One-Step estimation procedure in univariate and multivariate GLMs with categorical explanatory variables

Alexandre Brouste (LMM), Christophe Dutang (UGA),
Lilit Hovsepyan (LMM) & Tom Rohmer (INRAE)

JSS, Occimath, 2024, Toulouse

LMM
Laboratoire Manceau
de Mathématiques
Le Mans Université

Introduction

- GLMs in univariate and multivariate contexts
  - Estimated via the maximum likelihood estimator (MLE)
    - usually asymptotically efficient
    - time-consuming: with Newton-Raphson type algorithms, particularly with large datasets or numerous variables

Introduction

- GLMs in univariate and multivariate contexts
  - Estimated via the maximum likelihood estimator (MLE)
    - usually asymptotically efficient
    - time-consuming: with Newton-Raphson type algorithms, particularly with large datasets or numerous variables
- In the *univariate* scenario:
  - Closed-form estimator (CFE): fast to be computed, not always efficient

Introduction

- GLMs in univariate and multivariate contexts
  - Estimated via the maximum likelihood estimator (MLE)
    - usually asymptotically efficient
    - time-consuming: with Newton-Raphson type algorithms, particularly with large datasets or numerous variables
- In the *univariate* scenario:
  - Closed-form estimator (CFE): fast to be computed, not always efficient
  - One-step closed-form estimator (OS-CFE): fast to be computed, asymptotically efficient

Introduction

- GLMs in univariate and multivariate contexts
  - Estimated via the maximum likelihood estimator (MLE)
    - usually asymptotically efficient
    - time-consuming: with Newton-Raphson type algorithms, particularly with large datasets or numerous variables
- In the *univariate* scenario:
  - Closed-form estimator (CFE): fast to be computed, not always efficient
  - One-step closed-form estimator (OS-CFE): fast to be computed, asymptotically efficient
- In the *multivariate* scenario:
  - Inference for margins (IFM), (Xu 1996, Joe 1997, 2005)
  - MLE-IFM vs OSCFE-IFM

Notation for univariate GLMs

$\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ observation sample. Y_i, $i \in I$, independent r.v.s belong to the one-parameter exponential family of probability measures valued in $\Lambda \subset \mathbb{R}$.

$$\log \mathcal{L}(\boldsymbol{\beta}, \phi \mid \boldsymbol{Y}) = \sum_{i=1}^{n} \frac{\lambda_i(\boldsymbol{\beta}) Y_i - b\left(\lambda_i(\boldsymbol{\beta})\right)}{a(\phi)} + \sum_{i=1}^{n} c(Y_i, \phi),$$

Notation for univariate GLMs

$\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ observation sample. Y_i, $i \in I$, independent r.v.s belong to the one-parameter exponential family of probability measures valued in $\Lambda \subset \mathbb{R}$.

$$\log \mathcal{L}(\boldsymbol{\beta}, \phi \,|\, \boldsymbol{Y}) = \sum_{i=1}^{n} \frac{\lambda_i(\boldsymbol{\beta}) Y_i - b\left(\lambda_i(\boldsymbol{\beta})\right)}{a(\phi)} + \sum_{i=1}^{n} c(Y_i, \phi),$$

$a : \mathbb{R} \to \mathbb{R}$, $b : \Lambda \to \mathbb{R}$ and $c : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}$ are fixed real-valued measurable functions, ϕ is the dispersion parameter.
The parameters $\lambda_1, \ldots, \lambda_n$ depend on $\boldsymbol{\beta} \in B \subset \mathbb{R}^p$.
Theoretical moments of Y_i are:

$$\mathbf{E}_{\boldsymbol{\beta}} Y_i = b'(\lambda_i(\boldsymbol{\beta})) = \mu_i \quad \text{and} \quad \mathbf{Var}_{\boldsymbol{\beta}} Y_i = b''(\lambda_i(\boldsymbol{\beta})) a(\phi) = V(\mu_i) a(\phi),$$

where $V : \mu \mapsto V(\mu) = b'' \circ (b')^{-1}(\mu)$ is the variance of μ.

Notation for univariate GLMs

Linear predictors and the link function is noted respectively by η_i and g in

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \eta_i, \quad \text{for all } \boldsymbol{\beta} \in B,$$

where g is a twice continuously differentiable and bijective function from $b'(\Lambda)$ to $\mathbb{R}$.

Notation for univariate GLMs

Linear predictors and the link function is noted respectively by η_i and g in

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \eta_i, \quad \text{for all } \boldsymbol{\beta} \in B,$$

where g is a twice continuously differentiable and bijective function from $b'(\Lambda)$ to $\mathbb{R}$.

The parameter $\boldsymbol{\beta} \in B \subset \mathbb{R}^p$ is unknown and should be estimated. Classically, the MLE $\widehat{\boldsymbol{\beta}}_n$ for $\boldsymbol{\beta}$ is defined by

$$(\widehat{\boldsymbol{\beta}}_n, \widehat{\phi}_n) = \arg \max_{(\boldsymbol{\beta},\phi) \in B \times \mathbb{R}_*^+} \log \mathcal{L}(\boldsymbol{\beta}, \phi \mid \boldsymbol{Y}).$$

$$S_n(\widehat{\boldsymbol{\beta}}_n) := \frac{\partial}{\partial \boldsymbol{\beta}} \log \mathcal{L}(\widehat{\boldsymbol{\beta}}_n, \phi \mid \boldsymbol{Y}) = 0$$

Under the regularity conditions (Fahrmeir, L. & Kaufmann, H. (1985)) the MLE $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ asymptotically exists.

Notation for univariate GLMs

As soon as the MLE is unique, that is to say there is no over-parametrization in the model, we have

$$\mathcal{I}_n^{T/2}(\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow[n \to +\infty]{L} \mathcal{N}_p\left(\mathbf{0}_p, I_p\right),$$

where $\mathcal{I}_n(\boldsymbol{\beta})$ is the Fisher Information matrix, $\mathcal{I}_n^{1/2}\mathcal{I}_n^{T/2} = \mathcal{I}_n$, and I_p is the identity matrix of $\mathbb{R}^{p \times p}$.

Notation for univariate GLMs

As soon as the MLE is unique, that is to say there is no over-parametrization in the model, we have

$$\mathcal{I}_n^{T/2}(\beta)(\widehat{\beta}_n - \beta) \xrightarrow[n\to+\infty]{L} \mathcal{N}_p\left(\mathbf{0}_p, I_p\right),$$

where $\mathcal{I}_n(\beta)$ is the Fisher Information matrix, $\mathcal{I}_n^{1/2}\mathcal{I}_n^{T/2} = \mathcal{I}_n$, and I_p is the identity matrix of $\mathbb{R}^{p\times p}$.

But Newton-Raphson type algorithm can be time-consuming when having large number of variables/modalities or sample size.

We aim for fast computable and asymptotically efficient estimators.

Binary dummy variables

When the explanatory variables are only categorical, it can be encoded using binary dummies, where observations $(x_i^{(j+1)})_i$ take values in a finite set $\{v_{j,1}, \dots, v_{j,d_j}\}$

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

Binary dummy variables

When the explanatory variables are only categorical, it can be encoded using binary dummies, where observations $(x_i^{(j+1)})_i$ take values in a finite set $\{v_{j,1}, \dots, v_{j,d_j}\}$

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

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

Binary dummy variables

The vector of linear predictors $\boldsymbol{\eta} = (\eta_i)_{i=1,\dots,n}$ can be rewritten as

$$\boldsymbol{\eta} = X\boldsymbol{\beta}.$$

Binary dummy variables

The vector of linear predictors $\boldsymbol{\eta} = (\eta_i)_{i=1\dots,n}$ can be rewritten as

$$\boldsymbol{\eta} = X\boldsymbol{\beta}.$$

Redundancies of the matrix X implies the model to be non identifiable.

Thus, we need to impose linear conditions on $\boldsymbol{\beta}$ by a contrast matrix R: $R\boldsymbol{\beta} = 0$. We also can consider a restricted parameter $\tilde{\boldsymbol{\beta}}$ for which the model is identifiable. Hence, there exists a matrix $\tilde{X}$ related to R, such that

$$\boldsymbol{\eta} = \tilde{X}\tilde{\boldsymbol{\beta}}.$$

Binary dummy variables

The vector of linear predictors $\boldsymbol{\eta} = (\eta_i)_{i=1\ldots,n}$ can be rewritten as

$$\boldsymbol{\eta} = X\boldsymbol{\beta}.$$

Redundancies of the matrix X implies the model to be non identifiable.
Thus, we need to impose linear conditions on $\boldsymbol{\beta}$ by a contrast matrix R: $R\boldsymbol{\beta} = 0$. We also can consider a restricted parameter $\tilde{\boldsymbol{\beta}}$ for which the model is identifiable. Hence, there exists a matrix $\tilde{X}$ related to R, such that

$$\boldsymbol{\eta} = \tilde{X}\tilde{\boldsymbol{\beta}}.$$

Let's define the vector $\boldsymbol{\eta}^{\star} = (h_j)_{j=1,\ldots,d}$ constituted with the d distinct values of $\boldsymbol{\eta}$. There exists a matrix $\tilde{Q}$ related to R, such that

$$\boldsymbol{\eta}^{\star} = \tilde{Q}\tilde{\boldsymbol{\beta}}.$$

CFE and OS-CFE

The proposed (A. Brouste et al. (2020), (2022)) closed-form estimator of the restricted parameter is

$$\tilde{\beta}_n^{CFE} = (\tilde{Q}^T\tilde{Q})^{-1}\tilde{Q}^T g(\overline{\boldsymbol{Y}}_{\boldsymbol{n}}), \quad g(\overline{\boldsymbol{Y}}_{\boldsymbol{n}}) = \begin{pmatrix} g(\overline{Y}_n^1) & \dots & g(\overline{Y}_n^d) \end{pmatrix}^T$$

where

$$\overline{Y}_n^k = \frac{\sum\limits_{i=1;\eta_i=h_k}^{n} Y_i}{m_k}, m_k = \#\{i \in \{1,\dots,n\}; \eta_i = h_k\}.$$

OS-CFE

$$\tilde{\beta}_n^{\text{OS-CFE}} = \tilde{\beta}_n^{CFE} + \tilde{\mathcal{I}}_n(\tilde{\beta}_n^{CFE})^{-1}\tilde{S}_n(\tilde{\beta}_n^{CFE})$$

Asymptotic results

We showed recently that

Asymptotic results

(Brouste, A., Dutang, C., Hovsepyan, L. and Rohmer, T. (2023))

$$\sqrt{n}(\tilde{\beta}_n^{CFE} - \tilde{\beta}) \xrightarrow[n\to+\infty]{L} \mathcal{N}_{p^\star}\left(\mathbf{0}_{p^\star}, a(\phi)(\tilde{Q}^T\tilde{Q})^{-1}\tilde{Q}^T\Sigma^{-1}(\tilde{\beta})\tilde{Q}(\tilde{Q}^T\tilde{Q})^{-1}\right),$$

$$\sqrt{n}(\tilde{\beta}_n^{\text{OS-CFE}} - \tilde{\beta}) \xrightarrow[n\to+\infty]{L} \mathcal{N}_{p^\star}\left(\mathbf{0}_{p^\star}, \tilde{\mathcal{I}}^{-1}(\beta)\right).$$

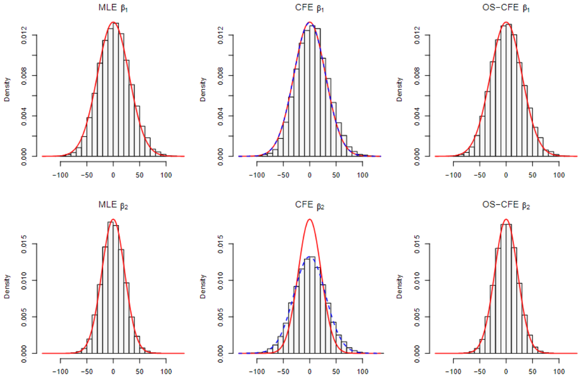
where $\tilde{S}_n$ and $\tilde{\mathcal{I}}$ are the restricted score vector and Fisher information,

$$\tilde{\mathcal{I}}(\beta) = \tilde{Q}\Sigma\tilde{Q}^T a(\phi)^{-1}$$

Monte-Carlo simulations

- Single effects Gamma-GLM, n $= 10^4$, fixed sample size:

Computation time	MLE	CFE	OS-CFE
Gamma	393.659	23.564	25.198



Introduction to dataset

The Covea Affinity dataset under study is composed of 76,446 claim amounts ranging from 4 to 33,531 EUR.

Three covariates have been selected from the 124 available for the pricing of the guarantee

- vehicle brand with $d_2 = 2$ modalities,
- pricing segment with $d_3 = 6$ modalities,
- age class with $d_4 = 8$ modalities.

	CFE	OS-CFE	MLE
Time (s)	0.01	0.01	0.30

Notation for multivariate GLMs

Let the sample $\boldsymbol{Y} = (\underline{\boldsymbol{Y}}_1, \ldots, \underline{\boldsymbol{Y}}_n)$ be composed of $\mathbb{R}^s$-valued independent random vectors. Each vector $\underline{\boldsymbol{Y}}_i = (Y_{i,1}, \ldots, Y_{i,s})$ has marginals $Y_{i,j}$, with natural parameters λ_{ij} linked to parameters $\boldsymbol{\beta}_j$.

The likelihood $\mathcal{L}_{ij}$ for $Y_{i,j}$ is given by:

$$\log \mathcal{L}_{ij}(\boldsymbol{\beta}_j, \phi_j \mid y_{i,j}) = \frac{\lambda_{ij}(\boldsymbol{\beta}_j) y_{i,j} - b_j(\lambda_{ij}(\boldsymbol{\beta}_j))}{a_j(\phi_j)} + c_j(y_{i,j}, \phi_j).$$

Notation for multivariate GLMs

Let the sample $\boldsymbol{Y} = (\underline{\boldsymbol{Y}}_1, \ldots, \underline{\boldsymbol{Y}}_n)$ be composed of $\mathbb{R}^s$-valued independent random vectors. Each vector $\underline{\boldsymbol{Y}}_i = (Y_{i,1}, \ldots, Y_{i,s})$ has marginals $Y_{i,j}$, with natural parameters λ_{ij} linked to parameters $\boldsymbol{\beta}_j$.

The likelihood $\mathcal{L}_{ij}$ for $Y_{i,j}$ is given by:

$$\log \mathcal{L}_{ij}(\boldsymbol{\beta}_j, \phi_j \mid y_{i,j}) = \frac{\lambda_{ij}(\boldsymbol{\beta}_j) y_{i,j} - b_j(\lambda_{ij}(\boldsymbol{\beta}_j))}{a_j(\phi_j)} + c_j(y_{i,j}, \phi_j).$$

GLMs relate the expected value $\mathbb{E} Y_{i,j} = b_j'(\lambda_{ij}(\boldsymbol{\beta}_j))$ to the predictors η_{ij} via link functions g_j:

$$g_j(\mathbb{E} Y_{i,j}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j = \eta_{ij}.$$

Here, $\boldsymbol{x}_{ij}$ are vectors determined by m_j <u>deterministic</u> explanatory variables.

Copula and Sklar's theorem

In this setting, the variables $Y_{i1}, \ldots, Y_{is}$ constituting $\underline{\boldsymbol{Y}}_i$ are not assumed independent. We consider a parametric copula for the joint distribution of $(Y_{i1}, \ldots, Y_{is})$:

Sklar's Theorem (1959):

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_s)$ be an s-dimensional random vector with c.d.f. $\boldsymbol{F}$ and continuous marginal c.d.f.s $F_1, \ldots F_s$. Then there exists a unique function $C : [0,1]^s \to [0,1]$ such that:

$$\boldsymbol{F}(\boldsymbol{y}) = C\{F_1(y_1), \ldots, F_s(y_s)\}, \qquad \boldsymbol{y} = (y_1, \ldots, y_s) \in \mathbb{R}^s.$$

▷ The so called copula C characterize the dependence between the components of $\boldsymbol{Y}$.

IFM approach

Let $\alpha_j = (\beta_j, \phi_j)$. The log-likelihood of $y = (\underline{y}_1, \ldots, \underline{y}_n)$ can be written as:

$$\log \mathcal{L}(\alpha, \theta \mid y) = \sum_{i=1}^{n} \log c_\theta(F_1(y_{i,1} \mid \alpha_1), \ldots, F_s(y_{i,s} \mid \alpha_s)) + \sum_{j=1}^{s} \sum_{i=1}^{n} \log \mathcal{L}_{ij}(\alpha_j \mid y_{i,j}).$$

Estimation:

- MLE approach: $\hat{\xi} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_s, \hat{\theta})$ is solution of

$$\left(\frac{\partial \log \mathcal{L}}{\partial \alpha_1}, \ldots, \frac{\partial \log \mathcal{L}}{\partial \alpha_s}, \frac{\partial \log \mathcal{L}}{\partial \theta}\right)(\xi) = 0.$$

- IFM approach: $\hat{\xi} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_s, \hat{\theta})$ is solution of

$$\left(\frac{\partial \log \mathcal{L}_1}{\partial \alpha_1}, \ldots, \frac{\partial \log \mathcal{L}_s}{\partial \alpha_s}, \frac{\partial \log \mathcal{L}}{\partial \theta}\right)(\xi) = 0.$$

One-Step Closed-form IFM (OSCFE-IFM) estimator

- OSCFE-IFM approach:
  - For $\boldsymbol{\beta}_j$, the One-Step Closed Form Estimator (Brouste et al. 2023) is given by:

$$\hat{\beta}_j^\star = (Q_j^T Q_j)^{-1} Q_j^T g_j(\bar{Y}_{\cdot,j}), \quad \hat{\boldsymbol{\beta}}_j = \hat{\beta}_j^\star + \mathcal{I}_j(\hat{\beta}_j^\star)^{-1} S_j(\hat{\beta}_j^\star)$$

  Here, $\hat{\beta}_j^\star$ is a consistent, mean-based estimator, $\mathcal{I}_j$ represents the Fisher Information, and S_j the score function for the jth marginal.

  - $\hat{\phi}_j = \arg\max_\phi \log \mathcal{L}_j(\hat{\boldsymbol{\beta}}_j, \phi; y_{1,j}, \ldots, y_{n,j})$
  - Determine $\hat{\theta}$ by solving:

$$\frac{\partial \log \mathcal{L}}{\partial \theta}(\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_s, \theta) = 0.$$

▷ The OSCFE-IFM approach $(\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_s, \hat{\theta})$ ensures consistency, asymptotic Gaussian behavior, and equivalence to the standard IFM.

Monte-Carlo simulations

100 simulations of the gamma-GLM model with single effects only,
2 response variables, 15 parameters to estimate, $n = 10^5$

Spearman ρ	Copula type	Theo. θ	Mean $\hat{\theta}$		Sd $\hat{\theta}$	
			IFM	**OSCFE-IFM**	**IFM**	**OSCFE-IFM**
0.4	Clayton	0.758	0.758	0.758	0.007	0.007
	Frank	2.610	2.613	2.613	0.021	0.021
	Gumbel	1.382	1.382	1.382	0.004	0.004
	Normal	0.416	0.416	0.416	0.002	0.002
0.8	Clayton	3.188	3.187	3.187	0.018	0.018
	Frank	7.902	7.901	7.902	0.033	0.033
	Gumbel	2.582	2.582	2.582	0.009	0.009
	Normal	0.814	0.813	0.813	0.001	0.001

Monte-Carlo simulations

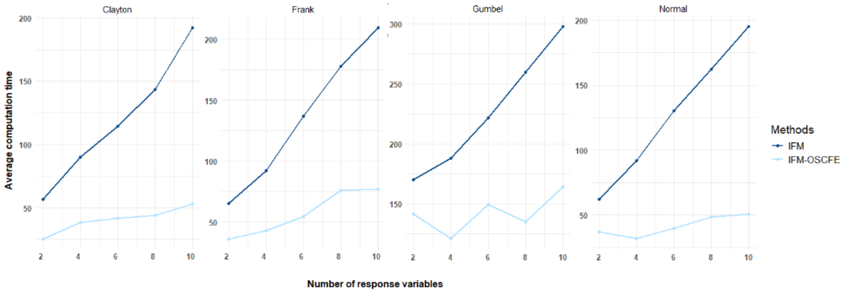


Fig. 1: Copula parameter θ average computation time (sec.) for 4 copula types, $\rho = 0.8$, 100 simulations, 2 explanatory variables with 20 modalities and $n = 10^5$ observations for $s = 2$ to 10 response variables.

Fisher-Scoring algorithms are time-consuming, so

- in case of univariate GLMs
  - CFE is faster to be computed but not efficient
  - OS-CFE is asymptotically efficient as well as fast estimator

Fisher-Scoring algorithms are time-consuming, so

- in case of univariate GLMs
  - CFE is faster to be computed but not efficient
  - OS-CFE is asymptotically efficient as well as fast estimator
- in case of multivariate GLMS:
  - IFM is a consistent estimator but remains time-consuming (Brouste et al. 2023)
  - The OSCFE-IFM approach is consistent, with marginal estimations that are closed-form and asymptotically efficient. On simulated data, the OSCFE-IFM solution closely matches the IFM while significantly reducing computation times.

Thanks!