Factores de mejora para la población española

29 de octubre de 2017

Índice

1. Datos

Se han considerados edades entre 40 y 95 años y el periodo 1975-2015 (que son los años disponibles en el INE). Dado que la agregación para las edades avanzadas es diferente en el periodo 1975-1990 y 1991-2015, se han analizado los datos de la seguiente forma:

- Edades 40 a 89 años: periodo 1975-2015
- Edades 90 a 95 años: periodo 1991-2015

2. Notación

- Definimos $m_c(t,x)$ como la tasa cruda de mortalidad a la edad x en el año t.

$$m_c(t,x) = \frac{\text{Número de muertos durante el años } t \text{ y la edad cumplida es } x}{\text{Población media en el año } t \text{ y de edad cumplida } x}$$

  La población media correspondría a la *Población estacionaria* en el INE.

- La tasa de mortalidad, $q(t,x)$ es la probabilidad de que un individuo de edad x en el año t muera entre t y $t+1$.
- La fuerza de mortalidad, $\mu(t,x)$ es la tasa instantánea de muerte en el momento t para individuos de exactamente edad x en el instante t.

En general, se satisface:

1. $q(x,t) \approx 1 - exp(-m(x,t))$

3. Metodología

La metodología para el cálculo de los factores de mejora es la siguiente:

1. Estimación y proyección de $q(x,t)$ mediante modelos estocásticos de mortalidad
2. Una vez que se conocen las $q(x,t)$ proyectadas (en este caso estamos proyectando hasta 2027), definimos los factores de mejora proyectados para la edad x
3. Se calculan los factores de mejora promedio sobre los modelos utilizados

4. Modelos

Como hemos comentando, el primer paso del proceso es suavizar y proyectar las $q(x,t)$ mediante modelos estocásticos. Para reducir el riesgo de modelo (dado que ningún modelo ajusta de manera óptima para todas las edades y todos los años), hemos utilizado tres de los modelos más populares en este ámbito y hemos promediado los resultados. Los modelos utilizdos son: Lee-Carter (Lee and Carter 1992), Cairns-Blake-Dowd (Cairns et al. 2006) y Psplines (Currie et al. 2004). Estos corresponde a los modelos M1, M5 y M4 usados en Cairns et al. (2009), y que son modelos estándar en la estimación y proyección de tasas de mortalidad.

Los tres modelos se han utilizado para suavizar $q(x,t)$ para las edades de 40 a 89 años. Para las edades de 90 a 95 años se ha utilizado solo el Lee-Carter, ya que el CBD no ajusta bien para edades avanzadas y el Psplines necesita un rango más amplio de edades para que se pueda utilizar

Hipótesis estadísticas

En la literatura de los modelos estocásticos de mortalidad vemos que algunos modelos intentan modelizar $\mu(x,t)$ y otros $q(x,t)$. Esto no es un problema ya que anteriormente hemos visto la relación entre ambas cantidades. Dependiendo de qué distribución estemos utilizando, el número de muertos, $d_{x,t}$ sigue una distribución Poisson o Binomial:

- Caso Poisson;

$$d_{x,t} \sim P(e^c_{x,t}\mu(x,t))$$

  Donde $e^c_{x,t}$ es la población media en el año t

- Caso Binomial:

$$d_{x,t} \sim B(l_{x,t}, q(x,t))$$

  Donde $l_{x,t}$ es la población al inicio del año (corresponde a la variable *Supervivientes* en el INE)

4.1. Lee-Carter

Lee and Carter (1992) propuso originalmente el siguiente modelo basado en la distribución de Poisson:

$$log(\mu(x,t) = \alpha_x + \beta_x \times \kappa_t$$

Pero (Currie 2013) sugiere que se obtiene un mejor ajuste si se utiliza la distribución Binomial, es decir:

$$logit(q(x,t)) = \alpha_x + \beta_x \times \kappa_t + error$$

La estimación de todos los parámetros se hace mediante máxima verosimilitud.
Las proyecciones se hacen asumiendo un proceso ARIMA(p,d,q) para κ_t. Se han probado varias opciones y el modelo que mejor se ajusta en términos de AIC es un paseo aleatorio con drift. Por lo tanto, asumimos que los κ_t futuros se comportan como:

$$\kappa_{t+1} = \kappa_t + drift + \epsilon_{t+1} \quad \epsilon_{t+1} \sim N(0, \sigma^2)$$

4.2. CBD

Cairns et al. (2006) introdujo un modelo con dos factores para la fuerza de mortalidad:

$$logit(q(x,t)) = \kappa_t^1 + \kappa_t^2(x - \overline{x}) + error$$

DE nuevo, utilizamos la distribución Binomial. Las proyecciones de los parámetros del tiempo se hacen asumiendo un paseo aleatorio bivariante con drift para κ_t^1 y κ_t^2.

4.3. Pspline

Currie et al. (2004) propuso el uso de splines con penalizaciones (Eilers and Marx 1996) para la estimación de la fuerza de mortalidad. Esta es una técnica bien establecida como método de suavizado en el contexto de los modelos lineales generalizados. Las principales características son:

- Usa una base de B-splines para la regresión
- Modificar la verosimilitud mediante una penalización sobre los coeficientes de regresión.

El modelo se expresa como:

$$log(\mu(x,t)) = \sum_j \sum_j \theta_{ij} B_{ij}(x,t)$$

donde $B_{ij}(x,t)$ Es la base para la regresión que tiene en cuenta simultáneamente el efecto de la edad y del año, y se construye a parrtir del producto de Kronecker para tener en cuenta a la vez el efecto de la edad y del año, y θ_{ij} son los coefficientes que han de ser estimados (mediante máxima verosimilitud penaizada). La penalización que se impone a los coefficientes controla la suavidad de los datos ajustados, y depende de dos parámetros (uno que controla la suavidad a lo largo de las edades, y el otro a lo largo de los años) y que son seleccionados mediante el criterio BIC. Las predicciones se obtienen extendiendo las bases de B-splines y reajustando el modelo.

La Figura 1 muestra como no hay un único modelo que sea el que proporciona un mejor ajuste para todas las edades.

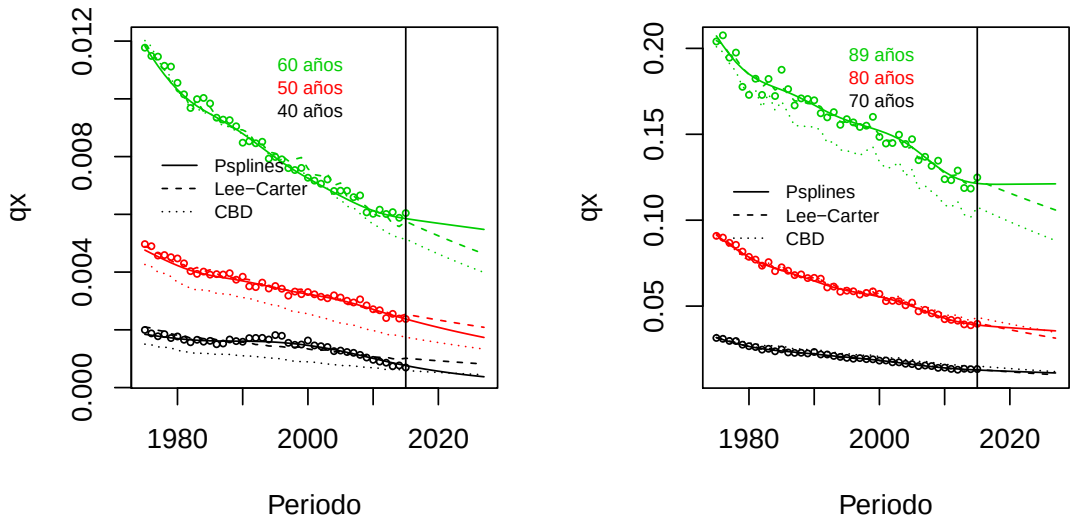

Figura 1: Ajuste y proyección de $q(x,t)$ para distintas edades utilizando los tres modelos propuestos

4.4. Cómo estimar la incertidumbre

En el caso del modelo Lee-Carter y CBD, no existe una expresión explícita para las estimaciones de los parámetros del modelo, por lo tanto, la incertidumbre sobre dichos parámetros se cunatifica mediante técnicas bootstrap. En particular, se ha utilizado bootstrap semiparamétrico propuesto por Brouhns et al. (2005). El procedimiento es el siguiente: B muestras (en este caso B=1000) del número de muertos $d_{x,t}^b$, $b = 1, \ldots, B$, son generadas a partir de una distribución de Poisson con media $d_{x,t}$ (el número de muertos observados para cada edad y año). Cada muestra bootstrap se utiliza para reestimar el modelo y obtener B parámetros estimados. Con esas estimaciones generan trayectorias proyectadas, las cuales tienen en cuenta el error de predicción y el error de modelo.

En el caso de los Psplines, es posible obtener expresiones explícitas para las estimaciones de los parámetros, y por lo tanto es inmediato calcular la incertidumbre sobre dichos parámetros tanto en el ajuste como en la proyección. En todos los casos se calculan las $q(x,t)$ proyectadas estresadas al 99,5 %

Bondad de ajuste del método propuesto

Para comprobar que efectivamente el promedio de las q_x estimadas por los tres modelos es una vuena aproximación a las q_x crudas, hemos calculado el R^2 entre los logaritmos de ambas cantidades (se ha optado por hacerlo sobre los logaritmos en vez de sobre los datos sin transformar al ser las qx valores entre 0 y 1, para los cuales no tiene mucho sentido el cálculo de R^2):

$$R^2 = 1 - \frac{\sum_{x=40}^{95}\sum_{t=1975}^{2015}(log(q_{x,t}) - log(q_{x,t}^{media}))^2}{\sum_{x=40}^{95}\sum_{t=1975}^{2015}(log(q_{x,t}) - \overline{log(q_{x,t})})^2},$$

Donde

$$q_{x,t}^{media} = \frac{q_{x,t}^{LC} + q_{x,t}^{CBD} + q_{x,t}^{Pspline}}{3}$$

Obtenemos $R^2 = 99{,}8\,\%$, lo que nos indica que la metodología utilizada es óptima.

5. Factores de mejora

Una vez que se han suavizado y proyectado las $q(x,t)$ se calculan los factores de mejora proyectados por cada modelo como:

$$\lambda_x = 1 - \left(\frac{\hat{q}(x, final)}{\hat{q}(x, inicial)} \right)^{1/(final - inicial + 1)}$$

En este caso hemos utilizado como último año proyectado el 2027 y el inicial es 2016, de modo que $final - inicial + 1 = 12$ (el número de años proyectados). Finalmente, se calculan los factores de mejora promedio sobre los modelos utilizados. Con estos factores de mejora calcularíamos las $q(x,t)$ proyectadas:

$$\hat{q}(x, f) = q(x, 2015)(1 - \lambda_x)^{f - 2015}$$

donde f es el año para el cual queremos predecir la $q(x, f)$. Para el cálculo de los factores de mejora estresados al 99,5 % procedemos de forma similar, promediando los factores de mejora calculados a partir de las $q(x,t)$ estimadas por cada modelos y estresadas al 99.5 %. La Figura 2 muestra el *Best estimate* (promediado sobre los tres modelos) y el resultado al estresarlo al 99.5 %.

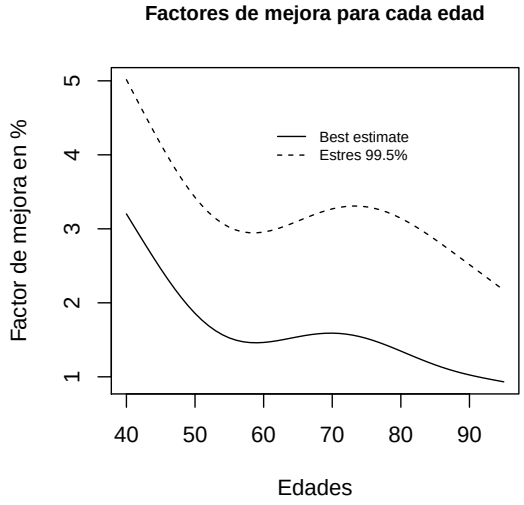

Figura 2: Factores de mejora para cada edad promediados sobre los tres modelos y estresados al 99.5 %

Finalmente, la Figura 3 muestra las $q(x,t)$ crudas y proyectadas utilizando los factores de mejora calculados para todas las edades.

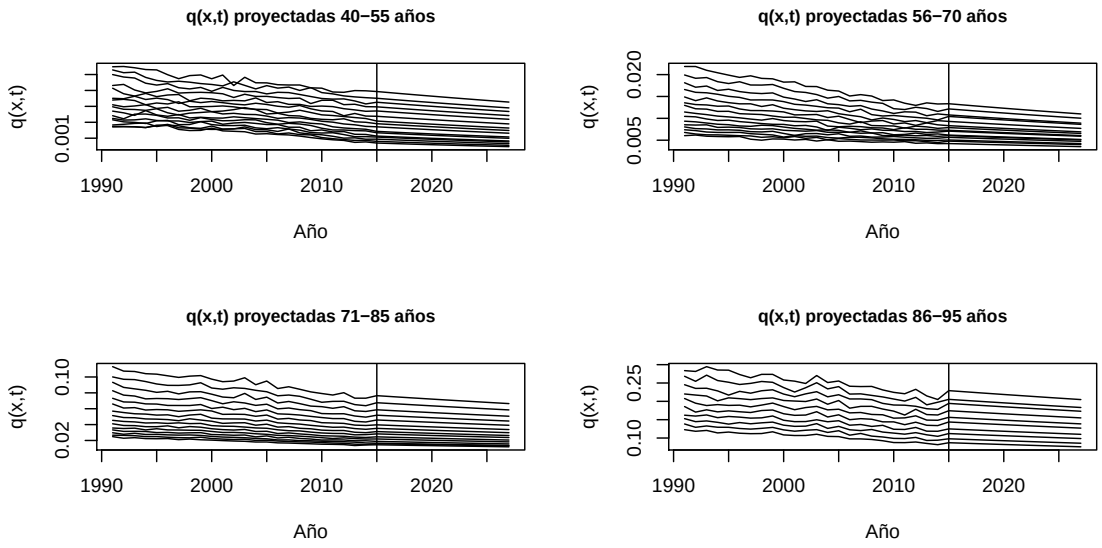

Figura 3: $q(x,t)$ cruda (hasta 2015) y proyectada (hasta 2017) utilizando los factores de mejora estimados para todas las edades

5.1. Factores de mejora anuales

Otra posibilidad es el cálculo de los factores de mejora anuales. En este caso, una vez proyectadas las q_x, se calcula la mejora de la mortalidad entre dos años consecutivos como:

$$\lambda_{x,t} = 1 - \left(\frac{\hat{q}(x,t)}{\hat{q}(x,t-1)} \right)$$

En este caso obtedríamos una matriz de factores de mejora por edad y entre años. De modo que una vez calculados estos factores de mejora, las q_x en un año futuro f se obtendrían :

$$\hat{q}(x,f) = q(x,2015) \times \lambda_{x,2016} \times \ldots \times \lambda_{x,f-1} \times \lambda_{x,f}$$

5.2. Factores de mejora para edades jóvenes y avanzadas

El cálculo de los factores de mejora para las edades jóvenes y avanzadas no es una tarea sencilla. En el caso de las edades jóvenes, existe mucha variabilidad en las q_x, y en el caso de lsa edades avanzadas, en la mayoría de los casos no se tiene suficiente masa de datos para hacer estimaciones fiables. Ante estas dificultades se han hecho ls siguientes propuestas

Edades jóvenes

Se propone el mantener constante el factor de mejora a los 40 años, 3.2 %, para las edades inferiores a 40 años. Para ello nos hemos basado en el cáculo de los factores de mejora anuales entre 1991 y 2015 para las edades de 15 a 39 años, y se ha calculado la media:

$$\frac{\sum_{x=15}^{39}\sum_{t=1991}^{2015}\lambda_{x,t}}{600} = 0{,}317 \approx 3{,}2\,\%.$$

Por lo tanto, es razonable asumir que el factor de mejora se mantiene constante e igual alde 40 añospara edades inferiores.

Edades avanzadas

El caso de las edades avanzadas es más complicado ya que a partir de ciertas edades no se disponen de datos fiables para el cálculo de las q_x. En este caso se ha optado por proyectar los factores de mejora calculados (hasta los 95 años) anteriormente. En la Figura 2 se puede apreciar que los factores de mejora para las las edades de 85 a 95 años decrece de forma lineal. Utilizando esta propiedad, ajustamos una línea recta a los 11 factores de mejora para las edad entre 85 a 95. Es decir:

$$y = \alpha + \beta x + \epsilon$$

donde: $y = (\lambda_{85}, \lambda_{86}, \ldots, \lambda_{95})$, $x = c(86, 86, \ldots, 95)$. Se obtiene $\hat{\alpha} = 3{,}06$ y $\hat{\beta} = -0{,}022$, de modo que el factor de mejora para cualquier edad, e, entre 95 y 120 se obtiene como:

$$\lambda_e = 3{,}06 - 0{,}022 \times e$$

Para el los factores de mejora estresados procedemos de igual forma y obtenedmos $\lambda_e^{99{,}5\,\%} = 8{,}69 - 0{,}068 \times e$

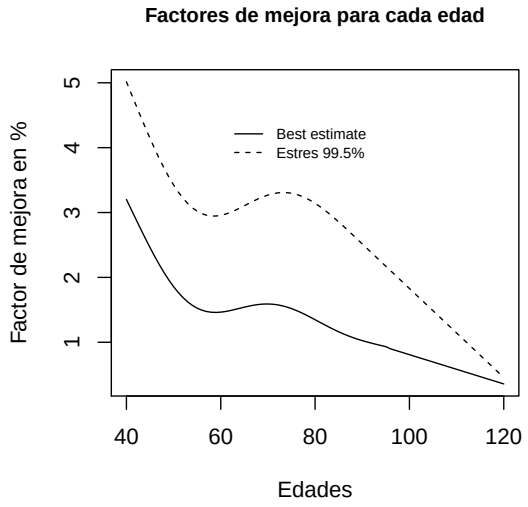

Figura 4: Factores de mejora para cada edad promediados sobre los tres modelos y estresados al 99.5 % hasta la edad de 120 años

Referencias

Brouhns, N., Denuit, M., Van Keilegom, I.: Bootstrapping the Poisson Log-Bilinear Model for Mortality Forecasting. Scandinavian Actuarial Journal, **3**, 212-224 (2005).

Brouhns, N., Denuit, M., Vermunt, J.K.: A Poisson Log-Bilinear Regression Approach to the Construction of Projected Life Tables. Insurance: Mathematics and Economics, **31**, 373-393. (2002)

Cairns, A.J.G., Blake, D., and Dowd, K.: A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. Journal of Risk Insurance. **73**, 687-718 (2006).

Currie, I. and Durban. M.: Flexible smoothing with P-splines: a unified approach. Statistical Modelling. **4**, 333 – 349 (2002).

Currie, I., Durban. M. and Eilers, P.H.C.: Smoothing and forecasting mortality rates. Statistical Modelling. **4**, 279-298 (2004).

Currie, I. D.: Smoothing constrained generalized linear models with an application to the Lee-Carter model. Statistical Modelling. **13**, 69-93 (2013).

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A., and Balevich, I.: A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. North American Actuarial Journal. **13**, 1-35

Eilers, P.H.C., Currie, I. and Durban. M.: Fast and compact smoothing on large multidimensional grids. Computational Statisticas & Data Analysis. 50, 61 - 76 (2006).

Haberman S., and Renshaw A. Age-Period-Cohort Parametric Mortality Rate Projections. Insurance: Mathematics and Economics, **45**, 255-270 (2005).

Lee, R. and Carter, L.: Modelling and forecasting U. S. mortality. Journal of the American Statistical Association. **87**, 659-671 (1992).

Eilers, P.H.C. and Marx, B.D.: Flexible smoothing with B-splines and penalties. Statistical Science. **11**, 89 – 121 (1996).