

The Lee-Carter Model

POP 502 / ECO 572/ SOC 532 • SPRING 2017

We review the Lee-Carter approach to forecasting mortality. This topic is not covered in the textbook, but their 1992 JASA article is very clear. The most distinctive feature of their approach is the use of a stochastic process to model uncertainty about the future.

The Mortality Surface

Lee and Carter seek to summarize and age-period surface of log-mortality rates $\log m_{xt}$ in terms of vectors \mathbf{a} and \mathbf{b} along the age dimension and \mathbf{k} along the time dimension such that

$$\log m_{xt} = a_x + b_x k_t + e_{xt}$$

with restrictions such that the b 's are normalized to sum to one and the k 's sum to zero, so the a 's are average log rates.

The vector \mathbf{a} can be interpreted as an average age profile of mortality, the vector \mathbf{k} tracks mortality changes over time, and the vector \mathbf{b} determines how much each age group changes when k_t changes. When k_t is linear on time each age group changes at its own exponential rate, but this is not a requirement of the model. The error term reflects age-period effects not captured by the model.

Estimation using SVD

Lee and Carter estimated the a 's, b 's and k 's with U.S. mortality data from 1933 to 1987 using least squares. Specifically, they estimate \mathbf{a} by averaging log-rates over time and \mathbf{b} and \mathbf{k} via a singular value decomposition (SVD) of the residuals, essentially a method for approximating a matrix as the product of two vectors. In a second step they adjusted the k 's so they predict the correct total number of deaths each year, but this step is not essential and I have skipped it.

The online computing logs show that one can reproduce their calculations quite closely using data from the Human Mortality Database. I used the 5x1 U.S. life table for both sexes, extracting the death rates for the standard five-year age groups up to 85+ for the years 1933 to 1987. I used the published rates up to age 80-84, but for the open-ended age group 85+ I combined ages 85-89 up to 110+ using l_{85}/T_{85} .

Figure 1 shows how well the model fits U.S. mortality in 1933 and 1987, the two extremes of the range, showing the familiar shape of mortality by age and larger relative declines at younger ages, and reproduces part of their figure 4.

Figure 1: Lee-Carter Fits for 1933 and 1987

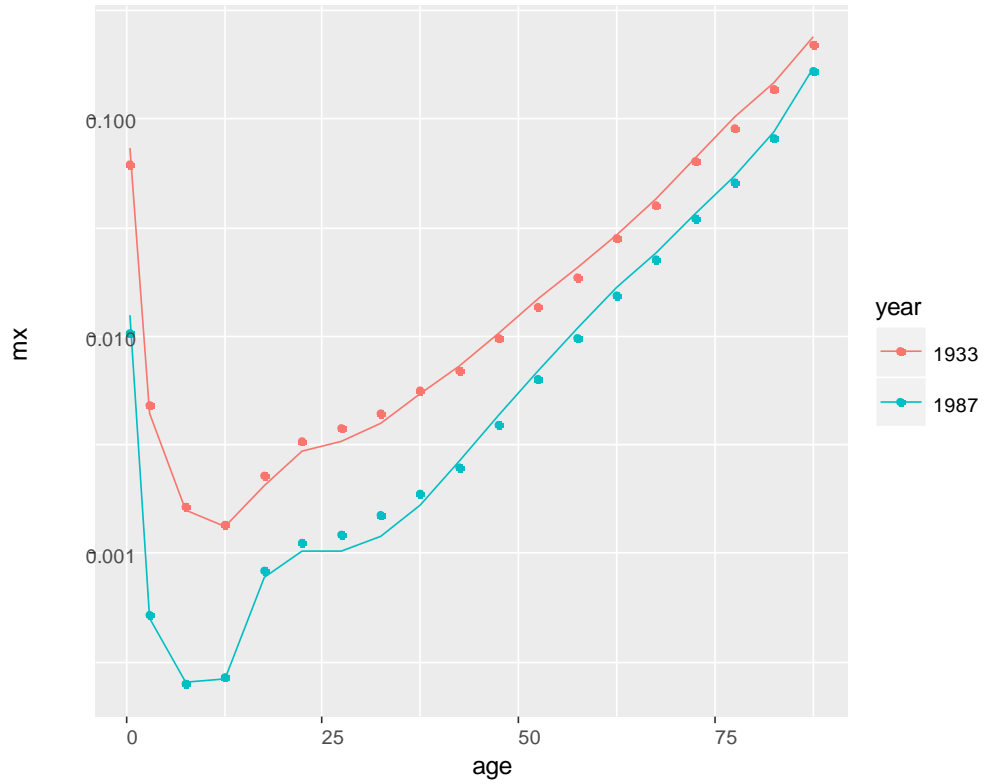
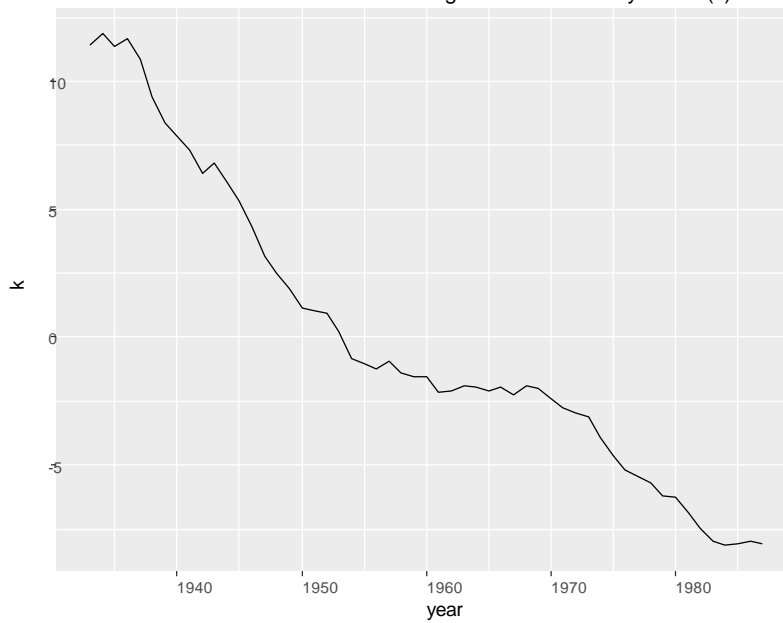


Figure 2 shows the steady decline of k over time, and reproduces part of their figure 2. (My estimates of k average zero for 1933-1987, but the trajectory is essentially the same as in the paper.)

Figure 2: The Mortality Trend (k)



The basic data used in the original paper consisted of rates up to age 85+, but Lee and Carter recognized that a large fraction of the U.S. population survives to age 85 (no less than 30% of both sexes combined, and 39% of females, in 1987), so they extended the model to older ages up to 105+. The **a** extension was based on work by Coale and Guo, and Coale and Kisker, showing that after age 80 mortality increases at a linearly declining rate, rather than the constant rate in a Gompertz model. The **b** schedule was simply kept constant after age 85. See their Table 1. These values are available on my website in a text file called LeeCarter . dat.

The Time Series Model

The second distinguishing feature of the Lee-Carter approach is that, having reduced the time dimension of mortality to a single index k_t , they use statistical time series methods to model and forecast this index. In their application to U.S. mortality they discovered that, except for the flu epidemic of 1918, the index behaves like a simple random walk with drift, where

$$k_t = k_{t-1} + d + e_t$$

where d is the drift, estimated as -0.365 , and the e_t are independent error terms with variance v , estimated as 0.652^2 . Note that the k 's are not independent; it is successive differences (or innovations) that are independent.

The variance of k_t increases with the forecast horizon t , as you might expect. Using the law of iterated expectations it is easy to show that starting from a fixed value k_0 at time t_0 , the variance of k_t is

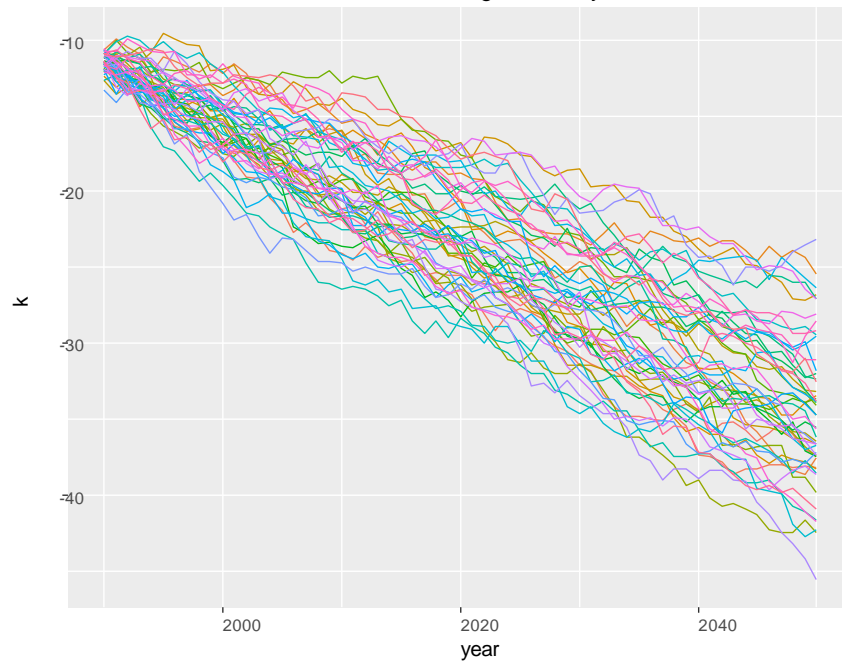
$$\text{var}(k_t) = (t - t_0)v.$$

This is important because it gives us the standard error of a forecast.

Simulating the Random Walk

Perhaps the best way to understand the stochastic nature of the projection is to do a bit of simulation. In the computing logs we set a 50-year horizon and generate 50 random trajectories, starting with a value of $k_0 = -11.05$, which is my estimate of k_{1989} . The results are shown in Figure 3. The key thing to note is how our uncertainty regarding the level of mortality increases as the projection horizon gets longer. (We could add a shaded area to represent the 95% confidence region $k_t \pm 1.96\sqrt{(t - t_0)0.652}$, but it would barely be visible behind all the lines.)

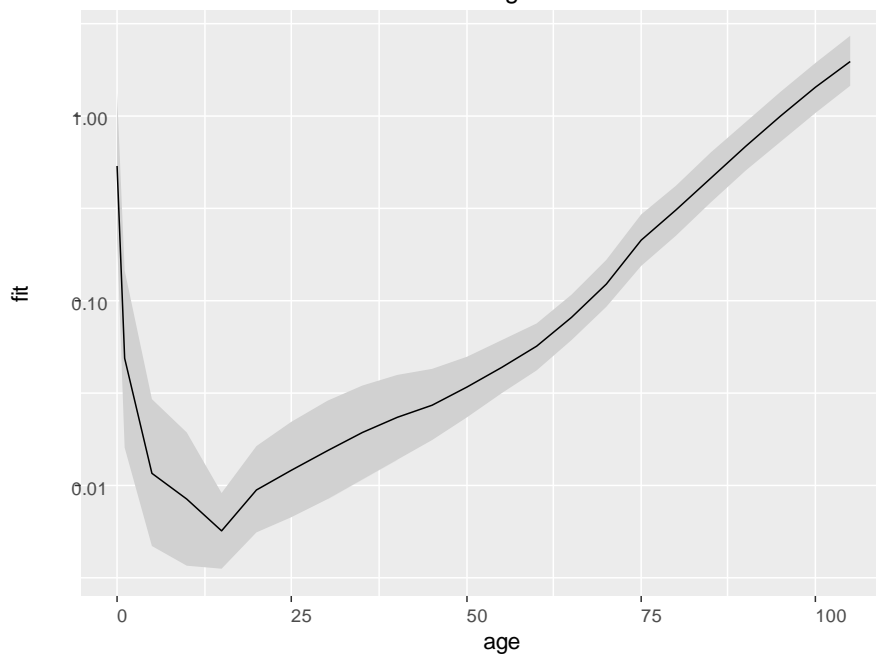
Figure 3. Fifty Random Walks



Forecasting Age-Specific Mortality

Once we have a forecast for k we combine it with the vectors \mathbf{a} and \mathbf{b} to produce a forecast of age-specific mortality. Figure 4 shows a forecast for 2050 using the published values of \mathbf{a} and \mathbf{b} with an estimated $k = -33.3$, which has a standard deviation of 5.09. (Can you reproduce these values?) The figure also shows a region bounded by the upper and lower 95% confidence bands.

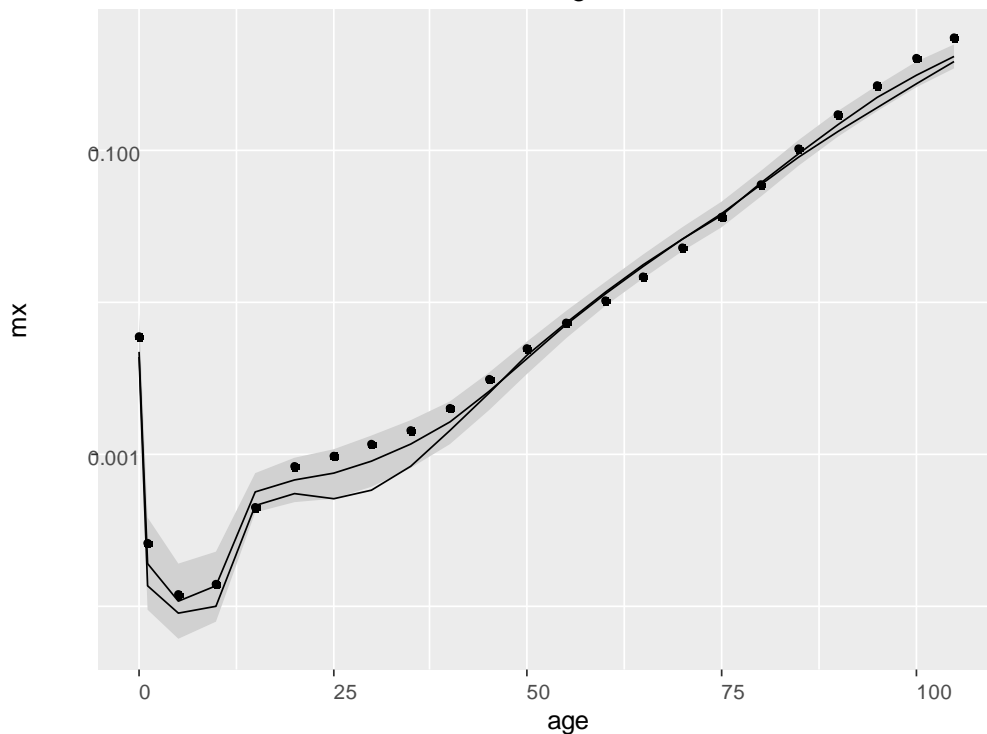
Figure 4: Forecast for 2050 with 95%



I tried to follow the same steps as in the original paper with one exception. Apparently Lee and Carter use the \mathbf{a} and \mathbf{b} schedules up to age 80-84 and close the life table using a variant of the Coale-Guo method, where the rate at 105-109 is about 0.72 and the slope declines linearly between 80-85 and 105-109. In reality the rates at 105+ are not that high, and one gets simpler and better forecasts using the published schedules for all ages, which is what I have done here.

If you forecast a recent year and then compare predicted and observed rates you will discover that the forecast is good but a bit too optimistic around ages 20 to 50. This occurs because the age pattern of mortality at the start of the forecast in 1989 already differed from the model values. The solution is to use the actual 1989 rates instead of the vector \mathbf{a} to reflect the age pattern and reset k to zero while keeping \mathbf{b} unchanged. This produces much better results, as you can see from Figure 5, which compares the rates observed in 2013 with forecasts made as of 1989 with the average \mathbf{a} and jumping from the actual 1989 rates. At first Lee and Carter worried about giving too much weight to a single year, but eventually concluded that it was better to update the age-pattern.

Figure 5: Observed and Forecast for



The next step in the forecast is to construct a full life table from the age-specific mortality rates. This can be done using standard techniques, so I'll skip the details. I find that the simple assumption of a constant risk in each age interval works well enough for most purposes. The forecast for 2050 has an expectation of life of 84.3 years with 95% confidence limits of 80.7 and 87.7 (in close agreement with Table 4 in the paper). The Social Security Administration (SSA) has their own data and they produce forecasts that are usually more pessimistic than the Lee-Carter estimates. For 2050 the SSA predicted a life expectancy of 80.2 with lower and upper bounds of 77.9 and 83.8. These differences were already apparent at the time the paper was published and Lee and Carter comment on them. They express concern that the SSA will be unprepared for the high dependency ratios that will accompany life expectancies substantially above their forecasts.