UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Maestría en Estadística Aplicada
Trabajo de grado
Marzo 2021

Proyección de la esperanza de vida en Colombia: aproximación a través de métodos bayesianos

Life expectancy forecast in Colombia: approach through Bayesian methods

Juan Carlos Castillo Abril[a]
juan.castilloa@usantotomas.edu.co

Wilmer Darío Pineda Ríos[b]
wilmerpineda@usantotomas.edu.co

Resumen

La esperanza de vida al nacer es uno de los principales indicadores utilizados en el análisis y proyección demográfica, con el propósito de sintetizar las condiciones de vida y otras dimensiones sociales de un país o territorio. Esta medida representa la duración promedio de vida de una generación ficticia sometida a condiciones de mortalidad observadas en un año determinado de estudio. El propósito de este trabajo es estudiar la dinámica de la esperanza de vida en Colombia y obtener un de esta, para un periodo de 10 años a partir de 2019.

***Palabras clave*:** Esperanza de vida, modelos lineales generalizados, modelos para datos de conteo (Poisson, Binomial Negativo), Demografía (Tasas de mortalidad), estadística bayesiana.

Abstract

Life expectancy at birth is one of the main indicators used in demographic analysis and projection, with the purpose of synthesizing the living conditions and other social dimensions of a country or territory. This measure represents the average life span of a fictitious generation subjected to conditions of mortality observed in a given year of study. The purpose of this work is to study the dynamics of life expectancy in Colombia and obtain a forecast of it, for a period of 10 years from 2019.

***Keywords*:** Life expectancy, generalized linear models, models for counting data (Poisson, Negative Binomial), Demography (Mortality rates), bayesian statistics.

1. Introducción

Algo innegable es que el crecimiento poblacional en el mundo se ha acelerado, principalmente desde mediados del siglo XX y principios del siglo XXI, donde la población mundial alcanzó cerca de 7000 millones de habitantes. Colombia no ha sido ajena a esta dinámica, ya que entre finales del siglo XVIII y comienzos del siglo XX, paso de tener 900 mil habitantes a unos 4 millones de personas (Acosta 2014). Para Noviembre de 2018 según el DANE [1] estaríamos conformados por 45.5 millones de habitantes. El

[a] Estudiante de Maestría en Estadística Aplicada
[b] Director de Tesis. Magíster en Estadística y candidato a Ph.D
[1] Departamento Administrativo Nacional de Estadística

crecimiento de la población ha estado acompañado de un cambio en la distribución por edades y un aumento en el promedio de edad de la población, mostrando síntomas de envejecimiento (Acosta 2014).

Se estima que los países de más alto desarrollo se encuentran en etapas donde se observa un aumento de la esperanza de vida a más de 80 años y cuyas principales causas de muerte se limitan a aquellas ocasionadas por la debilidad de la vejez. Colombia es un país con características propias de ciclos avanzados y primarios de la transición epidemiológica. Mientras que a inicios del siglo XX el país tenía como principal causa de muerte, los factores de riesgo como enfermedades de tipo infecciosa y parasitaria, en los últimos años se está moviendo hacia enfermedades relacionadas con el sistema circulatorio y los cánceres, propias de las edades más avanzadas. No obstante, las causas externas como los homicidios y accidentes de transporte terrestre aún se encuentran dentro de las principales causas de muerte (Acosta 2014).

Para Horiuchi (1991) una población envejece cuando se presentan tasas de fecundidad bajas y una reducción de las tasas de mortalidad. De acuerdo con McKcown (1976) el incremento del ingreso permitió unas mejores condiciones de vida y nutrición para la población, traduciéndose en el aumento de la esperanza de vida. Colombia es un país con un aumento progresivo de la población y su expectativa de vida (producto de la caída en de la tasa de mortalidad) y la reducción de la base de la pirámide poblacional (o envejecimiento de la población), los cuales tienen consecuentes cambios sobre su perfil epidemiológico.

Una amplia investigación realizada por expertos del Instituto de Métrica y Evaluación de la Salud (IHME) en la Universidad de Washington, Estados Unidos, reveló que la esperanza de vida ha aumentado en promedio 6 años durante el último cuarto de siglo en el mundo, incluso en los países más pobres. El estudio, publicado en la prestigiosa revista médica The Lancet, incluyó información de 188 países.

En Colombia se han logrado progresos sustanciales en salud en los últimos 25 años, por lo que ha aumentado la esperanza de vida y ha reducido las cargas de salud de enfermedades mortales y lesiones como enfermedades cardíacas, accidentes cerebrovasculares y accidentes de tráfico. En 2016 la esperanza de vida en Colombia subió hasta llegar a 74.38 años. Ese año la esperanza de vida de las mujeres fue de 78 años, mayor que la de los hombres que fue de 70.85 años. Colombia ha ascendido en el listado de los 192 países por Esperanza de Vida publicado por la Oragnización Mundial de la Salud, pasando de ocupar el puesto 91 en 2015 al 89 en 2016. Esto quiere decir que se sitúa aproximadamente en la parte media del ranking de países por esperanza de vida. Si miramos la evolución de la Esperanza de Vida en Colombia en los últimos años, vemos que ha subido respecto a 2015 en el que fue de 74.2 años, al igual de lo que ocurre respecto a 2006, en el que estaba en 72.53 años.

La importancia del estudio de estos cambios demográficos y epidemiológicos radica en que su entendimiento es concluyente en las decisiones y planeación de políticas públicas de salud, pensiones y fiscales, así como en la prevención de los principales factores de riesgo a los que está expuesta la población.

El propósito del trabajo es pronosticar la esperanza de vida en Colombia en un horizonte de 10 años, aunque trabajos como el de (Florez 2020) se han desarrollado para la zona norte del país, en donde usaron modelos jerárquicos bayesianos para realizar mapeos de mortalidad para la región caribe, no obstante, no se presentan pronósticos de la esperanza de vida para todo el territorio nacional, como se considera en el presente trabajo, que toma como referencia trabajos similares aplicados sobre otros países; se considera entonces, que al desarrollar esta perspectiva para la nación, implicaría un aporte significativo al estudio demográfico colombiano.

Para (Benito 2008) la estadística bayesiana es el único enfoque en el que se hace un uso explícito de la probabilidad para cuantificar la incertidumbre de la inferencia. Se trata de un proceso de aprendizaje iterativo en el que se alcanzan conclusiones sobre un fenómeno (distribución posterior) a partir del conocimiento previo sobre el sistema (distribución previa) y de nuevas evidencias (información proveniente de los datos). Es decir, que los resultados obtenidos se pueden utilizar para la actualización de lo que se conoce sobre el sistema y además incluirlos en estudios posteriores.

Este trabajo tiene una breve introducción en la primera sección, en la segunda sección se desarrolla el marco teórico, en donde se revisan algunos conceptos necesarios para cumplir con los objetivos propuestos,

en la sección 2.1 se abordan tópicos básicos y se explica lo que se entiende por tabla de mortalidad. En la sección 2.2 se describe de qué fuente de información obtendremos las tasas de mortalidad para Colombia. En la sección 2.3 se explica cómo se caracterizan los efectos de edad, sexo y tiempo y en la sección 2.4 como se cuantifican las interacciones entre los efectos. En el apartado 2.5 se definen los modelos que se ajustaran a los datos. Se planean los objetivos en la sección 3, por su parte en la 4 se describe la metodología. En la sección 5 se presenta los datos y escalas para Colombia, por su parte en la sección 6 se muestran los resultados y en la sección 7 se exponen las conclusiones, por último, se exhiben las referencias bibliográficas.

2. Marco teórico

El concepto de esperanza de vida al nacer implica conocer la estructura probabilística de la dinámica de la mortalidad de una población objetivo. En este sentido, la tabla de mortalidad resulta una herramienta adecuada para describir esta estructura. Una tabla de mortalidad se interpreta como un modelo que representa la distribución de probabilidad del tiempo de sobrevivencia esperado de los miembros de un grupo determinado. Para su construcción, se compila la información del número de individuos de la población, sexo, grupos de edad expuestas al riesgo de muerte (cohorte) y momento en el cual es observado este riesgo (Zarruk 2012).

En la fase de análisis y definición del modelo para la esperanza de vida, se tomará en cuenta las interacciones entre edad, sexo y tiempo. En la mayoría de datos demográficos, hay interacciones que requieren atención en la etapa de modelamiento. En esta sección, se prestará atención de forma particular en modelar cómo los distintos patrones de edad y sexo cambian durante el tiempo.

Las variables interactúan cuando la naturaleza de la relación entre una variable y el resultado de interés depende del nivel de una o más variables. Si por ejemplo la relación entre edad y mortalidad difiere entre mujeres y hombres, entonces diremos que hay interacción entre edad y sexo.

El objetivo principal de este estudio es pronosticar la esperanza de vida en Colombia para los años 2019-2028. Para esto, se define el estimador que describe las tasas de mortalidad global de la población colombiana por edad (a), sexo (s) y tiempo (t). La entidad oficial de estadística en Colombia es el Departamento Administrativo Nacional de Estadísticas (DANE). El DANE se encarga de todo tipo de estadísticas producidas para el país, incluidos los datos de entrada para este modelo, que corresponden al número de muertes observadas y conteos de población expuesta al riesgo de morir, denominadas exposiciones, cada una de estas cantidades discriminadas por edad y sexo. En la sección 2.2 (tasas de mortalidad) se describe detalladamente como la organización **Latin American Human Mortality** a partir de los datos del DANE conforma la fuente de datos para Colombia, correspondiente a los recuentos de muertes y población, por edad-sexo para el periodo de estudio. Para obtener las proyecciones de esperanza de vida, en primer lugar, se supondrá que el número de muertes siguen una distribución Poisson, ya que es la distribución más simple que se puede usar para datos de conteo, que coloca su masa en el conjunto de enteros no negativos. Sus probabilidades dependen de un solo parámetro, la media $\mu > 0$.

La función másica de probabilidad de una variable aleatoria con distribución Poisson se define como, $p(y;\mu) = e^{-\mu}\frac{\mu^y}{y!}$ para $y = 0, 1, 2, \ldots$, la cual pertenece a la familia exponencial donde además se tiene que $E(y) = var(y) = \mu$. La distribución Poisson es unimodal con moda igual a la parte entera de μ. Su sesgo se describe por medio de la siguiente expresión $E\left(\frac{(y-\mu)^3}{\sigma^3}\right) = \frac{1}{\sqrt{\mu}}$.

La distribución Poisson es un caso particular de la familia exponencial, que abarca distribuciones estándar como la normal, binomial y la misma Poisson. Debido a que, en este caso, la distribución de la variable respuesta es no normal y adicionalmente la función que conecta el componente aleatorio y el predictor lineal es el logaritmo natural, entonces es conveniente utilizar un Modelo Lineal Generalizado (GLM) loglineal que tiene como función de enlace el logaritmo natural para una respuesta Poisson (datos de conteo). Debido a que, el GLM extiende el alcance del modelo lineal a (1) Distribuciones de respuesta

No Normales y (2) funciones de enlace de la media igualadas al predictor lineal.

En el modelo lineal generalizado para variables aleatorias Poisson, la variable aleatoria Y tiene una distribución Poisson, es decir, para un valor $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}$ se tiene:

$$Y|\boldsymbol{x} \sim Poisson(\exp(\boldsymbol{c^t x} + d))$$

Por tanto, el valor de Y dado $\boldsymbol{x}$, se distribuye Poisson con parámetro $\exp(\boldsymbol{c^t x} + d)$ en donde $\boldsymbol{c} \in \mathbb{R}^{\boldsymbol{n}}$ y $d \in \mathbb{R}$, para que las operaciones sean compatibles.

Ahora bien, por propiedades de la distribución Poisson, (explicado arriba) el logaritmo de su valor esperado se puede modelar utilizando una combinación lineal de los parámetros desconocidos, de donde, el logaritmo es la función de enlace canónica.

Un concepto fundamental y cada vez más utilizado, es el de estadística bayesiana, que como afirma (Pineda 2018) se fundamenta en el teorema de Bayes, que se expresa utilizando la probabilidad condicional:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \propto P(B|A)P(A)$$

Donde $P(A)$ es la probabilidad previa (a priori), $P(B|A)$ es la verosimilitud asociada a la previa y $P(A|B)$ es la probabilidad posterior (a posteriori). Del Teorema de Bayes entonces, obtenemos estimaciones basadas en el conocimiento interno del conjunto de datos. La metodología bayesiana específica un modelo de probabilidad, el cual contendrá de alguna manera conocimiento previo acerca de un parámetro que está siendo investigado, acondicionando el modelo de probabilidad para realizar el ajuste de los supuestos. En la sección (2.6) se explicara con mayor detalle las herramientas utilizadas en este trabajo para desarrollar la inferencia bayesiana.

2.1. Tablas de mortalidad

Como dice (Zarruk 2012) una tabla de mortalidad puede interpretarse como un modelo que representa la distribución estadística del tiempo de sobrevivencia esperado de los miembros un grupo determinado. Para la elaboración de una la tabla de mortalidad, en esencia, se necesita compilar información del número y edad de las personas expuestas al riesgo de muerte, así como sus edades al momento de la muerte.

En la práctica, con el fin de poder clasificar a los individuos según su edad se recurre a dos conceptos asociados a esta

- **Edad actuarial** Se asigna edad x al individuo que tiene la edad comprendida en el intervalo $[x - 1/2, x + 1/2]$.
- **Edad entera alcanzada** Se asigna edad x al individuo que tiene la edad comprendida en el intervalo $[x, x + 1)$.

La estructura clásica de una tabla de mortalidad está constituida por las siguientes columnas,

x	l_x	d_x	q_x	p_x	e_x

El siguiente ejemplo ha sido tomado de la resolución número 1555 de 2010 de la superintendencia financiera de Colombia (SFC) y presenta un fragmento de tabla de mortalidad de rentistas mujeres. Las tablas vigentes se encuentran en el sitio web (Resolución 1555 de 2010, SFC)

x	l_x	d_x	q_x	p_x	e_x
15	1 000 000	272	0,000272	0,999728	70
⋮	⋮	⋮	⋮	⋮	⋮
20	998 570	311	0,000311	0,999689	65,1
⋮	⋮	⋮	⋮	⋮	⋮
40	998 699	863	0,000864	0,999136	45,7
41	987 836	926	0,000937	0,999063	44,7
42	986 910	994	0,001007	0,998993	43,7
43	985 916	1 070	0,001085	0,998915	42,8
44	984 846	1 152	0,00117	0,99883	41,8
45	983 694	1 242	0,001263	0,998737	40,9
⋮	⋮	⋮	⋮	⋮	⋮
60	949 454	4 082	0,004299	0,995701	27
⋮	⋮	⋮	⋮	⋮	⋮
106	1828	878	0,480306	0,519694	1,4
107	950	492	0,517895	0,482105	1,3
108	458	256	0,558952	0,441048	1,1
109	202	121	0,59901	0,40099	0,9
110	81	81	1	0	0,5

- x edad del individuo, $0 \leq x \leq \omega$, donde ω es la edad maximal (100+).
- l_x número de sobrevivientes a la edad x o exposiciones, asumiendo que se toma una cohorte inicial de l_0 recién nacidos.
- $d_x = l_x - l_{x+1}$ número de personas que fallecen entre las edades x y $x+1$.
- $q_x = d_x/l_x$ probabilidad de que un individuo de edad x no sobreviva a la edad $x+1$.
- $p_x = 1 - q_x = l_{x+1}/l_x$ probabilidad de que un individuo de edad x sobreviva hasta a la edad $x+1$.
- $e_x = 0,5 + \frac{\sum_{\{x_i:x_i>x\}}^{\omega} l_{x_i}}{l_x}$ esperanza de vida a la edad x. Corresponde al número de años esperados de vida para una persona de edad x, es decir, al número de años promedio que vivirá la persona después de los x años ya alcanzados.

Asumiendo que inicialmente se tenía un grupo hipotético de 1.000.000 mujeres de 15 años de edad, de acuerdo con esta tabla, a la edad 40 sólo quedan 998.699 individuos vivos, o $l_{40} = 998.699$. Dado que mueren 1.242 a la edad 45, o $d_{45} = 1,242$, de los 983.694 sobrevivientes, la probabilidad de morir a esta edad es $q_{45} = d_{45}/l_{45} = 0,001263$, que también se puede interpretar como una tasa de $1,263$ muertes por cada mil; $e_{60} = 0,5 + \frac{\sum_{\{x_i:x_i>60\}}^{\omega} l_{x_i}}{l_{60}} = 27$ significa que, en promedio, se espera que las mujeres aseguradas de edad 60 vivan 27 años más, es decir, hasta la edad 87. También se observa que a la edad 110 aún sobreviven 197 personas y que este subgrupo mueren antes de alcanzar la edad 111. Al límite de la tabla lo denotaremos como w , en este caso $w = 111$. (Zarruk 2012).

Tomando los valores presentados en la tabla de mortalidad de rentistas mujeres, sea por ejemplo $x = 106$, el calculo de la esperanza de vida a la edad x esta determinado como sigue:

$$\begin{aligned} e_{106} &= 0,5 + \frac{\sum_{\{x_i:x_i>106\}}^{\omega} l_{x_i}}{l_{106}} \\ &= 0,5 + \frac{950 + 458 + 202 + 81}{1828} \end{aligned}$$

$$= 0,5 + \frac{1691}{1828}$$
$$= 1,4$$

2.2. Tasas de mortalidad

Los datos utilizados para el caso colombiano, han sido tomados de la Base de Datos de Mortalidad Humana de América Latina (LAHMD) (Latin American Human Mortality Database) [www.lamortalidad.org] proyecto inspirado en The Human Mortality Database [www.lamortalidad.org] y es el resultado del trabajo conjunto de la Profesora Beatriz Piedad Urdinola de la Universidad Nacional de Colombia, Departamento de Estadística de Bogotá y el Profesor Bernardo Lanza Queiroz CEDEPLAR, Belo Horizonte, Brasil, financiado por The Population Association of America (www.popassoc.org) y la Dirección de Investigación de la Universidad Nacional de Colombia-Bogotá (DIB www.dib.unal.edu.co). En la actualidad, la base de datos contiene información detallada sobre la mortalidad de siete países de América Latina: Argentina, Brasil, Colombia, Chile, Ecuador, México y Perú. Toda la información está desglosada por edad, sexo, región y causa de muerte. Adicionalmente, existe información sobre la literatura académica sobre el estudio de la mortalidad para estos mismos países.

La base de datos para Colombia correspondiente a las defunciones, tomada de LAHMD se conformó a partir de la información suministrada, por parte del Departamento Administrativo Nacional de Estadísticas (DANE, www.dane.gov.co) que es la entidad oficial encarga de producir todo tipo de estadísticas para el país, incluidos censos y registros civiles. Los registros de mortalidad por edad y sexo de 1970 a 1978 se transcribieron del Informe del DANE titulado "Registro de defunciones en Colombia 1970-1978". Publicado por el DANE y el Fondo de las Naciones Unidas para la Infancia (UNICEF) en 1987. Desde el año 1979 el DANE proporcionó registros individuales de defunción en formato electrónico, los cuales cubren toda la población nacional registrada por lugar de residencia e incluyen datos sobre la región y la causa de muerte. De 1979 a 1996 el DANE recopiló los registros de defunción siguiendo la Clasificación Internacional de Enfermedades, novena edición (CIE IX), publicada en 1977 por la Organización Mundial de la Salud (OMS) y cuyo fin era clasificar las enfermedades, afecciones y causas externas de enfermedades y traumatismos, con objeto de recopilar información sanitaria útil relacionada con defunciones, enfermedades y traumatismos (mortalidad y morbilidad). Desde el año 1997 hasta el último disponible, el DANE registra e informa la causa de la muerte exactamente como en la versión CIE-X, y así se continúa haciendo. En cuanto a los datos de población por edad y sexo (exposiciones) fueron proporcionados electrónicamente por el DANE al proyecto LAHMDB en números agregados por edades individuales.

Adicionalmente la información se complementó con los micro datos de mortalidad suministrados por el DANE en su sitio de Estadísticas Vitales (COLOMBIA - Estadísticas Vitales - EEVV - 2017 - 2018) [microdatos.dane.gov.co] esta información esta segmentada por sexo y por grupos de edad $0, [1-4], [5-9], \ldots, 85+$. Los datos para Latinoamérica en general muestran niveles importantes de subregistro, los registros para Colombia no son la excepción, por tanto, para evitar sobreestimaciones, se debería aplicar métodos para corrección de los datos, cabe aclarar que en este trabajo no se hizo dicha corrección y se tomaron los datos directamente de LAHMD.

Cuando se trata estadísticas de mortalidad, es una práctica común separar al grupo etáreo $[0-4]$ años correspondiente a infantes, en dos grupos de edad, el primero hace referencia a la edad 0 (bebes menores a 1 año) y el otro, al grupo de edad $[1-4]$ años. Por su parte, el grupo de edad 85+ se refiere a todas las personas con edad mayor o igual a 85 años, este grupo es usualmente pequeño.

Es común, en estudios demográficos identificar la población mayor, lo cual equivale a obtener la cantidad de personas que llegan a edades más avanzadas, producto del denominado proceso de envejecimiento poblacional. La obtención de la proporción de mayores, entre otras cosas, permite medir el impacto de las tendencias demográficas, con lo cual se pueden desarrollar previsiones tempranas, que se consoliden en medidas para afrontar los retos que una sociedad con una significativa participación de población mayor supone.

Una forma de cuantificar la participación de nuestra población mayor, es utilizar la información suministrada por el censo de población y vivienda realizado en 2018, el cual consistió en contar y caracterizar las personas residentes en Colombia, así como las viviendas y los hogares del territorio nacional. A través del censo, el país obtuvo datos de primera mano sobre el número de habitantes, su distribución en el territorio y sus condiciones de vida permitiendo al país ser más preciso en la toma de decisiones y orientar políticas públicas en aspectos sociales (DANE 2018). Las figuras (1 - 2), muestran cómo está compuesta la población en Colombia (proporción de personas mayores, jóvenes, proporción de mujeres y de hombres, entre otros).

Para (Urdinola 2015) el proceso de envejecimiento poblacional, se refleja en las cifras de porcentaje de adultos mayores. En los países desarrollados, que ya completaron la transición demográfica, la proporción de personas mayores de 65 años es de 16 % (Naciones-Unidas 2006). Mientras que, en Colombia, donde la transición está en proceso, esta proporción es de 9.1 % (DANE 2018) y se proyecta que para 2050 alcance 17.5 % (Celade 2009).

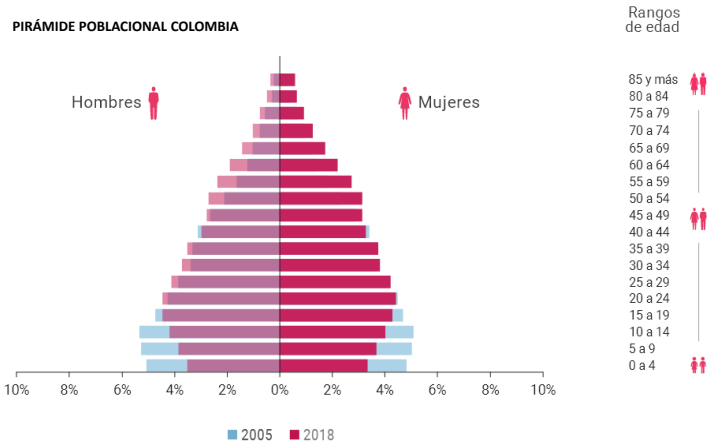

Figura 1: Pirámide Poblacional DANE-CENSO NACIONAL 2018

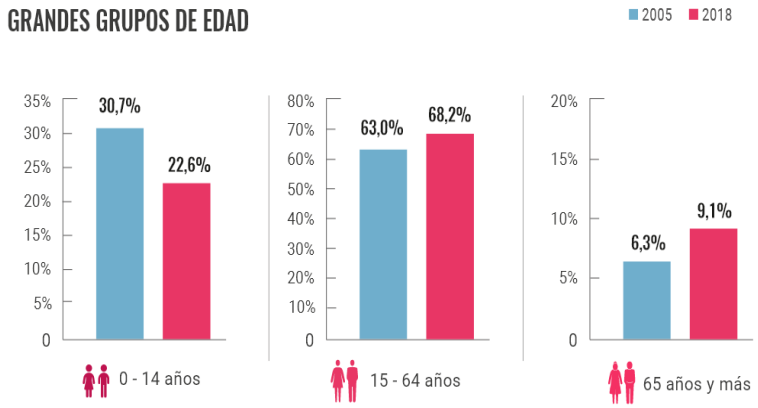

Figura 2: Grandes grupos etarios DANE-CENSO NACIONAL 2018

2.3. Efectos edad, sexo y tiempo

Las tasas de mortalidad varían en general para grupos diferenciados por sexo y cohorte. Estos efectos, así como su evolución a través del tiempo, deben ser identificados y caracterizados a partir de un análisis descriptivo y exploratorio.

Las siguientes son características básicas de estos efectos:

- **Efecto edad** La tasa de mortalidad es alta para la edad 0, desciende en el rango de edad $[0-4]$ y es creciente para los rangos de edad $[5-85+)$.
- **Efecto sexo** Según el DANE, la esperanza de vida al nacer se incrementa a través del tiempo, siendo mayor para las mujeres que para los hombres, esto debido entre otros factores, al conflicto armado interno. Esto implica que las tasas de mortalidad observada en los hombres son más altas con respecto a las mujeres.
- **Tiempo** Las tasas de mortalidad tienden a disminuir a través del tiempo.

En la práctica estos efectos no son del todo claros. Por lo tanto es necesario definir un estimador para caracterizar la variabilidad global de la mortalidad a partir de la variabilidades relativas dada por los efectos edad, sexo y tiempo.

Sea m_{ast} un estimador de la tasa de log mortalidad para el grupo de edad a, sexo s y año t, donde:

$$a = 1, \ldots, A$$

Cuando el subíndice a toma el valor de 1 corresponde al primer grupo etario 0, de la misma manera cuando se toma el valor 2 representa al segundo grupo etario $[1-4]$ y así sucesivamente hasta que a toma el valor de 85+, por tanto, el valor de A representa el ultimo grupo de edad. Recordando que en la sección 2.2 se menciona que la información viene segmentada en los grupos de edad $0, [1-4], [5-9], [10-14], \ldots, 85+$.

$$s = 1, 2$$

Con 1 = Femenino y 2 = Masculino.

$$t = 1, \ldots, T$$

En este caso, el subíndice t representa los años de estudio correspondientes a la información utilizada en el ajuste del modelo, que inicia en 1970 y va hasta 2018, entonces cuando t toma el valor de 1 corresponde al año 1970, cuando es 2 se asigna el valor de 1971 y así sucesivamente, hasta que t toma el valor de 2018, es así, que el valor de T corresponderá al año 2018.

$$\lambda_0 = \frac{1}{2AT} \sum_{a=1}^{A} \sum_{s=1}^{2} \sum_{t=1}^{T} m_{ast}$$

El efecto edad es

$$\lambda_a^{edad} = \frac{1}{2T} \sum_{s=1}^{2} \sum_{t=1}^{T} m_{ast} - \lambda_0$$

El efecto sexo es

$$\lambda_s^{sexo} = \frac{1}{AT} \sum_{a=1}^{A} \sum_{t=1}^{T} m_{ast} - \lambda_0$$

El efecto tiempo es

$$\lambda_t^{tiempo} = \frac{1}{2A}\sum_{a=1}^{A}\sum_{s=1}^{2} m_{ast} - \lambda_0$$

Estos estimadores permitirán describir estadísticamente las interacciones entre los efectos anteriormente definidos y confirmar las características básicas expuestas en los resultados del DANE.

2.4. Interacciones

En datos demográficos es común observar interacciones entre efectos. Podemos cuantificar estas interacciones extendiendo la descomposición utilizada para caracterizar los efectos edad, sexo y tiempo. Para ello definimos tres nuevos estimadores que permitirán capturar estas interacciones.

La interacción edad-sexo es definida como

$$\lambda_{as}^{edad:sexo} = \frac{1}{T}\sum_{t=1}^{T} m_{ast} - \lambda_0 - \lambda_a^{edad} - \lambda_s^{sexo}$$

La interacción edad-tiempo es

$$\lambda_{at}^{edad:tiempo} = \frac{1}{2}\sum_{s=1}^{2} m_{ast} - \lambda_0 - \lambda_a^{edad} - \lambda_t^{tiempo}$$

La interacción sexo-tiempo es

$$\lambda_{st}^{sexo:tiempo} = \frac{1}{A}\sum_{a=1}^{A} m_{ast} - \lambda_0 - \lambda_s^{sexo} - \lambda_t^{tiempo}$$

2.5. Modelos

El objetivo principal de este estudio es pronosticar la esperanza de vida en Colombia para los años 2019-2028. Para esto se estima y pronostica $\boldsymbol{\gamma}_{ast}$ que describe las tasas de mortalidad global de la población Colombiana por edad, sexo y tiempo. Los datos de entrada son las muertes observadas y_{ast} y las correspondientes exposiciones $\boldsymbol{w}_{ast}$. Como la distribución Poisson es frecuentemente usada para eventos de conteo que ocurren en el tiempo, se supone entonces, que el número de muertes sigue esta distribución.

$$y_{ast}|\boldsymbol{\gamma}_{ast}, \boldsymbol{w}_{ast} \overset{ind}{\sim} \text{Poisson}(\boldsymbol{\gamma}_{ast}\boldsymbol{w}_{ast})$$

2.5.1. Modelo para las tasas de mortalidad

Para la transformación logarítmica de las tasas de mortalidad se asume una distribución normal definida como

$$\begin{aligned}\log \boldsymbol{\gamma}_{ast}|\boldsymbol{\beta}^0, \boldsymbol{\beta}^{edad}, \boldsymbol{\beta}^{sexo}, \boldsymbol{\beta}^{tiempo}, \boldsymbol{\beta}^{edad:sexo}, \boldsymbol{\beta}^{sexo:tiempo}, \boldsymbol{\beta}^{edad:tiempo}, \sigma \\ \overset{ind}{\sim} \text{N}(\boldsymbol{\beta}^0 + \boldsymbol{\beta}_a^{edad} + \boldsymbol{\beta}_s^{sexo} + \boldsymbol{\beta}_t^{tiempo} + \boldsymbol{\beta}_{as}^{edad:sexo} + \boldsymbol{\beta}_{st}^{sexo:tiempo} + \boldsymbol{\beta}_{at}^{edad:tiempo}, \sigma) \end{aligned} \quad (1)$$

La transformación logarítmica de las tasas de mortalidad presenta la ventaja de considerar datos negativos permitiendo un espectro más amplio de datos.

La ecuación (1) contiene los tres principales efectos (edad, sexo, tiempo) así como los tres interacciones de segundo orden entre los efectos. Cualquier variación que quede después de tener en cuenta los efectos principales y las interacciones de segundo orden se modela como ruido blanco (normalmente distribuido).

Bajo el contexto bayesiano, la ecuación (1) representa la distribución previa de la log mortalidad ($\log \gamma_{ast}$).

2.5.2. Modelo para el efecto edad

Para la distribución previa del efecto edad, se considera el siguiente modelo de tendencia local

$$\beta_a^{edad}|\boldsymbol{\alpha}, \psi, \tau \overset{ind}{\sim} \begin{cases} \mathrm{N}(\alpha_a + \psi, \tau^2) & \text{si la edad } a \text{ se refiere a la etapa de bebe, } a \in [0, 1 \text{ año}) \\ \mathrm{N}(\alpha_a, \tau^2) & \text{en otro caso} \end{cases} \tag{2}$$

Es común, que en ejercicios que pretenden ajustar modelos a un conjunto de datos, surja una pregunta casi de manera natural, ¿por qué la elección del modelo? Ahora bien, ya que además se está desarrollando una aproximación bayesiana, la pregunta con toda seguridad se transforma en, ¿por qué la elección de la a priori? Entonces como una manera de argumentar esta elección, se desarrolla la presente sección, que pretende justificar la utilización de un modelo de tendencia local representado en la ecuación (2), como previa para los efectos de la edad.

Un resultado con casi dos siglos de evidencia, confirmado por años de verificación empírica, conocido mundialmente como el modelo de Gompertz, establece que: "Las tasas de mortalidad para los grupos de edad mayor, aumentan linealmente a medida que se aumenta la edad". Del anterior resultado, se puede obtener la siguiente proposición equivalente: "Las tasas de mortalidad presentan una tendencia creciente, pero localizada en los grupos de mayor edad". Adicionalmente, para los grupos de edad más jóvenes, también se localiza una tendencia decreciente (en las tasas de mortalidad), derivando en una generalización del conocido modelo, dando cuenta que los efectos de la edad a menudo tienen tendencias persistentes hacia arriba o hacia abajo, haciéndolo un modelo de tendencia, y como depende donde se localice, en la literatura especializada se le denomina modelo de tendencia local.

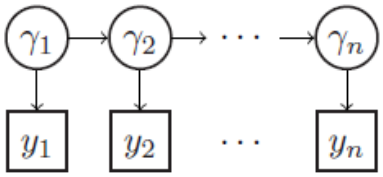

Figura 3: En un modelo de tendencia local. El γ_i representa el valor esperado de y_i. El modelo tiene dos tipos de errores: los que afectan a γ_i, y por lo tanto cambian permanentemente el valor esperado, y los que solo afectan a y_i, por lo tanto, son transitorios

Un modelo de tendencia local estándar, tiene la forma:

$$\begin{aligned} y_i &\sim \mathrm{N}(\gamma_i, \tau^2), \\ \gamma_i &\sim \mathrm{N}(\gamma_{i-1}, \omega^2). \end{aligned}$$

Sin embargo, como se puede detallar en la ecuación (2), se presenta una modificación al modelo de tendencia local estándar. Debido a que se agrega la covariable ψ al primer nivel del modelo, que se aplica solo a los bebés (niños menores a un (1) año), por tanto para este grupo de edad tomará el valor de uno y cero para el resto. La variable ψ en la ecuación (2) mide la mortalidad adicional experimentada por los bebés, en comparación con lo que se esperará de la mortalidad en otros niños.

Por otro lado, tomando como referencia la Tasa de Mortalidad Infantil (TMI) definido como la razón de defunciones a la edad de 0 a 1 año, frente a los nacimientos del mismo periodo o de manera equivalente, la TMI es el número de defunciones de niños menores de 1 año por cada 1.000 nacidos vivos para un año dado, en un determinado país, territorio o área geográfica. Esta tasa además de ser un indicador efectivo en describir las condiciones de mortalidad, es muy eficiente en capturar diferentes problemas de bienestar social y de desarrollo socioeconómico de cualquier población, que se asocia a las mejoras en capital físico (por ejemplo, infraestructura y hospitales) y humano (como la educación de los padres) que debe hacer una sociedad por mejorar sus condiciones de vida.

En el estudio de Urdinola desarrollado en 2011, se afirma que Colombia aún se encuentra lejos de los niveles alcanzados por países desarrollados (en cuanto a TMI). En particular, para el año de análisis, 1993, en el país la TMI oficial fue de 20 por cada mil niños nacidos vivos, mientras que la de países desarrollados se sitúan alrededor de 5 por cada mil niños nacidos vivos. Para 2019 y según datos del DANE, en Colombia la TMI se ubicó en la preocupante cifra de 11 por cada mil niños nacidos vivos. Por tanto, bajo la luz de esta argumentación, tiene todo el sentido la modificación al modelo de tendencia local estándar, que se matematiza en la inclusión de ψ en la ecuación (2), con el objeto de capturar la mortalidad adicional experimentada por los bebés, confirmado por indicadores como la TMI presentes en Colombia.

En cuanto a la distribución de los términos que representan el nivel (α), la tendencia (δ) se tienen las siguientes:

$$\alpha_a|\alpha_{a-1},\ldots,\alpha_1,\delta_{a-1},\ldots,\delta_1,\omega_\alpha \overset{ind}{\sim} \mathrm{N}(\alpha_{a-1}+\delta_{a-1},\omega_\alpha^2), \tag{3}$$

$$\delta_a|\delta_{a-1},\ldots,\delta_1,\omega_\delta \overset{ind}{\sim} \mathrm{N}(\delta_{a-1},\omega_\delta^2). \tag{4}$$

La distribución seleccionada para las desviaciones estándar τ, ω_α y ω_δ es una Semi-t, la cual es derivada de la distribución t de Student en donde se toman valores absolutos de las variables. Específicamente se toma una "trunc-half-t (7, 1, 5.408)"lo cual se traduce, como una distribución semi-t truncada con 7 grados de libertad y escala 1.

Para la selección los parámetros de la distribución Semi-t truncada, (Gelman 2006) propone que el parámetro de escala de la previa debería seguir una distribución semi-t (4, 0, 1). Por otro lado, consultando directamente al profesor Bryant y después de detallar los datos del caso colombiano, considera que: "una semi-t con 7 grados de libertad es un poco más fuerte, que la sugerencia que hace (Gelman 2006) de 4 grados de libertad. En la práctica, la elección de los grados de libertad tiene un impacto muy pequeño en el resultado final. Sin embargo, el uso de valores menores a 7 puede ralentizar la convergencia considerablemente, porque el muestreador puede atascarse en valores donde la varianza es grande".

En cuanto a la elección de 1 para la escala busca descartar valores completamente inverosímiles. Finalmente el valor de 5.408 es el cuantil 0.999 de una distribución semi-t con escala 1 y 7 grados de libertad. En otras palabras, se trunca la distribución en el cuantil 0.999, teniendo entonces muy poco efecto en el resultado final.

2.5.3. Modelo para el efecto tiempo

Para los efectos del tiempo, usamos el modelo de tendencia local estándar (explicado en la sección anterior). Asumiendo la siguiente estructura:

$$\beta_t^{tiempo}|\boldsymbol{\alpha},\tau \overset{ind}{\sim} \mathrm{N}(\alpha_t,\tau^2), \tag{5}$$

$$\alpha_t|\alpha_{t-1},\ldots,\alpha_1,\delta_{t-1},\ldots,\delta_1,\omega_\alpha \overset{ind}{\sim} \mathrm{N}(\alpha_{t-1}+\delta_{t-1},\omega_\alpha^2), \tag{6}$$

$$\delta_t|\delta_{t-1},\ldots,\delta_1,\omega_\delta \overset{ind}{\sim} \mathrm{N}(\delta_{t-1},\omega_\delta^2). \tag{7}$$

Nuevamente en esta sección como en la anterior, se ha omitido el uso de superíndices o subíndices en las ecuaciones (5, 6 y 7), para resaltar el hecho de encontrarse ante la previa del tiempo. Adicionalmente para τ, ω_α y ω_δ, se selecciona como previa una Semi-t, "trunc-half-t (7, 1, 5.408)" que se traduce como una distribución semi-t truncada con 7 grados de libertad y escala 1 y en donde el término "truncada" indica que no se permite que el valor sea superior a 5.408.

La notación de las distribuciones semi-t, son las siguientes:

$$(\tau, \omega_\alpha, \omega_\delta) \sim t^+(7, 1, 5.408)$$

Para el caso colombiano, se encontró que la versión estándar del modelo de tendencia local funcionó bien, ya que fue necesario tener una fuente adicional de error, descrita en la Ecuación 6, debido a que, en Colombia, la mortalidad cambia con menos fluidez. Se sospecha que esto se debe a que los niveles de mortalidad a largo plazo tienden a cambiar lenta y continuamente, en respuesta al cambio lento y continuo en los determinantes de la tendencias a largo plazo, como la tecnología, las instituciones y las condiciones ambientales.

2.5.4. Modelo para la interacción edad-tiempo

La previa para la interacción edad-tiempo, fue propuesta con el ánimo de capturar los patrones que son evidentes en la interacción edad-tiempo obtenida como resultado de la descomposición de la tasa de log mortalidad (Figura 11 – Sección de Resultados). En donde, se evidencia como cada grupo de edad tiene su propia serie de tiempo y además cada una de estas series temporales sigue un modelo de tendencia local simplificado,

$$\beta_{at}^{edad:tiempo} | \boldsymbol{\alpha}, \tau \overset{ind}{\sim} \mathrm{N}(\alpha_{at}, \tau^2), \tag{8}$$

$$\alpha_{at}^{edad:tiempo} | \alpha_{a,t-1}, \ldots, \alpha_{a,1}, \delta_{a,t-1}, \ldots, \delta_{a,1}, \omega_\alpha \overset{ind}{\sim} \mathrm{N}(\alpha_{a,t-1} + \delta_{a,t-1}, \omega_\alpha^2), \tag{9}$$

$$\delta_{at}^{edad:tiempo} | \delta_{a,t-1}, \ldots, \delta_{a,1}, \phi, \omega_\delta \overset{ind}{\sim} \mathrm{N}(\phi\delta_{a,t-1}, \omega_\delta^2), \tag{10}$$

$$\frac{\phi - 0.8}{1 - 0.8} \overset{ind}{\sim} Beta(2, 2), \tag{11}$$

$$(\tau, \omega_\alpha, \omega_\delta) \overset{ind}{\sim} t^+(7, 0.5^2, 2.704), \tag{12}$$

El modelo de tendencia local de las ecuaciones (8),(9) y (10) difiere de los modelos de tendencia local anteriores, debido a que se incluye un término de amortiguación ϕ. El término de amortiguación significa que, en lugar de una caminata aleatoria ordinaria, los términos de tendencia $\delta_{a,t-1}$ siguen a un paseo aleatorio amortiguado. Una caminata aleatoria amortiguada difiere de una caminata aleatoria ordinaria en que cada paso tiende a ser más pequeño que el anterior. La disminución en el tamaño del paso se rige por ϕ, que toma un valor entre 0 y 1. Cuando ϕ está cerca de 0, el tamaño del paso disminuye rápidamente, y cuando ϕ está cerca de 1, disminuye lentamente. Cuando ϕ es igual a 1 exactamente, no hay amortiguación, y el modelo vuelve a una caminata aleatoria ordinaria (Bryant 2018).

Los modelos de tendencia local amortiguados se basan en el principio de que no hay alza o tendencia a la baja continúa indefinidamente. Para la mayoría de las series de tiempo, este principio es confirmado por la evidencia. Los estudios empíricos sobre el rendimiento de los modelos de series de tiempo generalmente encuentran que los modelos en los que las tendencias están amortiguadas ofrecen pronósticos más precisos que los modelos donde las tendencias no están amortiguadas (Bryant 2018).

En nuestro caso la interacción edad-tiempo para las tasas de mortalidad, la amortiguación parece particularmente apropiada. Las tasas de mortalidad humana tienen un perfil de edad característico, que se repite, con variaciones, en muchas poblaciones. La previa para el parámetro de amortiguamiento restringe a ϕ a pertenecer al intervalo $[0.8, 1]$.

Aunque el parámetro ϕ suele estar cerca de 1, se toma el valor de 0.8, ya que es uno de los valores predeterminados para el argumento 'lower' de la función 'ets' en el paquete R 'forecast' de Hyndman y sus colegas, se asume que la elección de valores predeterminados funciona bien en la práctica. Por otro lado, la elección de la Beta(2,2) se basa en el 'boundary avoiding prior' descrito en p317 de Gelman et al, del libro Bayesian Data Analysis 3rd Edition.

El 0.5^2 (en lugar de 1^2) se debe a que se trata de una interacción en lugar de un efecto principal. Se supone que las interacciones son más pequeñas que los efectos principales, siguiendo el enfoque de [Priors-Gelman] y en cuanto al truncamiento es solo para evitar problemas de cálculo (2.704) es el cuantil 0.999.

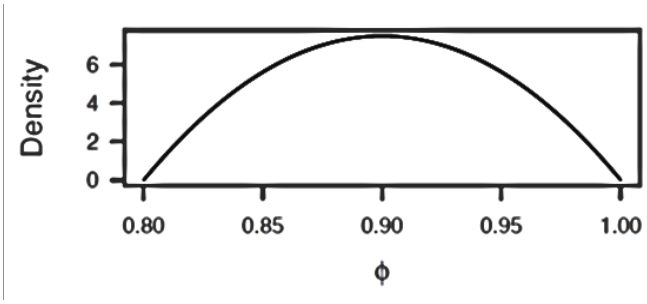

Figura 4: Previa para el parámetro de amortiguación ϕ. La previa es una distribución beta con parámetros de forma 2 y 2, con una transformación para que se encuentren en el intervalo entre 0,8 y 1.

2.5.5. Modelo para la interacción sexo-tiempo

Para los efectos de la interacción sexo-tiempo, usamos el modelo de tendencia local estándar (explicado en secciones anteriores). Asumiendo la siguiente estructura:

$$\beta_{st}^{sexo:tiempo}|\boldsymbol{\alpha}, \tau \overset{ind}{\sim} \mathrm{N}(\alpha_{st}, \tau^2), \tag{13}$$

$$\alpha_{st}^{sexo:tiempo}|\alpha_{s,t-1}, \ldots, \alpha_{s,1}, \omega_\alpha \overset{ind}{\sim} \mathrm{N}(\alpha_{s,t-1}, \omega_\alpha^2), \tag{14}$$

$$(\tau, \omega_\alpha) \overset{ind}{\sim} t^+(7, 0.5^2, 2.704) \tag{15}$$

Resultó que la interacción sexo-tiempo, fue necesaria para obtener buenos pronósticos, en Colombia.

2.5.6. Modelos para los otros términos

$$\beta^0 \overset{ind}{\sim} \mathrm{N}(0, 1), \tag{16}$$

$$\beta_s^{sexo} \overset{ind}{\sim} \mathrm{N}(0, 1), \tag{17}$$

Utilizamos un modelo de tendencia local estándar para la interacción edad-sexo. Todos los términos de desviación estándar reciben nuestra habitual a priori débilmente informativa semi-t truncada.

2.6. Inferencia Bayesiana

(Bryant 2018) afirma que, para inferir cantidades desconocidas de los datos observados usando estadística bayesiana, también usamos distribuciones de probabilidad como en la "estadística clásica". En términos

matemáticos, nuestro modelo probabilístico para los datos y las cantidades desconocidas equivale a una distribución de probabilidad gigante,

$$p(desconocidas, datos)$$

que se lee como "la distribución de probabilidad conjunta de las desconocidas y los datos". Es tradicional, en los análisis bayesianos, descomponer la distribución de probabilidad conjunta en dos términos,

$$p(desconocidas, datos) = p(desconocidas)p(datos|desconocidas)$$

El primer término, $p(desconocidas)$, es la distribución de probabilidad de las cantidades desconocidas, y se denomina a priori. El segundo término, $p(datos|desconocidas)$, es la distribución de probabilidad condicional de los datos, dado un valor para las desconocidas. Este término se conoce como la verosimilitud. La verosimilitud resume cualquier información sobre las cantidades desconocidas, contenida en los datos disponibles. El paso de inferencia en un análisis bayesiano consiste en derivar la distribución de probabilidad condicional de las desconocidas, dados los datos,

$$p(desconocidas|datos)$$

Esta distribución se denomina "distribución posteriori". Es el resultado principal de un análisis bayesiano. "Priori" y "posteriori" son las palabras en latín para "previa" y "posterior". Una forma tradicional de definir las distribuciones previa y posterior, es que la distribución previa describe las creencias del analista antes de ver los datos, y la distribución posterior describe las creencias del analista después de ver los datos. Este tipo de definiciones dan una impresión engañosa de la práctica estadística bayesiana moderna. De hecho, la mayoría de los modeladores formulan sus distribuciones previas, después de ver los datos. Además, la mayoría de los modeladores no intentan utilizar distribuciones las previa y posterior para describir sus propias creencias, sino más bien para describir alguna versión de "lo que es razonable creer, dada la información disponible".

La tarea de derivar la distribución posterior puede ser técnicamente exigente, revisemos brevemente el computo, resumen y transformaciones utilizadas para llegar a las distribuciones posterior. Por medio de la combinación de fórmulas individuales que componen nuestro modelo probabilístico (previas), podemos derivar una expresión matemática para la distribución posterior, sin embargo, en la mayoría de los casos, este resultado es difícil de usar, ya que, a partir de este no se puede derivar fácilmente probabilidades o medidas de resumen. Cuando se presenta esta situación, los estadísticos bayesianos utilizan la simulación por computadora para generar una gran muestra de realizaciones de la distribución posterior (Bryant 2018).

La forma estándar de obtener una muestra de una distribución posterior es utilizar un conjunto de técnicas conocidas como "Markov Chain Monte Carlo" (MCMC) (Cadenas de Markov Monte Carlo) (Bryant 2018). La idea básica de la técnica MCMC es comenzar con un valor que se encuentre en algún lugar dentro del rango posible, y luego generar una serie o "cadena" en la que cada nuevo valor se genera aleatoriamente dado el valor anterior. Los nuevos valores son generados de tal manera que, en cada iteración, la cadena tiende a moverse hacia valores que tienen altas probabilidades posterior. En la mayoría de las aplicaciones, los valores iniciales se eligen mediante algún tipo de aproximación, y no son una extracción genuina de la distribución posterior. Cuando las reglas para generar nuevos valores están configuradas correctamente y el hecho de que cada iteración es aleatoria, significa que una cadena eventualmente olvida su punto de partida. Si descartamos las primeras extracciones en la cadena, lo que se conoce como quemado (burn-in), entonces el resto de las realizaciones serán representativas de la distribución posterior. En problemas reales, necesitamos métodos para evaluar el rendimiento del MCMC. La técnica más común es correr múltiples cadenas, a partir de diferentes valores iniciales, y buscar el punto en que todos parecen haber sido extraídos de la misma distribución, asumiendo de esta manera que tenemos la distribución correcta. Cadenas que parecen ser extraídas de la misma distribución

se denominan convergentes, para juzgar dicha convergencia los estadísticos bayesianos han desarrollado medidas formales de las cuales se pude concluir si la cadena se pude llamar convergente.

Todos los modelos considerados en este trabajo de grado, vienen implementados en los paquetes R (demest - dembase) desarrollados por la profesora Zhang y el profesor Bryant. Según los autores, no se necesita tener un conocimiento profundo en MCMC para utilizar los métodos implementados en los paquetes, para más detalles los autores han construido un sitio web [www.bayesiandemography.com] también se puede consultar el website del libro guía de este trabajo [www.bdef-book.com].

Aunque por lo general, las expresiones que representan las distribuciones posteriores son algo complicadas. Normalmente, no se necesita toda esta complejidad, basta con utilizar algunas medidas de resumen (la moda, la mediana y la media), para responder a casi cualquier clase de pregunta que surja en un estudio de tipo bayesiano. El enfoque estándar es dar una estimación puntual, como una moda, mediana o media, además de uno o más "intervalos de credibilidad"; en donde un $X\,\%$ intervalo de credibilidad para una cantidad desconocida, es un par de números que contiene $X\,\%$ de la distribución posterior para esa cantidad. El $95\,\%$ es el valor más común para un intervalo de credibilidad, en este trabajo, adicionalmente se utilizaran intervalos de credibilidad del $50\,\%$.

Después de ejecutar la simulación MCMC hasta lograr la convergencia y de acumular una muestra de la distribución posteriori, se pasa a calcular medidas de resumen. Por otro lado, usando el hecho de que, para una muestra suficientemente grande, se tiene que:

$$\textit{Una medida resumen para la distribución posterior.} \approx \textit{La misma medida resumen calculada en la muestra posterior.}$$

La capacidad de hacer inferencias fácilmente sobre cantidades derivadas, es extremadamente útil en la práctica. De hecho, para muchas aplicaciones, es una de las ventajas cruciales de los métodos bayesianos. La distribución predictiva posterior, es el valor esperado del modelo especificado, ponderando los posibles valores del parámetro por su densidad posterior. Por ejemplo, una distribución predictiva posterior podría describir los recuentos de muertes que observaríamos si de alguna manera pudiéramos reproducir la historia, manteniendo la misma posibilidad de tener una muerte. Técnicamente, hay dos formas equivalentes de realizar pronósticos con Modelos bayesianos. Una forma es incluir las cantidades pronosticadas como parte de las cantidades desconocidas en el modelo probabilístico conjunto, y generar selecciones aleatorias para estas, como parte de extracciones de la distribución posterior. La otra forma es no incluir las cantidades pronosticadas en la distribución posterior, generar extracciones de la distribución posterior primero, y luego se generan extracciones para las cantidades pronosticadas a partir de su distribución predictiva posterior como ya se mencionó. Ambas formas de pronóstico están bien para el pronóstico a corto plazo. Para el pronóstico a largo plazo, la primera forma de pronóstico podría encontrar problemas computacionales, en el sentido de que las cadenas que parten de diferentes valores iniciales pueden ser de lenta convergencia. Todos los pronósticos, incluyendo sus medidas de incertidumbre, son necesariamente aproximadas. En términos computacionales, los modelos se desarrollaron a través de los paquetes estadísticos R 4.1.2 (demest - dembase) para Windows, desarrollados por los profesores Zhang y Bryant, en donde por medio de actualizaciones derivadas del algoritmo Metropolis-Hastings el cual es un método de MCMC en particular, se obtienen los resultados que se presentan en este trabajo. Como detalle técnico, la ejecución se realizó en una maquina con procesador Intel(R) Core(TM) i7-1165G7 (2.80GHz/11th Gen) con 16 de RAM, en donde se requirieron de 15 a 20 horas.

2.6.1. El algoritmo Metropolis-Hastings

El algoritmo Metropolis-Hastings es una generalización de otro algoritmo, en donde, se toma como referencia una razón de razones r:

$$r = \frac{p(\theta^*|y)/J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/J_t(\theta^{t-1}|\theta^*)} \tag{18}$$

El algoritmo tiene una regla de salto (jumping rule), que consiste en simplemente tomar una muestra de la propuesta θ^*, de la distribución objetivo; es decir, $J(\theta^*|\theta) \equiv p(\theta^*|y)$, para todo θ. Entonces, la razón r en 18 es siempre exactamente 1, y las iteraciones θ^t son una secuencia de extracciones independientes de $p(\theta|y)$. En general, sin embargo, la simulación iterativa se aplica a problemas en los cuales el muestreo directo no es posible (Gelman 2014).

En este trabajo, se usan las actualizaciones del Metropolis-Hastings para estimar los parámetros de "razón" que en la salida de R se encuentran bajo el título *model.likelihood.rate*. Adicionalmente en la salida, se cuenta con una serie de medidas que resumen el desarrollo de las actualizaciones. Por ejemplo "Jump" mide la distancia entre la razón propuesta y la razón actual. "Acceptance" es la proporción de valores propuestos que se aceptan. (Normalmente queremos una tasa de aceptación entre 0.2 y 0.6). "Autocorr" mide la eficacia del algoritmo de Metropolis-Hastings; queremos que la auto correlación sea baja. La estimación de los parámetros posteriori, se lleva a cabo utilizando la función *estimateModel* que tiene los siguientes parámetros:

nBurnin: Gobierna la cantidad de iteraciones que *estimateModel* calcula antes de que comience a registrar los resultados. El período de quemado es necesario para asegurarse de que los cálculos se hayan alejado de la aproximación inicial, que generalmente es precisa.

nSim: Controla el número de iteraciones que se ejecutará en *estimateModel* mientras se construye la muestra real a partir de la distribución posterior.

nChain: El número de cadenas independientes que *estimateModel* utilizará al calcular sus resultados. *estimateModel* intenta utilizar el procesamiento en paralelo, por lo que, si se cuenta con una computadora que tiene al menos 4 núcleos, ejecutará las 4 cadenas al mismo tiempo.

nThin: cuando se construye la muestra posterior, el *estimateModel* retendrá $1/nThin$ de las iteraciones.

El tamaño total de la muestra posterior $= nSim * nChain/nThin$.

La forma estándar de evaluar si los cálculos de MCMC están dando la respuesta correcta es ejecutar varios cálculos independientes en paralelo el número se controla especificando los valores en la función *estimateModel* mediante el parámetro **nChain**. La estadística **Rhat** mide si las cadenas están dando respuestas similares, es decir el valor de **Rhat** es una medida, que nos indica si los cálculos han convergido en su resultado final, cuando se presenta una cercanía a 1, significa que las cadenas están dando respuestas similares e indica buena convergencia, como bien lo afirma Bryant.

2.6.2. Obtención de los pronósticos de la Esperanza de vida

Los pasos para obtener el pronóstico posteriori de la esperanza de vida, fueron los siguientes:

1. Para la a priori del tiempo, se introducen los valores de ω_δ y τ en las tiempo ecuaciones (6)–(7)–(8), generando resultados para β_t^{tiempo}

2. Para la interacción a priori edad-tiempo, se incorporan valores de ω_δ , τ y ϕ en las ecuaciones (9)–(10)–(11), y se obtienen valores para $\beta_{at}^{edad:tiempo}$

3. En la interacción sexo-tiempo, se generan los valores para $\beta_{st}^{sexo:tiempo}$ utilizando el procedimiento similar de generación $\beta_{st}^{edad:tiempo}$

4. Se insertan los valores de β_t^{tiempo}, β_{at}^{tiempo} y $\beta_{st}^{sexo:tiempo}$, además de los valores β^0, β^{edad}, β^{sexo}, $\beta^{edad:sexo}$ y σ en la ecuación (1), y de esta manera se generan los valores para γ_{ast}.

5. Se pronostica la esperanza de vida, a partir de las tasas de mortalidad γ_{ast}.

3. Objetivos

3.1. Objetivo general

Proponer un modelo de proyección para la esperanza de vida en Colombia desde 2019 hasta 2028, utilizando métodos bayesianos.

3.2. Objetivos específicos

1. Definir un método de predicción basado en la metodología bayesiana e implementarlo para Colombia.

2. Llevar a cabo un pronóstico de la esperanza de vida en Colombia, para el periodo comprendido entre 2019 y 2028, usando el modelo propuesto anteriormente.

3. Evaluar e identificar cuáles han sido los factores desde el punto de vista del contexto colombiano, que pudieran haber influenciado el aumento o disminución de la esperanza de vida.

4. Metodología

La metodología en el presente trabajo está compuesta por las siguientes etapas:

1. **Esperanza de vida en Colombia.** En este trabajo se modela la mortalidad, para Colombia trabajando con las variables edad, sexo y tiempo. Un tema importante de este trabajo es el estudio de tipo correlacional, donde se analizara las interacciones entre edad, sexo y tiempo. Las variables interactúan cuando la naturaleza de la relación entre una variable y el resultado de interés depende del nivel de una o más variables.

   Debido a que el concepto de esperanza de vida trae consigo, el conocimiento de la estructura probabilística de la mortalidad de la población estudiada, fue necesario iniciar recopilando la información relacionada con el número de individuos de la población expuestas al riesgo de morir, además de los recuentos de muertes, todo esto clasificado por edad y sexo, para los años de estudio, dicha información fue suministrada por la organización **Latin American Human Mortality** quien a su vez se ha valido de los datos del DANE y en parte de la UNICEF. Además de recopilar la información, como en todo estudio estadístico fue indispensable organizarla, de tal forma que contara con las variables necesarias para poder definir y calcular, la esperanza de vida y debido a que es uno de los principales indicadores que puede sintetizar las condiciones de vida y otras dimensiones sociales de un país, puede ser utilizado para el análisis y proyección demográfica.

2. **Efectos de Edad, Sexo y Tiempo.** En esta etapa se distinguen los efectos para la edad, sexo y tiempo, aunque a veces, los efectos no son tan claros, se requiere entonces un método, que ayude a determinarlos. Uno de estos métodos consiste en "descomponer" las estimaciones directas de las tasas de mortalidad. Inicialmente mediante análisis descriptivos y exploratorios, junto con el conocimiento de características básicas que tienen los efectos, es necesario la definición de un estimador con el cual es posible entre otras cosas, caracterizar la variabilidad global de la mortalidad partiendo de las relativas para los efectos de edad, sexo y tiempo.

   Los estimadores para cada uno de los efectos, permitirán la descripción estadística de las interacciones entre estos. Las interacciones entre efectos es algo común en estudios demográficos, que se pueden cuantificar haciendo una extensión de la descomposición que se utilizó en la caracterización de cada efecto, y de nuevo se definen estimadores que permitirán capturar dichas interacciones.

3. **Modelos.** Nuestro objetivo final para Colombia, es pronosticar la esperanza de vida para los años 2019-2028 teniendo información de 1970 a 2018. Para esto, se estima y pronostica γ_{ast}, que representan las tasas de log mortalidad por edad, sexo y tiempo. Los datos de entrada son las muertes y_{ast} y exposiciones ω_{ast}. La distribución asignada para el número de muertes es una Poisson, la cual es frecuentemente ajustada en el caso de eventos de conteo. Por su lado, para la transformación logarítmica de las tasas de mortalidad se asume una normal esta representa la distribución previa de la log mortalidad, la cual entre varias ventajas considera datos negativos ampliando el espectro de los datos, la ecuación de este modelo contiene los tres efectos principales y las interacciones de segundo orden entre los principales. De manera similar se definen las previas para los efectos edad, tiempo y sexo, además de las respectivas interacciones entre estos efectos.

4. **Pronósticos.** Utilizando todos los datos observados para 1970-2018 para entrenar el modelo y pronosticamos la esperanza de vida para un período de 10 años de 2019-2028. Para realizar los pronósticos con estos modelos bayesianos, se utilizaron actualizaciones derivadas del algoritmo Metropolis-Hastings como un caso particular de los métodos MCMC, este algoritmo se explicó minuciosamente en la sección 2.6.1.

Cada una de las etapas de la metodología se explicó de forma detallada en las secciones dos (etapa uno de la metodología) , secciones 2.3 y 2.4 para describir los efectos e interacciones (etapa dos), para explicar la etapa tres se desarrolló la sección 2.5 y la inferencia bayesiana, junto con los pronósticos se plasmó en la sección 2.6 (etapa cuatro).

5. Datos y escalas

Se modelará la mortalidad para Colombia en función de 3 dimensiones: Edad, sexo y tiempo.

Se dice que existe interacción entre variables, cuando la relación ente la variable y el resultado de interés depende del nivel de una o más variables. Particularizando para 2 variables quedaría:

Se dice que 2 variables X e Y, interactúan cuando la relación entre la variable X y el resultado de interés Z, es distinto o cambia para cada uno de los niveles de la variable Y.

Por ejemplo si la relación entre edad (variable X) y mortalidad (resultado Z) difiere o es distinta para los niveles de la variable sexo (variable Y) (con niveles: mujer (Y_1) y hombre (Y_2)) entonces se dirá que existe interacción entre edad (X) y sexo (Y). Las interacciones que existan en los datos demográficos se les deben prestar atención y por tanto deben ser modelados por parte del investigador; en este trabajo, se prestara atención de forma particular, en modelar como los distintos patrones de edad y sexo cambian durante el tiempo.

En la presente sección describiremos de manera general el conjunto de datos que conforman las tasas de mortalidad y ojearemos un concepto matemático muy frecuente en estudios demográficos, la función logarítmica.

5.1. Tasas de mortalidad

La información aquí utilizada, está conformada por dos conjuntos de datos, por un lado, el conteo anual de muertes y adicionalmente las exposiciones anuales, que se refiere a la cantidad de personas expuestas al riesgo de morir, tanto las exposiciones como los conteos de muertes están desagregados por grupos de edad $0, 1-4, 5-9, 10-14, \ldots, 85+$ y por sexo, para el periodo 1970-2018. Los grupos de edad 0 y $1-4$ se obtienen al particionar la información del grupo de edad $0-4$, realizar este procedimiento es muy común cuando se trabaja con estadísticas de mortalidad. Una de las razones por la cual se trata a los infantes (edad 0) separadamente de los otros niños, es porque la tasa de mortalidad para los infantes tiende a ser más alta que la de los otros niños. Otro grupo usualmente pequeño y característico cuando

se trabajan estadísticas de mortalidad, es el de 85+ que se refiere a todas las personas con o por encima de 85 años de edad.

La estimación directa de la tasa de mortalidad se obtiene al dividir los conteos de muerte entre la exposición, usualmente cuando se trabajan estadísticas de mortalidad y se grafican en una escala ordinaria, se ocultan las diferencias entre valores pequeños, por tanto, más adelante utilizaremos una escala logarítmica. La figura 5 muestra como varían las tasas de mortalidad sobre una escala original (ordinaria), por ejemplo, en 1970 estas se encuentran entre 0.0006563056 (para 10-14 años) hasta 0.1866837275 (para edades 85+).

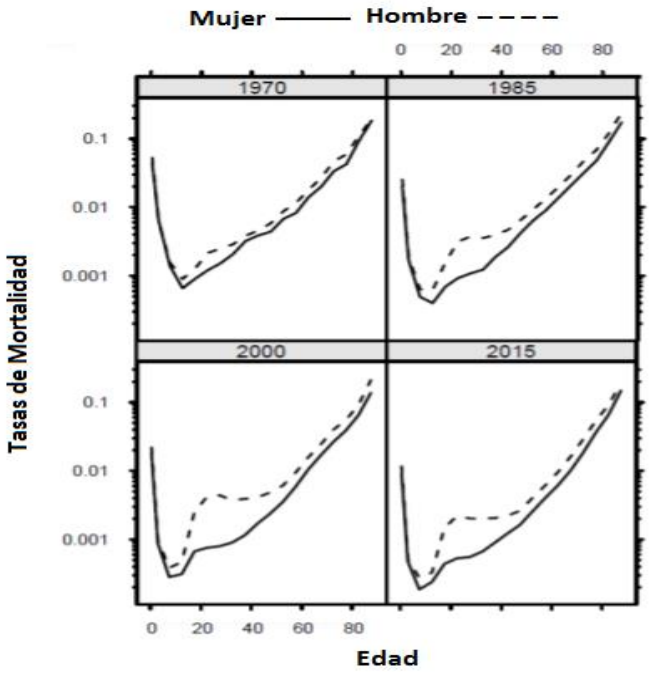

Figura 5: Estimación directa (conteos de muertes/exposición) de las tasas de mortalidad especificando por edad en Colombia, sobre una escala original (ordinaria). Las líneas son trazadas a través del centro de cada grupo de edad.

5.2. Función Log

Ya en la sección (2) habíamos explicado como el log de la media es el parámetro natural para la distribución Poisson, y el enlace log es el enlace canónico para un GLM Poisson. El modelo loglineal Poisson es entonces:

$$\log(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}$$

La función log posee facultades interesantes, a la hora de tratar dos tipos de diferencias, que se usan frecuentemente, la relativa y la absoluta; concretamente si evaluamos la función log en valores pequeños de x, tendríamos que una diferencia absoluta de x sobre la escala log equivale a una diferencia relativa de $x * 100\,\%$ sobre la escala original.

A continuación, estudiaremos los efectos de las escalas sobre las tasas de mortalidad. Para esto, revisemos algunas definiciones preliminares.

Sea m_{ast} la estimación directa de la tasa de Log mortalidad para el grupo de edad a, sexo s y año t, donde $a = 1, 2, \ldots, A$, $s = 1, 2$ y $t = 1, 2, \ldots, T$. Además sea M_{ast} la estimación directa de la tasa de mortalidad en escala original, para el grupo de edad a, sexo s y año t, donde $a = 1, 2, \ldots, A$, $s = 1, 2$ y $t = 1, 2, \ldots, T$. Si fijamos un año t_j de la misma manera se fija el sexo en mujeres ($s = 1$), entonces se entiende por **diferencia relativa** el siguiente cociente: $\frac{m_{a_i 1 t_j} - m_{a_{i-1} 1 t_j}}{m_{a_{i-1} 1 t_j}}$ de forma análoga para la tasa de mortalidad en escala original, se tendría $\frac{M_{a_i 1 t_j} - M_{a_{i-1} 1 t_j}}{M_{a_{i-1} 1 t_j}}$. Por otro lado la **diferencia absoluta** seria la siguiente diferencia para la tasa de Log mortalidad $m_{a_i 1 t_j} - m_{a_{i-1} 1 t_j}$ y de la misma forma para la tasa de mortalidad en escala original $M_{a_i 1 t_j} - M_{a_{i-1} 1 t_j}$.

La escala Log amplifica las **diferencias relativas** en lugar de **diferencias absolutas**. Ya que sobre una escala Log, por ejemplo, la diferencia entre 0.001 y 0.002 es equivalente a la diferencia entre 1 y 2.

La figura 6 ilustra los efectos de una transformación Log, usando estimaciones directas de las tasas de mortalidad para las mujeres de Colombia en 2014. **Por debajo** de 80 años, las diferencias absolutas entre grupos de edad son pequeñas y las diferencias relativas son moderadas. **Por encima** de 80 años las diferencias absolutas son grandes, **mientras** las diferencias relativas **de nuevo** son moderadas. El panel (a) **resalta** las diferencias absolutas, mientras el panel (b) resalta las diferencias relativas.

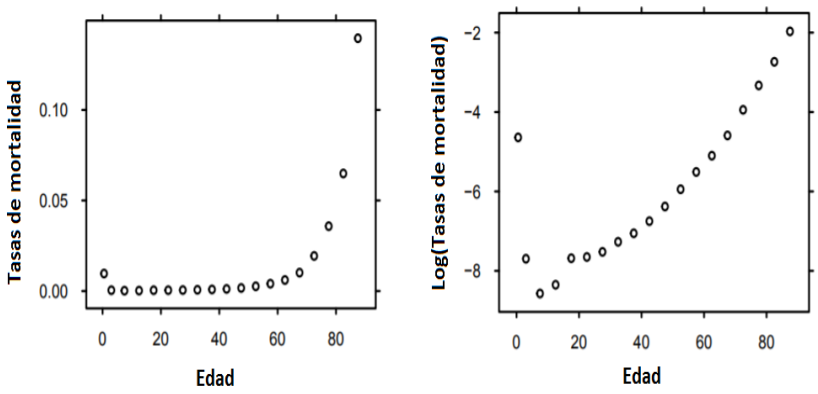

Figura 6: Estimación directa (conteos de muertes/exposición) de las tasas de mortalidad para las mujeres de Colombia en 2014, en escala original (izquierda) y escala log (derecha).

Nótese que el eje vertical del Panel (b) muestra valores después de la transformación a la escala logarítmica, mientras el eje vertical de la figura 5 muestra valores antes de la transformación. Ambos tipos de anotación son comunes con la escala log.

Por ejemplo, suponga que, cuando medimos sobre la escala log, la tasa de mortalidad para el grupo A (realmente el log(tasa de mortalidad) del grupo A) es 0.1 más alta que la tasa de mortalidad para el grupo B (realmente el log(tasa de mortalidad) del grupo B) (es decir la diferencia absoluta es de 0.1). Si volvemos a convertir a la escala original (las tasas de mortalidad), encontraremos que el valor resultante para el grupo A es casi exactamente $0.1 * 100\,\% = 10\,\%$ más alto que el valor para el grupo B (es decir la diferencia relativa).

Grupo Etareo	Log(Tasa Mortalidad)	Dif Absoluta	Tasa Mortalidad	Dif Relativa
0	-6		0,00247875	
20	-6,1	0,1	0,00224287	0,10517092
40	-6	0,1	0,00247875	0,10517092
60	-5,9	0,1	0,00273944	0,10517092
80	-5,8	0,1	0,00302755	0,10517092
100	-5,7	0,1	0,00334597	0,10517092
120	-5,6	0,1	0,00369786	0,10517092

Tabla 1: Comparación escalas logarítmica y original.

6. Resultados

6.1. Esperanza de Vida

La esperanza de vida son los años restantes que se podría esperar que un individuo viva, si los patrones de mortalidad existentes al momento de su nacimiento no cambian en el transcurso de la vida. La esperanza de vida es usualmente calculada para individuos al momento de su nacimiento (edad 0), pero se puede calcular para individuos de cualquier edad. Nosotros podemos, por ejemplo, calcular la esperanza de vida a los 65 años, indica la cantidad de años que un individuo podría esperar vivir desde el momento en el cual cumple la edad 65.

Los demógrafos distinguen entre la esperanza de vida de "periodo" y "cohorte". Con esperanza de vida de periodo, las tasas de mortalidad se refieren a un año particular, tal como 2014. Con esperanza de vida de cohorte, las tasas de mortalidad son las experimentadas por una cohorte actual. La esperanza de vida de cohorte para la cohorte nacida en 2014, por ejemplo, usaría las tasas de 2014 para la edad 0, usaría las tasas de 2015 para la edad 1, las tasas de 2016 para la edad 2 y así sucesivamente. En este trabajo, cuando nos referimos a esperanza de vida, significa esperanza de vida de periodo. La forma de calcular la esperanza de vida se revisó en la sección (2.1).

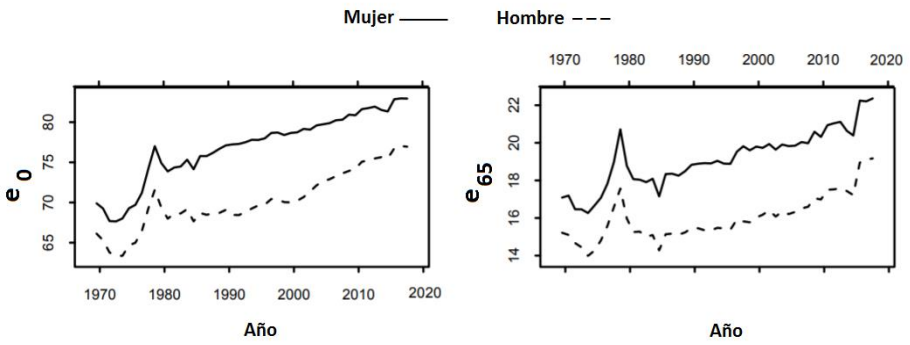

Figura 7: Esperanza de vida para las edades 0 (e_0) y 65 (e_{65}) en Colombia.

Las expectativas de vida en Colombia, calculadas utilizando las estimaciones directas de las tasas de mortalidad, se muestran en la Figura 7. El panel izquierdo muestra la esperanza de vida a los 0 años, y el panel derecho muestra la esperanza de vida a los 65 años. Como es típico en las poblaciones modernas, la esperanza de vida femenina es mayor que la esperanza de vida masculina al nacer y en las edades más avanzadas. La esperanza de vida a los 65 años, para mujeres y hombres, no mostró una tendencia clara hasta alrededor de 1985, después de lo cual aumentó constantemente.

6.2. Efectos de Edad, Sexo y Tiempo

Examinando las tasas de mortalidad en la Figura 5 a simple vista, es fácil ver, en términos generales, cómo la mortalidad varía con la edad, el sexo y el tiempo. En otras palabras, es fácil identificar un efecto de edad, un efecto de sexo y un efecto de tiempo:

- **Efecto Edad.** La mortalidad es alta a los 0 años antes de caer a niveles muy bajos. Luego sube constantemente hasta los 85 años o más.
- **Efecto sexo.** La mortalidad es más baja, en general, para las mujeres que para los hombres.
- **Efecto Tiempo.** Las tasas de mortalidad han tenido una tendencia descendiente con el tiempo.

Por lo tanto, es útil tener un método formal para identificar posibles efectos. Uno de estos métodos es "descomponer" las estimaciones directas de las tasas. Al igual que con la Figura 5, trabajamos con una versión logarítmica de las estimaciones directas. El efecto edad en la edad a describe cuánto difiere la tasa promedio de mortalidad logarítmica para la edad a, de la tasa promedio general de mortalidad logarítmica. Como los efectos de la edad son todos relativos al promedio general, ellos suman cero. La variabilidad en los efectos de la edad representa la contribución de la edad a la variabilidad general de las tasas de mortalidad logarítmica. Los efectos sexo y tiempo se definen de manera similar.

Los resultados de la descomposición se muestran en la Figura 8 La descomposición confirma, por ejemplo, que las tasas femeninas son más bajas que las masculinas, y que las tasas de mortalidad han estado disminuyendo con el tiempo. Al observar las escalas verticales de los tres gráficos (λ_a^{edad},λ_s^{sexo},λ_t^{tiempo}) también tenemos una idea de la importancia relativa de los tres efectos. Los efectos estimados de la edad son los mayores de los tres (-2.5 a 4) y por tanto, los más importantes. Los efectos del tiempo, a su vez, tienen aproximadamente tres veces el rango de los efectos sexo.

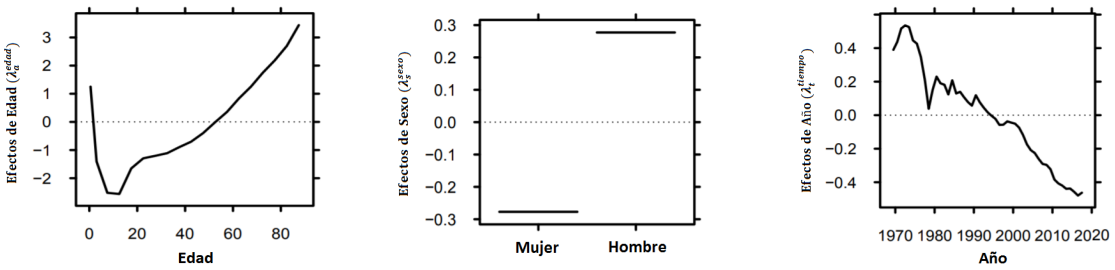

Figura 8: Efectos de edad λ_a^{edad}, sexo λ_s^{sexo} y tiempo λ_t^{tiempo} obtenidos al descomponer las estimaciones directas de las tasas de mortalidad logarítmicas.

6.3. Interacciones

Mirando de cerca la Figura 5, parece que las diferencias entre mujeres y hombres en la mortalidad no son constantes en todo el rango de edad. Claramente las mujeres tienen menor mortalidad por encima de los 15 años y por debajo de los 70 años, pero no es tan claro este comportamiento en los grupos de edad más bajos o más altos. Al describir la relación entre mortalidad, edad y sexo, podemos distinguir entre un efecto de edad, un efecto de sexo y una "interacción" de edad y sexo. La interacción edad-sexo captura la forma en que las diferencias de sexo varían en el rango de edad o, de manera equivalente, cómo las diferencias de edad varían entre los sexos.

Del mismo modo, observando la Figura 5, parece que las tasas de mortalidad no han disminuido al mismo ritmo en todos los grupos de edad. Podemos distinguir entre un efecto de edad, un efecto de tiempo y

una interacción de edad y tiempo. La interacción edad-tiempo captura el hecho de que la disminución de la mortalidad ha sido más rápida en edades más tempranas que en edades más avanzadas.

Las interacciones son comunes en los datos demográficos y, a menudo, lo suficientemente grandes como para incluirlas en nuestros modelos. De hecho, una parte sustancial del proceso de construcción de un modelo demográfico consiste en buscar interacciones y decidir cómo manejarlas.

Podemos cuantificar las interacciones extendiendo la técnica de descomposición que utilizamos con los efectos de la edad, el sexo y el tiempo. Una vez más, la estrategia básica es restar los promedios y ver qué queda. Por ejemplo, el efecto de interacción para la edad a y el sexo s es igual a la tasa promedio de mortalidad logarítmica para la edad a y el sexo s, menos la tasa promedio general de mortalidad logarítmica, menos el efecto de la edad para la edad a y menos el efecto del sexo para el sexo s.

La interacción edad-sexo es definida como

$$\lambda_{as}^{edad:sexo} = \frac{1}{T}\sum_{t=1}^{T} m_{ast} - \lambda_0 - \lambda_a^{edad} - \lambda_s^{sexo}$$

La interacción edad-tiempo es

$$\lambda_{at}^{edad:tiempo} = \frac{1}{2}\sum_{s=1}^{2} m_{ast} - \lambda_0 - \lambda_a^{edad} - \lambda_t^{tiempo}$$

La interacción sexo-tiempo es

$$\lambda_{st}^{sexo:tiempo} = \frac{1}{A}\sum_{a=1}^{A} m_{ast} - \lambda_0 - \lambda_s^{sexo} - \lambda_t^{tiempo}$$

Los resultados de aplicar estas técnicas a los datos colombianos son graficados en la figura 9 a 11. El panel (9) muestra el perfil de edad residual para las mujeres, después de considerar los efectos de la edad y el efecto sexo para las mujeres. El hecho de que la interacción sea positiva hasta los 15 años significa que las tasas de mujeres son más altas de lo esperado en estas edades, dado el perfil de edad promedio y el efecto sexo promedio para las mujeres.

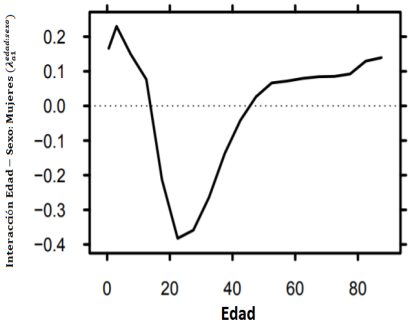

Figura 9: Interacción Edad-Sexo: Mujeres ($\lambda_{a1}^{edad:sexo}$)

La interacción sexo-tiempo en el panel (10) se interpreta de manera similar a la interacción edad-sexo. Mide el residual, después de tener en cuenta los efectos principales de sexo y tiempo. (En terminología estadística, un "efecto principal" es un efecto que involucra solo una dimensión, como un efecto de edad o un efecto sexo, en oposición a una interacción, que involucra dos o más dimensiones.) Los resultados en la

Figura 10 implican que las tasas de mortalidad de las mujeres fueron relativamente altas, en comparación con los hombres, en 1970; que la brecha se redujo hasta aproximadamente 1990; y que la brecha se ha mantenido estable hasta aproximadamente 2003 y desde entonces ha vuelto a aumentar. Estos resultados se pueden verificar mediante una inspección cuidadosa de los perfiles de edad bruta de mujeres y hombres en la Figura 5.

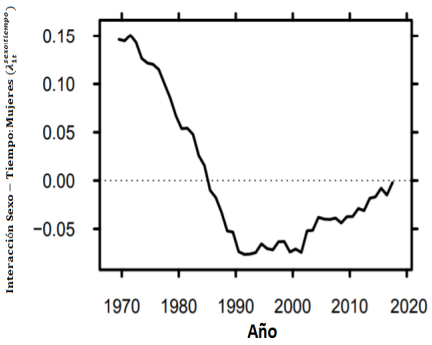

Figura 10: Interacción Sexo-Tiempo: Mujeres ($\lambda_t^{sexo:tiempo}$)

Finalmente, la interacción edad-tiempo estimada en la Figura 11 confirma que las tasas de mortalidad han disminuido más rápido para los jóvenes que para los viejos, aunque el efecto es más marcado en los extremos de la distribución por edades.

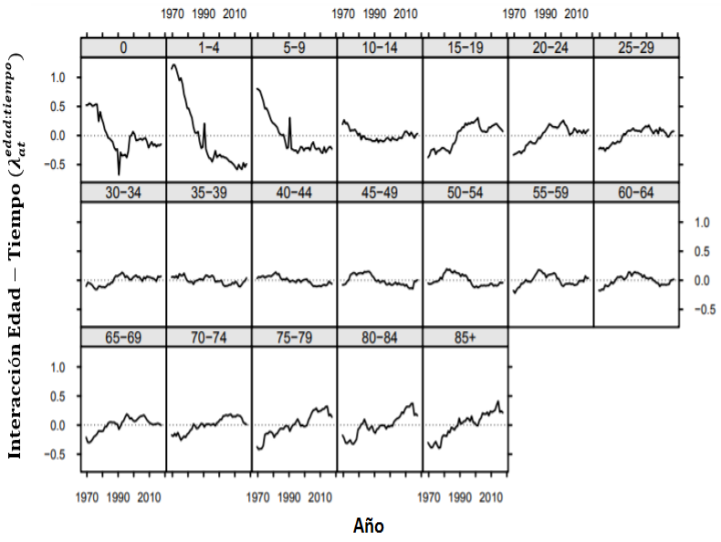

Figura 11: Interacciones edad-tiempo ($\lambda_{at}^{edad:tiempo}$) obtenidas al descomponer las estimaciones directas del log de las tasas de mortalidad.

Podríamos extender la descomposición para estimar las interacciones de tercer orden entre edad, sexo y tiempo. De hecho, una interacción edad-sexo-tiempo es visible en las tasas brutas de la figura 5. Entre 1985 y 2015, el perfil de edad para hombres, pero no para mujeres, desarrolla una joroba alrededor de los 20 años. Esta joroba alrededor de los 20 años es una característica de las tasas de mortalidad en muchos países, y se conoce como la joroba de accidente. Para describir este fenómeno, se necesitaría una interacción de tercer orden. Pero en su lugar pasaremos a la construcción de modelos.

6.4. Distribuciones Posteriores

Como ya se discutió en la sección (2.6) en este trabajo se utilizó el algoritmo Metropolis-Hastings, que en términos generales pertenece a una familia de métodos de simulación de cadenas de Markov que son útiles (entre muchas aplicaciones) para muestrear distribuciones bayesianas posteriores. En el software desarrollado por los profesores Zhang y Bryant, que ya se mencionó en secciones pasadas, cuenta con la función *estimateModel* del paquete *demest* que básicamente realiza la estimación de tasas, conteos, probabilidades o medias de una data demográfica. Después de especificar el modelo a ser ajustado que se define en el argumento *model* (función *estimateModel*) en nuestro caso particular, responde a las definiciones de las a priori explicadas en la sección (2.5). Se pasa a definir los argumentos, los valores definidos a continuación para el caso colombiano, fueron sugeridos por el profesor Bryant, basado en su experiencia y como co-creador del paquete.

nBurnin: Número de iteraciones descartadas antes de que comience el registro de resultados. Lo que llamamos en la sección (2.6.1) el periodo de calentamiento. En el caso colombiano se decidió que tuviera un valor de 50.000.

nSim: Número de iteraciones realizadas durante el registro, es decir controla el número de iteraciones que se ejecutará en *estimateModel* mientras se construye la muestra real a partir de la distribución posterior. Para este trabajo su valor es de 50.000.

nChain: Número de cadenas independientes a utilizar. Para este caso particular tomó el valor de 4, por si se cuenta con una computadora que tiene al menos 4 núcleos, ejecutará las 4 cadenas al mismo tiempo.

nThin: Intervalo de aclareo, es decir, cuando se construye la muestra posterior, la función *estimateModel* retendrá $(1/nThin)$ de las iteraciones. En nuestro caso fue de 200.

Finalmente, el tamaño total de la muestra posterior, $nSim * nChain/nThin$.

La definición de los argumentos se plasma en la función *estimateModel* y los resultados, se guardan en el objeto *model_with_interact.pred* para obtener una descripción algebraica (como lo llama Bryant); de las a posteriores que se obtuvieron, se llama a *showModel* sobre los resultados de *predictModel* lo obtenido es para el autor del texto guía lo mismo que la distribución posterior. Es decir, se invoca:

showModel("model_with_interact.pred")

```
          y[i] ~ Poisson(rate[i] * exposure[i])
   log(rate[i]) ~ N((Intercept) + age[j[i]] + sex[j[i]] + year[j[i]] + age:sex[j[i]] + age:year[j[i]] + sex:year[j[i]], sd^2)

-- values for '(Intercept)' held constant --

-- values for 'age' held constant --

-- values for 'sex' held constant --

        year[j] = level[j] + error[j]
       level[j] = level[j-1] + trend[j-1] + errorLevel[j]
       trend[j] = trend[j-1] + errorTrend[j]
       level[0] = N(0, 0^2)
       trend[0] = N(0, 1)
  errorLevel[j] ~ N(0, scaleLevel^2)
  errorTrend[j] ~ N(0, scaleTrend^2)
     scaleLevel ~ trunc-half-t(7, 1, 5.408)
     scaleTrend ~ trunc-half-t(7, 1, 5.408)
        error[j] ~ N(0, scaleError^2)
     scaleError ~ trunc-half-t(7, 1, 5.408)

-- values for 'age:sex' held constant --

  age:year[k,1] = level[k,1] + error[k,1]
      level[k,1] = level[k-1,1] + trend[k-1,1]
      trend[k,1] = damp * trend[k-1,1] + errorTrend[k,1]
      level[0,1] = 0
      trend[0,1] ~ N(0, 1)
  dampTransform = (damp-0.8)/(1-0.8)
  dampTransform ~ Beta(2,2)
errorTrend[k,1] ~ N(0, scaleTrend^2)
     scaleTrend ~ trunc-half-t(7, 1, 2.704)
     error[k,1] ~ N(0, scaleError^2)
     scaleError ~ trunc-half-t(7, 0.5^2, 2.704)

  sex:year[k,1] = level[k,1] + error[k,1]
      level[k,1] = level[k-1,1] + errorLevel[k,1]
      level[0,1] = N(0, 0^2)
errorLevel[k,1] ~ N(0, scaleLevel^2)
     scaleLevel ~ trunc-half-t(7, 0.5^2, 2.704)
     error[k,1] ~ N(0, scaleError^2)
     scaleError ~ trunc-half-t(7, 0.5^2, 2.704)

             sd ~ trunc-half-t(7, 1, 5.408)
```

Figura 12: Distribución posterior.

Como se puede detallar, de la salida anterior, como resultado permanecen invariantes: Intercepto, Edad, Sexo, interacción edad-sexo. Para el resto de efectos, se obtienen sus correspondientes distribuciones, como ejercicio, se hace explicita la distribución de la interacción edad-tiempo, las demás se deducen usando la "misma lógica".

Para Bryant, realmente hay dos formas de describir la distribución posterior, una es dar una descripción algebraica (esto es lo que hace *showModel*) y la otra es dar una muestra numérica de esa distribución; esto se hace usando la función *fetch* de donde se obtienen las estimaciones de los parámetros para los años históricos.

```
------------------------------------------------------------
model:
y ~ Poisson(mean ~ (age + sex + year)^2)
dimensions: age, sex, year
------------------------------------------------------------
y:
Object of class "Counts"
dimensions: age, sex, year
n cells: 1862,  n missing: 0, integers: TRUE, n zeros: 0, median: 3791.5
------------------------------------------------------------
MCMC statistics:
nBurnin: 50000,  nSim: 50000,  nChain: 4,  nThin: 200,  nCore: 4,  nIteration: 1000
Metropolis-Hastings updates:
                      jump acceptance autocorr
model.likelihood.rate 0.025      0.547    0.055
parameters:
                                Rhat                 Est.                    N
                                 med  max    n       min    med  max
model.likelihood.rate              1 1.01   25   0.00019 0.0044 0.25   1862
model.prior.mean                1.02 1.07   25      -8.6   -5.5 -1.4   1862
model.prior.sd                  1.01         1              0.09           1
model.hyper.age.scaleLevel      1.16         1              0.25           1
model.hyper.age.scaleTrend      1.11         1              0.23           1
model.hyper.age.coef            1.21         1               2.1           1
model.hyper.age.scaleError      1.64         1              0.19           1
model.hyper.year.scaleLevel     1.01         1             0.043           1
model.hyper.year.scaleTrend     1.08         1             0.005           1
model.hyper.year.scaleError        1         1            0.0067           1
model.hyper.age:sex.scaleLevel  1.01         1              0.15           1
model.hyper.age:sex.scaleError  1.01         1             0.025           1
model.hyper.age:year.scaleTrend 1.04         1             0.013           1
model.hyper.age:year.damp       1.05         1              0.84           1
model.hyper.age:year.scaleError 1.07         1            0.0022           1
model.hyper.sex:year.scaleLevel 1.01         1             0.018           1
model.hyper.sex:year.scaleError    1         1             0.003           1
------------------------------------------------------------
```

Figura 13: Resumen MCMC.

De la salida anterior, podemos detallar varios resultados. En la sección *MCMC Statistics* se confirman los valores asignados en los argumentos asociados a las cadenas de Markov que como ya sabemos son *nBurnin*, *nSim*, etc. También podemos apreciar los resultados de la actualización del algoritmo Metropolis-Hastings (sección Metropolis-Hastings updates) entre los cuales tenemos un ***jump*** $= 0.025$; informándonos la distancia entre la tasa propuesta y la actual. Además, tenemos una ***acceptance*** $= 0.547$; nos indica que la proporción de valores propuestos que se aceptaron fue del 54.7 % que se encuentra entre los rangos recomendados, ya que entre 0.2 y 0.6, es lo que normalmente se desea. Por ultimo en esta sección encontramos la ***autocorr*** $= 0.055$; indicándonos una auto-correlación baja e informando un algoritmo MH eficaz.

La última sección y no menos importante, de esta salida es la de "parameters" en donde se presentan las estimaciones de los parámetros para los años históricos específicamente en la subsección 'Est.' Indicando tres medidas resumen mínimo (min), mediana (med) y máximo (max) que como ya explicamos en la sección (2.6) debido a que las expresiones que representan las distribuciones posteriores son algo complicadas, basta con utilizar algunas medidas de resumen, para responder a casi cualquier clase de pregunta que surja en un estudio de tipo bayesiano, en este caso recurrimos al min, med y max. Una subsección de **parameters** crucial es la conformada por la columna **Rhat** que básicamente nos indica si las cadenas asociadas a las distribuciones posteriores de los parámetros, convergieron o no convergieron, en este caso, ya que contamos con valores de 1 (o muy cercanos a uno) significa que las cadenas están dando respuestas similares e indica buena convergencia.

6.5. Obteniendo pronósticos de Esperanza de Vida

Tomamos los datos de mortalidad colombianos asumiendo el conjunto de datos de entrenamiento como aquel que se extiende desde 1970 hasta 2018. Ajustamos el modelo al conjunto de datos de entrenamiento. Luego, pronosticamos las tasas de mortalidad γ_{ast} durante el período 2019-2028. Es decir que utilizamos todos los datos observados desde 1970-2018 para entrenar el modelo y pronosticamos la esperanza de vida para un período de 10 años de 2019-2028.

La Figura 14 presenta pronósticos de la esperanza de vida al nacer para ambos sexos. No hay señales de errores de salto para los pronósticos. En el año 2018, la esperanza de vida al nacer observada de la población es de 82 años para las mujeres, 77 años para los hombres. Los pronósticos continúan a partir de estos valores.

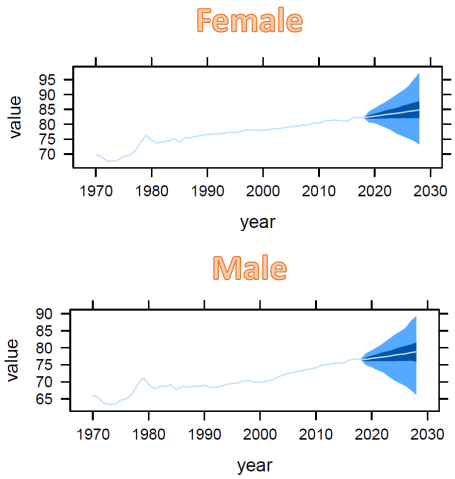

Figura 14: Pronósticos de la Esperanza de vida al nacer para 2019 – 2028 del modelo. Las líneas blancas muestran las medianas posteriores y las bandas azules los intervalos de credibilidad al 95 %.

Predecir adecuadamente el proceso de envejecimiento de la población es una preocupación no sólo de países desarrollados sino también de naciones en desarrollo, que observan con preocupación el acelerado proceso de transición demográfica de los últimos 50 años. La inquietud se origina en las repercusiones económicas y sociales que se derivan de una población que envejece. Una de estas repercusiones, que además es considerada como central por muchos analistas, es el sostenimiento de las personas mayores tema que no da espera, esta situación inspira estudios que analizan fenómenos como, la proyección del gasto fiscal en pensiones a través de la estimación de la esperanza de vida, basada en pronósticos de tasas de mortalidad (Reyes 2010).

Year	Mujeres	Hombres
2019	84	78
2020	85	79
2021	87	80
2022	88	81
2023	89	82
2024	90	84
2025	91	85
2026	93	86
2027	95	88
2028	97	89

Tabla 2: Pronosticos bayesianos de la esperanza de vida al nacer para Colombia.

Según (Urdinola 2015) la necesidad de financiar la creciente población mayor, implica grandes cargas al fisco de las naciones. Los compromisos pensionales pueden no reflejar su verdadera magnitud a largo plazo, dependiendo del sistema implementado. El valor presente de los beneficios esperados de los programas públicos de pensiones se conoce como la deuda pública implícita de pensiones (BancoMundial 1994). Analizando las prestaciones devengadas hasta 2050, un estudio del Fondo Monetario Internacional proyecta para países desarrollados como Francia y Alemania, que esta deuda excederá su PIB; mientras que, en países en desarrollo, como algunos del este de Europa y América Latina, la deuda equivaldría a dos y hasta tres veces el PIB (Knowledge-SMU 2007).

En el artículo de Urdinola desarrollado en 2015, las consecuencias de este fenómeno (envejecimiento poblacional) dependerán de las medidas que se adopten para afrontar dichos retos. (Naciones-Unidas 2007) plantea que para mantener estable la transferencia de recursos, la población en edad productiva tendrá que soportar mayores cargas, en forma de impuestos y contribuciones, a menos que se acelere el crecimiento económico sostenidamente. Las pautas de consumo, inversión y ahorro también se verán afectadas por el envejecimiento, y como consecuencia, se producirán cambios en la demanda de bienes y servicios.

Que el porcentaje de personas mayores, sea cada vez más representativo, modifica diferentes aspectos en las sociedades. Según el paper de Lee-Manson de 2010, encuentran que el envejecimiento poblacional, reflejado en bajas tasas de crecimiento de la población, causan disminución en la tasa de ahorro y, paralelamente, aumento en el consumo por habitante. Es decir, la disminución en la mortalidad puede modificar los patrones de consumo y ahorro. Por su parte desde la visión de Lee en su trabajo de 1976, sostiene que el tamaño de la población y la estructura etaria son determinantes importantes de la condición socioeconómica, ya que afectan el proceso de formación de capital humano y, por este camino, eventualmente pueden llegar a afectar el ingreso (Urdinola 2015).

En Colombia, el crecimiento de la población mayor comienza a evidenciar sus primeros síntomas; medir el impacto de las tendencias demográficas permite hacer previsiones desde ahora que se consoliden en medidas para afrontar los retos que una sociedad envejecida supone (Urdinola 2015).

Desde el punto de vista de la calidad predictiva, se ha decidido ajustar el modelo tomando como insumo los datos desde 1970 hasta 2008 y en seguida, se realizan predicciones hasta 2018 (periodo en el cual se conoce los verdaderos valores), y luego se calculan dos métricas para evaluar dichas predicciones, a saber: (i) error absoluto, (ii) Intervalo de Score (intervalo de puntuación).

El error absoluto mide el rendimiento de las estimaciones puntuales generadas por un pronóstico. Mediante el uso de medianas posteriores de la esperanza de vida como estimación puntual. Los errores absolutos son las diferencias absolutas entre las estimaciones puntuales y los valores observados.

Sea $\hat{e}_{ast}$ la mediana posteriori de la esperanza de vida en la edad a, el sexo s y el año t. Sea e_{ast} la esperanza de vida observada en la edad a, sexo s y el año t. Entonces el error absoluto sería:

$$|\hat{e}_{ast} - e_{ast}|$$

Se considera que un modelo tiene una buena calidad predictiva si este error es menor al 30 %.

Por su parte, el score del intervalo premia los intervalos estrechos, pero también penaliza los intervalos que no contienen el valor observado. Para la esperanza de vida en la edad a, sexo s y año t, sea $\hat{Q}_{ast}(p)$ denota el cuantil $100p\,\%$ de la distribución posterior, donde p es una probabilidad dada. Los límites inferior y superior del intervalo de credibilidad $(1-\alpha)\times 100\,\%$ están dados por $\hat{A}_{ast} = \hat{Q}_{ast}(\alpha/2)$ y $\hat{B}_{ast} = \hat{Q}_{ast}(1-\alpha/2)$. El score del intervalo es

$$(\hat{B}_{ast} - \hat{A}_{ast}) + \frac{2}{\alpha}(\hat{A}_{ast} - e_{ast})I(e_{ast} < \hat{A}_{ast}) + \frac{2}{\alpha}(e_{ast} - \hat{B}_{ast})I(e_{ast} > \hat{B}_{ast})$$

Aquí $I(c)$ es una función indicadora, que toma el valor 1 si se cumple la condición c dada entre paréntesis, y 0 en caso contrario.

El primer término en el score del intervalo es igual a la longitud del intervalo de confianza. El segundo término da una penalización si el valor observado es más pequeño que el límite inferior del intervalo de confianza. El tercer término da una penalización si el valor observado es mayor que el límite superior del intervalo de confianza.

A manera de síntesis, el error absoluto se enfoca en una estimación puntual particular, en cambio el score del intervalo se enfoca en un intervalo de un tamaño particular.

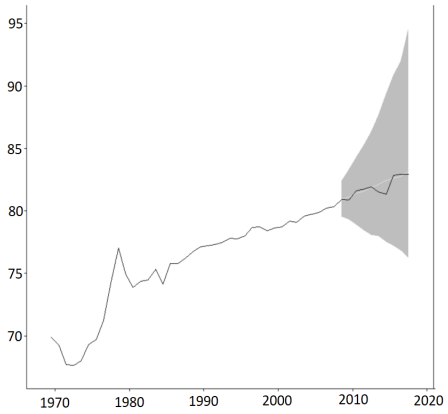

Figura 15: Calidad predictiva del modelo

El error absoluto fue de 0.29 años, en cuanto al score de intervalo se obtuvo un valor de 9.63, que según John Bryant un buen modelo debe tener valores menores a 15.

7. Conclusiones

En este proyecto se estableció como objetivo general, proponer un modelo de proyección para la esperanza de vida en Colombia desde 2019 hasta 2028, utilizando métodos bayesianos. Para ello, se examinó una manera de estimación demográfica y de pronóstico, inspirado en la metodología propuesta por los profesores John Bryant y Junni L. Zhang, desarrollada en su libro ***Bayesian demographic estimation and forecasting***.

Los resultados muestran que:

- En este trabajo, se pudo evidenciar que, para capturar todas las tendencias en grandes conjuntos de datos demográficos, es necesario ajustar modelos complicados, con muchas interacciones. Sin embargo, existen formas de mantener manejable esta complejidad, utilizando los que los autores denominan construcción de un modelo pieza por pieza, usando descomposiciones y gráficos para guiar la construcción de cada pieza.

- Construir series de datos largas para pronósticos desagregados puede ser complicado, aunque utilizar recursos alternativos informativos puede compensar los datos limitados en un principio. Los paquetes de R desarrollados por los profesores John Bryant y Junni L. Zhang facilitan la configuración, y ejecución de modelos como se hizo en este trabajo, con estos paquetes fue posible producir modelos razonables de forma rápida y sencilla. En este trabajo utilizamos los principalmente los paquetes dembase y demest, los cuales proporcionan herramientas para hacer inferencias sobre tasas, probabilidades, medias y recuentos de clasificación cruzada, utilizando métodos bayesianos. El software y los métodos se encuentran en [https://github.com/StatisticsNZ/demest] – [https://rdrr.io/github/StatisticsNZ/dembase/].

- En conclusión, el trabajo realizado en (Bryant 2018) y este trabajo tienen buenos resultados que muestran de que la metodología propuesta por los profesores Bryant y Zhang, funciona y es aplicable al caso colombiano.
- Como próximos trabajos enmarcados en esta metodología de modelamiento pieza por pieza, desarrollada por Bryant y Zhang, se sugieren dos aspectos que se consideran interesantes, el primero sería incluir efectos que tengan en cuenta fenómenos como las pandemias, que según expertos los presentaremos más seguidos de aquí en adelante. El segundo aspecto y no menos interesante, que se pude usar para desarrollar un trabajo futuro, sería evaluar la esperanza de vida por regiones; usando una metodología similar a la desarrollada en el presente trabajo.
- Para el caso colombiano se logró identificar el subregistro, como un factor que pudo haber influenciado el aumento de la esperanza de vida, para próximos trabajos sería interesante realizar la corrección de los datos.

Acknowledgments

First of all, I would like to acknowledge to God, who allowed me to successfully complete this rewarding experience, in an academic, professional and personal way. I am particularly grateful to Professor Wilmer Pineda, for his help, time and dedication in guiding me through this experience. I wish to thank my family and especially to my dear brother Iván Darío, colleague and friend. To my beloved Jasbleidy and Carlos Iván, my "principal components". My special thanks are extended to Professor John Bryant, author of the guide book on which this work is based, who, despite the distance, has always guided me and never stopped answering my questions at the risk of sounding very obvious, researchers like Professor Bryant enrich and strengthen applied statistics.

Referencias

Acosta, K. (2014), *Cambios recientes en las principales causas de mortalidad en Colombia*, Banco de la República.

Andrew Gelman, A. J. (2008), 'A weakly informative default prior distribution for logistic and other regression models', *Institute of Mathematical Statistics* .

AsíVamosEnSalud (2021), 'Tasa de mortalidad infantil – georeferenciado', https://www.asivamosensalud.org/indicadores/poblaciones-vulnerables/tasa-de-mortalidad-infantil-georeferenciado.

BancoMundial (1994), *Averting the old age crisis. Policies to protect the old and promote growth*, Oxford University Press, Inc.

Benito, R. (2008), 'Ventajas de la estadística bayesiana frente a la frecuentista: ¿por qué nos resistimos a usarla?', *Revista Ecosistemas* **27**(2), 136–139.

Bryant, Z. (2018), *Bayesian demographic estimation and forecasting*, Chapman and Hall/CRC.

Celade (2009), *Estimaciones y proyecciones vigentes*, División de población. Santiago de Chile: Comisión Económica para América Latina y el Caribe CEPAL. Centro Latinoamericano y Caribeño de Demografía.

DANE (2007), 'Proyecciones de población, 2005-2020', *Departamento Administrativo Nacional de Estadística* (1), 1–224.

DANE (2018), 'Censo nacional de población y vivienda. 2018-colombia', https://www.dane.gov.co/index.php/estadisticas-por-tema/demogr fia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018.

Debon, A. (2003), 'Graduación de tablas de mortalidad. aplicaciones actuariales', *Universitat de Valencia* (1), 1–230.

Florez, K. (2020), *Atlas de mortalidad de la región Caribe colombiana 2008-2015: mapeo y análisis desde el enfoque bayesiano*, Universidad del Norte, Barranquilla Colombia.

Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models', *Bayesian Analysis* (3), 515–533.

Horiuchi, S. (1991), 'Assessing the effects of mortality reduction on population ageing', *Population bulletin of the United Nations* **31**(31), 38–51.

Knowledge-SMU (2007), 'Managing retirement risk in an ageing world: The global picture', http://knowledge.smu. edu.sg/article.cfm?articleid=1074.

Lee-Mason (2010), 'Some macroeconomic aspects of global population aging', http://ssrn.com/abstract=1532916.

Lee, R. (1976), 'Demographic forecasting and the rasterlin hypothesis', *Population and Development Review* (2), 459–468.

McKcown, T. (1976), *Modern rise of population*, Edward Arnold (Plubisher) Ltd.

Naciones-Unidas (2006), *Demographic yearbook*, Division of the Economic and Social Affairs.

Naciones-Unidas (2007), *Estudio económico y social mundial 2007: el desarrollo en un mundo que envejece*, Departamento de Asuntos Económicos y Sociales.

Pineda, W. (2018), *Estadística Bayesiana*, Universidad Santo Tomás.

Reyes, A. (2010), 'Una aproximación al costo fiscal en pensiones como consecuencia del envejecimiento de la población en colombia y el efecto de la sobre-mortalidad masculina', *Universidad Nacional de Colombia* (1), 1–38.

Urdinola, P. (2011), 'Determinantes socioeconómicos de la mortalidad infantil en colombia', *Revista Colombiana de Estadística* (3), 39–72.

Urdinola, P. (2015), *Aplicaciones en demografía*, Universidad Nacional de Colombia.

Zarruk, Villegas, O. (2012), 'Tablas de mortalidad', *Universidad Externado de Colombia* (5).