O regressão Bell para dados de mortalidade

Silvio Cabral

Setembro/2020

Conteúdo

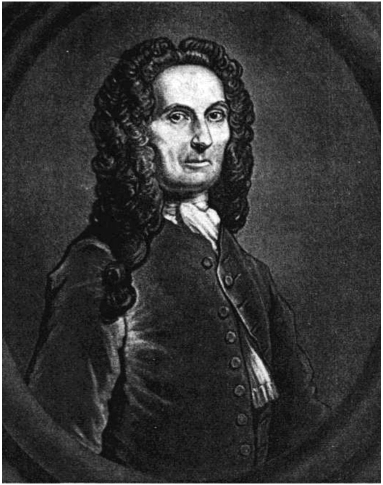
Conteúdo

Contexto histórico

Há 200 anos, a primeira versão dos seguros de vida nasceu entre as empresas e seus funcionários. O passivo dependia do número de mortes ocorridas entre as vidas seguradas a cada ano (Dickson et al. 2013). Desde então, a modelagem da mortalidade tornou-se um tópico de interesse científico comercial e geral, buscando desenvolver uma teoria quantitativa, também chamada de lei, que conduzisse o envelhecimento da população, sua mortalidade e sua expectativa de vida; que são informações necessárias para definir e precificar os seguros de vida, pensões e previdências, e também realizar projeções de populações.

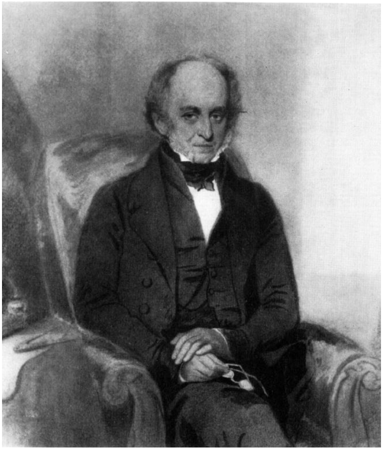
de Moivre

- A primeira lei de mortalidade foi desenvolvida em 1725 pelo matemático francês Abraham de Moivre
- Dados fornecidos por John Graunts



Gompertz

- Cerca de 100 anos depois surgiu uma das mais famosas leis, a Lei de Gompertz 1825;
- Abordagem biológica da modelagem matemática;
- Força de mortalidade exponencialmente crescente;
- Mortalidade causada unicamente pela degradação do corpo.



Makeham

- Em 1867 Makeham modificou o modelo proposto por Gompertz;
- Captura o padrão de mortalidade não derivada da vitalidade.

APRIL 1892.] 1

JOURNAL

OF THE

INSTITUTE OF ACTUARIES.

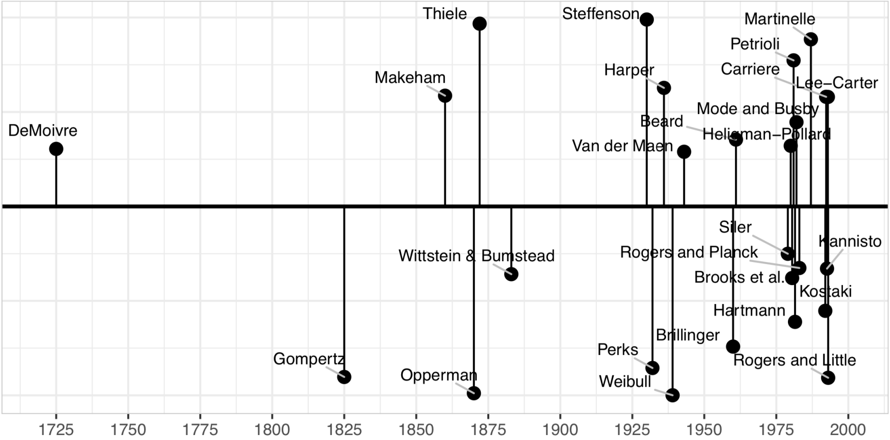
The late William Matthew Makeham.

BY the death of Mr. William Matthew Makeham, so feelingly alluded to by the President of the Institute at the opening of the present session, the scientific literature of life contingencies has lost one of its most brilliant names. A reference to the pages of this *Journal* will convey some impression of the importance and variety of the subjects he dealt with from time to time, but the productions of such a master mind cannot be fitly judged by their length or number. In some of his shorter contributions the highest level of scientific ability and invention is reached, and in all—with scarcely a single exception—are exhibited remarkable powers of insight and analysis, combined with a lucid and luminous mode of exposition, rarely found in association save in intellects of the highest calibre. The value of Makeham's writings will endure for all time, and his will always be a foremost place among those early exponents of actuarial science who guided it into the channels along which it has steadily progressed.

Although Makeham's name is so closely associated with his formula of graduation, or "Law of Mortality", as it is indifferently termed, there are other eminent achievements of his which will immediately present themselves to the minds of readers of this *Journal*. The concentration of thought and mathematical

VOL. XXX. B

Timeline



Thiele
Steffenson
Martinelle
Petrioli
Makeham
Harper
Carriere
Lee–Carter
Mode and Busby
DeMoivre
Beard
Heligman–Pollard
Van der Maen
Siler
Kannisto
Wittstein & Bumstead
Rogers and Planck
Brooks et al.
Kostaki
Hartmann
Brillinger
Perks
Gompertz
Rogers and Little
Opperman
Weibull
1725
1750
1775
1800
1825
1850
1875
1900
1925
1950
1975
2000

Alguns modelos

- El-Gohary, Alshamrani e Al-Otaibi 2013 apresentou uma **generalização** da distribuição associada ao modelo Gompertz 1825, que possui função hazard na curva de crescente, constante, decrescente ou com formato de banheira, dependendo do parâmetro de forma;
- Jafari, Tahmasebi e Alizadeh 2014 apresentou o modelo Beta-Gompertz, que possui os modelos Beta-exponencial e Gompertz generalizada como casos especiais;
- Abd El-Bar 2018 propôs uma **extensão** da distribuição associada ao modelo Gompertz-Makeham, chamada de Gompertz-Makeham transmutada. Este modelo possui função hazard capaz de capturar o comportamento crescente, crescente-constante e em forma de banheira.

Conteúdo

Tempo de vida e função de sobrevivência

Seja T uma variável aleatória que representa o **tempo de vida** de um indivíduo até a sua morte, definida no espaço de probabilidade $(\mathbb{R}_+, \mathcal{B}, \mathbb{P})$, onde $\mathbb{R}_+$ representa seu espaço amostral, $\mathcal{B}$ representa o conjunto de eventos, e $\mathbb{P} : \mathcal{B} \to [0,1]$ representa a função que atribui probabilidades aos eventos (Durrett 2019).
Essa variável pode ser caracterizada pela função de sobrevivência

$$S(t) = \mathbb{P}(T > t),$$

que mensura a probabilidade do indivíduo sobreviver além do tempo t (Colosimo 2001).

Força de mortalidade

Definindo

$$F(t) = \mathbb{P}(T \leq t),$$

e assumindo que F possui uma densidade $F' = f$, a força de mortalidade no momento t é definida como

$$\mu(t) := \lim_{\varepsilon \downarrow 0} \frac{\mathbb{P}(t < T < t + \varepsilon | T > t)}{\varepsilon} = \frac{f(t)}{S(t)},$$

que é uma medida de mortalidade instantânea, que também pode ser representada por

$$\mu(t) = \frac{\partial}{\partial t} \left[-\log S(t) \right].$$

Vale notar que $\varepsilon \mu(t)$ que pode ser interpretado como a probabilidade de morte em um incremento ε de tempo, dada a sobrevivência até o tempo t.

Função de vida média residual

Dado o contexto de mortalidade, além do interesse na função de sobrevivência e na força de mortalidade, também há o interesse na função de vida média residual (FVMR), que é definida como

$$\mathcal{M}(t) = \mathbb{E}(T - t | T > t) = \frac{1}{S(t)} \int_0^{\infty} S(x + t) dx.$$

essa função tem o papel de mensurar o tempo esperado de vida de um indivíduo, dado que ele sobreviveu até o tempo t (Marshall e Olkin 2007). Essa função também pode caracterizar a distribuição de T, como mostrado em Gupta 1975.

Estimação

Aqui, consideraremos que D_t é o número de mortes em um dado intervalo $[t, t+1)$, para $t = 0, \ldots, m$, com $\mathbb{E}(D_t) = \mu(t; \boldsymbol{\theta})E_t$, onde $\mu(t; \boldsymbol{\theta})$ representa a força de mortalidade na idade t e E_t representa o número de pessoas com idade t expostas ao risco de morte. Também iremos considerar $\boldsymbol{D} = (D_0, \ldots, D_m)'$ e $\boldsymbol{E} = (E_0, \ldots, E_m)'$.

Lembrando que cada modelo está associado a um respectivo vetor de parâmetro $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)' \in \Theta$, o qual desejamos estimar

Distribuição Poisson

Este método de estimação é um dos mais utilizados para a modelagem de mortalidade (Brillinger et al. 1986), nele consideramos que D_t possui distribuição Poisson. Devido a esse pressuposto temos que $\mathbb{E}(D_t) = \mu(t;\boldsymbol{\theta})E_t$ e $\mathbb{VAR}(D_t) = \mu(t;\boldsymbol{\theta})E_t$. Tomando $\boldsymbol{D}$ e $\boldsymbol{E}$ definidos anteriormente, obtemos que a função de verossimilhança do vetor $\boldsymbol{\theta}$ é dada por

$$L(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{E}) = \prod_t \frac{\lambda(t,\boldsymbol{\theta})^{D_t} e^{-\lambda(t,\boldsymbol{\theta})}}{D_t!}$$

onde $\lambda(t,\boldsymbol{\theta}) = \mu(t;\boldsymbol{\theta})E_t$. Dessa forma a função de log-verossimilhança vem a ser

$$\ell(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{E}) = \log L(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{E}) = \sum_t \left[D_t \log \lambda(t,\boldsymbol{\theta}) - \lambda(t,\boldsymbol{\theta})\right],$$

e a estimativa de $\boldsymbol{\theta}$ será obtida maximizando a log-verossimilhança, isto é

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{E})$$

Superdispersão

$$\mathbb{VAR}(D_t) > \mathbb{E}(D_t)$$

"Embora as premissas de Poisson e binomial sejam amplamente encontradas na literatura atuarial, raramente são válidas. Os dados de mortalidade geralmente mostram mais variação do que o permitido por esses modelos. Este é o fenômeno da superdispersão" - Macdonald, Richards e Currie 2018

Distribuição Binomial Negativa

Um modelo alternativo ao Poisson para dados de contagem é o binomial negativo, nele consideramos que D_t possui distribuição binomial Negativa. Devido a esse pressuposto temos que $\mathbb{E}(D_t) = \lambda(t, \boldsymbol{\theta}) = \mu(t; \boldsymbol{\theta}) E_t$ e $\mathbb{VAR}(D_t) = \lambda(t, \boldsymbol{\theta}) + \lambda(t, \boldsymbol{\theta})^2 / \phi$.

Para este modelo, temos que

$$f(D_t | \boldsymbol{\theta}, \phi) = \frac{\Gamma(D_t + \phi)}{\Gamma(D_t + 1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda(t, \boldsymbol{\theta})} \right)^{\phi} \left(\frac{\lambda(t, \boldsymbol{\theta})}{\phi + \lambda(t, \boldsymbol{\theta})} \right)^{D_t}$$

onde Γ é a função gama e $\phi > 0$.

Tomando $\boldsymbol{D}$ e $\boldsymbol{E}$ definidos anteriormente, obtemos que a função de verossimilhança do vetor $(\boldsymbol{\theta}, \phi)'$ é dada por

$$L(\boldsymbol{\theta}, \phi | \boldsymbol{D}, \boldsymbol{E}) = \prod_t \frac{\Gamma(D_t + \phi)}{\Gamma(D_t + 1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda(t, \boldsymbol{\theta})} \right)^{\phi} \left(\frac{\lambda(t, \boldsymbol{\theta})}{\phi + \lambda(t, \boldsymbol{\theta})} \right)^{D_t}$$

onde $\lambda(t, \boldsymbol{\theta}) = \mu(t; \boldsymbol{\theta}) E_t$.

Distribuição Binomial Negativa

Como $\log\left(\frac{\Gamma(D_t+\phi)}{\Gamma(\phi)}\right) = \sum_{j=0}^{D_t-1}\log(j+\phi)$ se D_t é inteiro (Cameron e Trivedi 2013; Svetliza e Paula 2003), temos que a função de log-verossimilhança é dada por

$$\begin{aligned}
\ell(\boldsymbol{\theta},\phi|\boldsymbol{D},\boldsymbol{E}) &= \log L(\boldsymbol{\theta},\phi|\boldsymbol{D},\boldsymbol{E}) \\
&= \sum_t \left[\sum_{j=0}^{D_t-1}\log(j+\phi) + D_t\log\lambda(t,\boldsymbol{\theta})\right. \\
&\quad \left. -(D_t+\phi)\log(\phi+\lambda(t,\boldsymbol{\theta})) + \phi\log\phi\right]
\end{aligned}$$

Logo, a estimativa de $(\boldsymbol{\theta},\phi)$ será obtida maximizando a log-verossimilhança, isto é

$$\left(\widehat{\boldsymbol{\theta}},\widehat{\phi}\right) = \arg\max_{(\boldsymbol{\theta},\phi)}\ell(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{E})$$

Contudo, problemas de não-convergência são comuns, como citado em Tai e Noymer 2018

Distribuição Bell

Um variável aleatória Z tem distribuição Bell se sua função de massa de probabilidade é dada por

$$\mathbb{P}(Z = z) = \exp\left(1 - e^{W_0(\eta)}\right) \frac{W_0(\eta)^z B_z}{z!}, \quad z = 0, 1, 2, 3, \ldots$$

onde $W_0(\cdot)$ é a função de Lambert, $\eta > 0$ e $B_z = e^{-1} \sum_{k=0}^{\infty} k^z/k!$ são os números Bell. O número Bell B_z é o z-ésimo momento de uma distribuição Poisson com parâmetro 1.

Começando com $B_0 = B_1 = 1$, os primieiro números Bell são $B_2 = 2$, $B_3 = 5$, $B_4 = 14$.

Temos que $\mathbb{E}(Z) = \eta$ e $\mathbb{VAR}(Z) = \eta[1 + W_0(\eta)]$. Como $W_0(\eta) > 0$ para $\eta > 0$, temos que $\mathbb{VAR}(Z) > \mathbb{E}(Z)$, o que implica que a distribuição Bell pode ser adequada para modelar dados de contagem com super-dispersão.

Distribuição Bell

Propriedades interessantes

1. É uma distribuição de um-parâmetros;
2. Pertence a família exponencial de distribuições de um-parâmetro;
3. A distribuição de Poisson não está aninhada na família Bell, mas para pequenos valores do parâmetro a distribuição de Bell se aproxima da distribuição de Poisson;

Para mais detalhes, ver Castellares, Ferrari e Lemonte 2018.

Distribuição Bell

Assumindo que os números de mortes D_t são gerados por uma distribuição Bell com parâmetro $\eta_t = \eta_t(\boldsymbol{\theta}) = \mu(t;\boldsymbol{\theta})E_t = \lambda(t;\boldsymbol{\theta})$. Tomando $\boldsymbol{D}$ e $\boldsymbol{E}$, a função de verossimilhança para esta distribuição assume forma

$$L(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{E}) = \prod_t \exp\left(1 - e^{W_0(\lambda(t;\boldsymbol{\theta}))}\right) \frac{W_0(\lambda(t;\boldsymbol{\theta}))^{D_t} B_{D_t}}{D_t!}.$$

E a função de log-verossimilhança, a menos de contantes, pode ser expressa como

$$\ell(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{E}) = \sum_t \left[D_t \log W_0(\lambda(t;\boldsymbol{\theta})) - e^{W_0(\lambda(t;\boldsymbol{\theta}))}\right].$$

Logo, a estimativa de $\boldsymbol{\theta}$ será obtida maximizando a log-verossimilhança, isto é

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{E})$$

Conteúdo

Força de mortalidade e sobrevivência

Este modelo foi construído a partir da força de mortalidade, onde Gompertz supõe que é **exponencialmente crescente**, ou seja,

$$\mu(t;\boldsymbol{\theta}) = ae^{bt},$$

onde $t > 0$ e $\boldsymbol{\theta} = (a,b)' \in \mathbb{R}_+^2$. Dessa forma, podemos derivar as outras funções de interesse a partir da relação $\mu(t) = -d\log S(t)/dt$, pois daí obtemos que

$$S(t;\boldsymbol{\theta}) = \exp\left\{-\int_0^t \mu(t;\boldsymbol{\theta})dx\right\} = \exp\left\{-\frac{a\left(e^{bt}-1\right)}{b}\right\}, \qquad t > 0.$$

Função de vida média residual

A forma fechada da vida média residual na idade t para o modelo Gompertz 1825 é dada por

$$\mathcal{M}(t;\boldsymbol{\theta}) = \frac{1}{b}\exp\left\{\frac{a}{b}e^{bt}\right\} E_1\left(\frac{a}{b}e^{bt}\right), \qquad t > 0,$$

onde $E_1(z) = \int_1^\infty t^{-1}e^{-zt}\,dt$ é a função integro-exponencial (Milgram 1985).

Estimação - Binomial Negativa

Para a estimação do modelo, devemos apenas plugar a função μ da lei Gompertz na função λ, e fazer a maximização de verossimilhança. Isto é, considerarmos que $\mathbb{E}D_t = \lambda(t;\boldsymbol{\theta}) = ae^{bt}E_t$, e então maximizar a função de log-verissimilhança

$$
\begin{aligned}
\ell(\boldsymbol{\theta}, \phi | \boldsymbol{D}, \boldsymbol{E}) &= \log L(\boldsymbol{\theta}, \phi | \boldsymbol{D}, \boldsymbol{E}) \\
&= \sum_t \left[\sum_{j=0}^{D_t - 1} \log(j + \phi) + D_t \log \lambda(t, \boldsymbol{\theta}) \right. \\
&\quad \left. - (D_t + \phi) \log(\phi + \lambda(t, \boldsymbol{\theta})) + \phi \log \phi \right]
\end{aligned}
$$

com respeito a $\boldsymbol{\theta} = (a, b, \phi)'$

Estimação - Bell

De maneira análoga, tomamos $\mathbb{E}D_t = \lambda(t;\boldsymbol{\theta}) = ae^{bt}E_t$, e maximizamos a função de log-verossimilhança

$$\ell(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{E}) = \sum_t \left[D_t \log W_0(\lambda(t;\boldsymbol{\theta})) - e^{W_0(\lambda(t;\boldsymbol{\theta}))}\right]$$

com respeito a $\boldsymbol{\theta} = (a,b)'$, para obtermos as estimativas de verossimilhança.

Conteúdo

Força de mortalidade e sobrevivência

Neste modelo, Makeham 1860a modificou o modelo Gompertz 1825 acrescentando uma contante c na sua força de mortalidade, resultando numa função *hazard* dada por

$$\mu(t; \boldsymbol{\theta}) = ae^{bt} + c, \qquad t > 0,$$

onde $\boldsymbol{\theta} = (a, b, c)' \in \mathbb{R}_+^3$ (note que se tomarmos $c = 0$ o modelo será reduzido ao modelo Gompertz 1825). Utilizando a da relação $\mu(t) = -d \log S(t)/dt$, podemos derivar facilmente a função de sobrevivência do modelo Makeham:

$$S(t; \boldsymbol{\theta}) = \exp\left\{-\int_0^t \mu(x)dx\right\} = \exp\left\{-\frac{a\left(e^{bt} - 1\right)}{b}\right\} e^{-ct}, \qquad t > 0.$$

Função de vida média residual

A forma fechada da vida média residual na idade t para o modelo Makeham 1860b é

$$\mathcal{M}(t;\boldsymbol{\theta}) = \frac{1}{b}\exp\left\{\frac{ae^{bt}}{b}\right\}\left(\frac{ae^{bt}}{b}\right)^{\frac{c}{b}}\Gamma\left(-\frac{c}{b},\frac{a}{b}e^{bt}\right),$$

onde

$$\Gamma(u,x) = \int_x^{\infty} t^{u-1}e^{-t}dt, \qquad u \in \mathbb{R}.$$

é a função Gamma-Imcompleta superior

Estimação - Binomial Negativa

Para a estimação do modelo, devemos apenas plugar a função μ da lei Gompertz na função λ, e fazer a maximização de verossimilhança. Isto é, considerarmos que $\mathbb{E}D_t = \left(ae^{bt} + c\right) E_t = \lambda(t; \boldsymbol{\theta})$,, e então maximizar a função de log-verissimilhança

$$\begin{aligned}
\ell(\boldsymbol{\theta}, \phi | \boldsymbol{D}, \boldsymbol{E}) &= \log L(\boldsymbol{\theta}, \phi | \boldsymbol{D}, \boldsymbol{E}) \\
&= \sum_t \left[\sum_{j=0}^{D_t - 1} \log(j + \phi) + D_t \log \lambda(t, \boldsymbol{\theta}) \right. \\
&\quad \left. - (D_t + \phi) \log(\phi + \lambda(t, \boldsymbol{\theta})) + \phi \log \phi \right]
\end{aligned}$$

com respeito a $\boldsymbol{\theta} = (a, b, c, \phi)'$

Estimação - Bell

De maneira análoga, tomamos $\mathbb{E}D_t = \lambda(t;\boldsymbol{\theta}) = \left(ae^{bt} + c\right) E_t$, e maximizamos a função de log-verossimilhança

$$\ell(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{E}) = \sum_t \left[D_t \log W_0(\lambda(t;\boldsymbol{\theta})) - e^{W_0(\lambda(t;\boldsymbol{\theta}))}\right]$$

com respeito a $\boldsymbol{\theta} = (a,b,c)'$, para obtermos as estimativas de verossimilhança.

Ahmed MT Abd El-Bar. "An extended gompertz-makeham distribution with application to lifetime data". Em: *Communications in Statistics-Simulation and Computation* 47.8 (2018), pp. 2454–2475.

David R Brillinger et al. "The natural variability of vital rates and associated statistics". Em: *Biometrics* 42.4 (1986), pp. 693–734.

Fredy Castellares, Silvia LP Ferrari e Artur J Lemonte. "On the Bell distribution and its associated regression model for count data". Em: *Applied Mathematical Modelling* 56 (2018), pp. 172–185.

E Colosimo. "Análise de sobrevivência aplicada". Em: *46 Reuniao anual da regiao brasileira da sociedade International de biometria* (2001).

A Colin Cameron e Pravin K Trivedi. *Regression analysis of count data*. Vol. 53. Cambridge university press, 2013.

David CM Dickson et al. *Actuarial mathematics for life contingent risks*. Cambridge University Press, 2013.

Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019.

Alshamrani El-Gohary, Ahmad Alshamrani e Adel Naif Al-Otaibi. "The generalized Gompertz distribution". Em: *Applied Mathematical Modelling* 37.1-2 (2013), pp. 13–24.

Benjamin Gompertz. "XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS &c". Em: *Philosophical transactions of the Royal Society of London* 115 (1825), pp. 513–583.

Ramesh C Gupta. "On characterization of distribution by conditional expectation". Em: *Communications in Statistics-Theory and Methods* 4.1 (1975), pp. 99–103.

Ali Akbar Jafari, Saeid Tahmasebi e Morad Alizadeh. "The beta-Gompertz distribution". Em: *Revista Colombiana de Estadistica* 37.1 (2014), pp. 141–158.

William Matthew Makeham. "On the law of mortality and the construction of annuity tables". Em: *Journal of the Institute of Actuaries* 8.6 (1860), pp. 301–310.

William Matthew Makeham. “On the law of mortality and the construction of annuity tables”. Em: *Journal of the Institute of Actuaries* 8.6 (1860), pp. 301–310.

MS Milgram. “The generalized integro-exponential function”. Em: *Mathematics of computation* 44.170 (1985), pp. 443–458.

Albert W Marshall e Ingram Olkin. *Life distributions*. Vol. 13. Springer, 2007.

Angus S Macdonald, Stephen J Richards e Iain D Currie. *Modelling Mortality with Actuarial Applications*. Cambridge University Press, 2018.

Carolina F Svetliza e Gilberto A Paula. “Diagnostics in nonlinear negative binomial models”. Em: *Communications in Statistics-Theory and Methods* 32.6 (2003), pp. 1227–1250.

Tzu Han Tai e Andrew Noymer. “Models for estimating empirical Gompertz mortality: With an application to evolution of the Gompertzian slope”. Em: *Population ecology* 60.1-2 (2018), pp. 171–184.