Age Heaping in Population Data of Emerging Countries

Andres Barajas Paz, Andrew Cairns, Torsten Kleinow

Heriot-Watt University, Edinburgh

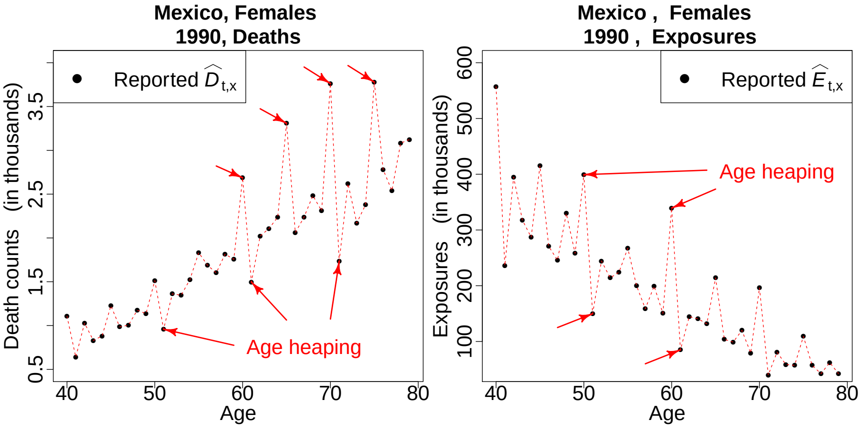
ab108@hw.ac.uk

Fourteenth International Longevity Risk and Capital Markets Solutions Conference, Amsterdam, September, 2018

Overview

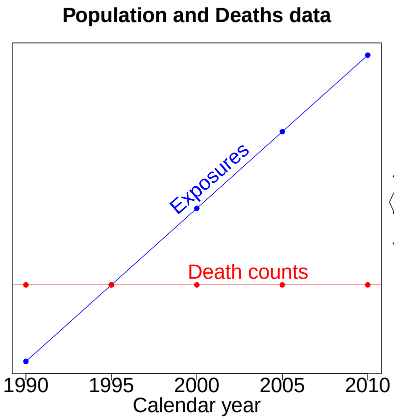
- Motivation
- Main Objective
- MLE and Bayesian approaches
- Model and Notation
- Results
- Conclusions
- Forthcoming research

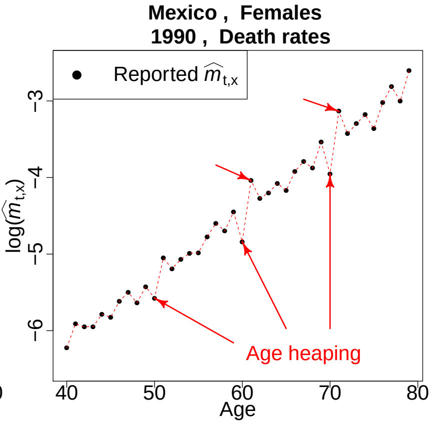
Motivation.



Age Heaping occurs when people misreport age.

Motivation.


Population and Deaths data
Exposures
Death counts
1990
1995
2000
2005
2010
Calendar year


Mexico , Females
1990 , Death rates
Reported $\widehat{m}_{t,x}$
$\log(\widehat{m}_{t,x})$
−3
−4
−5
−6
Age heaping
40
50
60
70
80
Age

Motivation.

- Mortality analyses $\xrightarrow{\text{Good quality mortality data}}$ HMD.

However, in many other countries population and deaths data can be somewhat unreliable.

- Population $\xrightarrow{\text{Misreporting of age}}$ census.
- Deaths $\xrightarrow{\text{Misreporting of age}}$ deaths data.

Main Objective

- Develop mortality models for countries where their population data is affected by age heaping.

Application: Reported data $\rightarrow$ Smoothed HMD $\rightarrow$ International Reinsurance.

MLE and Bayesian approaches

We design a model taking into account two dimensional data. Hence, we consider the data by age x and across cohorts $y = t - x$.

- First approach $\rightarrow$ MLE
- Second approach $\rightarrow$ Bayesian framework,

MLE and Bayesian approaches

For any cohort y we denote by $|y|$ the number of ages available for this cohort, that is, $n_y = |y|$ is the length of cohort y in our data set. The corresponding set of ages x is denoted by $\mathcal{X}_y$.

$$E_{x,y} \xrightarrow{\text{Age heaping}} \widehat{E}_{x,y}$$

$$D_{x,y} \xrightarrow{\text{Age heaping}} \widehat{D}_{x,y}$$

Approximate log-likelihood function

$$D_{x,y} \sim Poisson\Big(m_{x,y}E_{x,y}\Big),$$

$$m_{x,y} = \exp\left[a_y + b_y\left(x - \overline{x}\right) + c_y\left(\left(x - \overline{x}\right)^2 - \sigma_x^2\right)\right],$$

$$\ell(\theta) = \sum_{x,y} \widehat{D}_{x,y} \log\left(m_{x,y}\widehat{E}_{x,y}\right) - m_{x,y}\widehat{E}_{x,y} + C.$$

where,

$$\theta = \{\underline{a}, \underline{b}, \underline{c}\}$$

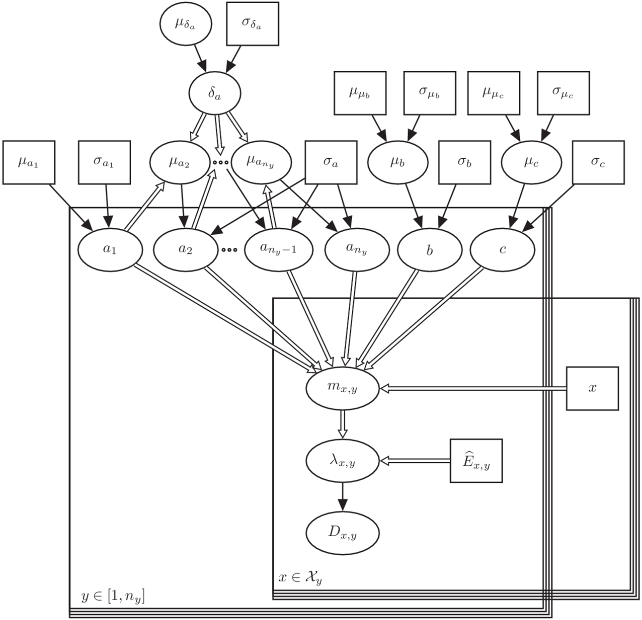
Penalised log-likelihood function

$$\ell\ell p(\theta) = \ell(\theta) - \lambda_1 p(\underline{a}) - \lambda_2 p(\underline{b}) - \lambda_3 p(\underline{c}),$$

$$p(\xi_y) = \sum_{\widetilde{y}=2}^{n_y-1} \left(\triangle^2 \xi_y\right)^2,$$

where $\triangle^2 \xi_y$ is the second order difference of ξ_y, and λ_1, λ_2 and λ_3 are the smoothing parameters.

Directed Acyclic Graph



$$\ell(\theta) = \sum_{x,y} \widehat{D}_{x,y} \log\left(m_{x,y}\widehat{E}_{x,y}\right) - m_{x,y}\widehat{E}_{x,y} + C.$$

where, $\theta = \{\underline{a}, \underline{b}, \underline{c}, \delta_a, \mu_b, \mu_c\}$

Prior distributions

$a_{y+1}|a_y \sim N\left(a_y + \delta_a, \sigma_a^2\right)$, $a_1 \sim N(0, 0.01)$, $\delta_a \sim N(\mu_{\delta_a}, \sigma_{\delta_a}^2)$

$b_y \sim N(\mu_b, \sigma_b^2)$ **iid,** $\mu_b \sim N(\mu_{\mu_b}, \sigma_{\mu_b}^2)$ **iid**,

$c_y \sim N(\mu_c, \sigma_c^2)$ **iid,** $\mu_c \sim N(\mu_{\mu_c}, \sigma_{\mu_c}^2)$ **iid**.

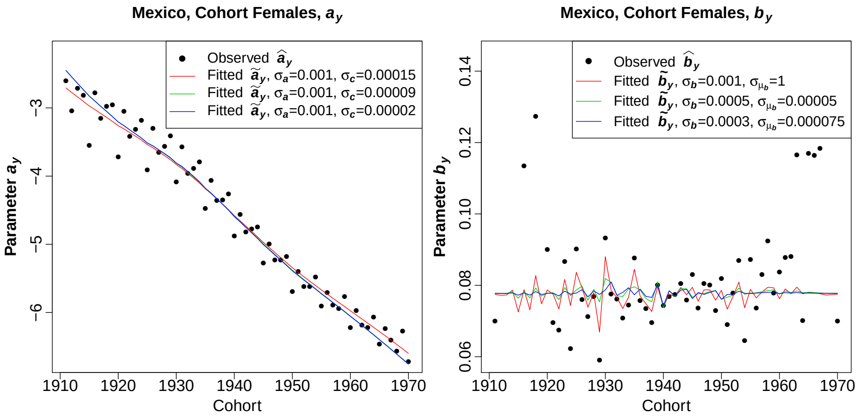
Full log posterior $\log\left(\pi\left(\theta\right)\right)$

$$
\begin{aligned}
&\propto \left[\sum_{x,y} \widehat{D}_{x,y} \log\left(m_{x,y}\widehat{E}_{x,y}\right) - m_{x,y}\widehat{E}_{x,y}\right] \\
&- \frac{1}{2}\left[\left(\sum_{y=1}^{n_y-1} \log\left(2\pi\sigma_a^2\right) + \frac{(a_{y+1} - (a_y + \delta_a))^2}{\sigma_a^2}\right) + \log\left(2\pi\sigma_{a_1}^2\right) + \frac{a_1^2}{\sigma_{a_1}^2}\right] \\
&- \frac{1}{2}\left[\sum_{y=1}^{n_y} \log\left(2\pi\sigma_b^2\right) + \frac{(b_y - \mu_b)^2}{\sigma_b^2}\right] - \frac{1}{2}\left[\sum_{y=1}^{n_y} \log\left(2\pi\sigma_c^2\right) + \frac{(c_y - \mu_c)^2}{\sigma_c^2}\right] \\
&- \frac{1}{2}\left[\log\left(2\pi\sigma_{\delta_a}^2\right) + \frac{(\delta_a - \mu_{\delta_a})^2}{\sigma_{\delta_a}^2}\right] - \frac{1}{2}\left[\log\left(2\pi\sigma_{\mu_b}^2\right) + \frac{(\mu_b - \mu_{\mu_b})^2}{\sigma_{\mu_b}^2}\right] \\
&- \frac{1}{2}\left[\log\left(2\pi\sigma_{\mu_c}^2\right) + \frac{(\mu_c - \mu_{\mu_c})^2}{\sigma_{\mu_c}^2}\right].
\end{aligned}
$$

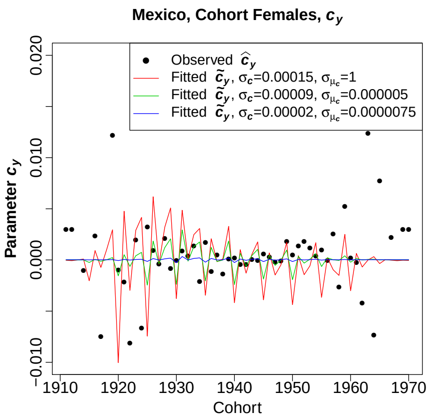
HERIOT WATT UNIVERSITY

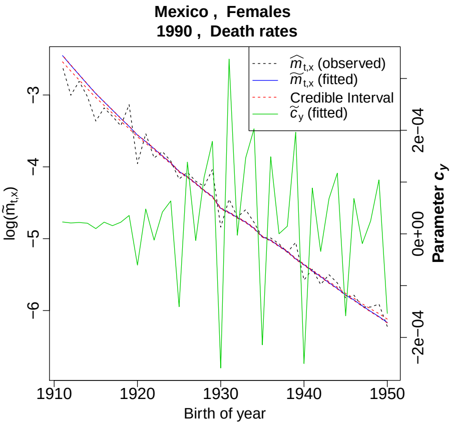
Actuarial Research Centre
Institute and Faculty of Actuaries

Parameters $\underline{a}$, $\underline{b}$ M-H, Mexico.

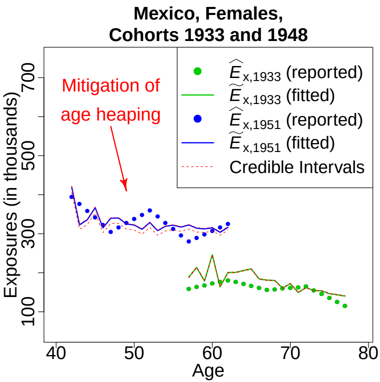


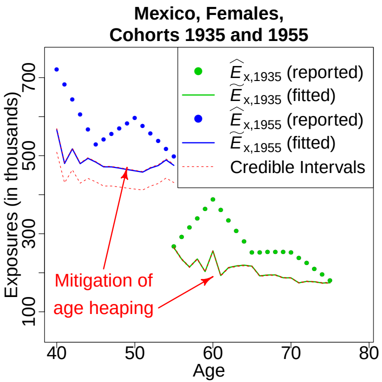
Parameter $\underline{c}$ M-H, Mexico.





Fitted exposures $\widetilde{E}_{x,y}$, Mexico.

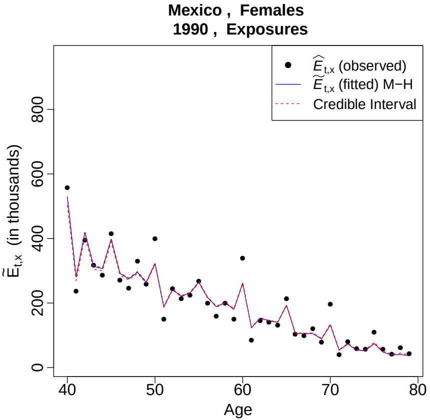


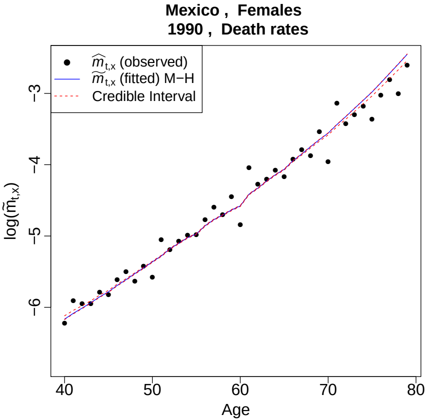


$$\textbf{Fitted exposures } \widetilde{E}_{x,y} = \frac{\widehat{D}_{x,y}}{\widetilde{m}_{x,y}} = \frac{\textbf{Reported deaths}}{\textbf{Fitted Force of Mortality}}$$

$$\textbf{where } \widetilde{m}_{x,y} = \exp\left[\widetilde{a}_y + \widetilde{b}_y(x - \overline{x}) + \widetilde{c}_y\left((x - \overline{x})^2 - \sigma_x^2\right)\right].$$

Fitted exposures $\widetilde{E}_{t,x}$, Mexico 1990.





Conclusions

- Smooth time series $\underline{c} \xrightarrow{\text{Reduce age heaping}} m_{t,x}$ and population. However, we do not want to smooth too much because it would destroy the natural volatility from the data.
- This model improves the quality of the Mexican data by reducing age heaping across all cohorts.
- The remaining volatility in the fitted exposures comes from the death counts.

Forthcoming research

- Include constraints on death counts to reduce the volatility in the fitted exposures.
- We will collaborate with HMD to see how their approach can be adapted to Mexican data for producing complete life table series, which is also relevant to international reinsurance.

Thank You!

Questions?