Exercises in Pension Systems

University of Copenhagen, Autumn 2021

13 September – 19 September (Week 2)

Purpose

There are three exercises for this week. The first exercise considers the Lee-Carter model and consists primarily of coding. You will learn how to estimate the model and use it for forecasting mortality. In the second exercise we explore how person-years of exposure are approximated and recreate HMD's estimates through raw data on births, deaths, and population counts. The third exercise deals with life disparity measures and their interpretation.

Exercise 1: The Lee-Carter model

In this exercise, we consider the model of Lee and Carter (1992) given by

$$\log m(x,t) = \log \mu(x,t) + \varepsilon_{x,t} = \alpha_x + \beta_x \kappa_t + \varepsilon_{x,t}, \tag{1}$$

where μ expresses the (true) underlying mortality rate. We suppose that occurrence-exposure data is available for time periods $t \in \{1,\dots,T\}$ and ages $x \in \{x_{\min},\dots,x_{\max}\}$. We typically say that T is the *jump-off* year, i.e. the last year of historic data, after which mortality is to be projected.

The R6-class `mortClass` estimates the Lee-Carter model under the assumption of Poisson distributed death counts

$$D(x,t) \mid E(x,t) \overset{\text{indep.}}{\sim} \text{Poisson}(E(x,t)\mu(x,t)), \tag{2}$$

The model is estimated when the class is initialized.

Question 1.0.1: Execute the following lines of code

```
HMDobj <- HMDdatClass$new(Otxtfile = 'Data/USA_Deaths_1x1.txt'
                        , Etxtfile = 'Data/USA_Exposures_1x1.txt')
OEdata <- HMDobj$getOEdata(timelim = c(1950,2019))
mort   <- mortClass$new(OEdata)
```

The `mortClass` estimates Lee-Carter models based on the data available in `OEdata`. You can extract the estimated parameters with `mortClass$par`.

Exercise 1.1: Estimating Lee-Carter by least-squares

In the original work of Lee and Carter (1992), the model (1) was estimated by SVD to minimize

$$\sum_{x,t} \left(\log m(x,t) - \alpha_x - \beta_x \kappa_t\right)^2 . \tag{3}$$

In this exercise, we will implement an algorithm that estimates the model using this approach. Estimation follows a two-stage procedure:

Stage 1 - Fitting

i) Calculate $\hat{a}_x$ as the average over time of $\log m(x,t)$ for each age, that is

$$\hat{a}_x = \frac{1}{T} \sum_{t=1}^{T} \log m(x,t)$$

ii) Compute the singular value decomposition of $\log m(x,t) - \hat{a}_x$. You can use the `svd`-function in R for this. $\hat{\beta}_x$ and $\hat{\kappa}_t$ are given by the left and right singular vectors of the SVD after normalizing so that $\sum_x \hat{\beta}_x = 1$ and $\sum_t \hat{\kappa}_t = 0$.

Stage 2 - Adjustment
The fitted death rates will in general not lead to the actual number of deaths, that is

$$\sum_x D(x,t) \neq \sum_x E(x,t) \exp(\hat{\alpha}_x + \hat{\beta}_x \hat{\kappa}_t), \quad \forall t \tag{4}$$

Lee and Carter suggested a correction to $\hat{\kappa}$ such that the RHS and the LHS in (4) are equal.

Question 1.1.1: Implement a function that estimates the Lee-Carter model using the two-stage procedure described above. (*Hint:* For the second stage, the function `lc.par2mu` computes a mortality surface based on Lee-Carter parameters.)

Question 1.1.2: Verify (e.g. by plotting) that the parameters you estimate are similar to those found in Question 1.0.1.

Question 1.1.3: What happens if you run your SVD implementation on Danish data instead of US data? If your code breaks down, suggest a simple solution so that the SVD works with the Danish data.

Exercise 1.2: Alternative parameter interpretations

The Lee-Carter model is invariant under the transformation

$$\{\alpha_x, \beta_x, \kappa_t\} \mapsto \{\alpha_x - \beta_x c, \beta_x / d, d(\kappa_t + c)\}, \tag{5}$$

for any $c \in \mathbb{R}$ and $d \in \mathbb{R} \setminus \{0\}$. The parameter constraints proposed by Lee and Carter, $\sum_x \beta_x = 1$ and $\sum_t \kappa_t = 0$, ensure identification but any parameter set satisfying (5) is equally good from a statistical point of view.

Depending on context, one specific parameter set might be preferable over another. For instance the identification constraints used by Lee and Carter make α equal to the empirical average of $\log m$, simplifying estimation and giving α a direct interpretation. For forecasting purposes, an alternative normalization might be more suitable. Suppose we forecast the time-varying index κ_t by a random walk with drift

$$\kappa_{T+h} = \kappa_{T+h-1} + \theta + \omega_{T+h}, \tag{6}$$

for some horizon $h \in \mathbb{N}_+$ with $\omega_{T+h} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

Question 1.2.1: Write down a new set of parameters $\{\widetilde{\alpha}_x, \widetilde{\beta}_x, \widetilde{\kappa}_t\}$ so that

- $\widetilde{\beta}_x = -\frac{\partial}{\partial h} \log \bar{\mu}(x, T+h)$ where $\bar{\mu}(x, T+h) = \exp(\widetilde{\alpha} + \widetilde{\beta}(\widetilde{\kappa}_{T+h-1} + \theta))$ denotes the median projection (no noise).
- $\widetilde{\kappa}_T = 0$.

Why might this parametrization be useful?

Question 1.2.2: Denote by $\{\hat{\alpha}_x, \hat{\beta}_x, \hat{\kappa}_t\}$ the parameters from `mort`. Verify that $\hat{\mu}(x,t) = \widetilde{\mu}(x,t)$ for all x and t where $\hat{\mu}$ and $\widetilde{\mu}$ denote the mortality surfaces given parameters $\{\hat{\alpha}_x, \hat{\beta}_x, \hat{\kappa}_t\}$ and $\{\widetilde{\alpha}_x, \widetilde{\beta}_x, \widetilde{\kappa}_t\}$ respectively. (*Hint:* You may find the function `all.equal` useful.)

Exercise 1.3: Forecasting and simulation

In this exercise, we will make stochastic projections using the Lee-Carter methodology. We assume throughout that projections follow a random walk with drift (6).

Question 1.3.1: Supplied with "data" $\{\kappa_t\}_{t=1,\dots,T}$, what are the ML estimators of θ and σ^2 under (6)?

From the lectures, we know that projections of κ are made conditionally on the value in the jump of year so $\kappa_{T+h} | \kappa_T \sim \mathcal{N}(\kappa_T + h\hat{\theta}, h\hat{\sigma}^2)$. Consequently, we can write the median forecast as

$$\bar{\mu}(x, T+h) = \exp(\hat{\alpha}_x + \hat{\beta}_x(\hat{\kappa}_T + h\hat{\theta}) = \bar{\mu}(x, T)\exp(\hat{\beta}_x h\hat{\theta}), \tag{7}$$

while a 95%-CI based on the innovation noise is given by $\bar{\mu}(x, T+h)\exp(\pm 1.96\sqrt{h}\sigma\hat{\beta}_x)$. There are other sources of uncertainty besides noise from the innovation process, e.g. parameter uncertainty.

Question 1.3.2: Write down an equation characterizing the lower and upper 95%-CI of μ including also the parameter uncertainty on $\hat{\theta}$. (*Hint:* What is the distribution of $\hat{\theta}$ in terms of the true parameters?)

Question 1.3.3: Implement a function in R that takes a κ-vector as input and outputs the following five projections: median, lower-CI w.o. parameter uncertainty, upper-CI w.o. parameter uncertainty, lower-CI w. parameter uncertainty, upper-CI w. parameter uncertainty. Plot an estimated κ-series and the five projections for $h = 100$.

Question 1.3.4: Implement a function in R that takes a κ-vector as input and outputs n stochastic realizations of the κ-process with forecasting horizon h. Make a few (e.g. 50) simulations and superimpose them on the plot from the previous question. Do the simulated paths behave like expected?

Question 1.3.5: Simulate 10,000 samples of $e_c(0, T+1)$ for males and females using the Lee-Carter model. Plot the two distributions (e.g. a density plot). Comment on the output. (*Hint:* You can use the function `surf2lifeexp` to compute life expectancies based on a `surf`-object. You can create a `surf`-object from an array with `mu2surf`.)

Exercise 2: Recreating HMD exposure-to-risk

Consider data on 1×1 format. We assume in the following that death rates are constant over these squares and that the population is closed. The total number of deaths among those aged $[x, x+1)$ at time $[t, t+1)$ is

$$D(x,t) = D_L(x,t) + D_U(x,t), \tag{8}$$

where D_L and D_U denote the number of lower- and upper-triangle deaths respectively. The corresponding number of people exposed-to-risk-of-death measured in terms of person-years is

$$E(x,t) = E_L(x,t) + E_U(x,t), \tag{9}$$

where E_L and E_U are exposures in the lower and upper triangle respectively. The objective of this exercise is to approximate E_L and E_U.

Let b be a time at birth within a cohort, i.e., $0 \le b \le 1$, and suppose that births are distributed according to some distribution with pdf $p(b)$, mean $\bar{b}$ and variance σ^2. See Figure 1 below.

In practice, we estimate the birth distribution using information on births by calendar months. Births are then assumed to be uniformly distributed within each month (or annually if data by month is unavailable). Therefore, define discrete intervals $0 = b_0 < b_1 < \cdots < b_{12} = 1$, where b_i are month endpoints expressed as a proportion of the year, e.g. $b_1 = 31/365$. We define the fraction of births within each subinterval as

$$\hat{p}(j) = \int_{b_{j-1}}^{b_j} p(b) \, \mathrm{d}b, \quad j = 1, \ldots, 12, \tag{10}$$

and note that $\sum_{j=1}^{12} \hat{p}(j) = 1$. For example, $\hat{p}(1)$ is the fraction of births in January. The empirical density function is then defined by

$$p(b) = \frac{\hat{p}(j)}{b_j - b_{j-1}}, \quad b_{j-1} < b \le b_j. \tag{11}$$

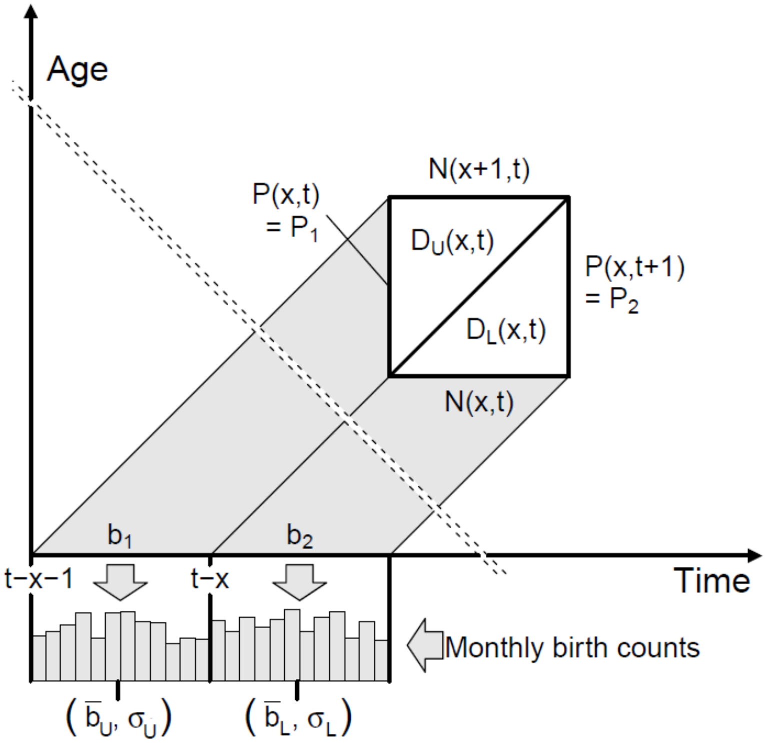

Figure 1: Available data on a Lexis square. $N(x,t)$ denotes the individuals who attain exact age x in calendar year t, while $P(x,t)$ denotes the population aged x at the beginning of calendar year t.

Question 2.1: Use (11) to determine $\bar{b}$ and σ^2 in terms of available data $\hat{p}(j)$.

Since period exposures cover the experience of two cohorts we differentiate the birth distribution by a subscript $i \in \{L, U\}$ denoting quantities corresponding to the lower and upper triangle respectively.

Question 2.2: How much exposure is contributed per survivor entering at b in the lower triangle? How much exposure is contributed per survivor exiting at b in the upper triangle? Determine s_i, the average contribution per survivor to exposure, for $i \in \{L, U\}$.

Assume that deaths in a triangle are distributed with pdf

$$f_L(a,b) = C_L p_L(b-a), \quad 0 \le a \le b < 1, \tag{12}$$

$$f_U(a,b) = C_U p_U(b-a+1), \quad 0 \le b \le a < 1, \tag{13}$$

where C_i is some constant chosen such that f_i integrates to 1.

Question 2.3: Determine C_i for $i \in \{L, U\}$.

Define z_i as the average number of years lived per death in the lower or upper triangle.

Question 2.4: Why is z_i equivalent to the average amount of lost exposure per death in a triangle?

Question 2.5: Show that $z_i = \frac{s_i^2+\sigma_i^2}{2s_i}$ for $i \in \{L, U\}$.

Question 2.6: Show that the estimated exposure in the lower and upper triangle are given by

$$E_L = (1-\bar{b}_L)P_2 + \left(\frac{1-\bar{b}_L}{2} - \frac{\sigma_L^2}{2(1-\bar{b}_L)}\right) D_L, \tag{14}$$

$$E_U = \bar{b}_U P_1 - \left(\frac{\bar{b}_U}{2} - \frac{\sigma_U^2}{2\bar{b}_U}\right) D_U. \tag{15}$$

Question 2.7: What does $E(x,t)$ reduce to when births and birthdays occur uniformly over the year for all cohorts?

Question 2.8: Using raw data for Danish births, deaths and population counts, recreate HMD's exposure-to-risk estimate for females aged 65 in 2020, that is find $E(65, 2020)$ and verify that it corresponds to HMD's estimate. You can use the following commands to load the data and set the month endpoints.

```
> # Set your wd
> # setwd("dir/")

> # Load pop, births and deaths
> DNK_pop_raw    <- read.csv("Data/DNK_Pop_raw.txt")
> DNK_births_raw <- read.csv("Data/DNK_Births_raw.txt")
> DNK_deaths_raw <- read.csv("Data/DNK_Deaths_raw.txt")

> # Define month endpoints as proportions of the year
> b.j <- cumsum(c(0,31,28,31,30,31,30,31,31,30,31,30,31))/365
```

Exercise 3 (BONUS): Life disparity

Let $\mu(x,t)$ denote the force of mortality at age x at time t. Denote by $e(x,t) = \int_x^\infty S(u,t)/S(x,t)\,\mathrm{d}u$ the (period) remaining life expectancy with $S(x,t) = \exp(-I(x,t))$ and $I(x,t) = \int_0^x \mu(u,t)\,\mathrm{d}u$. Further, to ease notation, let a *dot* over a variable denote its derivative wrt. time, e.g. $\dot{\nu}(x,t) = \frac{\partial}{\partial t}\nu(x,t)$.

Question 3.1: Show that

$$\dot{e}(0,t) = \int_0^\infty \rho(u,t)w(u,t)\,\mathrm{d}u, \tag{16}$$

where $\rho(x,t) := -\frac{\partial}{\partial t}\log\mu(x,t)$ and $w(x,t) := \mu(u,t)S(u,t)e(u,t)$. What's the interpretation of ρ and w?

Define the life table entropy conditioned on surviving to age x as

$$H(x,t) := \frac{\int_x^\infty I(u,t)S(u,t)\,\mathrm{d}u}{\int_x^\infty S(u,t)\,\mathrm{d}u}. \tag{17}$$

Question 3.2: Show that

$$H(x,t) = \frac{e^\dagger(x,t)}{e(x,t)},$$

where $e^\dagger(x,t) = \frac{1}{S(x,t)}\int_x^\infty w(u,t)\,\mathrm{d}u$. How can we interpret $e^\dagger$ and H?

Question 3.3: Show that

$$e^\dagger(x,t) = \frac{1}{S(x,t)}\int_x^\infty S(u,t)[I(u,t) - I(x,t)]\,\mathrm{d}u.$$

Question 3.4: Working from the definition of $e^\dagger$ (i.e., not the expression from Q.3.3) show that

$$\dot{e}^\dagger(0,t) = \int_0^\infty \rho(u,t)w(u,t)[I(u,t) + H(u,t) - 1]\,\mathrm{d}u,$$

and use it to conclude that $\dot{H}(0,t)$ satisfies

$$\dot{H}(0,t) = \frac{1}{e(0,t)}\int_0^\infty \rho(u,t)w(u,t)h(u,t)\,\mathrm{d}u,$$

where $h(x,t) = I(x,t) + H(x,t) - 1 - H(0,t)$.

Assume that $\rho(x,t) > 0$ for all x and t and that $H(0,t)$ is decreasing over time.

Question 3.5: Prove that there exists a unique age x_0 that separates positive and negative contributions from $\rho(x,t)$ to $H(0,t)$. How can we find x_0? (*Hint:* Consider the endpoints of $x \mapsto h(x,t)$ and show that it is an increasing function).

Question 3.6: Figure 2 shows $e(0,t)$ and $x_0(t)$ for $t = 1900,\ldots,2019$ for Swedish females. Interpret the plot.

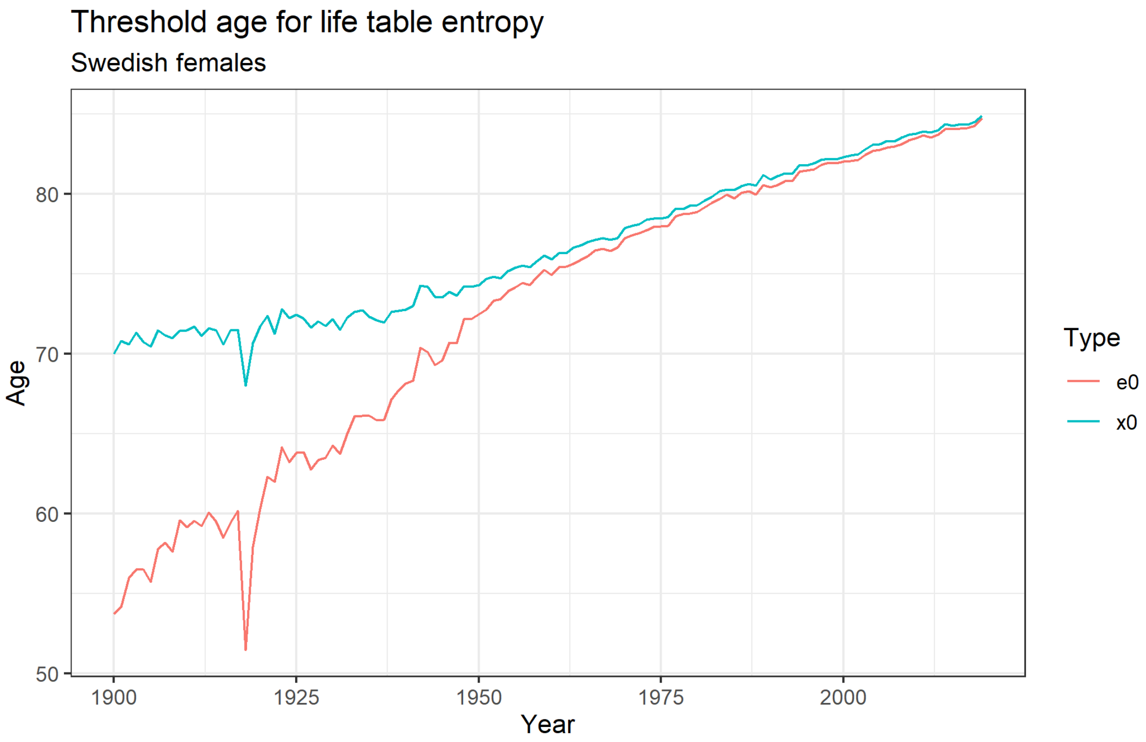

Figure 2: Threshold age for the entropy H for females in Sweden over time.