Un estimateur explicite pour le GLM à variables explicatives catégorielles

–Conférence de clôture du programme PANORisk–

Alexandre Brouste[1], Christophe Dutang[2], Tom Rohmer[3]

[1]LMM, Le Mans
[2]CEREMADE, Université Paris-Dauphine
[3]Inrae Toulouse

15 novembre 2021

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Summary

Summary

Parcours PANORisk

nov 2016- nov 2018

Post-doctorat, Le Mans Université, PANORisk

- Brouste, A., Dutang, C., Rohmer, T.
  Closed form Maximum Likelihood Estimation for Generalized Linear Models in the case of categorical explanatory variables : application to insurance loss modelling
  2019, Computational Statistic

- A. Brouste, C. Dutang, V. Dessert & Rohmer, T.,
  E. Gales, P. Golhen, W. Lekeufack & B. Milleville
  Solvency tuned premium for a composite loss distribution
  Workpaper 2018, disponible sur HAL

Dans la continuité de ces travaux

Un article soumis, un package R en developpement

Brouste, A., Dutang, C., Rohmer, T.
A closed-form alternative estimator for GLM with categorical explanatory variables
Article soumis, 2021

Brouste, A., Dutang, C., Rohmer, T.
glmtools - an R Package to compute closed-form estimator for Generalized Linear Models with categorical explanatory variables
Dépôt prochain + article, 2022

Summary

parametric assumption: One-parameter exponential family

Consider the sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, composed of independent random variables (but not i.i.d).

- For $i = 1, \ldots, n$, the distribution of Y_i belongs to the one dimensional exponential family with parameter $\lambda_i \in \Lambda \subset \mathbb{R}$:

the log-density or the log p.m.f. of y_i is assumed to be

$$\log L(\lambda_i | \underline{y}) = (\lambda_i y_i - b(\lambda_i))/a(\phi) + c(y_i, \phi), \quad y_i \in \mathbb{Y} \subset \mathbb{R}, \tag{1}$$

and $-\infty$ if $y_i \notin \mathbb{Y}$, where

- The parameters $\lambda_1, \ldots, \lambda_n$ depend on a finite-dimensional parameter $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^p$ and depend on deterministic exogenous variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$: $\lambda_i = \lambda(\boldsymbol{\vartheta}; \boldsymbol{x}_i)$
- ϕ dispersion parameter

Direct computations lead to

$$b'(\lambda_i) = E_{\boldsymbol{\vartheta}}(Y_i) \quad \text{and} \quad b''(\lambda_i)a(\phi) = Var_{\boldsymbol{\vartheta}}(Y_i).$$

A link function between natural parameter and explanatory variables

Generalized linear models assume:

1. $Y_1, \dots, Y_n$ are independent observations with density or p.m.f. (1)
2. a linear predictor w.r.t. explanatory variables

$$\langle \boldsymbol{x}_i, \boldsymbol{\vartheta} \rangle = \eta_i = \vartheta_1 + x_i^{(2)} \vartheta_2 + \ldots + \vartheta_p x_i^{(p)}$$

3. a link function g:

$$g(E_{\vartheta}(Y_i)) = g(b'(\lambda_i)) = \eta_i,$$

where g is a twice continuously differentiable and injective on $b'(\Lambda)$.

A link function between natural parameter and explanatory variables

Generalized linear models assume:

1. $Y_1, \ldots, Y_n$ are independent observations with density or p.m.f. (1)
2. a linear predictor w.r.t. explanatory variables

$$\langle \boldsymbol{x}_i, \boldsymbol{\vartheta} \rangle = \eta_i = \vartheta_1 + x_i^{(2)}\vartheta_2 + \ldots + \vartheta_p x_i^{(p)}$$

3. a link function g:

$$g(E_{\boldsymbol{\vartheta}}(Y_i)) = g(b'(\lambda_i)) = \eta_i,$$

where g is a twice continuously differentiable and injective on $b'(\Lambda)$.

In other words, $\lambda_i = \lambda(\boldsymbol{\vartheta}, \boldsymbol{x}_i) = \ell(\eta_{\boldsymbol{x}_i}) = \ell(\langle \boldsymbol{x}_i, \boldsymbol{\vartheta} \rangle)$ with $\ell = (b')^{-1} \circ g^{-1}$.
Canonical case: $\ell = id$, i.e. $g = (b')^{-1}$

Score equations for MLE

Let us compute the log-likelihood of $\underline{\boldsymbol{y}} = (y_1, \dots, y_n)$:

$$\log L(\boldsymbol{\vartheta} \mid \underline{\boldsymbol{y}}) = \sum_{i=1}^{n} \frac{y_i \ell(\eta_i) - b\left(\ell(\eta_i)\right)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi), \tag{2}$$

Here, the vector of the parameters $\boldsymbol{\vartheta}$ is unknown.

If the model is identifiable,

- the sequence of maximum likelihood estimators $(\widehat{\boldsymbol{\vartheta}}_n)_{n \geq 1}$ defined by $\widehat{\boldsymbol{\vartheta}}_n = \arg\max_{\boldsymbol{\vartheta} \in \Theta} L(\boldsymbol{\vartheta} \mid \underline{\boldsymbol{y}})$ asymptotically exists and is consistent [Fahrmeir & Kaufmann 1985].
- The maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\vartheta}}_n$, if it exists, is the solution of the non linear system

$$S_j(\boldsymbol{\vartheta}) = 0 \Leftrightarrow \frac{1}{a(\phi)} \sum_{i=1}^{n} x_i^{(j)} \ell'(\eta_i) \left(y_i - b'\left(\ell(\eta_i)\right)\right) = 0, \quad j = 1, \dots, p, \tag{3}$$

Iterative Re-weighted Last Square procedure to get $\widehat{\boldsymbol{\vartheta}}_n$.

Categorical explanatory variables.

Consider the general case where all m explanatory variables are categorical, that is for $j = 1, \ldots, m$ every observations $(x_i^{(j+1)})_i$ takes values in a finite set $\{v_{j,1}, \ldots, v_{j,d_j}\}$. $x_i^{(j+1)}$ needs to encoded using binary dummies as follows

$$x_i^{(j+1),k} = 1_{\{x_i^{(j+1)} = v_{j,k}\}}, \quad k \in \{1, \ldots, d_j\}.$$

Categorical explanatory variables.

Consider the general case where all m explanatory variables are categorical, that is for $j = 1, \dots, m$ every observations $(x_i^{(j+1)})_i$ takes values in a finite set $\{v_{j,1}, \dots, v_{j,d_j}\}$. $x_i^{(j+1)}$ needs to encoded using binary dummies as follows

$$x_i^{(j+1),k} = 1_{\{x_i^{(j+1)} = v_{j,k}\}}, \quad k \in \{1, \dots, d_j\}.$$

$$\begin{aligned}
g\left(\mathbf{E}_{\boldsymbol{\vartheta}} Y_i\right) =& \vartheta^{(1)} + \sum_{j=2}^{m+1} \sum_{k=1}^{d_j} x_i^{(j),k} \vartheta_k^{(j)} && \text{Intercept and single effect} \\
&+ \sum_{j_2<j_3} \sum_{k_2,k_3} x_i^{(j_2),k_2} x_i^{(j_3),k_3} \vartheta_{k_2,k_3}^{(j_2,j_3)} && \text{Double effect} \\
&+ \sum_{j_2<j_3<j_4} \sum_{k_2,k_3,k_4} x_i^{(j_2),k_2} x_i^{(j_3),k_3} x_i^{(j_4),k_4} \vartheta_{k_2,k_3,k_4}^{(j_2,j_3,j_4)} && \text{Triple effect} \\
&+ \dots \\
&+ \sum_{k_2,\dots,k_{m+1}} x_i^{(2),k_2} \dots x_i^{(m+1),k_{m+1}} \vartheta_{k_2,\dots,k_{m+1}}^{(2,\dots,m+1)}, && \text{All crossed effect}
\end{aligned}$$

identifiability contraint

Because of redundancies in the linear predictors and we must impose a contrast matrix $R \in \mathbb{R}^{q \times p}$, in order to identify the unknown parameters, namely

$$R\boldsymbol{\vartheta} = 0.$$

MLE estimator

$$\widehat{\boldsymbol{\vartheta}}_n = \arg \max_{\boldsymbol{\vartheta} \in \Theta | R\boldsymbol{\vartheta} = \mathbf{0}} \mathcal{L}(\boldsymbol{\vartheta} \mid \boldsymbol{Y}),$$

identifiability contraint

Because of redundancies in the linear predictors and we must impose a contrast matrix $R \in \mathbb{R}^{q \times p}$, in order to identify the unknown parameters, namely

$$R\boldsymbol{\vartheta} = 0.$$

MLE estimator

$$\widehat{\boldsymbol{\vartheta}}_n = \arg \max_{\boldsymbol{\vartheta} \in \Theta \,|\, R\boldsymbol{\vartheta} = \mathbf{0}} \mathcal{L}(\boldsymbol{\vartheta} \,|\, \boldsymbol{Y}),$$

Typical examples of contrast vectors in the case of 1-categorical explanatory variable:

No-intercept model	$R = (1, 0, \ldots, 0)$	$i.e.$	$\vartheta^{(1)} = 0;$	(C_0)
Model without factor 1	$R = (0, 1, \ldots, 0)$	$i.e.$	$\vartheta_1^{(2)} = 0;$	(C_1)
Zero-sum condition	$R = (0, 1, \ldots, 1)$	$i.e.$	$\sum_{j=1}^{d} \vartheta_j^{(2)} = 0.$	(C_Σ)

Pro-cons

+ $\widehat{\boldsymbol{\vartheta}}_n$ is an efficient consistent estimator

- Time-consuming IWLS algorithm for high dimensional problem

- R propose a limited choice of distribution/link function.

Reformulation

the linear predictor η_{x_i} simplifies in the following way

$$g\left(\mathbb{E}_{\boldsymbol{\vartheta}} Y_i\right) = \eta_{x_i} = \vartheta_{\mathbf{1}} + \sum_{j} \vartheta_{k_j}^{(j)} + \sum_{j_2 < j_3} \vartheta_{k_2,k_3}^{(j_2,j_3)} + \cdots + \vartheta_{k_2,\ldots,k_{m+1}}^{(2,\ldots,m+1)} := \eta_{k_2,\ldots,k_{m+1}},$$

Let $\boldsymbol{\eta} = \left(\eta_{k_2,\ldots,k_{m+1}}\right)_{k_2,\ldots,k_{m+1}}$ and define the matrix Q as

$$\boldsymbol{\eta} = Q\boldsymbol{\vartheta}.$$

dimension	$Q =$	terms
$m = 1$	$(1_{d_2},$	Intercept
	$I_{d_2})$	Single effect
$m = 2$	$(1_{d_3 d_2},$	Intercept
	$1_{d_3} \otimes I_{d_2}, I_{d_3} \otimes 1_{d_2},$	Single effect
	$I_{d_3 d_2})$	Double effect
$m = 3$	$(1_{d_4 d_3 d_2},$	Intercept
	$1_{d_4 d_3} \otimes I_{d_2}, 1_{d_4} \otimes I_{d_3} \otimes 1_{d_2}, I_{d_4} \otimes 1_{d_3 d_2},$	Single effect
	$1_{d_4} \otimes I_{d_3 d_2}, I_{d_4} \otimes 1_{d_3} \otimes I_{d_2}, I_{d_4 d_3} \otimes 1_{d_2},$	Double effect
	$I_{d_4 d_3 d_2}),$	Triple effect

Table: Examples of Q matrix for 1, 2 or 3 variables

Closed-form estimator

CFE estimator

$$\widetilde{\vartheta}_n = (Q'Q + R'R)^{-1} Q' g(\bar{Y}),$$

where $g(\bar{Y}) = \left(g(\bar{Y}^{k_2,\ldots,k_{m+1}})\right)_{k_2,\ldots,k_{m+1}}$, with the mean values

$$\bar{Y}^{k_2,\ldots,k_{m+1}} = \frac{1}{m_{k_2,\ldots,k_{m+1}}} \sum_{i=1;\eta_{x_i}=\eta_{k_2,\ldots,k_{m+1}}}^{n} Y_i$$

and the frequencies

$$m_{k_2,\ldots,k_{m+1}} = \#\{i; \eta_{x_i} = \eta_{k_2,\ldots,k_{m+1}}\}.$$

Nice properties I

Full model (with all crossed effect)

As soon as the matrix R is such that $Q'Q + R'R$ is definite-positive, we have

$$\widetilde{\boldsymbol{\vartheta}}_n = \widehat{\boldsymbol{\vartheta}}_n,$$

that is, the CFE is the MLE.

Nice properties I

Full model (with all crossed effect)

As soon as the matrix R is such that $Q'Q + R'R$ is definite-positive, we have

$$\widetilde{\boldsymbol{\vartheta}}_n = \widehat{\boldsymbol{\vartheta}}_n,$$

that is, the CFE is the MLE.

Example, in the model,

$$g\left(\mathbb{E}_{\boldsymbol{\vartheta}} Y_i\right) = \vartheta^{(1)} + \sum_{k=1}^{d_2} x_i^{(2),k} \vartheta_k^{(2)} + \sum_{l=1}^{d_3} x_i^{(3),l} \vartheta_l^{(3)} + \sum_{k=1}^{d_2} \sum_{l=1}^{d_3} x_i^{(2),k} x_i^{(3),l} \vartheta_{k,l}^{(2,3)}.$$

type	*ref. category (1^{st} modality)*	*No intercept, no single-variable dummy*
contrast	$\vartheta_{(2),1} = \vartheta_{(3),1} = 0$; $\forall l,\ \vartheta_{1l} = 0$; $\forall k,\ \vartheta_{k1} = 0$	$\vartheta_0 = 0$; $\forall l,\ \vartheta_{(3),l} = 0$; $\forall k,\ \vartheta_{(2),k} = 0$
	$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$

Nice properties II

Restricted Model (without the all crossed effect)

As soon as the matrix R is such that $Q'Q + R'R$ is definite-positive, we have

$$\widetilde{\boldsymbol{\vartheta}}_n \xrightarrow{\text{a.s.}} \boldsymbol{\vartheta},$$

Moreover $\widetilde{\boldsymbol{\vartheta}}_n$ is asymptotically normal.

- the CFE is not the MLE
- The MLE is asymptotically efficient but time consuming for high dimensional datasets
- $\widetilde{\boldsymbol{\vartheta}}_n$ is very fast
- $\widetilde{\boldsymbol{\vartheta}}_n$ does not depend of the distribution of Y_i

Gamma GLM for 2 categorical explanatory variables with single-effect only

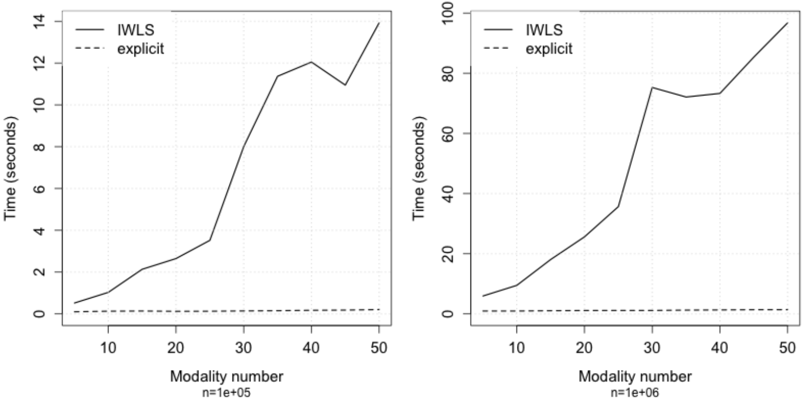

Figure: Computation time for gamma response with log link function (average over 5 runs)

Example, assurance dataset

Loss frequency modeling

- $n = 764428$ claims
- 7 explanatory categorical variables

Modèle	Méthode	AIC	Temps	parameters
GLM Poisson (offset, lien log) effets simples	MLE	461	12.4 s	43
	CFE	1113	1.2 s	43
GLM Gamma (lien log) effets simples	MLE	1076k	2.2s	43
	CFE	1079k	0.2s	43
GLM Poisson (offset, lien log) effets mixtes 8x6x2	MLE	512	57 s	96
	CFE	512	0.6 s	96
GLM Poisson (offset, lien log) effets mixtes 13x10x4	MLE	—	—	520
	CFE	439	4.1 s	520

Perspectives

1. Model selection

2. Multivariate model

GLM multivariés

Considérons $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ des vecteurs aléatoires indépendants et non identiquement distribuées: $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id})$ avec:

1. $Y_{i1}, \ldots, Y_{id}$ ne sont pas indépendant
2. $Y_{i1}, \ldots, Y_{id}$ sont de type exponentielle de paramètre naturel $\lambda_{i1}, \ldots, \lambda_{id}$:

$$\log \mathcal{L}_j(\lambda_{ij}|y_{ij}) = \frac{\lambda_{ij} T_j(y_i) - b_j(\lambda_{ij})}{\phi_j} + c_j(y_i, \phi_j).$$

3. $\lambda_{ij} = \lambda_j(\boldsymbol{x}_i, \boldsymbol{\vartheta}_j)$

$\rightarrow$ Il existe des fonctions de liens $g_1, \ldots, g_d$ telles que

$$g_j(\mathbb{E}(T(Y_{ij}))) = \eta_{ij} = \langle \boldsymbol{x}_i, \boldsymbol{\vartheta}_j \rangle, \quad j = 1, \ldots, d, \quad i = 1, \ldots, n.$$

- Hypothèse paramétrique sur les marges
- Hypothèse paramétrique sur la copule

The Sklar's theorem

Sklar's Theorem (1959)

Soit $X = (X_1, \ldots, X_d)$ un vecteur aléatoire de dimension d avec f.d.r. F and f.d.r. marginales $F_1, \ldots F_d$ supposées continus. Alors il existe une unique fonction $C : [0,1]^d \to [0,1]$ telle que:

$$F(x_1, \ldots, x_d) = C\{F_1(x_1), \ldots, F_d(x_d)\}, \qquad (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

- La copule C caractérises la structure de dépendence du vecteur aléatoire X.
- La copule C peut être exprimée de façon unique par:

$$C(u_1, \ldots, u_d) = F\{F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\}, \qquad (u_1, \ldots, u_d) \in [0,1]^d.$$

Copules classiques

- Independence copula:

$$C^{\Pi}(u_1, \ldots, u_d) = \prod_{j=1}^{d} u_j;$$

- Normal copulas

$$C_{\Sigma}^{N}(u_1, \ldots, u_d) = \Phi_{d,\Sigma}\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\};$$

- Gumbel–Hougaard copulas:

$$C_{\theta}^{GH}(u_1, \ldots, u_d) = \exp\left(-\left[\sum_{j=1}^{d}\{-\log(u_j)\}^{\theta}\right]^{1/\theta}\right), \quad \theta \geq 1;$$

- Clayton copulas:

$$C_{\theta}^{Cl}(u_1, \ldots, u_d) = \left(\sum_{j=1}^{d} u_j^{-\theta} - d + 1\right)^{-1/\theta}, \quad \theta > 0.$$

Estimation paramétrique dans le Modèle multivarié

Supposons que la copule des vecteurs aléatoires $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}$ soit C_θ. Posons c_θ la densité de la copule C_θ, c.-à-d.:

$$c_\theta\left(u_1, \ldots, u_d\right) = \frac{\partial^d C_\theta\left(u_1, \ldots, u_d\right)}{\partial u_1 \ldots \partial u_d}.$$

La densité de $\boldsymbol{Y}_i$ est

$$f(\boldsymbol{Y}_i|\boldsymbol{\vartheta}, \theta) = c_\theta(F_1(Y_{i1}|\boldsymbol{\vartheta_1}), \ldots, F_d(Y_{id}|\boldsymbol{\vartheta_d})) \prod_{j=1}^{d} f_j(Y_{ij}|\boldsymbol{\vartheta_j}).$$

Ainsi la log-vraisemblance de $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}$ peut-être réécrite comme

$$\sum_{i=1}^{n} \log \mathcal{L}(\boldsymbol{\vartheta}, \theta|\boldsymbol{Y}_i) = \sum_{i=1}^{n} \log c_\theta(F_1(Y_{i1}|\boldsymbol{\vartheta_1}), \ldots, F_d(Y_{id}|\boldsymbol{\vartheta_d})) + \sum_{j=1}^{d} \sum_{i=1}^{n} \log \mathcal{L}_j(\boldsymbol{\vartheta_j}|Y_{ij})$$

Inference for Margins(IFM) estimation

1. Estimation des paramètres marginales en utilisant le CFE dans les modèles univariés

2. Maximisation la log-likelihood multivariée conditionellement à $\widetilde{\boldsymbol{\vartheta}}_{\boldsymbol{j}}$

$$\tilde{\theta}^{IFM,\gamma} = \arg\max_{\theta} \sum_{i=1}^{n} \log c_{\theta^{\gamma}}(F_1(Y_{i1}|\widetilde{\boldsymbol{\vartheta}}_{\boldsymbol{1}}), \ldots, F_d(Y_{id}|\widetilde{\boldsymbol{\vartheta}}_{\boldsymbol{d}})).$$

3. Normalité asymptotique et convergence p.s. pour l'estimateur 'IFM'. (travaux en cours)