Lee-Carter forecasting of life expectancy by age, sex, and country in 2020 and 2021 and assessment of the model fit in 2019 with different reference periods

This repository contains the codes used by the study

"What reference period should be used in the estimation of life expectancy losses due to the pandemic of 2020-21? An experiment on 38 mortality series." by Vladimir M. Shkolnikov and Dmitry A. Jdanov

The estimation in the paper is carried out in two steps. The first step in Predict-mortality-8.R calculates the life table values. The outputs of this R code serve as inputs for the second step in *Calc_CI_for_ex_from_predctedLTs-2.R* which calculates the confidence intervals for the values estimated in the first step. Short description of the two steps follows.

Predict-mortality-8.R

The code builds the Lee-Carter model (Lee and Carter 2009, Booth et al. 2006) over a specific reference (fitting) period (e.g. 2000-2019 or 2010-2019). Then it calculates the model and the observed life expectancies and their age-specific components for the year 2019 (the last year before the COVID-19 pandemic) for assessment of the model fit in this year. Finally, the code predicts age-specific death rates, age-specific life expectancy values, and other life table quantities for the pandemic years 2020 and 2021 (or another prediction period) by applying the Lee-Carter model.
All calculations use standard country-sex-specific mortality and population-exposure files from the Human Mortality Database (HMD at www.mortality.org).

Please note that the code depends on the following R packages: *rstudioapi*, *demography* (Demography 2022), and *forecast*.

The file contains two functions:

--- *LTabC* (in a separate file *LTabC.R*) for computing a complete life table with parameters:
    mx - a vector of age-specific mortality rates for ages 0, 1, 2, …, 100+
    sex - m (males), f (females), or b (both sexes)
    popname - name of the population

--- *Predict_Mort_LC* (in the file *Predict-mortality-8.R*) for prediction of annual mortality for a given prediction period for a sex and a country
  Parameters of the function (see also example with multiple countries below):
    PathData - Path to data files ending with "/"
    CNTR - country abbreviation (as in the HMD)
    s - sex 1-f 2-m
    Mortfile - name of the mortality file 1x1
    Popfile - name of the population or pop exposure file 1x1
    Yretro1 - first year of the retrospective period
    Yretro2 - last year of the retrospective period
    Ypredict1 - first year of the target period
    Ypredict2 - last year of the target period
    maxAge - maximal age for prediction and all mortality and life table columns

Input Data

Organization of the input data with multiple years, countries, etc. is specified in the input file *HMD_Mx_Px_data.csv*. This file should be located in the working directory (where the code Predict-

mortality-8.R is located). The lines of the file determine the sequence of calculations. Each line provides parameters for calculation concerning one country, one sex, one retrospective period, and the prognosis period.

In each line of the file, the comma-separated fields are:

   country (e.g. Australia),

   code (e.g. AUS),

   s – sex (1-females, 2-males),

   yretro1 – the first year of the reference period (e.g. 2000),

   yretro2 – the last year of the reference period (e.g. 2019),

   ypredict1 – the first year of the prediction period (e.g. 2000),

   ypredict2 – the last year of the prediction period,

   popfile – HMD file of population exposures by single-year ages (e.g. Exposures_1x1.txt),

   mortfile – the name of HMD file of death rates by single-year ages (e.g. Mx_1x1.txt),

   pathdata – path to input data files (e.g. Input_Data/).

The example file *HMD_Mx_Px_data.csv* determines calculations for males in 38 HMD populations with the retrospective period 2000-2019 and prognosis period 2020-2021 with all the population-exposure and death-rates files located in the subfolder *Input_Data/*.

Country-, sex-, and year-specific files of population exposures and death rates by single-year ages. These files should be located in a subfolder (or subfolders) of the working directory which are specified in the field *pathdata* in the *HMD_Mx_Px_data.csv*.

All calculations by Predict-mortality-8.R utilize this data. These standard structured .txt files (e.g. Exposures_1x1.txt and Mx_1x1.txt) files from the HMD contain age- and sex-specific population exposures and death rates for selected countries. For example, AUS.Exposures_1x1.txt contains population exposures in Australia across years from 1921 to 2019 for females, males, and both sexes by ages 0, 1, 2, …, 110+, and AUS.Mx_1x1.txt contains death rates in Australia for the same dimensions. Although in the input data ages run up to 110+, the actual calculations use 100+ as the highest age group.

Output Data

The Predict-mortality-8.R code produces two output files. Names of the files are specified manually in the *write.csv* commands in the code Predict-mortality-8.R.

The first output provides data for evaluation of the model fit in the last year of the reference period. Our example: output file Dev4-ex-retro2000-19_2019m.csv. The name of this file designates something like "Deviations. life expectancy values. Reference period 2000-2019". The file has the following comma-separated fields:

   CNTR – country code (e.g. AUS),

   SEX – sex (f or m),

   Yretro1 – the first year of the reference period (e.g. 2000),

   Yretro2 – the last year of the reference period (e.g. 2019),

   Mean_dxx_obs – the mean (concerning age) of the observed d(x)*x values in Yretro2 (e.g. 2019),

   Mean_dxx_fit – the mean of the model d(x)*x values,

   RMSD – the root mean squared deviation between the observed and the model d(x)*x values in Yretro2 (e.g. 2019),

   e0_obs – the life expectancy at birth observed in Yretro2 (e.g. 2019),

   e0_fit – the model life expectancy at birth in Yretro2 (e.g. 2019),

   e15_obs, e15_fit, e60_obs, e60_fit, e80_obs, e80_fit, e90_obs, and e90_fit – the observed and the model life expectancies at ages 15, 60, 80, and 90 years.

Dev4-ex-retro2000-19_2019m.csv contains the observed and the model life expectancies for the year 2019 for 38 HMD populations corresponding to the input data described above.

The other output file provides the Lee-Carter forecasted life tables for the years 2020 and 2021. Our example *LTabs4-retro2000-19_2020-21m.csv: The filename designates* "Life tables. Reference period 2000-2019. Prediction for 2020 and 2021. Males". The file has the following comma-separated fields:

   Popx – country code (e.g. AUS),

   YEAR – year (e.g. 2020 or 2021),

   Sexx – sex (m or f),

   x – age (0, 1, 2, …, 99, 100+),

   nx – width of the age interval,

   ax – share of the age interval [x, x+1) lived by those who are dying in this interval,

   mx – central death rate in the age interval [x, x+1) and 100+,

   qx – probability of dying in the age interval [x, x+1), equals 1 for 100+,

   lx – survival to age x out of the radix=100000,

   dx – deaths in the age interval [x, x+1) and 100+,

   Lx – person-years lived within the age interval [x, x+1) and 100+,

   Tx – person-years lived at age x and all older ages,

   ex – life expectancy at age x,

   mx_l - lower 95%CI (uncertainty of prediction) for the death rate mx,

   mx_u – upper 95%CI (uncertainty of prediction) for the death rate mx,

   mx_s – standard error for the death rate mx.

LTabs4-retro2000-19_2020-21m.csv provides forecasted life tables for 2020 and 2021 for males in 38 populations.

Calc_CI_for_ex_from_predctedLTs-2.R

The second R-code calculates confidence limits (uncertainty of prediction) for the predicted age-specific life expectancies. By default, it carries out 2000 simulations. The latter number is based on a preliminary analysis. The number of iterations can be changed by changing the variable Niter in the code.

Each simulation is a calculation of the same life table from normally distributed m(x) values. The code uses the output file of predicted life tables (e.g. *LTabs4-retro2000-19_2020-21m.csv*) generated by *Predict-mortality-8.R*) as an input. It calculates the mean and the 2.5% and 97.5% percentiles of the simulated age-specific life expectancies. Finally, the code adds these CIs to the input file (e.g. LTabs4-retro2000-19_2020-21m.csv generated by *Predict-mortality-8.R*) and saves the result as the output file (e.g. LTabs5-retro2000-19_2020-21m.csv).

Note that the code carries out quite laborious calculations. For our example, the calculation of the CIs for a set of 38 populations (38 LTs times 2 years times 2000 simulations = 144 thousand life tables) lasts 8-10 minutes in my notebook.

Input Data

File with predicted life tables produced by Predict-mortality-8.R (e.g. LTabs4-retro2000-19_2020-21m.csv). The file should be located in the working directory.

The content of the file is described above.

Output Data

The *Calc_CI_for_ex_from_predctedLTs-2.R* code produces one output file (e.g. *LTabs4-retro2000-19_2020-21m.csv*). It has the same fields as the input file plus three additional quantities:

EE – mean simulated life expectancy at age x,

EElo – lower simulated 95%CI for EE,

EEhi – upper simulated 95%CI for EE.

References

Lee R.D. and Carter L.R. Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87(419): 659–671, September 1992. doi:10.1080/01621459.1992.10475265.

Booth H., Hyndman R.J., Tickle L., De Jong P. Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. Demographic Research, 15, 2006: 289-310.

Demography: Forecasting Mortality, Fertility, Migration and Population Data. 2022. https://cran.r-project.org/web/packages/demography/index.html accessed 15.12.2022