Leverage closed-form MLE for multivariate regression models: GLM-trees and actuarial applications

Antoine Burg, Université Paris Dauphine (CEREMADE), SCOR

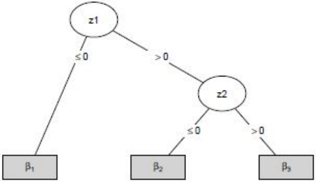
Joint work with Christophe Dutang, LJK, Grenoble INP - UGA.

Insurance Data Science, 2024/06/18

Outline

Motivation - Use of Generalized Linear Models (GLM)-trees

The GLM-based tree algorithm [RZ13] is a recursive split of the dataset based on partitioning variables, similar to other tree algorithm, e.g. CART.
A GLM is fitted at each node on a set of explanatory variables.



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

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

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

Advantages: Automatic partitioning → enables classification + increased accuracy.

Drawbacks: Numerical optimization of the MLE → high computational time.

Motivation - Reinterpretation of Age-Period-Cohort (M3) model

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

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

- **Interpretation as a GLM** ... with *logit* or *cloglog* link [Cur13] (binomial assumption)

$$g(\mathbb{E}[q_{x,t}]) = \alpha_x + \kappa_t + \gamma_{t-x}$$

...**with categorical variables**

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

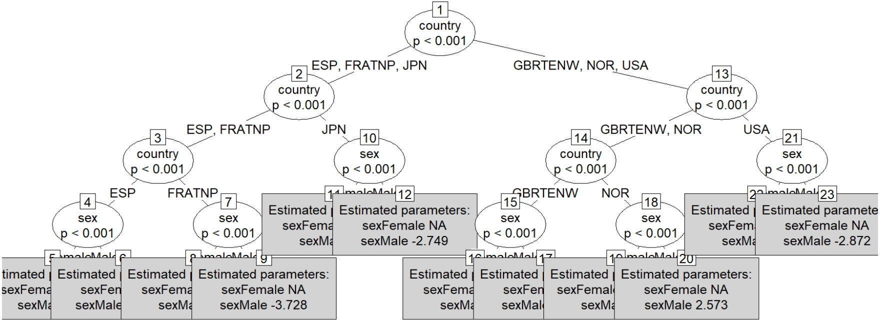
$z_i^{(2)}$ stands for year and takes value in $[t_1; t_{max}]$: $z_i^{(2),k} = 1_{z_i^{(2)} = t_k}$

$$g(\mathbb{E}[q_{x,t}]) = \sum_{k=x_1}^{x_{max}} z_i^{(1),k}\alpha_k + \sum_{k=t_1}^{t_{max}} z_i^{(2),k}\kappa_k \qquad \text{intercept and single effect}$$

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

Use case - classification based on all-cause mortality

- Data source: Human Mortality Database
- Explicative variables
  - Age: 50-84
  - Years: 1990-2019
- Partitioning variables
  - Countries: France, Japan, Norway, Spain, UK, US
  - Sex: Male or Female
- R functions glmtree (*partykit*) and glm (*stats*)



Limitations

- **Computation time:** 39s
- **Convergence:** numerical optimization may fail to converge to a solution
  → known issue with glm.fit in R when data are already well partitioned.
- **Potential solutions:** find closed-form formulas for the regression estimators.

(Multivariate) GLM - Special case of a single categorical variable

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

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

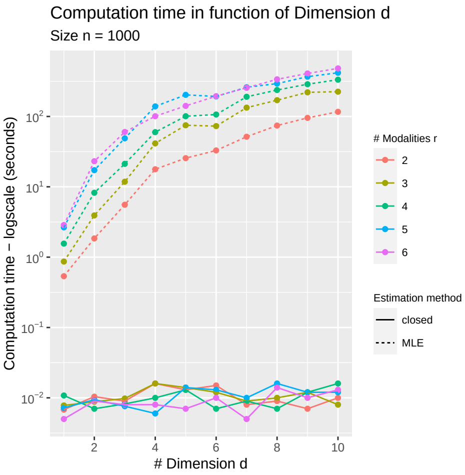
Simplified score formula:
If we denote $\boldsymbol{b}_k = (\beta_k^{(1)}, \dots, \beta_k^{(d)}) \in \mathbb{R}^d$

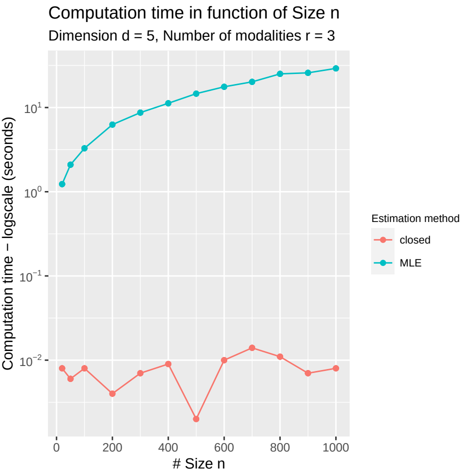
$$\begin{aligned} s(\boldsymbol{\beta}) &= \sum_{k=1}^{r} \left[\boldsymbol{J}h(\boldsymbol{b}_k)^\intercal \times \boldsymbol{J}\mu^{-1}(h(\boldsymbol{b}_k))^\intercal \overline{\boldsymbol{T}}(\boldsymbol{y})^{(k)} \right]_{k, \times d} \\ &\quad - \left[\boldsymbol{J}h(\boldsymbol{b}_k)^\intercal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

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

Particular cases: Multinomial and Dirichlet distributions. Univariate distributions were already looked at in [BDR20, BDR22].

Computation time - multinomial distribution





Age-Period-Cohort (M3) model for CoD (Multivariate GLM)

- **New MGLM framework based on APC model**

$$(D_{x,t}^{(1)}, \ldots, D_{x,t}^{(d)}) \sim \mathcal{M}_d(E_{x,t}, q_{x,t} = (q_{x,t}^{(1)}, \ldots, q_{x,t}^{(d)}))$$

$$g(\mathbb{E}\begin{pmatrix} q_{x,t}^{(1)} \\ \vdots \\ q_{x,t}^{(d)} \end{pmatrix}) = \begin{pmatrix} \alpha_x^{(1)} + \kappa_t^{(1)} + \gamma_{t-x}^{(1)} \\ \vdots \\ \alpha_x^{(d)} + \kappa_t^{(d)} + \gamma_{t-x}^{(d)} \end{pmatrix}$$

- **Estimate model parameters:**
If η is the single categorical variable estimator, $\theta = (\alpha_x^{(1)}, \ldots, \alpha_x^{(d)}, \kappa_t, \ldots, \gamma_c, \ldots)$ the vector of parameters of the APC model, such that $\theta = Q\eta$, and R a contrast matrix for identifiability, then

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

is an estimator of the model parameters.

- **Next step:** Find a smart way to partition the data according to common trends between causes-of-death.

Bibliography I

Alexandre Brouste, Christophe Dutang, and Tom Rohmer, *Closed-form maximum likelihood estimator for generalized linear models in the case of categorical explanatory variables: application to insurance loss modeling*, Computational Statistics **35** (2020), 689–724.

______, *A closed-form alternative estimator for glm with categorical explanatory variables*, Communications in Statistics-Simulation and Computation (2022), 1–17.

Iain D Currie, *Fitting models of mortality with generalized linear and non-linear models*, Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK (2013).

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.