A new look at mortality models with Multivariate Generalized Linear Models (MGLM)

Antoine Burg*, CEREMADE, SCOR

*Joint work with Christophe Dutang, LJK, Grenoble INP - UGA

SCOR chair on mortality research, 2024/04/04

How to model mortality by Causes-of-Deaths ?

- Choice of d causes-of-death, e.g. cardiovascular, neoplasms, accidental ...
- Multivariate observations :

$$D_{x,t} = (D_{x,t}^{(1)}, \dots, D_{x,t}^{(d)}) \qquad D_{x,t}^{(j)} \text{ count of deaths from cause } j,$$

$$q_{x,t} = (q_{x,t}^{(1)}, \dots, q_{x,t}^{(d)}) \qquad q_{x,t}^{(j)} \text{ probability of death from cause } j.$$

- In general, we use the notation $Y = (Y^{(1)}, \dots, Y^{(d)})$.

Objectives of this work:

- Develop a multivariate framework for CoD modelling
- Focus on obtaining closed-form formulas for faster computations

Outline

1. Theoretical results: Closed-form estimators of regression parameters
   - Closed-form formulas for Multinomial distribution
   - Closed-form formulas for Dirichlet distribution
2. Application to Actuarial modelling
   - Reinterpretation of the Age-Period-Cohort model
   - Mortality analysis

MGLM - Assumptions

- Distributional assumption on response variables $\boldsymbol{Y}_i$:
  $(\boldsymbol{Y}_i)_i$ are conditionnally independent given $\boldsymbol{\theta}$ and their distribution belongs to the exponential family :

$$dF(\boldsymbol{y};\boldsymbol{\theta}) = \exp\{(\boldsymbol{\theta}^\intercal \boldsymbol{T}(\boldsymbol{y}) - \kappa(\boldsymbol{\theta}))\omega + a(\boldsymbol{y};\omega)\}d\nu(\boldsymbol{y}).$$

- Structural assumption between $\boldsymbol{\mu_i} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{Y}_i] \in \mathbb{R}^d$ and $\boldsymbol{Z}_i\beta$:

$$\boldsymbol{\mu}_i = h(\boldsymbol{Z}_i\beta) \Leftrightarrow g(\boldsymbol{\mu}_i) = \boldsymbol{Z}_i\beta.$$

  where $g : \mathbb{R}^d \to \mathbb{R}^d$ is the link function, $h : \mathbb{R}^d \to \mathbb{R}^d$, $(h = g^{-1})$.
- Design assumption: for each response $\boldsymbol{y}_i$, the design matrix of observed covariates is block diagonal of the form

$$\boldsymbol{Z}_i = \begin{pmatrix} \boldsymbol{z}_i^{(1)} & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{z}_i^{(d)} \end{pmatrix} \in \mathbb{R}^{d\times q}, \quad \text{where} \quad \boldsymbol{z}_i^{(j)} = \left(z_{i,1}, \ldots, z_{i,p_j}\right).$$

MGLM - Special case of a single categorical variable

Assumption: the explanatory variable $\boldsymbol{z}_i$ is a single categorical variable valued in $\{v_1, \ldots, v_r\}$

$$\forall i = 1, \ldots, n, \quad z_{i,k} = 1_{z_i = v_k}$$

- z_i is the categorical variable of observation i
- v_k a label from $\mathbb{R}^r$, with only 0 but one 1 (e.g. $(1, 0, \ldots, 0)$)

Simplified score formula:

If we denote $\boldsymbol{b}_k = (\beta_k^{(1)}, \ldots, \beta_k^{(d)}) \in \mathbb{R}^d$

$$\begin{aligned} s(\beta) &= \sum_{k=1}^{r} \left[\boldsymbol{J}h(\boldsymbol{b}_k)^\intercal \times \boldsymbol{J}\mu^{-1}(h(\boldsymbol{b}_k))^\intercal \overline{\boldsymbol{T}}(\boldsymbol{y})^{(k)} \right]_{k,\times d} \\ &\quad - \left[\boldsymbol{J}h(\boldsymbol{b}_k)^\intercal \times \boldsymbol{J}\mu^{-1}(h(\boldsymbol{b}_k))^\intercal \nabla_{\boldsymbol{\theta}} \kappa(\mu^{-1}(h(\boldsymbol{b}_k))) \right]_{k,\times d} \overline{\omega}^{(k)} \\ &= \sum_{k=1}^{r} \overline{\omega_k} \left[\overline{\boldsymbol{T}}_k(\boldsymbol{y}) \right] \otimes \boldsymbol{e}_k - \sum_{k=1}^{r} \overline{\omega_k} \left[\nabla_{\boldsymbol{\theta}} \kappa(\boldsymbol{b}_k) \right] \otimes \boldsymbol{e}_k \text{(for the canonical link)} \end{aligned}$$

where $\overline{\boldsymbol{T}}_k(\boldsymbol{y}) = \sum_{i=1}^{n} \frac{\omega_i z_{i,k}}{\overline{\omega_k}} \boldsymbol{T}(\boldsymbol{y}_i)$, $\quad \overline{\omega_k} = \sum_{i=1}^{n} \omega_i z_{i,k}$.

Closed-form formulas for Multinomial distribution

Let $\boldsymbol{Y}_i$ follow a multinomial distribution $\boldsymbol{Y}_i \sim \mathcal{M}_d(m_i, \boldsymbol{p} = (p_1, \dots, p_{d-1}))$. We consider the GLM applied to $\tilde{\boldsymbol{Y}}_i = \boldsymbol{Y}_i / m_i$ which have the following characteristics of the exponential family:

$$\boldsymbol{\theta} = \begin{pmatrix} \log \frac{p_1}{1-\sum\limits_{j=1}^{d-1} p_j} \\ \vdots \\ \log \frac{p_{d-1}}{1-\sum\limits_{j=1}^{d-1} p_j} \\ 0 \end{pmatrix}, \quad \boldsymbol{T}(\tilde{\boldsymbol{y}}) = \begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_d \end{pmatrix}, \quad \kappa(\boldsymbol{\theta}) = -\log\left(1 - \sum_{j=1}^{d-1} p_j\right).$$

Applying previous results enable to directly obtain closed-form expressions, which are a special case of the MLE:

$$\forall k \in [1; r], \quad \begin{pmatrix} \beta_k^{(1)} \\ \vdots \\ \beta_k^{(d-1)} \end{pmatrix} = \left(\log\left(\frac{\overline{y_j^{n,k}}}{1 - \sum\limits_{j=1}^{d-1} \overline{y_j^{n,k}}} \right) \right)_{j \in [1:d]}$$

Closed-form formulas for Dirichlet distribution

The Dirichlet distribution generalizes the univariate Beta distribution. It belongs to the exponential family with 2 usual parametrizations.

- Common:

$$\boldsymbol{\theta} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{pmatrix}, \boldsymbol{T}(\boldsymbol{y}) = \begin{pmatrix} \log y_1 \\ \vdots \\ \log y_d \end{pmatrix}, \kappa(\boldsymbol{\theta}) = \sum_{j=1}^{d} \log\left(\Gamma(\theta_j)\right) - \log\Gamma\left(\sum_{j=1}^{d} \theta_j\right).$$

- Alternative:

$$\boldsymbol{\theta} = \begin{pmatrix} \phi\mu_1 \\ \vdots \\ \phi\mu_{d-1} \\ \phi \end{pmatrix}, \boldsymbol{T}(\boldsymbol{y}) = \begin{pmatrix} \log y_1/y_d \\ \vdots \\ \log y_{d-1}/y_d \\ 0 \end{pmatrix}, \kappa(\boldsymbol{\theta}) \quad = \sum_{j=1}^{d-1} \log\Gamma(\theta_j) + \log\Gamma\left(\frac{\theta_d - \sum_{j=1}^{d-1}\theta_j}{\log\Gamma(\theta_d)}\right)$$

The link between the two parametrizations is given by

$$\phi = \sum_{j=1}^{d} \alpha_j, \quad \mu_j = \mathbb{E}[Y_j] = \frac{\alpha_j}{\sum_{l=1}^{d} \alpha_l}.$$

Closed-form formulas for Dirichlet distribution

Theorem: New estimators for Dirichlet distribution parameters

Let $\boldsymbol{Y}$ follow a Dirichlet distribution $\boldsymbol{Y} \sim \text{Dir}(\boldsymbol{\mu}, \phi)$. The two estimators $\hat{\phi}_{j,l}$ and c converge almost surely towards ϕ and are asymptotically normal.

$$\widehat{\phi}_{j,l} = \frac{1}{\widehat{Cov}\left(Y^{(j)}, \log \frac{Y^{(j)}}{Y^{(l)}}\right)}, \quad \forall j, l \in [1; d], j \neq l.$$

$$\widehat{\phi}_{+,l} = \frac{1}{d-1} \sum_{j=1, j \neq l}^{d} \widehat{\phi}_{j,l}.$$

In addition, we also have $\widehat{\alpha}_j = \widehat{\mu}_j \widehat{\phi}_{+,l}$ with $\widehat{\mu}_j$ the empirical mean estimator of $\mathbb{E}[Y^{(j)}]$.

Elements of the proof:

- Use of Stein's identity
- Strong Law of Large Numbers
- Delta method

Closed-form formulas for Dirichlet distribution

Depending on the parametrization, the MGLM become:

$$\mathbb{E}_{\boldsymbol{\theta}}[Y] = h(\boldsymbol{Z}_i\beta) \Leftrightarrow \forall j \in [1; d-1], \quad \alpha_j = \phi h(\boldsymbol{Z}_i\beta) \quad \text{or} \quad \mu_j = h(\boldsymbol{Z}_i\beta).$$

with

$$\alpha_d = \phi - \sum_{j=1}^{d-1} \alpha_j, \quad \mu_d = 1 - \sum_{j=1}^{d-1} \mu_j.$$

Theorem: closed-form formulas for regression parameters

The MLE of $\beta_k^{(j)}$ is given by

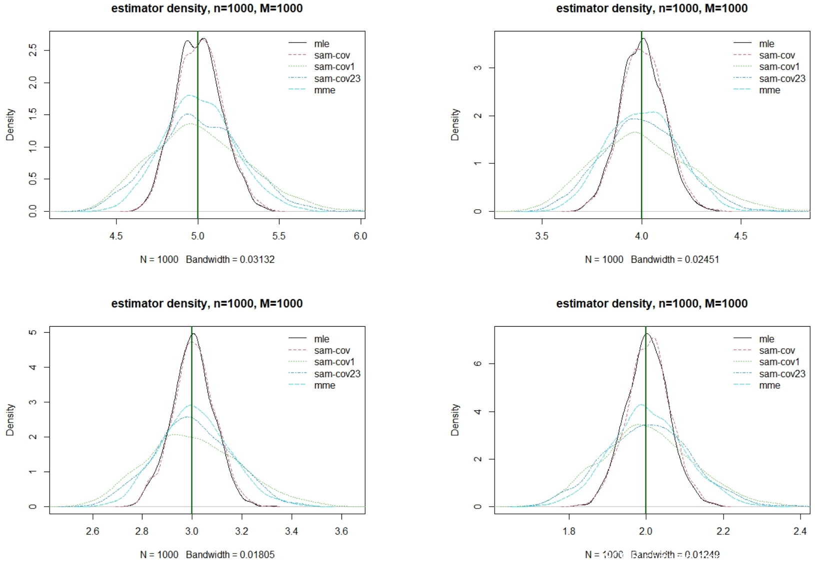
$$\forall k \in [1; r], \quad \forall j \in [1; d-1], \quad \beta_k^{(j)} = h^{-1}\left(\psi^{-1}\left(\overline{\ln y_{n,k}^{(j)}} + \psi(\widehat{\phi}_{+,l})\right)\right) \qquad (\textit{common})$$

$$\forall k \in [1; r], \forall j \in [1; d-1], \quad \beta_k^{(j)} = \frac{\overline{\xi}_{k,j} - \frac{1}{\widehat{\phi}_{+,l}}}{\overline{\log\left(\frac{y_k}{y_d}\right)}^{(n)}} \qquad (\textit{alternative})$$

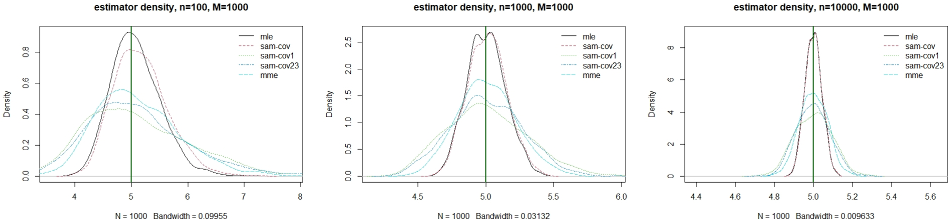
given the value of ϕ estimated by $\widehat{\phi}_{+,l}$.

Dirichlet distribution: estimation of parameters $\boldsymbol{a} = \boldsymbol{\mu}\phi$ - Simulated Data

Example: Dirichlet distribution with parameter $\boldsymbol{x} = (5, 4, 3, 2)$



Influence of sample size n



n	MLE	ϕ^+	$\phi_{1,d}$	$\phi_{2,3}$	MME
100	4.41	0.3	0.26	0.10	0.35
1000	4.72	0.43	0.26	0.26	0.36
10000	10.54	2.00	0.94	0.89	1.93

Table 1: Computation time (ms) for M=1000 simulations

Outline

1. Theoretical results: Closed-form estimators of regression parameters
   - Closed-form formulas for Multinomial distribution
   - Closed-form formulas for Dirichlet distribution
2. Application to Actuarial modelling
   - Reinterpretation of the Age-Period-Cohort model
   - Mortality analysis

Standard Actuarial Modelling

- Distributional assumption for death counts at age x and year t

$$D_{x,t} \sim Poisson(E_{x,t}\mu_{x,t})$$

  or other frequency distributions (binomial, negative binomial ...)

- Model for the force of mortality $\mu_{x,t}$

$$\log(\mu_{x,t}) = \alpha_x + \beta_x \kappa_t \qquad (Lee - Carter)/(M1)$$
$$\log(\mu_{x,t}) = \alpha_x + \kappa_t + \gamma_{t-x} \qquad (Age - Period - Cohort)/(M3)$$

  Cf [BDV02, RH06, CBD$^+$09, Cur16] for more details about these models.

- Numerical optimization of the log-likelihood

$$\mathcal{L}(\alpha, \beta, \kappa) = \sum_{x,t} [D_{x,t}(\alpha_x + \beta_x\kappa_t) - E_{x,t}\exp(\alpha_x + \beta_x\kappa_t)] + K$$

$\Longrightarrow$ **Limitations:**

Designed for a single population

Numerical optimization may slow computations down

Use Case 1: Reinterpretation of Age-Period-Cohort model

- **Classical formulation of APC Model for All-Causes mortality**

$$D_{x,t} \sim \text{Poisson}(E_{x,t}\mu_{x,t}), \quad \log(\mu_{x,t}) = \alpha_x + \kappa_t + \gamma_{t-x}$$
$$D_{x,t} \sim \mathcal{B}(E_{x,t}, q_{x,t})$$

- **New MGLM framework**

$$(D_{x,t}^{(1)}, \ldots, D_{x,t}^{(d)}) \sim \mathcal{M}_d(E_{x,t}, q_{x,t} = (q_{x,t}^{(1)}, \ldots, q_{x,t}^{(d)}))$$
$$\text{logit}\, \mathbb{E}[q_{x,t}] = \alpha_x + \kappa_t + \gamma_{t-x} := \eta_{x,t}$$

  where α_x, κ_t and γ_{t-x} are here vectors of size d.

- **Objective: identify coefficients of the model**
  Values of $\eta_{x,t}$ based on the previous results for single categorical variables.
  Next step: general case of several categorical variables to identify values of α_x, κ_t and γ_{t-x}.

Use Case 1: Reinterpretation of Age-Period-Cohort model

- **Categorical variables**

  $z_i^{(1)}$ stands for age and takes value in $[x_1; x_{max}]$: $z_i^{(1),k} = 1_{z_i^{(1)}=x_k}$

  $z_i^{(2)}$ stands for year and takes value in $[t_1; t_{max}]$: $z_i^{(2),k} = 1_{z_i^{(2)}=t_k}$

  Cohort terms γ_c are actually interaction terms between age and year

- **GLM with 2 categorical variables**

$$g(\mathbb{E}[Y_i]) = \theta_0 + \sum_{k=1}^{x_{max}} z_i^{(1),k}\alpha_k + \sum_{k=1}^{t_{max}} z_i^{(2),k}\kappa_k \qquad \text{intercept and single effect}$$

$$+ \sum_{j<l}\sum_{k,k'} z_i^{(1),k} z_i^{(2),k'}\gamma_{k,k'} \qquad \text{double effect}$$

$\theta = (\theta_0, \alpha_1, \ldots, \alpha_{x_{max}}, \kappa_1, \ldots, \kappa_{t_{max}}, \gamma_1, \ldots, \gamma_{x_{max}\cdot t_{max}})$ parameters vector

Use Case 1: Reinterpretation of Age-Period-Cohort model

- Following [BDR22], we write $\eta = Q\theta$ all the possible values taken by the linear predictor above.

  We also introduce the contrast matrix R that ensure that the model is identifiable with linear conditions $R\theta = 0$.
- If $\tilde{\eta}$ is an estimator of $g(\mathbb{E}[Y])$, they show that $\tilde{\theta}$ is a possible estimator for θ:

$$\tilde{\theta} = \left(Q^\intercal Q + R^\intercal R\right)^{-1} Q^\intercal \tilde{\eta}$$

- Example: we take x= 50..51 and t = 2020..2021. Then
  $\eta = (\eta_{50,2020}, \eta_{50,2021}, \eta_{51,2020}, \eta_{51,2021})$,
  $\theta = (\theta_0, \alpha_{50}, \alpha_{51}, \kappa_{2020}, \kappa_{2021}, \gamma_{50,2020}, \gamma_{50,2021}, \gamma_{51,2020}, \gamma_{51,2021})$,

$$Q = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}, R = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}$$

  corresponding to identifiability constraints $\theta_0 = 0$, $\sum_t \kappa_t = 0$, $\sum_c \gamma_c = 0$, $\sum_c c\gamma_c = 0$ and $\gamma_{50,2020} = \gamma_{51,2021}$.

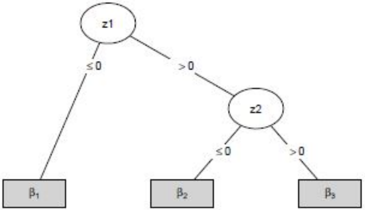
Use Case 2: Mortality analysis

Motivation: Use of GLM-trees algorithms, which require a lot of computations.

The GLM-based tree algorithm [RZ13] consists of splitting the dataset recursively based on a set of partitioning variables and of fitting a GLM on a set of explanatory variables to observations in each node.

Main steps are:

1. Fit the GLM on the current sample
2. Asses parameter stability for each partition variable
3. Choose the best splitting point
4. Repeat



Example from [SHZ19]:GLM tree with 2 partition variables

$$g(E(Y_i)) = \begin{cases} \beta_1 & \text{if } z_1 \leqslant 0 \\ \beta_2 & \text{if } z_1 > 0 \text{ and } z_2 \leqslant 0 \\ \beta_3 & \text{if } z_1 > 0 \text{ and } z_2 > 0 \end{cases}$$

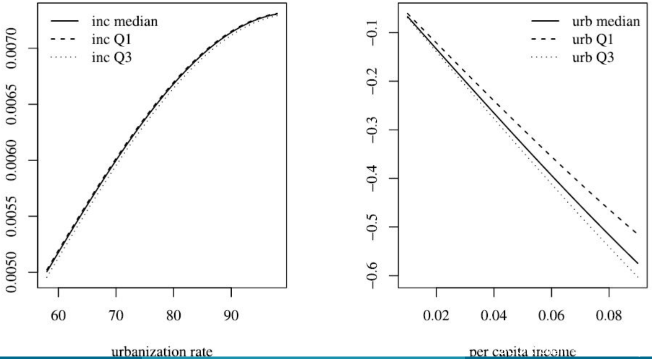
Use Case 2: Mortality analysis

An example from the literature (cf [CN23]).

Model: Beta regression used for univariate Covid-19 mortality rates in Brasil.

$$g_1(\mu_t) = \sum_{i=1}^{p} \beta_i x_{ti} \quad g_2(\phi_t) = \sum_{j=1}^{q} \gamma_j z_{tj}.$$

Results: after variable selection, influence of urbanization and per capita income identified.



Use Case 2: Mortality Analysis

Toy example: from open sources databases US county level.

- Causes-of-deaths data from the Center for Disease Control and Prevention:[1]
  - ICD-10 classification,
  - 4 main causes selected: neoplasms, cardiovascular, infectious and other.
- Explicative variables from US Census Bureau tables[2] :
  - table B19049 for **median income household** (with corrected inflation as of 2022) for different age groups: <25, 25-44, 45-64 and $65+$.
  - table B15003 provides **educational attainment** for the population 25 years and over.
  - table B01001 provides **population estimates** which are used as a measure of exposition.
- Transformation into single categorical variables:
  - median income: 3 levels equally distributed
  - educational attainment: consider concentration of highly educated people (bachelor or more) within the county.

[1] https://wonder.cdc.gov/Deaths-by-Underlying-Cause.html
[2] https://data.census.gov/table/

Use Case 2: Mortality analysis

β parameters	2019				2021			
Profile (age 65-74)	Infectious	Neoplasms	CVD	Other	Infectious	Neoplasms	CVD	Other
Income low % 0-15	-6.03	-4.89	-4.89	-6.98	-6.11	-4.88	-4.73	-5.11
Income low % 15- 20	-6.01	-4.91	-5.02	-6.44	-5.99	-4.88	-4.88	-4.99
Income low % 20-25	-5.93	-4.89	-5.07	-5.95	-5.96	-4.86	-4.95	-4.96
Income low % 25+	-6.02	-4.93	-5.19	-5.79	-5.96	-4.88	-5.05	-4.96
Income high % 15-20	NA	-5.47	-5.64	NA	-6.11	-5.34	-5.23	-6.65
Income high % 20-25	-6.41	-5.29	-5.56	-6.2	-6.85	-5.12	-5.05	- 5.48
Income high % 25+	-6.48	-5.16	-5.56	-6.02	-6.47	-5.12	-5.47	-5.38

Table 2: Zoom on some Regression parameters values

Next steps:

- isolate single effects of each variable,
- assess model performance and select variables,
- use GLM-trees for better accuracy.

Conclusions and Perspectives

Mortality analysis:

- Closed-form estimators enable to use GLM-trees algorithm for more accurate results,
- Identify changes in causes-of-deaths mortality patterns before and after Covid-19 event, at with a high granularity (depending on the data).

APC model application:

- Faster (better?) fit of the model.
- High flexibility of the constraints due to the contrast matrix ; possibility of finer variable selection (age/time range, conditions on cohort...).
- Possibility to use GLM-trees for classification: compare countries, subpopulations ...
- Change of trend detection ?

Bibliography I

- Alexandre Brouste, Christophe Dutang, and Tom Rohmer, *A closed-form alternative estimator for glm with categorical explanatory variables*, Communications in Statistics-Simulation and Computation (2022), 1–17.

- Natacha Brouhns, Michel Denuit, and Jeroen K Vermunt, *A poisson log-bilinear regression approach to the construction of projected lifetables*, Insurance: Mathematics and economics **31** (2002), no. 3, 373–393.

- Andrew JG Cairns, David Blake, Kevin Dowd, Guy D Coughlan, David Epstein, Alen Ong, and Igor Balevich, *A quantitative comparison of stochastic mortality models using data from england and wales and the united states*, North American Actuarial Journal **13** (2009), no. 1, 1–35.

- Francisco Cribari-Neto, *A beta regression analysis of covid-19 mortality in brazil*, Infectious Disease Modelling **8** (2023), no. 2, 309–317.

- Iain D Currie, *On fitting generalized linear and non-linear models of mortality*, Scandinavian Actuarial Journal **2016** (2016), no. 4, 356–383.

- Arthur E Renshaw and Steven Haberman, *A cohort-based extension to the lee–carter model for mortality reduction factors*, Insurance: Mathematics and economics **38** (2006), no. 3, 556–570.

Bibliography II

Thomas Rusch and Achim Zeileis, *Gaining insight with recursive partitioning of generalized linear models*, Journal of Statistical Computation and Simulation **83** (2013), no. 7, 1301–1315.

Heidi Seibold, Torsten Hothorn, and Achim Zeileis, *Generalised linear model trees with global additive effects*, Advances in Data Analysis and Classification **13** (2019), no. 3, 703–725.