One-step closed-form estimator for generalized linear model with categorical explanatory variables

Alexandre Brouste, Christophe Dutang, Lilit Hovsepyan, Tom Rohmer

One-step closed-form estimator for generalized linear model with categorical explanatory variables

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

[1] *Laboratoire Manceau de Mathématiques, Le Mans Université, F-72000 Le Mans*
[2] *Université Grenoble Alpes, CNRS, Grenoble INP, LJK, F-38000 Grenoble*
[3] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet Tolosan*

Résumé. L'article propose une procédure d'estimation rapide et asymptotiquement efficace des paramètres des modèles linéaires généralisés à variables catégorielles. Des estimateurs explicites existent [2] pour ces modèles mais ils ne sont pas toujours asympotiquement efficaces, notamment pour les modèles à effets simples. La procédure proposée dans cet article est basée sur une approche one-step où une unique étape de la descente de gradient est effectuée sur la fonction de log-vraisemblance initialisée à partir des estimateurs explicites. Ce travail présente de manière succincte les résultats théoriques obtenus, les simulations effectuées et une application à la tarification en assurance automobile.

Mots-clés. modèles linéaires généralisés, procédure one-step, assurance.

Abstract. The article proposes a quick and asymptotically efficient estimation procedure for the parameters of generalized linear models with categorical variables. The article proposes a procedure for rapid and asymptotically efficient estimation of the parameters of generalized linear models with categorical variables. Explicit estimators exist for these models [2] but they are not always asymptotically efficient, especially for models with simple effects. The procedure proposed in this article is based on a one-step approach where a single gradient descent step is performed on the initial log-likelihood function initialized with the explicit estimators. This work succinctly presents the theoretical results obtained, the simulations performed, and an application to automobile insurance pricing.

Keywords. generalized linear models, one-step procedure, insurance.

1 Notations for GLMs with categorical variables

The observation sample $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ is composed of independent random variables valued in $\mathbb{Y} \subset \mathbb{R}$ where for $i \in I$, Y_i belongs to the one-parameter exponential family of probability measures valued in $\Lambda \subset \mathbb{R}$. In this setting, the log-likelihood $\log \mathcal{L}$ of the sample is

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

where $a : \mathbb{R} \to \mathbb{R}$, $b : \Lambda \to \mathbb{R}$ and $c : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}$ are fixed real-valued measurable functions and ϕ is the dispersion parameter. Here, $\lambda_1, \lambda_2, \ldots, \lambda_n$ are related to exogenous explanatory variables $\boldsymbol{x}_i$, $i = 1, \ldots, n$ by the relations

$$g(\mathbf{E}_{\boldsymbol{\vartheta}} Y_i) = \boldsymbol{x}_i^T \boldsymbol{\vartheta} =: \eta_i, \quad \text{for all } \boldsymbol{\vartheta} \in \Theta, \quad i = 1, \ldots, n, \tag{2}$$

or, equivalently,

$$b'(\lambda_i) = g^{-1}(\eta_i), \quad i = 1, \ldots, n.$$

The parameter $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^p$ and ϕ in (1) are to be estimated. The function g is called the link function in the regression framework.

Let us consider m categorical explanatory variables $(z_i^{(j)})_{i=1,\ldots,n}$, $j = 2, \ldots, m+1$, that take values in a finite set $\{v_{j,1}, \ldots, v_{j,d_j}\}$. Assuming values are unordered, we encode our explanatory variables using binary dummies as

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

These binary dummies can be used both in single-effect models or in cross-effect models. The equation (2) can be generally reconstructed in the following way:

$$\begin{aligned}
g\left(\mathbf{E}_{\boldsymbol{\vartheta}} Y_i\right) =& \vartheta_1 + \sum_{j=2}^{m+1} \sum_{k=1}^{d_j} x_i^{(j),k} \vartheta_k^{(j)} && \text{Intercept and single effect} \\
&+ \sum_{j_2<j_3} \sum_{k_2,k_3} x_i^{(j_2),k_2} x_i^{(j_3),k_3} \vartheta_{k_2,k_3}^{(j_2,j_3)} && \text{Double effect} \\
&+ \sum_{j_2<j_3<j_4} \sum_{k_2,k_3,k_4} x_i^{(j_2),k_2} x_i^{(j_3),k_3} x_i^{(j_4),k_4} \vartheta_{k_2,k_3,k_4}^{(j_2,j_3,j_4)} && \text{Triple effect} \\
&+ \ldots \\
&+ \sum_{k_2,\ldots,k_{m+1}} x_i^{(2),k_2} \ldots x_i^{(m+1),k_{m+1}} \vartheta_{k_2,\ldots,k_{m+1}}^{(2,\ldots,m+1)}, && \text{All crossed effect}
\end{aligned} \tag{3}$$

where indexes j_i are in $\{2, \ldots, m+1\}$ and k_j are in $\{1, \ldots, d_j\}$ for $j = 2, \ldots, m+1$.

In the setting of categorical explanatory variables, the model (3) is not identifiable and linear constraints have to be imposed. For this reason, we consider a restricted parameter $\tilde{\boldsymbol{\vartheta}}$ which is new from our previous studies [1, 2].

The MLE $(\widehat{\tilde{\boldsymbol{\vartheta}}}_n, \widehat{\phi}_n)$ for $(\tilde{\boldsymbol{\vartheta}}, \phi)$ satisfies

$$(\widehat{\tilde{\boldsymbol{\vartheta}}}_n, \widehat{\phi}_n) = \arg \max_{(\tilde{\boldsymbol{\vartheta}}, \phi) \in \tilde{\Theta} \times \mathbb{R}_*^+} \log \boldsymbol{\mathcal{L}}(\tilde{\boldsymbol{\vartheta}}, \phi \,|\, \boldsymbol{Y}). \tag{4}$$

It is worth mentioning that the MLE is asymptotically efficient generally but not in a closed form and its computation can be time-consuming for large samples or numerous explanatory variables.

2 One-step closed-form estimator

2.1 Closed-form Estimator

The closed-form estimator (CFE) of the restricted parameter is defined in [1, 2] by

$$\tilde{\boldsymbol{\vartheta}}_n^{CFE} = (\tilde{Q}^T\tilde{Q})^{-1}\tilde{Q}^T g(\overline{\boldsymbol{Y}}_{\boldsymbol{n}}).$$

where

$$g(\overline{\boldsymbol{Y}}_{\boldsymbol{n}}) = \begin{pmatrix} g(\overline{Y}^1{}_n) \\ \vdots \\ g(\overline{Y}^k{}_n) \\ \vdots \\ g(\overline{Y}^d{}_n) \end{pmatrix}, \quad \overline{Y}^k{}_n = \frac{\sum\limits_{i=1;\eta_i=h_k}^{n} Y_i}{m_k}, \quad m_k = \#\{i \in \{1,\dots,n\}; \eta_i = h_k\} \tag{5}$$

where h_k are the different possible values of the linear predictors defined in (2). Here, the matrix $\widetilde{Q}$ is related to the identifiability condition and to the structure of the model (3).

It is known that the CFE is asymptotically normal, namely

$$\sqrt{n}(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \tilde{\boldsymbol{\vartheta}}) \xrightarrow[n\to+\infty]{L} \mathcal{N}_{p^\star}\left(\mathbf{0}_{p^\star}, a\,(\phi)\,(\tilde{Q}^T\tilde{Q})^{-1}\tilde{Q}^T\Sigma^{-1}(\tilde{\boldsymbol{\vartheta}})\tilde{Q}(\tilde{Q}^T\tilde{Q})^{-1}\right) \tag{6}$$

where $\Sigma^{-1}(\boldsymbol{\vartheta})$ is defined in [2].

In general, for instance in single effect models, the CFE is not the MLE and is not asymptotically efficient. Hence, we consider a one-step version of $\tilde{\boldsymbol{\vartheta}}_n^{CFE}$ in the next subsection.

2.2 One-Step Closed-form Estimator

The One-Step Closed-form Estimator (OS-CFE) of $\tilde{\boldsymbol{\vartheta}}$ is defined as

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

where $\tilde{S}$ stands for the score $\tilde{S}(\tilde{\boldsymbol{\vartheta}}) = \nabla_{\tilde{\boldsymbol{\vartheta}}}\mathcal{L}((\tilde{\boldsymbol{\vartheta}}, \phi)\,|\,\boldsymbol{Y})$ and $\tilde{\mathcal{I}}_n$ stands for the Fisher Information matrix $\tilde{\mathcal{I}}_n(\tilde{\boldsymbol{\vartheta}}) = -\mathbf{E}\left(\nabla^2_{\tilde{\boldsymbol{\vartheta}}}\mathcal{L}(\tilde{\boldsymbol{\vartheta}}, \phi\,|\,\boldsymbol{Y})\right)$ which can be described in terms of the matrix $\tilde{Q}$.

It is worth emphasizing that the OS-CFE of $\tilde{\boldsymbol{\vartheta}}$ does not depend on the dispersion parameter ϕ by simplification. The main result is that the OS-CFE of the restricted parameter $\tilde{\boldsymbol{\vartheta}}$ is asymptotically equivalent to the MLE.

Theorem 1. *Under the regularity conditions, as soon as for all $j = 1,\dots,d$ the frequencies $\frac{m_j}{n} \to p_j$ as $n \to \infty$,*

$$\sqrt{n}(\tilde{\boldsymbol{\vartheta}}_n^{OS-CFE} - \hat{\tilde{\boldsymbol{\vartheta}}}_n) \xrightarrow[n\to+\infty]{P} 0.$$

It also means that the OS-CFE is asymptotically normal with an optimal asymptotic variance.

3 Monte Carlo illustrations

The performances on finite size samples of the aforementioned estimators (MLE, CFE, OS-CFE), in terms of computation times and asymptotic variances, are assessed with numerical examples. Monte Carlo simulations of samples for Poisson and Gamma GLMs are done.

In this numerical example the canonical setting is used (ℓ is the identity function) leading to a log link function for the Poisson distribution ($g(x) = \log(x)$) and the inverse link function for the Gamma distribution ($g(x) = 1/x$).

The sequence of OS-CFE naturally overcomes the performance of CFE in terms of asymptotic variance (see Figures 1 and 2). According to the other comparison in terms of computation time which is highlighted in Table 1, OS-CFE is almost 50 times faster than MLE to be computed for the dataset with the size of $n = 10^4$.

Computation time	MLE	CFE	OS-CFE
Poisson	848.07	9.05	17.73
Gamma	1601.44	10.65	31.61

Table 1: Total computation time (s) for Poisson and Gamma distributions

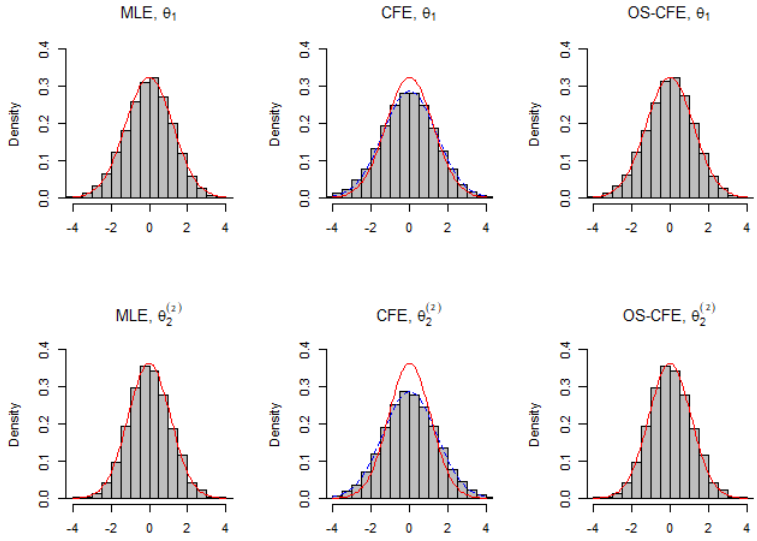

Figure 1: Histograms for the $B = 10^4$ simulations of the renormalized statistical errors of MLE, CFE, OS-CFE for the Poisson distribution with 2 categorical variables with $d_2 = 2$, $d_3 = 3$ for ϑ_1 and $\vartheta_2^{(2)}$. Red and blue lines are the theoretical Gaussian asymptotic densities respectively of the MLE (in red) and CFE (in blue).

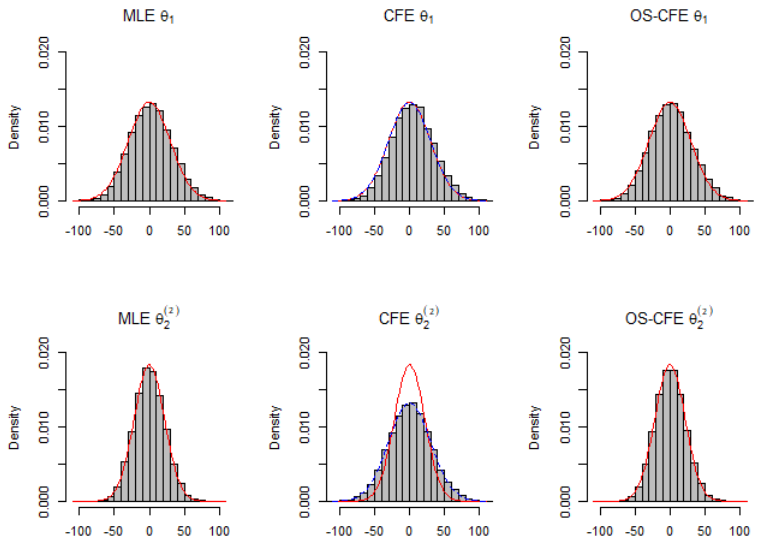

Figure 2: Histograms for the $B = 10^4$ Monte Carlo simulations of the renormalized statistical errors of MLE, CFE, OS-CFE for the Gamma distribution (canonical link) with 2 categorical variables with $d_2 = 2$, $d_3 = 3$ for ϑ_1 and $\vartheta_2^{(2)}$ and fixed $\phi = 8$. Red and blue lines are the theoretical Gaussian asymptotic densities respectively of the MLE (in red) and CFE (in blue).

4 Application to claim amounts in car insurance

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.

For confidentiality reasons, the modality values are not revealed.

The single effect models

$$g\left(\mathbf{E}_{\boldsymbol{\vartheta}} Y_i\right) = \vartheta_1 + \sum_{j=2}^{m+1} \sum_{k=1}^{d_j} x_i^{(j),k} \vartheta_k^{(j)},$$

is generally used by the insurers to model the claim amounts with (non-canonical) Gamma GLMs with a log link function ($g(x) = \log(x)$). The "reference category" linear contrast has been used. It is worth recalling that in this setting the MLE has no closed form and the

closed-form estimator is not efficient. In order to compute the log-likelihood, we use Equation (4) to fit the dispersion parameter.

The one-step estimator has been applied to the Covea dataset. It is almost 30 times faster than the MLE with similar estimation.

	CFE	OSCFE	MLE
ϑ_1	6.23	6.04	6.03
$\vartheta_2^{(2)}$	0.24	0.08	0.03
$\vartheta_2^{(3)}$	0.18	0.22	0.22
$\vartheta_3^{(3)}$	-0.48	0.04	-0.01
$\vartheta_4^{(3)}$	-0.07	0.08	0.09
$\vartheta_5^{(3)}$	0.06	0.18	0.19
$\vartheta_6^{(3)}$	0.20	0.21	0.22
$\vartheta_2^{(4)}$	-0.07	0.00	0.01
$\vartheta_3^{(4)}$	0.06	0.16	0.16
$\vartheta_4^{(4)}$	0.17	0.18	0.18
$\vartheta_5^{(4)}$	0.34	0.41	0.40
$\vartheta_6^{(4)}$	0.11	0.44	0.42
$\vartheta_7^{(4)}$	0.16	0.25	0.26
$\vartheta_8^{(4)}$	-0.01	0.34	0.33
$\log \mathcal{L}$	-554,868	-553,708	-553,685
Time (s)	0.01	0.01	0.30

Table 2: Values of $\widehat{\boldsymbol{\vartheta}}_n$, log-likelihood and total computation time (s) for CFE, OSCFE and MLE

5 Conclusion

Generalized linear models with single effects and single categorical explanatory variables are widely used in different applications (insurance, agriculture, biology). The classical iterative re-weighted least square calibration method is asymptotically efficient, but can be time consuming for large datasets. On the other hand, closed-form estimators proposed in [2] are faster to compute, but they are not asymptotically efficient.

In this paper we proposed fast and asymptotically efficient method for the calibration of GLMs with categorical explanatory variables based on the one-step procedure that can also be applied to single effect models. It is 30 times faster on the simulated and the Covea Affinity datasets compared with the classical methods.

Acknowledgments We would like to thank Héloïse Bertrand and Nicolas Villette for fruitful discussions on the topic. This research benefited from the support of the ANR project 'Effi-

cient inference for large and high-frequency data' (ANR-21-CE40-0021), the 'Chair Risques Émergents ou Atypiques en Assurance', under the aegis of Fondation du Risque, a joint initiative by Le Mans Université and MMA company, member of Covea group and the 'Chair Impact de la Transition Climatique en Assurance', under the aegis of Fondation du Risque, a joint initiative by Le Mans Université and Groupama Centre-Manche company, member of Groupama group.

References

[1] 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(2):689–724, 2020.

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