}}_{t,x}}, \quad \text{with} \quad q^A_{t,x} = \frac{A^D_{t,x}}{\frac{1}{2}(A^P_{t,x} + A^D_{t,x} + A^D_{t,x})},$$

where $A^i_{t,x}$ is the assured amount with $i \in \{\textbf{P}\text{rimo}, \textbf{U}\text{ltimo}, \textbf{D}\text{eceased}\}$, and $q^{\text{pop}}_{t,x}$ is an observed mortality rate in the population:

+ Through this definition of $q^A_{t,x}$ the correlation between mortality and salary/education/... is implicitly taken into account, since more weight is given to participants with large insured amounts;
– The observed $P_{t,x}$ can be very volatile for small portfolios, and there is no obvious distribution for the $P_{t,x}$'s. As a result, it is difficult to distinguish between individual mortality risk and uncertainty in portfolio factors;

Literature review

Including risk factors in Poisson regression

Gschlössl et al. [2011] model observed deaths directly:

$$D_i \sim \text{Poisson}(E_i \mu_i), \quad \text{with} \quad \ln \mu_i = \beta_0 + \beta_1 \ln \mu_i^b + \sum_{j=2}^{r+1} \beta_j x_{ij},$$

where μ_i^b is a smooth baseline mortality rate estimated from portfolio data, and the x_{ij} are other observed risk factors such as curtate duration of the policy, product type and amount insured:

- \+ A wide variety of risk factors can easily be included in this framework;
- – It is not trivial how to construct portfolio-specific mortality forecasts from the estimated model;
- – Continuous risk factors are either included in a linear way in the linear predictor or are converted to categorical variables. Structure in the data may be lost through this approach;

Portfolio mortality data

Individual vs aggregated observations

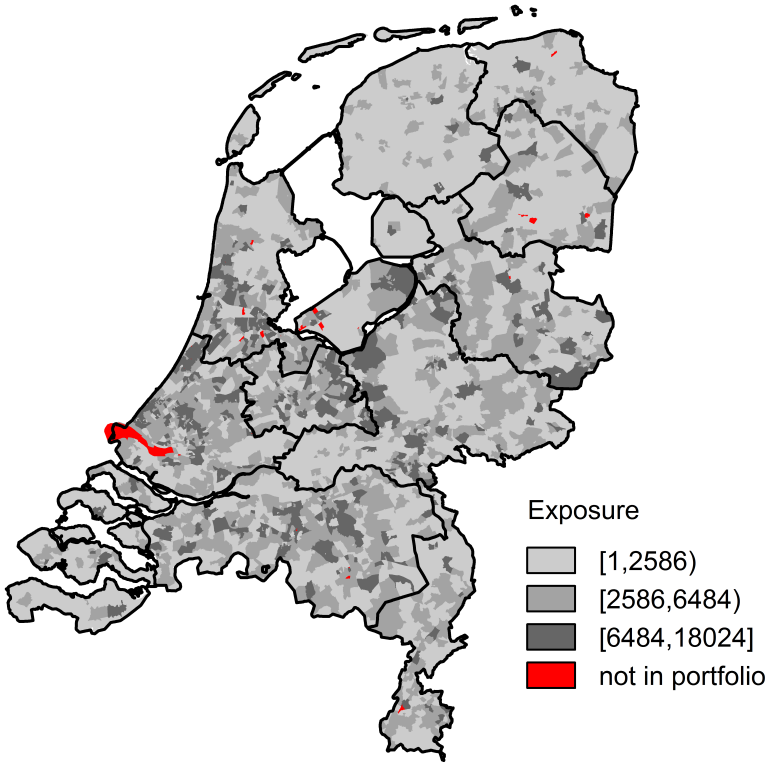
From the dataset we construct individual death and exposure-to-risk observations:

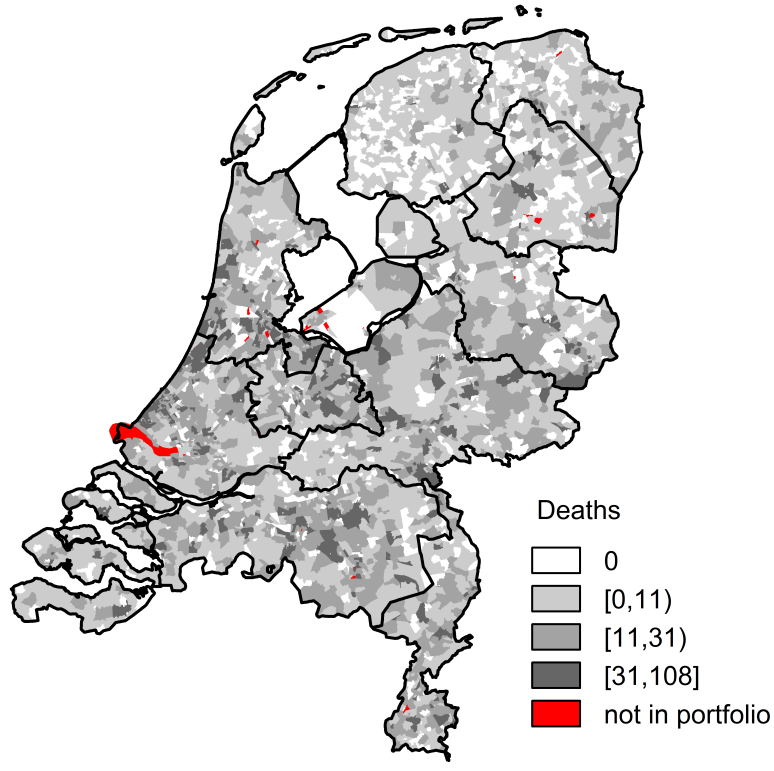
- δ_{tjx} is an indicator variable which equals 1 if participant j died in calendar year t at age x and 0 otherwise;
- $\tau_{tjx} \in [0, 1]$ is the fraction of the year lived by participant j in calendar year t at age x.

Define a risk profile as a unique combination of risk factors such as `Age` and `Sal`. From the individual observations (j) we construct death and exposure-to-risk observations at risk profile level (i) as follows:

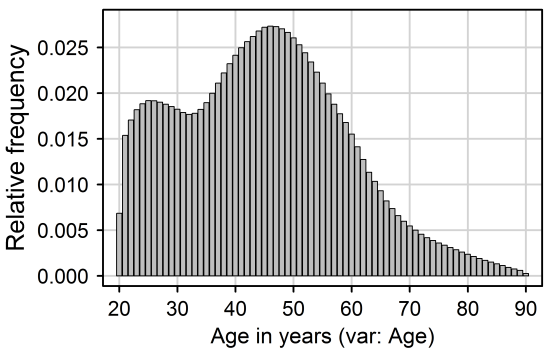
- $d_i = \sum_{t=2006}^{2011} \sum_{j=1}^{L_{t,x}} \sum_{x=x(j,t)}^{x(j,t)+1} \delta_{tjx} \cdot I(i, t, j, x)$;
- $\tilde{E}_i = \sum_{t=2006}^{2011} \sum_{j=1}^{L_{t,x}} \sum_{x=x(j,t)}^{x(j,t)+1} \tau_{tjx} \mu_{tjx}^{\text{pop}} \cdot I(i, t, j, x)$;
- $I(i, t, j, x)$ is an indicator variable that is 1 if in calendar year t participant j at age x belongs to risk profile i (thus also taking into account all other risk factors), and 0 otherwise.

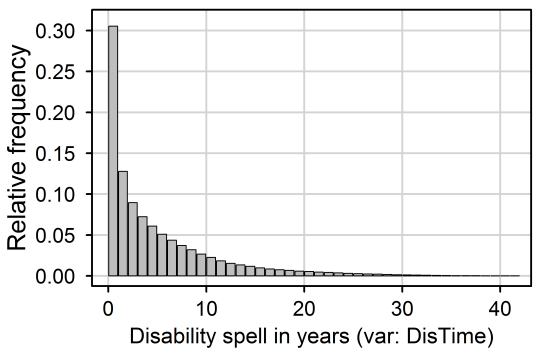
Observations in the dataset

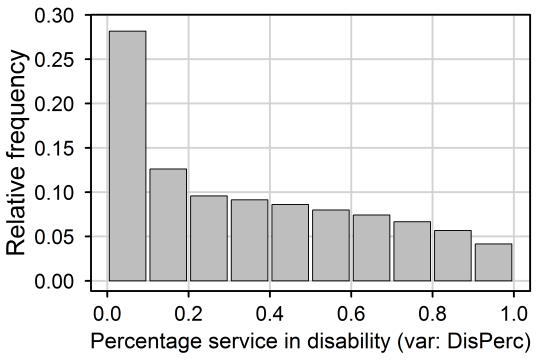


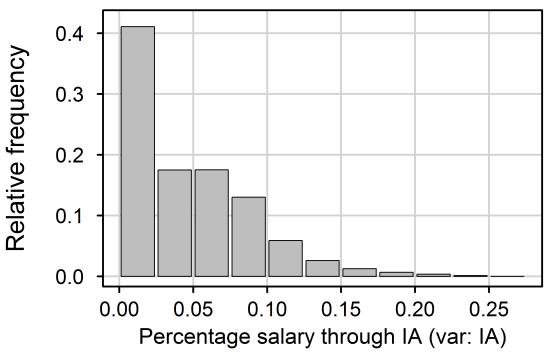


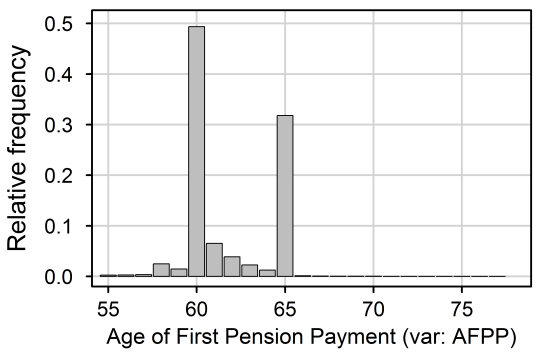
Risk factors in the dataset

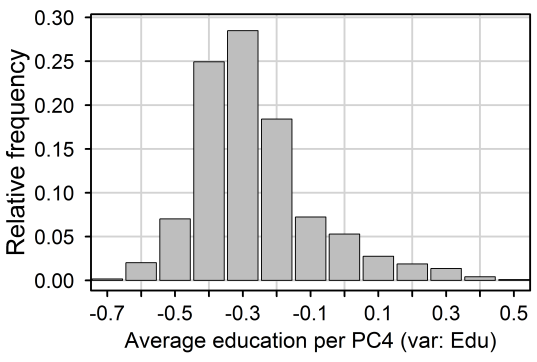












Poisson regression model

Generalized additive model

We assume the following model:

$$D_i \sim \text{Poisson}(\tilde{E}_i \eta_i),$$

and we use a Generalized Additive Model (GAM) to calibrate the η_i as:

$$\ln \eta_i = \beta_0 + \underbrace{\sum_{k=1}^{p} \beta_k x_{ik}^d}_{\text{categorical variables}} + \overbrace{\sum_{l=1}^{q} f_l(x_{il}^c)}^{\text{1D continuous variables}} + \underbrace{g(x_i^{\text{long}}, x_i^{\text{lat}})}_{\text{2D continuous variable}} .$$

The main advantage of GAMs (as opposed to e.g. GLMs) is that smooth effects can be estimated for the risk factors, instead of having to impose some relationship on beforehand

Strategy for working with large datasets

Binning the postal code effect

The final dataset has more than 22 million observations, which is too large for regular estimation procedures. We proceed as follows:

1. Estimate the effects for all risk factors except postal code;
   - Aggregate the observations over the different postal codes, thereby decreasing the size of the dataset.
2. Estimate a spatial effect on the residuals;
   - Take the estimated effect from step 1 as given, and aggregate the dataset over all risk factors except postal code.
3. Cluster similar postal code using the Fisher-Jenks binning method;
   - Minimize the variance *within* clusters, maximize the variance *between* clusters;
   - Consider different number of cluster, estimate GLMs with different numbers of clusters, and choose optimal number of clusters using BIC.
4. Estimate full model with all 1D risk factors and clustered postal code.

In-sample statistics and cross-validation tests

To compare different model specifications:

- We compute a variety of in-sample statistics such as the loglikelihood and information criteria (cAIC and BIC);
- We compute the log score (a proper scoring rule / cross validation test). This can be interpreted as the out-of-sample likelihood, see Czado et al. [2009].

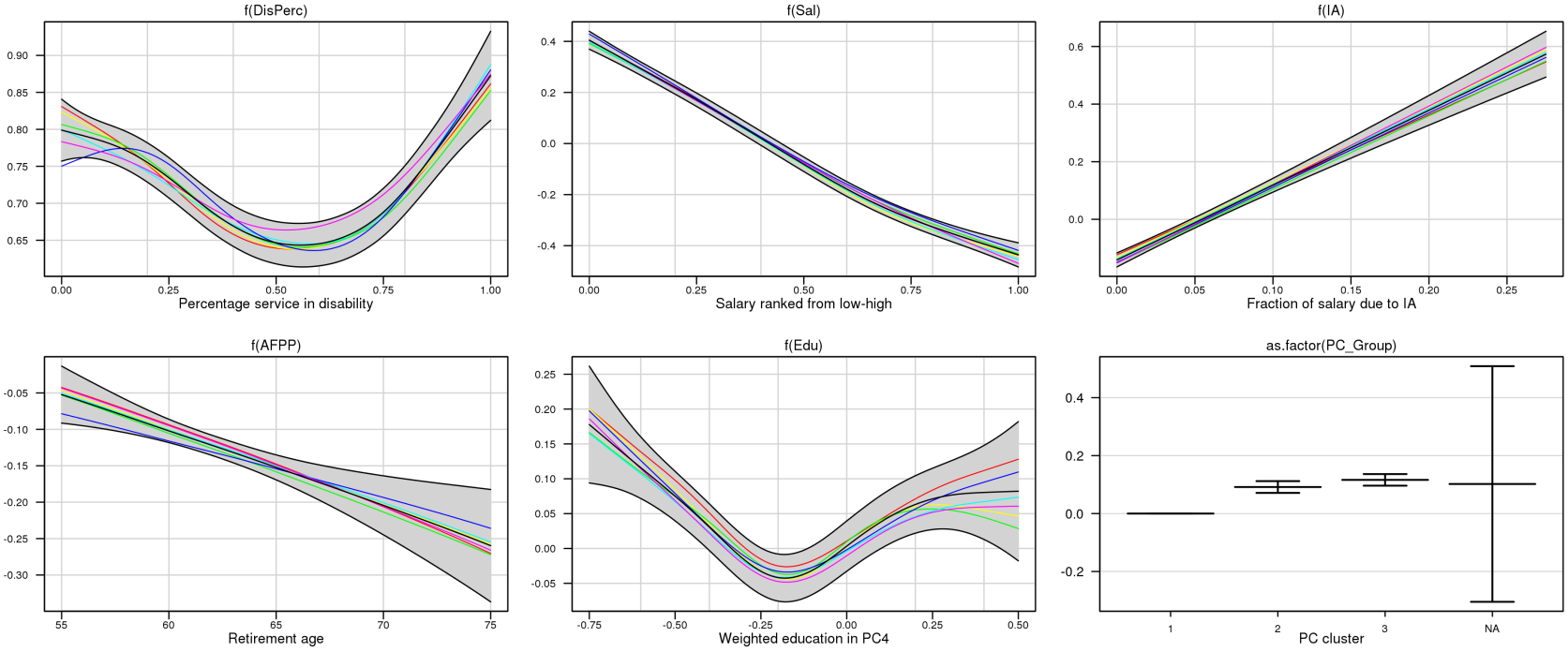
General observation (after calibration):

- Including more variables results in improves in-sample and out-of-sample statistics.

See van Berkum [2018] for more details.

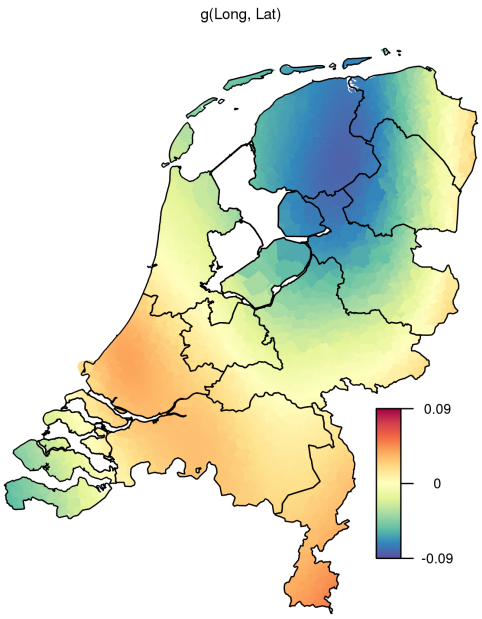
Parameter estimates

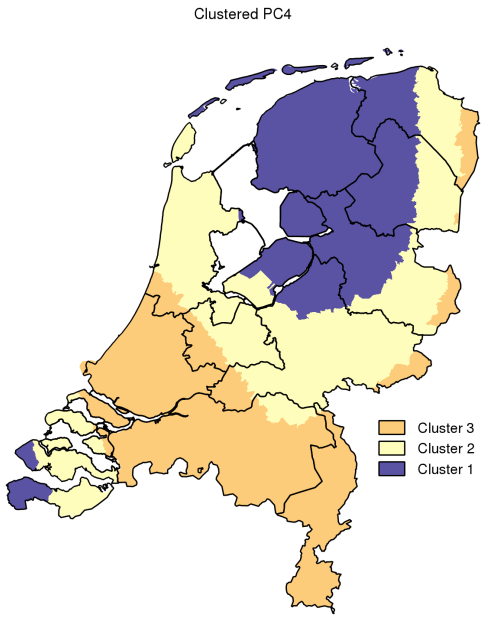
Model with risk factors `DisPerc`, `Sal`, `IA`, `AFPP`, `Edu` and `PC`



Parameter estimates

Model with risk factors `DisPerc`, `Sal`, `IA`, `AFPP`, `Edu` and `PC`





Remaining cohort life expectancy in 2017

Model with risk factors DisPerc, Sal, IA, AFPP, Edu and PC

PC	DisPerc	Sal	LE_{25}^{M}	LE_{25}^{F}	LE_{65}^{M}	LE_{65}^{F}
Lowest mortality (cluster 1)	No	0.90	**69.1**	71.2	**26.2**	28.8
		0.50	66.7	68.9	23.7	26.5
		0.10	63.5	66.0	20.6	23.6
	5%	0.90	62.7	65.2	19.9	22.9
		0.50	60.2	62.8	17.6	20.7
		0.10	56.7	59.5	14.8	18.0
Highest mortality (cluster 3)	No	0.90	68.2	70.3	25.2	27.9
		0.50	65.8	68.1	22.8	25.6
		0.10	62.6	65.1	19.7	22.7
	5%	0.90	61.8	64.3	19.0	22.0
		0.50	59.2	61.9	16.7	19.9
		0.10	**55.6**	58.5	**14.0**	17.2

NB: the risk factors IA, AFPP and Edu are assumed missing in calculating the above numbers.

Financial backtest

Our model is specified on observed numbers of death, but for a pension fund it is more relevant to accurately predict the (release of) the value of the liabilities

We assume the management of the pension fund at the beginning of the year 2011 wants to predict the value of the liabilities at the end of the year. Define the following variables:

- b_j is the annual pension benefit that is paid if participant j reaches the retirement age;
- a_j is an annuity (valued at Dec 31st 2011) that starts paying 1 unit at retirement age if participant j is then still alive;
- I_j is an indicator variable that is 1 if participant j is still alive at Dec 31st 2011 (given that participant j is alive at Jan 1st 2011), and 0 otherwise.

See van Berkum [2018] for details on how a_j is calculated.

Financial backtest

Define $p_{2011,j} = \exp[-\mu_{2011,j}]$ as the one-year survival probability for participant j:

- We define $\mu_{2011,j} = \mu^{\text{pop}}_{2011,x(j,2011)} \cdot \eta_j^{-2011}$
- We model uncertainty of participant j surviving the year 2011 using a Bernoulli($p_{2011,j}$) distributed r.v. $Y_{2011,j}$.

The stochastic value of the liabilities Γ on Dec 31st 2011 is then given by

$$\Gamma = \sum_{j=1}^{L_{2011}} (Y_{2011,j} \cdot b_j a_j + (1 - Y_{2011,j}) \cdot 0),$$

and the actual value of the liabilities at Dec 31st 2011 is given by

$$\widetilde{\Gamma} = \sum_{j=1}^{L_{2011}} \mathsf{I}_j \cdot b_j a_j.$$

Financial backtest

Confidence interval

The mean and variance of Γ are given by:

$$\mathbb{E}(\Gamma|\eta_j^{-2011}) = \sum_{j=1}^{L_{2011}} p_{2011,j} \cdot b_j a_j$$

$$\mathrm{Var}(\Gamma|\eta_j^{-2011}) = \sum_{j=1}^{L_{2011}} (b_j a_j)^2 \cdot p_{2011,j} \cdot (1 - p_{2011,j}),$$

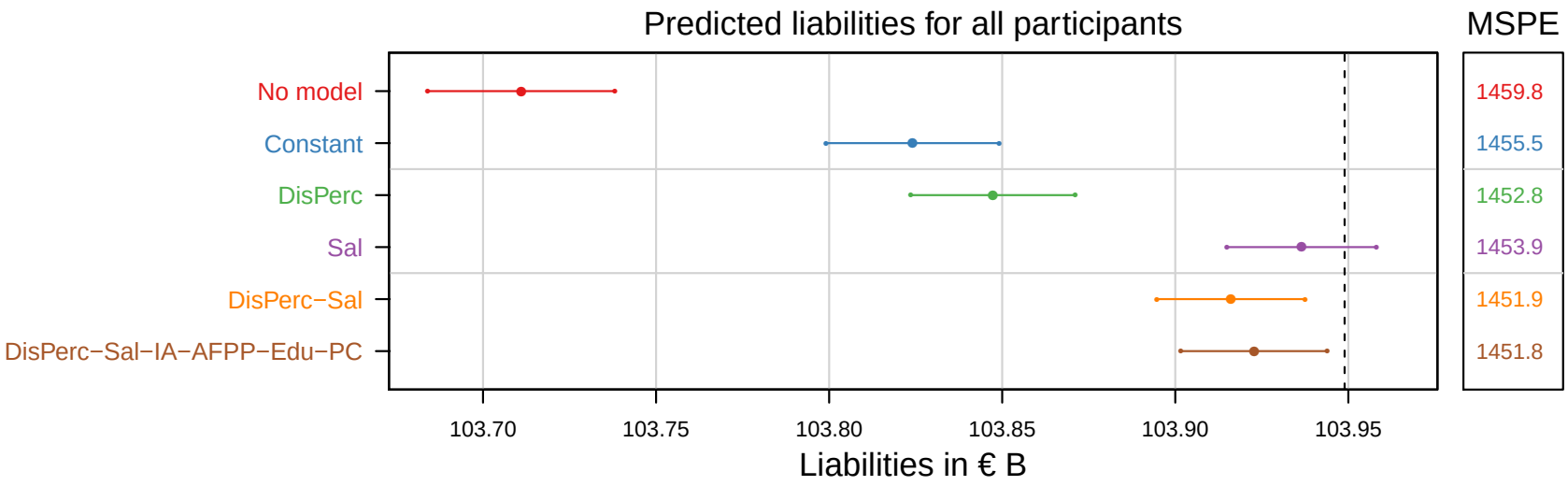
We shall compare the 90% confidence interval for Γ against the actual liabilities $\widetilde{\Gamma}$

We also calculate the **M**ean **S**quared **P**rediction **E**rror:

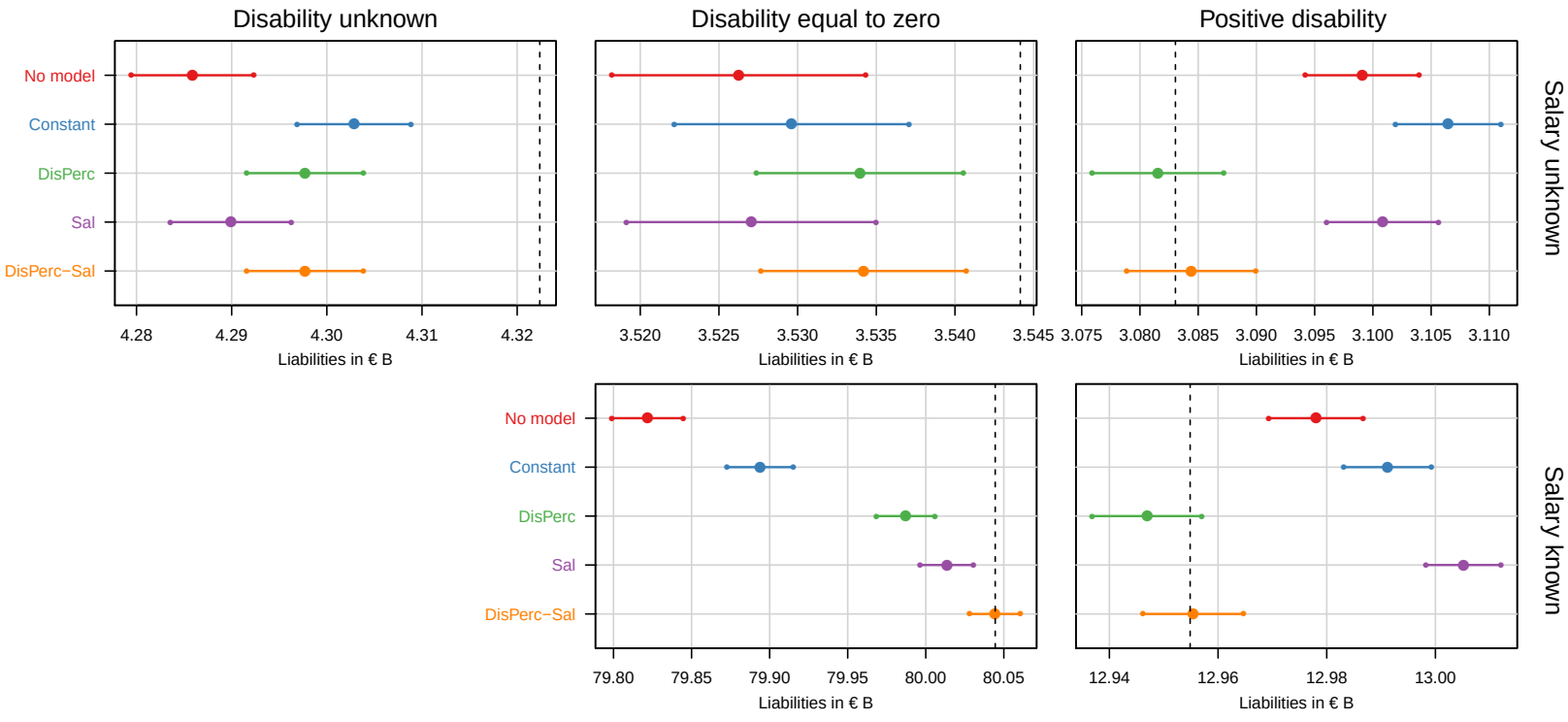
$$\begin{aligned}\mathrm{MSPE} &= \sum_{j=1}^{L_{2011}} (\mathrm{I}_{2011,j} \cdot b_j a_j - p_{2011,j} \cdot b_j a_j)^2 \Big/ \sum_{j=1}^{L_{2011}} b_j a_j \\ &= \sum_{j=1}^{L_{2011}} \underbrace{(b_j a_j)^2}_{\text{'weights'}} \underbrace{(\mathrm{I}_{2011,j} - p_{2011,j})^2}_{\text{'errors'}} \Big/ \underbrace{\sum_{j=1}^{L_{2011}} b_j a_j}_{\text{normalizing constant}} .\end{aligned}$$

Financial backtest

Mean Squared Prediction Error



Financial backtest



Conclusion

We have shown how to explain observed portfolio mortality using a wide variety of risk factors

From our model estimates we find that the following risk factors have a strong impact on mortality rates:

- Salary information;
- Disability information;
- An allowance for working at irregular hours.

Differences in remaining life expectancy at the retirement age may be more than 10 years!

For the purpose of accurately predicting the value of the liabilities, only salary and disability information seem to be crucial, and:

- Using only salary might lead to an appropriate level of liabilities for the portfolio;
- But, for subgroups the value of the liabilities might be less appropriate.

Recommendations

In general, when more risk characteristics are collected, more accurate risk profiles can be created:

- Pension funds and insurance companies should ensure their data warehouse systems are able to produce reliable and complete information based on *individual observations*.

Specifically for pensioners, few risk factors are known:

- Ideally, we would use accrued rights for pensioners, since these *are* known. However, accrued rights do not always provide accurate risk factors, because participants may have accrued rights at other pension funds;
- A solution might be to keep track of the last known salary (and part-time factor), such that for pensioners we are able to classify in terms of the distribution of normalized log salary;
- We have to be careful in estimating a single salary effect for both active participants and pensioners; we may have to distinguish between these two groups.

References

C. Czado, T. Gneiting, and L. Held. Predictive model assessment for count data. *Biometrics*, 65:1254 – 1261, 2009.

S. Gschlössl, P. Schoenmaekers, and M. Denuit. Risk classification in life insurance: methodology and case study. *European Actuarial Journal*, 1:23 – 41, 2011.

R. Plat. Stochastic portfolio specific mortality and the quantification of mortality basis risk. *Insurance: Mathematics and Economics*, 45: 123 – 132, 2009.

F. van Berkum. *Models for population-wide and portfolio-specific mortality*. PhD thesis, University of Amsterdam, March 2018. Available online at: http://hdl.handle.net/11245.1/7382ead9-cd57-4241-8a82-d4fedc0756bf.