A Closed-form Alternative Estimator for GLM with Categorical Explanatory Variables

Alexandre Brouste, Christophe Dutang, Tom Rohmer

A Closed-form Alternative Estimator for GLM with Categorical Explanatory Variables

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

[1]LMM, Le Mans Université, Avenue Olivier Messiaen, F-72085 Le Mans, ORCID: 0000-0001-6719-7432
[2]CEREMADE, CNRS, Université Paris-Dauphine, Université PSL, Place du Maréchal de Lattre de Tassigny, F-75016 Paris, ORCID: 0000-0001-6732-1501
[3]GenPhySE, Université de Toulouse, INRAE, ENVT, 24 chemin de borde rouge, F-31326 Castanet Tolosan, ORCID: 0000-0002-4751-2324

June 7, 2022

Abstract

The parameters of generalized linear models (GLMs) are usually estimated by the maximum likelihood estimator (MLE) which is known to be asymptotically efficient. But the MLE is computed using a Newton-Raphson-type algorithm which is time-consuming for a large number of variables or modalities, or a large sample size. An alternative closed-form estimator is proposed in this paper in the case of categorical explanatory variables. Asymptotic properties of the alternative estimator is studied. The performances in terms of both computation time and asymptotic variance of the proposed estimator are compared with the MLE for a Gamma distributed GLM.

Keywords: Regression models; explicit estimators; categorical explanatory variables; GLM ; asymptotic distribution

1 Introduction

Generalized linear models (GLMs) deal with regression models such that the distribution of the response variable belongs to the exponential family whose natural parameter is, up to a link function, a linear combination of explanatory variables. It is worth recalling that the distributions of exponential type include most of the classical discrete distributions (binomial, Poisson, *etc.*) and continuous distributions (Gaussian, gamma, inverse Gaussian, etc.), see, e.g., McCullagh & Nelder (1989).

The unknown parameters of GLMs are generally estimated by the maximum likelihood estimator (MLE). The asymptotic normality and efficiency of the MLE for the GLMs were studied in Fahrmeir & Kaufmann (1985). In a general framework, the MLE has no closed-form and is numerically computed by a gradient descent-type method known as the Fisher scoring which can be equivalently written as an iteratively re-weighted least square method (IWLS), see for example McCullagh & Nelder (1989, Chapter 2). This computation method is time-consuming for large datasets or for high dimensions.

The case of categorical explanatory variables is singular due to the non-identifiability of the model. Then, aforementioned results do not apply directly. Generally, linear identifiability conditions are imposed via a contrast matrix. In this setting, closed-form MLE have been exhibited in Brouste et al. (2020) for any distribution of exponential type under a supplementary assumption on the contrast matrix. Due to the closed-form formula, the parameters are quickly estimated and the methodology can handle large datasets and/or large number of modalities.

The alternative estimator proposed in this paper is motivated by theoretical arguments as well as practical motivations. Indeed, the MLE does not have an explicit solution in the case of two or more explanatory variables only used as single effects. Our proposed closed-form estimator makes possible to deal with large datasets or a large number of explanatory variables avoiding using the time-consuming IWLS algorithm.

Dealing with only categorical explanatory variables is of particular interest. For instance, in the insurance industry, policy pricing uses a finite number of risk group relying on categorical explanatory variables. Typically, motor insurance ratings rely on vehicle classification with a large number of modalities, e.g., the dataset `pg17trainpol` in **CASdatasets** by Dutang & Charpentier (2020), exhibits this feature with 1023 vehicle models for 101 vehicle brands. Another appealing example is Kadarmideen et al. (2000) where authors study disease, fertility and milk production in dairy cattle and consider 7530 levels for Herd-year-seasons effect. In both situation, having a fast efficient estimator is at stake.

In most situations, explanatory variables are used as single effect. Indeed, McCullagh & Nelder (1989) use single-effect models with the binomial distribution (Chapter 4) and with the Poisson distribution (Chapter 6); Lindsey (1997) also uses single-effect models for the Bernoulli distribution in Chapter 2. This is also particularly true in the actuarial field, where Denuit et al. (2020, Chap. 4) model claim count distributions, and Wuethrich & Merz (2021, Chap. 5) model claim size distributions with single-effect models.

Whatever the domain of application, finding closed-form solutions is at stake. In the literature of choice modeling, Lipovetsky & Conklin (2014) provide analytical formulae for multinomial logit model which permits the inference of the characteristics of the model's quality (standard errors of the utilities, choice probabilities, the residual deviance and pseudo-R-square), see also Lipovetsky (2015) for the special case of logit regression. Marley et al. (2016) show that Lipovetsky & Conklin (2014)'s analytical closed-form solutions and Frischknecht et al. (2014)'s normalized best-worst scores provides better fits to the aggregate choices in several best-worst choice data sets.

Furthermore, Lipovetsky et al. (2015) use Lipovetsky & Conklin (2014) to derive analytical formulas for determining the needed sample size for Best-Worst Scaling studies, which allows to test the formula assumption and demonstrate the soundness of the formula to be used and proposed rough empirical rule for determining the sample size for Best-Worst Scaling.

When building decision trees, having a closed-form estimator is also of particular interest. In that respect, Dutang & Guibert (2021) propose a fast estimation procedure based on Brouste et al. (2020) to fit GLM trees, which also makes possible to fit GLM forests on practical applications.

The paper is structured as follows. Notations are introduced in Section 2. In Section 3.1, the general setting and the asymptotic properties of the alternative estimator are presented. In Section 3.2, the case of two categorical explanatory variables (and also one explanatory variable) is described with several examples. We also focus on the partic-

ular case of single effect only. A simulation analysis is performed in Section 4 in order to benchmark the proposed estimator against the IWLS algorithm in terms of computation time and to compare the asymptotic variances on a Gamma distributed GLM with single effect only.

2 Notation

In the following, vectors of $\mathbb{R}^p$ or $\mathbb{R}^n$ are bolded, while the index i is reserved for the observations, while the indexes j, k, l are used for the explanatory variables.

In the GLM setting, the sample $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ is composed of independent random variables valued in $\mathbb{Y} \subset \mathbb{R}$. Each response Y_i belongs to a family of probability measures of one-parameter exponential type with respective parameters $\lambda_1, \dots, \lambda_n$ valued in $\Lambda \subset \mathbb{R}$.

That is, the log-likelihood $\mathcal{L}$ associated to the statistical experiment is

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

where $a : \mathbb{R} \to \mathbb{R}$, $b : \Lambda \to \mathbb{R}$ and $c : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}$ are known real-valued measurable functions and ϕ is the dispersion parameter, e.g., McCullagh & Nelder (1989, Section 2.2).

In the GLM setting, the parameters $\lambda_1, \dots, \lambda_n$ of Equation (1) depend on a finite-dimensional parameter $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^m$. A direct computation of theoretical moments leads to

$$b'(\lambda_i(\boldsymbol{\vartheta})) = \mathbf{E}_{\boldsymbol{\vartheta}} Y_i \quad \text{and} \quad b''(\lambda_i(\boldsymbol{\vartheta}))a(\phi) = \mathbf{Var}_{\boldsymbol{\vartheta}} Y_i.$$

for $i = 1, \dots, n$.

In the following, we consider deterministic exogenous variables $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, with $\boldsymbol{x}_i = (x_{i,1}, \dots, x_{i,p}) \in \mathbb{R}^p$ for $i = 1, \dots, n$, representing for categorical variable encoding. That is, $x_{i,j}$ is binary, typically indicating a label for a categorical variable (see Section 3 for the proper encoding for categorical explanatory variables).

Let g be a twice continuously differentiable and bijective function g from $b'(\Lambda)$ to $\mathbb{R}$. GLMs are defined by assuming the following relation between the expectation $\mathbf{E}_{\boldsymbol{\vartheta}} Y_i$ and the predictor

$$g(b'(\lambda_i(\boldsymbol{\vartheta}))) = \boldsymbol{x}_i^T \boldsymbol{\vartheta} = \eta_{\boldsymbol{x}_i}, \quad \text{for all } \boldsymbol{\vartheta} \in \Theta, \tag{2}$$

where $\eta_{\boldsymbol{x}_i}$ are the linear predictors. The parameter $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^p$ is unknown and has to be estimated and g is the so-called link function. Consequently, the (bijective) function $\ell = (b')^{-1} \circ g^{-1}$ maps linear predictors to parameters as $\lambda_i(\boldsymbol{\vartheta}) = \ell(\eta_{\boldsymbol{x}_i})$ which can be summarized as

$$X \times \Theta \xrightarrow{\langle \cdot, \cdot \rangle} D \underset{\ell}{\overset{\ell^{-1}}{\leftrightarrows}} \Lambda,$$

where D is the space of linear predictor and X the possible set of value of $\boldsymbol{x}_i$ for $i \in \{1, \dots, n\}$. In the special case $\ell(t) = t$, we talk of a canonical link function.

Consecutively, the log-likelihood (1) can be rewritten as

$$\log \mathcal{L}(\boldsymbol{\vartheta} \mid \boldsymbol{Y}, \boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \sum_{i=1}^{n} \frac{Y_i \ell(\eta_{\boldsymbol{x}_i}) - b\left(\ell(\eta_{\boldsymbol{x}_i})\right)}{a(\phi)} + \sum_{i=1}^{n} c(Y_i, \phi). \tag{3}$$

Since the covariates are supposed to be categorical in this paper, the linear predictors $\eta_{\boldsymbol{x}_i}$ takes d distinct values namely $\epsilon_1, \dots, \epsilon_d$. We consider the unique $d \times p$ matrix Q such that

$$\boldsymbol{\eta} = Q\boldsymbol{\vartheta}, \quad \boldsymbol{\eta} = (\epsilon_j)_{j=1,\dots,d}.$$

In the case of categorical explanatory variables, the model is not identifiable and linear identifiability conditions are imposed. Precisely, the MLE solves the optimization problem

$$\widehat{\boldsymbol{\vartheta}}_n = \arg \max_{\boldsymbol{\vartheta} \in \Theta | R\boldsymbol{\vartheta}=0} \mathcal{L}(\boldsymbol{\vartheta} \,|\, \boldsymbol{Y}), \tag{4}$$

where R is a contrast matrix that ensures the model to be identifiable. The identifiability condition is equivalent to the positive definiteness of the matrix Q^TQ+R^TR (see Appendix A.1). As mentioned in the introduction, the computation of the MLE is time consuming when it is not explicit.

3 General setting

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

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

These binary dummies can be used both in single-effect models or with cross-effect models. To take all possible GLM settings into account, we consider a GLM with predictor defined as

$$\begin{aligned}
g\left(\mathbf{E}_{\boldsymbol{\vartheta}} Y_i\right) =& \vartheta^{(1)} + \sum_{j=2}^{m+1} \sum_{k=1}^{d_j} x_i^{(j),k} \vartheta_k^{(j)} && \text{Intercept and single effect} \\
&+ \sum_{j_2<j_3} \sum_{k_2,k_3} x_i^{(j_2),k_2} x_i^{(j_3),k_3} \vartheta_{k_2,k_3}^{(j_2,j_3)} && \text{Double effect} \\
&+ \sum_{j_2<j_3<j_4} \sum_{k_2,k_3,k_4} x_i^{(j_2),k_2} x_i^{(j_3),k_3} x_i^{(j_4),k_4} \vartheta_{k_2,k_3,k_4}^{(j_2,j_3,j_4)} && \text{Triple effect} \\
&+ \dots \\
&+ \sum_{k_2,\dots,k_{m+1}} x_i^{(2),k_2} \dots x_i^{(m+1),k_{m+1}} \vartheta_{k_2,\dots,k_{m+1}}^{(2,\dots,m+1)}, && \text{All crossed effect}
\end{aligned} \tag{5}$$

where g is the link function and indexes $j_i \in \{2, \dots, m+1\}$ and $k_j \in \{1, \dots, d_j\}$ for $j = 2, \dots, m+1$.

In the m-variables case, the unknown parameter vector is

$$\boldsymbol{\vartheta} = \left(\vartheta^{(1)}, (\vartheta_k^{(j)})_{k,j}, (\vartheta_{k_2,k_3}^{(j_2,j_3)})_{k_2,k_3,j_2<j_3}, (\vartheta_{k_2,k_3,k_4}^{(j_2,j_3,j_4)})_{k_2,k_3,k_4,j_2<j_3<j_4}, \dots, (\vartheta_{k_2,\dots,k_{m+1}}^{(2,\dots,m+1)})_{k_2,\dots,k_{m+1}}\right)^T. \tag{6}$$

We introduce a specific notation based on Kronecker products $\otimes$ of ones vectors $\mathbf{1}$ and identity matrices I to denote cross effects of $j_1 < \dots < j_k$ variables

$$M_m^{(j_1,\dots,j_k)} = \mathbf{1}_{d_{m+1}\dots d_{j_k+1}} \otimes I_{d_{j_k}} \otimes \mathbf{1}_{d_{j_k-1}\dots d_{j_{k-1}}} \otimes I_{d_{j_{k-1}}} \otimes \dots \otimes I_{d_{j_1}} \otimes \mathbf{1}_{d_{j_1-1}\dots d_2}, \tag{7}$$

where $j_i \in \{2, \dots, m+1\}$ with the convention that $\mathbf{1}_0 = I_0 = 1$. When j_i, j_{i+1} are consecutive integers, i.e., $j_{i+1} = j_i + 1$, ones vectors disappear and simplifications occur $I_{d_{j_i}} \otimes I_{d_{j_{i+1}}} = I_{d_{j_i} d_{j_{i+1}}}$. There are two special cases: $M_m^{()} = \mathbf{1}_{d_{m+1} \dots d_2}$ is the ones vector and $M_m^{(2,\dots,m+1)} = I_{d_{m+1} d_3 d_2}$ is the identity matrix.

Using (7), we define the Q matrix as

$$
\begin{aligned}
Q = \big(& M_m^{()}, && \text{Intercept} \\
& M_m^{(2)}, \dots, M_m^{(d+1)}, && \text{Single effect} \\
& M_m^{(2,3)}, \dots, M_m^{(d,d+1)}, && \text{Double effect} \\
& M_m^{(2,3,4)}, \dots, M_m^{(d-1,d,d+1)}, && \text{Triple effect} \\
& \dots \\
& M_m^{(2,\dots,m+1)} \big). && \text{All crossed effect}
\end{aligned}
\tag{8}
$$

In other words, Q contains combinations of Kronecker products of ones-vector and identity matrix through matrices $M_m^{(.)}$ (7), see also Sunwoo (1996) which use Kronecker products for linear models. The total number of parameter is

$$p = 1 + \sum_{j=2}^{m+1} d_j + \sum_{j_2 < j_3} d_{j_2} d_{j_3} + \dots + d_2 \dots d_{m+1}.$$

Table 1 gives examples of such Q matrix for 1, 2 and 3 explanatory variables, whereas Table 8 in Appendix A.3 gives examples of $M_m^{()}$ and Q for 2, 3 and 4 explanatory variables.

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

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

Using the binary structures of dummies $x_i^{(j),k_j}$, i.e., for all $i = 1, \dots, n$,

$$\sum_{k_2,\dots,k_{m+1}} x_i^{(2),k_2} = \dots = \sum_{k_2,\dots,k_{m+1}} x_i^{(m+1),k_{m+1}} = 1,$$

we identify the (unique) m-uple $(k_2, \dots, k_{m+1})$ for the ith observation. Hence, the linear predictor η_{x_i} of Equation (5) simplifies in the following way

$$g\left(\mathbf{E}_{\boldsymbol{\vartheta}} Y_i\right) = \eta_{x_i} = \vartheta_1 + \sum_j \vartheta_{k_j}^{(j)} + \sum_{j_2 < j_3} \vartheta_{k_2,k_3}^{(j_2,j_3)} + \dots + \vartheta_{k_2,\dots,k_{m+1}}^{(2,\dots,m+1)} := \eta_{k_2,\dots,k_{m+1}}, \tag{9}$$

such that the ith observation belongs to the k_jth modality of the jth variable for $j = 2, \dots, m+1$.

Naturally, there are 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 $R\boldsymbol{\vartheta} = 0$. We assume, in the following, that the matrix R is such that $Q^TQ + R^TR$ is definite-positive and $\text{rank}(R) = q$. This contrast matrix also allows to built every sub-models of a GLM. In particular, the single effect case (no interaction) can be considered. In the R statistical software (R Core Team 2021), `glm` removes the first modality of each variable and the associated interaction terms (that corresponds to the second example of Table 3 below for two categorical explanatory variables).

To present our alternative estimator, using Equation (9), we introduce the absolute frequencies over the m explanatory variables by

$$m_{k_2,\ldots,k_{m+1}} = \sum_{i=1}^{n} x_i^{(2),k_2} \times \cdots \times x_i^{(m+1),k_{m+1}} = \#\{i \in \{1,\ldots,n\}; \eta_{\boldsymbol{x}_i} = \eta_{k_2,\ldots,k_{m+1}}\}, \quad (10)$$

and the cross-effect average responses by

$$\overline{\boldsymbol{Y}} = \begin{pmatrix} \overline{Y}_n^{1,\ldots,1} \\ \vdots \\ \overline{Y}_n^{k_2,\ldots,k_{m+1}} \\ \vdots \\ \overline{Y}_n^{d_2,\ldots,d_{m+1}} \end{pmatrix}, \overline{Y}_n^{k_2,\ldots,k_{m+1}} = \frac{\sum_{i=1}^{n} Y_i x_i^{(2),k_2} \ldots x_i^{(m+1),k_{m+1}}}{m_{k_2,\ldots,k_{m+1}}} = \frac{\sum_{i=1;\eta_{\boldsymbol{x}_i}=\eta_{k_2,\ldots,k_{m+1}}}^{n} Y_i}{m_{k_2,\ldots,k_{m+1}}}. \quad (11)$$

For the sake of clarity, we consider $m_{k_2,\ldots,k_{m+1}} > 0$. Nevertheless the results can be adapted to the case where $m_{k_2,\ldots,k_{m+1}} \geq 0$ in the same way of Brouste et al. (2020). In this paper, we propose the closed-form estimator defined by

$$\widetilde{\boldsymbol{\vartheta}}_n = (Q^TQ + R^TR)^{-1}Q^Tg(\overline{\boldsymbol{Y}}), \quad (12)$$

where Q is defined in (8), $g(\overline{\boldsymbol{Y}}) \in \mathbb{R}^p$ is the vector of g-transformation of average responses (11). Below, we exhibit the asymptotic properties of the alternative estimator $\widetilde{\boldsymbol{\vartheta}}_n$. It can be quickly computed by solving a linear system and can therefore be used to handle large datasets (see Section 4). We also investigate situations for which the alternative estimator is the MLE, i.e., $\widetilde{\boldsymbol{\vartheta}}_n = \widehat{\boldsymbol{\vartheta}}_n$.

3.1 Main results

Before stating theorems, using Equation (9), we introduce the vector of theoretical expectations

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{1,\ldots,1} \\ \vdots \\ \mu_{k_2,\ldots,k_{m+1}} \\ \vdots \\ \mu_{d_2,\ldots,d_{m+1}} \end{pmatrix}, \quad \mu_{k_2,\ldots,k_{m+1}} = g^{-1}(\eta_{k_2,\ldots,k_{m+1}}). \quad (13)$$

We also define the theoretical probabilities $p_{k_2,\ldots,k_{m+1}}$ as

$$\frac{m_{k_2,\ldots,k_{m+1}}}{n} \underset{n\to+\infty}{\longrightarrow} p_{k_2,\ldots,k_{m+1}} \in (0,1). \quad (14)$$

Theorem 1 gives the asymptotic distribution of $\widetilde{\boldsymbol{\vartheta}}_n$, whereas Theorem 2 gives sufficient conditions where the alternative estimator is MLE, i.e., $\widetilde{\boldsymbol{\vartheta}}_n = \widehat{\boldsymbol{\vartheta}}_n$.

Theorem 1. *Regardless of q, the proposed estimator $\widetilde{\boldsymbol{\vartheta}}_n$ is strongly consistent. Furthermore, $\widetilde{\boldsymbol{\vartheta}}_n$ is asymptotically normal, namely,*

$$\sqrt{n}\left(\widetilde{\boldsymbol{\vartheta}}_n - \boldsymbol{\vartheta}\right) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}_p\left(\mathbf{0}_p, Q_R \Sigma Q_R^T\right), \tag{15}$$

with $Q_R = (Q^TQ + R^TR)^{-1}Q^T$ and Σ is the diagonal matrix whose diagonal elements are

$$\sigma^2_{k_2,\ldots,k_{m+1}} = \frac{a(\phi)}{p_{k_2,\ldots,k_{m+1}}} b''((b')^{-1}(\mu_{k_2,\ldots,k_{m+1}}))(g'(\mu_{k_2,\ldots,k_{m+1}}))^2),$$

where $\mu_{k_2,\ldots,k_{m+1}}$ is defined in (13) and $p_{k_2,\ldots,k_{m+1}}$ in (14).

Proof. The proof is postponed in Appendix A.2. □

Remark 1. *It is worth emphasizing that, due to the non-identifiability and the imposed linear constraints, the matrix $Q_R\Sigma Q_R^T$ in (15) is singular.*

Let us define the matrix $R_\Delta = R_{-m} - R^{(2,\ldots,m+1)}Q_{-m}$, with R_{-m} and Q_{-m} the matrices respectively without the $d_2 \ldots d_{m+1}$ last columns of R and Q, and $R^{(2,\ldots,m+1)}$ the matrix with only the $d_2 \ldots d_{m+1}$ last columns of R. Let $q_{\min}$ the minimal number of conditions to get identifiability for GLM (5), that is $q_{\min} = p - \prod_{j=2}^{m+1} d_j$, i.e. $q = rank(R) \geq q_{\min}$.

Theorem 2. *For $q = q_{min}$, the identifiability condition is equivalent to*

$$\det(R_\Delta) \neq 0. \tag{16}$$

In that case, $\widehat{\boldsymbol{\vartheta}}_n = \widetilde{\boldsymbol{\vartheta}}_n$.

Proof. The proof is postponed in Appendix A.3. □

Remark 2. *For $q = q_{min}$, if Condition (16) is satisfied, $\widetilde{\boldsymbol{\vartheta}}_n = \widehat{\boldsymbol{\vartheta}}_n = \begin{pmatrix} Q \\ R \end{pmatrix}^{-1} \begin{pmatrix} g(\overline{\boldsymbol{Y}}) \\ \mathbf{0}_{q_{min}} \end{pmatrix}$. For $m = 1$, $p = 1 + d_2$ so $q_{min} = 1$. Hence the condition $det(R_\Delta) \neq 0$ is $\sum_{j=1}^{d_2} r_j - r_0 \neq 0$ where $R = (r_0, \ldots, r_{d_2})$. This is the condition proposed by Brouste et al. (2020).*

3.2 Two and one categorical explanatory variable(s)

We now focus on the case of two categorical explanatory variables for which the alternative estimator is generally not the MLE. Equation (5) simplifies to

$$g\left(\mathbf{E}_{\boldsymbol{\vartheta}} Y_i\right) = \vartheta^{(1)} + \sum_{k=1}^{d_2} x_i^{(2),k}\vartheta_k^{(2)} + \sum_{l=1}^{d_3} x_i^{(3),l}\vartheta_l^{(3)} + \sum_{k=1}^{d_2}\sum_{l=1}^{d_3} x_i^{(2),k}x_i^{(3),l}\vartheta_{k,l}^{(2,3)}. \tag{17}$$

Using Table 1, the Q matrix is

$$Q = (Q_{-2}, Q^{(2,3)}), \quad Q_{-2} = (\mathbf{1}_{d_2d_3}, \mathbf{1}_{d_3} \otimes I_{d_2}, I_{d_3} \otimes \mathbf{1}_{d_2}), \quad Q^{(2,3)} = I_{d_2d_3}. \tag{18}$$

The contrast matrix $R = (R_{-2}, R^{(2,3)})$ is given by

$$R_{-2} = \begin{pmatrix} r_{0,1} & r_{1,1}^{(2)} & \ldots & r_{d_2,1}^{(2)} & r_{1,1}^{(3)} & \ldots & r_{d_3,1}^{(3)} \\ \vdots & \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ r_{0,q} & r_{1,q}^{(2)} & \ldots & r_{d_2,q}^{(2)} & r_{1,q}^{(3)} & \ldots & r_{d_3,q}^{(3)} \end{pmatrix}, \quad R^{(2,3)} = \begin{pmatrix} r_{11,1} & \ldots & r_{d_2d_3,1} \\ \vdots & \ldots & \vdots \\ r_{11,q} & \ldots & r_{d_2d_3,q} \end{pmatrix}.$$

Here a general row number $q \geq 1 + d_2 + d_3$ is considered and $\text{rank}(R) = q$. The lower bound $q_{min} = 1 + d_2 + d_3$ is the minimal number of conditions to get identifiability for GLM (17). Let $R_\Delta = R_{-2} - Q_{-2}R^{(2,3)}$.

Corollary 1. *For $q = q_{min}$, the identifiability condition is given in Table 2. In that case, $\widehat{\boldsymbol{\vartheta}}_n = \widetilde{\boldsymbol{\vartheta}}_n$.*

Regardless of q, the proposed estimator $\widetilde{\boldsymbol{\vartheta}}_n$ is asymptotically normal with Σ is the diagonal matrix whose diagonal elements are given in Table 2.

Table 2: Parameters and identifiability

m	p	q_{min}	identifiability	theo. expectation	theo. covariance
$m=2$	$1+d_2+d_3$ $+d_2d_3$	$1+d_2+d_3$	$\det R_\Delta \neq 0$	$\mu_{k,l} = g^{-1}(\vartheta^{(1)} + \vartheta_k^{(2)}$ $+\vartheta_l^{(3)} + \vartheta_{k,l}^{(2,3)})$	$\sigma^2_{k,l} = \frac{a(\phi)}{p_{k,l}} b''((b')^{-1}(\mu_{k,l}))$ $\times (g'(\mu_{k,l}))^2$
$m=1$	$1+d_2$	1	$r_{0,1} \neq \sum_{k=1}^{d_2} r_{k,1}^{(2)}$	$\mu_j = g^{-1}(\vartheta^{(1)} + \vartheta_j^{(2)})$	$\sigma_j^2 = \frac{a(\phi)}{p_j} b''((b')^{-1}(\mu_j))(g'(\mu_j))^2$

Below, we exhibit two examples to illustrate the previous result. Example 1 presents the usual contrasts for two explanatory variables and $q = 1 + d_2 + d_3$. Example 2 gives an example for $q > 1 + d_2 + d_3$ where the MLE differs from the closed-form estimator: an over-contrasted matrix R which removes the interactions of the model, then which removes the interactions and the second variable.

Example 1 (Case $q = 1 + d_2 + d_3$ and R_Δ invertible)**.** *For the 3 examples of Table 3, we have rank$(R) = 1 + d_2 + d_3$ and the MLE $\widehat{\boldsymbol{\vartheta}}_n$ is equal to the alternative estimator $\widetilde{\boldsymbol{\vartheta}}_n$. Using Theorem 2.2 of Lu & Shiou (2002), Because $R_\Delta = R_{-2} - Q_{-2}R^{(2,3)}$ is invertible, we have*

$$\begin{pmatrix} Q \\ R \end{pmatrix}^{-1} = \begin{pmatrix} Q_{-2} & I_{d_2 d_3} \\ R_{-2} & R^{(2,3)} \end{pmatrix}^{-1} = \begin{pmatrix} -R_\Delta^{-1} R^{(2,3)} & R_\Delta^{-1} \\ I_{d_2 d_3} + Q_{-2} R_\Delta^{-1} R^{(2,3)} & -Q_{-2} R_\Delta^{-1} \end{pmatrix}.$$

Table 3: Three contrasts for two variables and full rank R matrix

type	*Zero-sum Condition*	*ref. category (1^{st} modality)*	*No intercept, no single-variable dummy*
contrast	$\sum_k \vartheta_k^{(2)} = \sum_l \vartheta_l^{(3)} = 0$ $\forall l, \sum_k \vartheta_{k,l}^{(2,3)} = 0$ $\forall k, \sum_l \vartheta_{k,l}^{(2,3)} = 0$	$\vartheta_1^{(2)} = \vartheta_1^{(3)} = 0$ $\forall l, \vartheta_{1,l}^{(2,3)} = 0$ $\forall k, \vartheta_{k,1}^{(2,3)} = 0$	$\vartheta^{(1)} = 0$ $\forall l, \vartheta_l^{(3)} = 0$ $\forall k, \vartheta_k^{(2)} = 0$
full rank R $d_2 = d_3 = 2$	$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$

Example 2 (Case $q > 1 + d_2 + d_3$, Model without interaction)**.** *Let us consider the case of model* (17) *without interaction. Indeed the contrast matrix is*

$$R = \begin{pmatrix} R_1 & \mathbf{0}_{2 \times 0_{d_2 d_3}} \\ \mathbf{0}_{1+d_2+d_2 \times d_2 d_2} & R_2 \end{pmatrix}, \tag{19}$$

with

$$R_1 = \begin{pmatrix} r_{0,1} & r_{1,1}^{(2)} & \dots & r_{d_2,1}^{(2)} & 0 & \dots & 0 \\ r_{0,2} & 0 & \dots & 0 & r_{1,2}^{(3)} & \dots & r_{d_3,2}^{(3)} \end{pmatrix}, \quad R_2 = I_{d_2 d_3},$$

where $r_{0,1}, r_{1,1}^{(2)}, \ldots, r_{d_2,1}^{(2)}$ *and* $r_{0,2}, r_{1,2}^{(3)}, \ldots, r_{d_3,2}^{(3)}$ *are taken so that* $Q^TQ + R^TR$ *is invertible.*

In this case rank$(R) = 2 + d_2d_3 \geq q_{min} = 1 + d_2 + d_3$ *and* $Q = (Q_{-2}, I)$*, we have*

$$Q^TQ + R^TR = \begin{pmatrix} Q_{-2}^TQ_{-2} + R_1^TR_1 & Q_{-2}^T \\ Q_{-2} & I + R_2^TR_2 \end{pmatrix} = \begin{pmatrix} Q_{-2}^TQ_{-2} + R_1^TR_1 & Q_{-2}^T \\ Q_{-2} & 2I_{d_2d_3} \end{pmatrix},$$

and using B.1

$$Q_R = \begin{pmatrix} (Q_{-2}^TQ_{-2} + R_1^TR_1)^{-1}Q_{-2}^T \\ \frac{1}{2}(I - Q_{-2}(Q_{-2}^TQ_{-2} + R_1^TR_1)^{-1}Q_{-2}^T) \end{pmatrix}.$$

Hence the alternative estimator $\widetilde{\boldsymbol{\vartheta}}_n$ *fails to be MLE because* $R\widetilde{\boldsymbol{\vartheta}}_n \neq 0$*. We perform a simulation analysis in Section 4 to compare asymptotic variances and computation times between the proposed alternative estimator and the asymptotically efficent MLE. An exception is the canonical Gaussian case with balanced plan, i.e.,* $m_{1,1} = \ldots = m_{d_2,d_3}$ *for which the alternative estimator is still the MLE.*

We finish this section with the case of a single categorical explanatory variable. The linear predictor (5) simplifies to an intercept and a single sum whereas the Q matrix is given at the bottom of Table 1. A similar corollary to Corollary 1 can derived where the minimum constraint number is $q = 1$, the identifiability condition and the variance components are given at the bottom of Table 2. In that case, the alternative is the MLE, as demonstrated by Brouste et al. (2020). The formula of matrices used for $\widetilde{\boldsymbol{\vartheta}}_n$ for the three usual contrasts (no intercept, no first-level, zero-sum) are given in Table 4, while the proofs are put in Appendix A.4.

Table 4: Three well known examples of contrasts

name	no-intercept	no first-level	zero-sum
R	$(1, \mathbf{0}_{d_2}^T)$	$(0, 1, \mathbf{0}_{d_2-1}^T)$	$(0, \mathbf{1}_{d_2}^T)$
$Q^TQ + R^TR$	$\begin{pmatrix} d_2+1 & \mathbf{1}_{d_2}^T \\ \mathbf{1}_{d_2} & I_{d_2} \end{pmatrix}$	$\begin{pmatrix} d_2 & 1 & \mathbf{1}_{d_2-1}^T \\ 1 & 2 & \mathbf{0}_{d_2-1}^T \\ \mathbf{1}_{d_2-1} & \mathbf{0}_{d_2-1} & I_{d_2-1} \end{pmatrix}$	$\begin{pmatrix} d_2 & \mathbf{1}_{d_2}^T \\ \mathbf{1}_{d_2} & I_{d_2} + \mathbf{1}_{d_2 \times d_2} \end{pmatrix}$
$\det(Q^TQ + R^TR)$	1	1	$(d_2)^2$
$\begin{pmatrix} Q \\ R \end{pmatrix}^{-1}$	$\begin{pmatrix} \mathbf{0}_{d_2}^T & 1 \\ I_{d_2} & -\mathbf{1}_{d_2} \end{pmatrix}$	$\begin{pmatrix} 1 & \mathbf{0}_{d_2-1}^T & -1 \\ 0 & \mathbf{0}_{d_2-1}^T & 1 \\ -\mathbf{1}_{d_2} & I_{d_2-1} & \mathbf{1}_{d_2-1} \end{pmatrix}$	$-\frac{1}{d_2}\begin{pmatrix} -\mathbf{1}_{d_2}^T & 1 \\ \mathbf{1}_{d_2 \times d_2} - d_2 I_{d_2} & -\mathbf{1}_{d_2} \end{pmatrix}$

4 Numerical illustrations

We make a simulation analysis to assess the performance advantage of the proposed estimator (12) compared the MLE computed by IWLS. We also compare the asymptotic variance of the two estimators based on simulated datasets. All computations are carried out with the R statistical software.

In our simulations, we consider a GLM with a gamma distribution. Namely, we assume a sample $Y_1, ..., Y_n$ of independent variables following a gamma distribution with a shape parameter $k > 0$ and rate parameter $\theta_i > 0$, the associated log-likelihood is given by (1) with

$$\lambda_i = -\frac{\theta_i}{k},\ a(\phi) = \phi,\ \phi = \frac{1}{k},\ b(\lambda_i) = -\log(-\lambda_i),\ c(y_i, \phi) = (\frac{1}{\phi} - 1)\log(y_i) - \log\Gamma(\frac{1}{\phi}) - \log(\frac{1}{\phi}).$$

We consider two explanatory variables $x_i^{(2)}$ and $x_i^{(3)}$ with d_2 and d_3 modalities. Given a parameter value $\boldsymbol{\vartheta}$, we assume a zero-sum condition

$$\vartheta_{d_2}^{(2)} = -\sum_{k=1}^{d_2-1} \vartheta_k^{(2)},\ \vartheta_{d_3}^{(3)} = -\sum_{l=1}^{d_3-1} \vartheta_l^{(3)}.$$

That corresponds to the following contrast matrix

$$R = \begin{pmatrix} 0 & \mathbf{1}_{d_2} & \mathbf{0}_{d_3} \\ 0 & \mathbf{0}_{d_2} & \mathbf{1}_{d_3} \end{pmatrix}.$$

The simulation procedure for a given sample size n consists of simulating explanatory variables, then of simulating the gamma variable and finally of estimating the MLE $\widehat{\boldsymbol{\vartheta}}_n$ and the proposed estimator $\widetilde{\boldsymbol{\vartheta}}_n$ The procedure is summarized as follows

1. For $i = 1, \ldots, n$, generate two equiprobable categorical variables $x_i^{(2)}$ and $x_i^{(3)}$ with d_2 and d_3 modalities;
2. Compute the linear predictor $\eta_i = \vartheta^{(1)} + \sum_{k=1}^{d_2} x_i^{(2),k} \vartheta_k^{(2)} + \sum_{l=1}^{d_3} x_i^{(3),l} \vartheta_l^{(3)}$ using Table 5;
3. Compute rate parameters for gamma distribution $\theta_i = -\frac{\ell(\eta_i)}{\phi}$;
4. **repeat** M times:
   (a) generate n responses $(Y_i)_i$ using $\lambda_i = -\frac{\theta_i}{k}$.
   (b) estimate the GLM with the two methods assuming

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

   (c) return computation times, the MLE $\widehat{\boldsymbol{\vartheta}}_n$ and the proposed estimator $\widetilde{\boldsymbol{\vartheta}}_n$.

   end repeat;
5. Compute the variance of the two estimators.

We first present the results of computation times for two link functions of the gamma distribution $g(x) = 1/x$ and $g(x) = \log(x)$. The computation time displayed is the average time over $M = 5$ runs for two sample sizes $n = 10^5$, $n = 10^6$ and $d_2 = d_3 = d$ with $d = 5, 10, 15, 20, 25, 30, 35, 40, 45, 50$. The parameter values are given in Table 5 and the dispersion parameter is fixed to $\phi = 8$.

Table 5: Parameter values of the gamma distribution used for computation times

link function		intercept	variables $j = 2, 3$	
$g(x)$	$\ell(x)$	$\vartheta^{(1)}$	$\vartheta_1^{(j)}, \ldots, \vartheta_{d-1}^{(j)}$	$\vartheta_d^{(j)}$
$1/x$	$-x$	$3d+1$	$1/d, \ldots, (d-1)/d$	$-(d-1)/2$
$\log(x)$	$-e^{-x}$	1	$1, 2, 1, 2, \ldots, 1$ or 2	$-3d/2+2$ or $-3(d-1)/2$

Computation times for both the MLE (denoted by IWLS) and the proposed estimator (denoted by explicit) are displayed in Figure 1 for the inverse link. We observe that the

IWLS has a computation time which increases almost linearly with the modality number d. In comparison, the proposed estimator's computation time is almost constant and significantly lower. Table 6 shows the ratio of the computation time of $\widetilde{\boldsymbol{\vartheta}}_n$ against the computation time of $\widehat{\boldsymbol{\vartheta}}_n$. Irrespective of the sample size, this ratio increases from 6 to 67 as d increases from 5 to 50. When considering the log link function, we observe similar results, yet the computation of $\widehat{\boldsymbol{\vartheta}}_n$ is more erratic and higher than for the canonical link, see Figure 2 and Table 6.

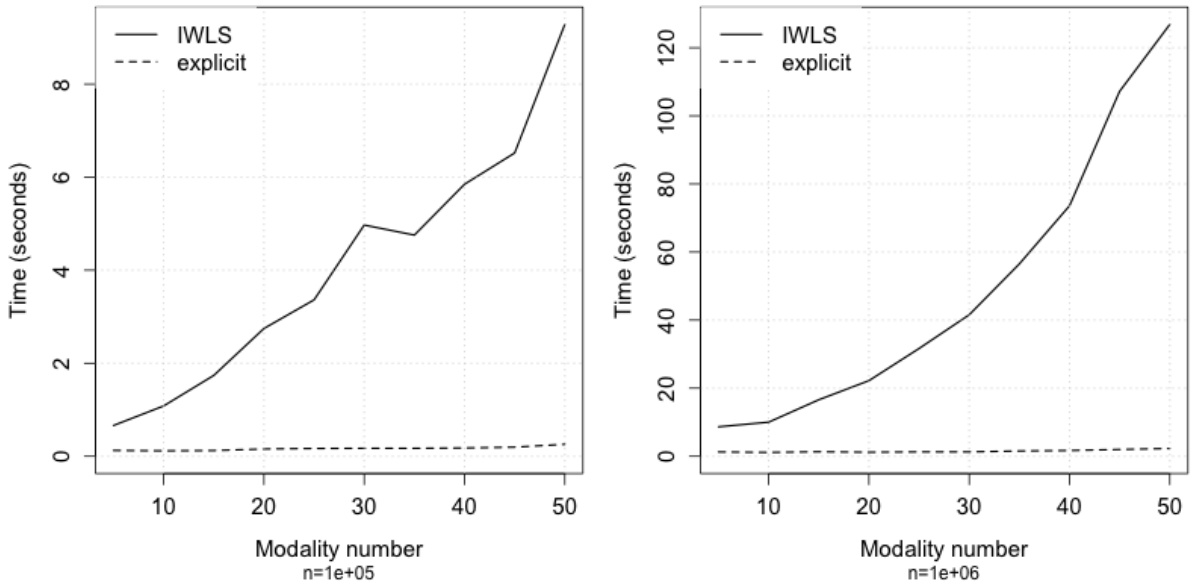


Figure 1: Computation time for gamma response with inverse link function (average over 5 runs)

Table 6: Time ratio of IWLS over explicit methods for sample size n and modality number d

	inverse link		log link	
	$n = 10^5$	$n = 10^6$	$n = 10^5$	$n = 10^6$
$d = 5$	5.26	6.84	5.12	6.23
$d = 10$	9.17	8.83	8.01	9.98
$d = 15$	14.09	12.59	15.49	17.10
$d = 20$	17.56	18.70	21.60	22.89
$d = 25$	20.04	24.75	27.71	31.39
$d = 30$	28.97	32.41	56.62	66.41
$d = 35$	27.74	36.12	74.51	57.60
$d = 40$	32.63	43.37	70.22	55.81
$d = 45$	33.27	53.65	60.28	61.01
$d = 50$	36.04	55.58	67.03	68.80

Now, we turn our attention to asymptotic variances for two link functions of the gamma distribution and link function $g(x) = 1/x$. Here, we consider $M = 10000$ replicates, a sample size $n = 10^3, 10^4$ and $d_2 = 2$ and $d_3 = 3$. True parameter vector is chosen to

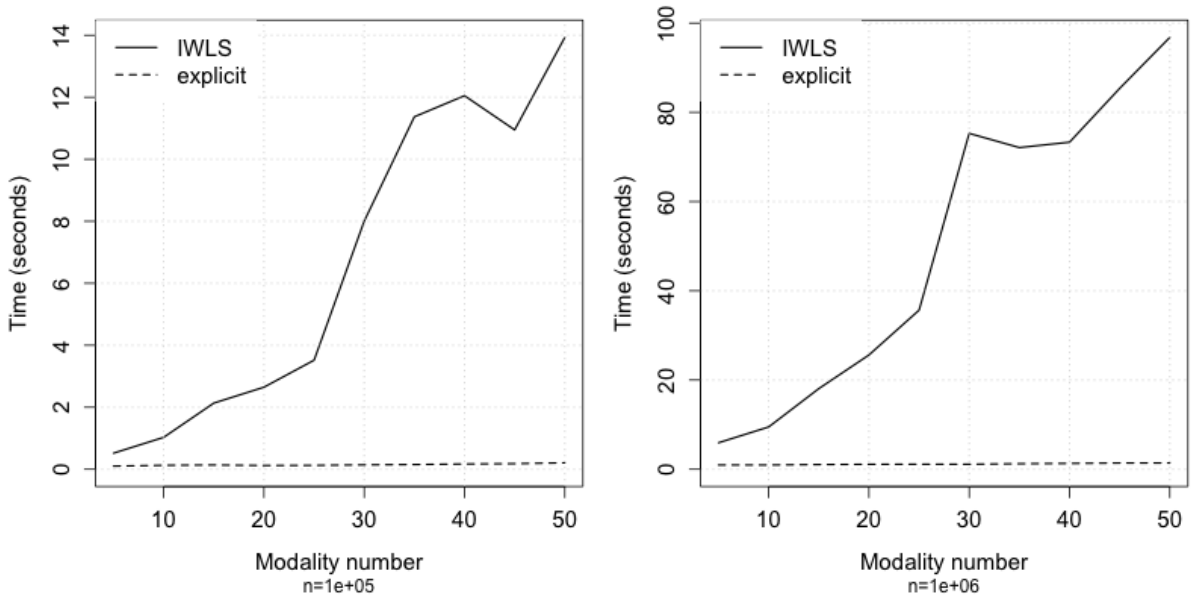


Figure 2: Computation time for gamma response with log link function (average over 5 runs)

guarantee the positivity of the parameter θ_i for the inverse link function, see Table 7. Again, the dispersion parameter is fixed to $\phi = 8$.

Table 7: Parameter values of the gamma distribution for asymptotic variance

link function		intercept	variable 2		variable 3		
$g(x)$	$\ell(x)$	$\vartheta^{(1)}$	$\vartheta_1^{(2)}$	$\vartheta_2^{(2)}$	$\vartheta_1^{(3)}$	$\vartheta_2^{(3)}$	$\vartheta_3^{(3)}$
$1/x$	$-x$	10	1	-1	2	3	-5
$\log(x)$	$-e^{-x}$	1	1	-1	2	3	-5

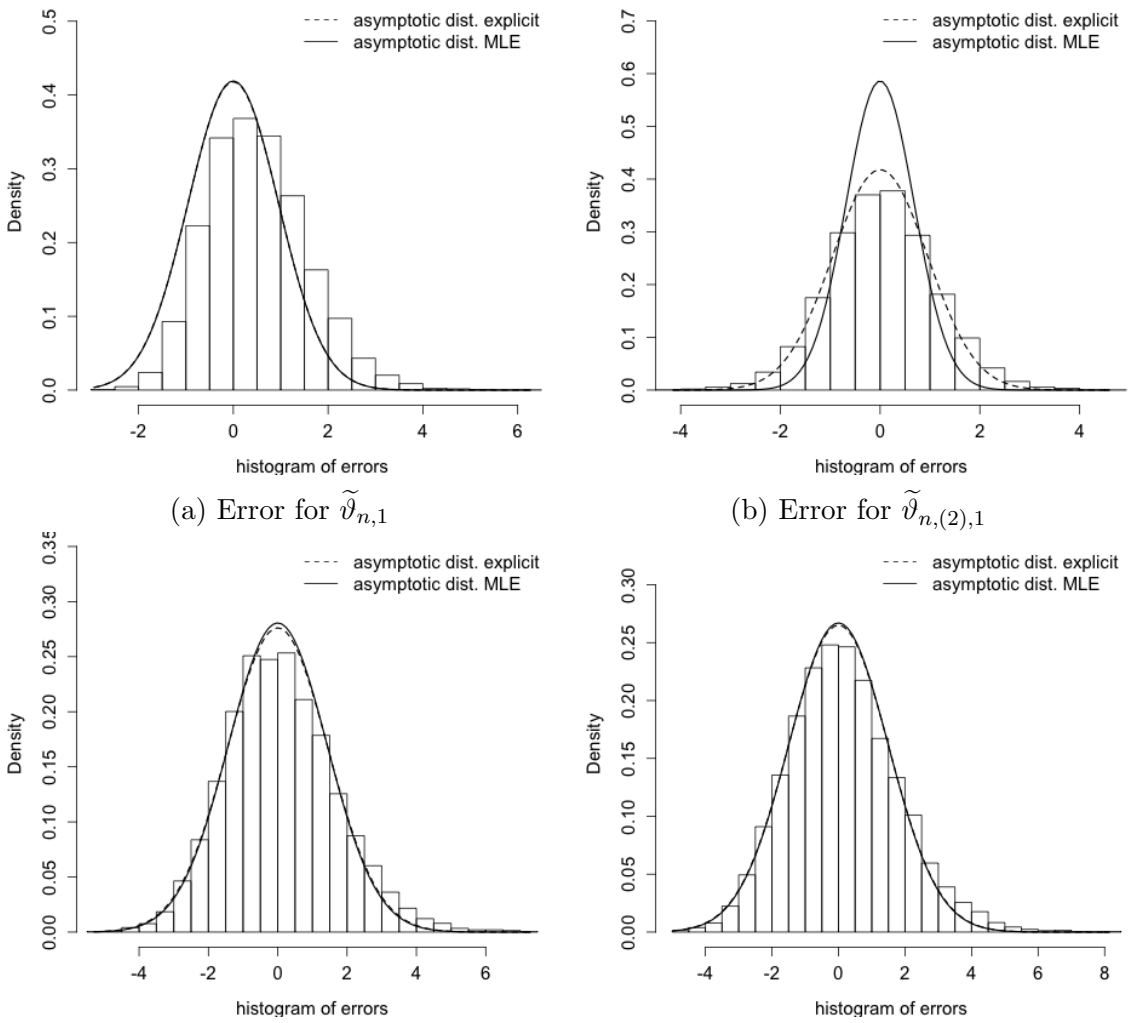
Figure 3 displays the asymptotic distribution of errors for $n = 1000$ of $\widetilde{\boldsymbol{\vartheta}}_n$ for the inverse link function. Figure 3a shows a small bias of $\widetilde{\boldsymbol{\vartheta}}_n$ when estimating the intercept, which is not present for other coefficients. Asymptotic variances between the MLE and the proposed estimator are very close except for $\vartheta_{(2),1}$. Indeed, Figure 3b shows a narrow distribution for MLE than for the proposed estimator. Similar conclusions can be drawn for $n = 10000$ and/or the log link function, as well as other distributions.

5 Conclusion

A closed form estimator for GLM with categorical explanatory variables has been presented. It is a fast computable alternative to the MLE, in particular in the practical case of GLM with single effect only. The asymptotic properties of this estimation procedure have been studied.

The closed-form estimator avoid using the IWLS algorithm which is time-consuming for a large number of variables or modalities. Numerical illustrations quantify the performances of our explicit estimator against the MLE in different GLM examples.

In order to handle the asymptotical non-efficiency of the proposed estimator (compared to the MLE), the Le Cam one-step procedure could be targeted in a further research.



(a) Error for $\widetilde{\vartheta}_{n,1}$

(b) Error for $\widetilde{\vartheta}_{n,(2),1}$

(c) Error for $\widetilde{\vartheta}_{n,(3),1}$

(d) Error for $\widetilde{\vartheta}_{n,(3),2}$

Figure 3: Histogram of parameter errors for gamma response with 2 variables ($n = 10^3$). Dashed (resp. solid) lines are theoretical asymptotic densities for $\widetilde{\boldsymbol{\vartheta}}_n$ (resp. $\widehat{\boldsymbol{\vartheta}}_n$).

Acknowledgements

This research benefited from the support of the ANR project 'Efficient Inference for large and high-frequency data' (ANR-21-CE40-0021), the 'Chair Risques Emergents ou atypiques en Assurance', under the aegis of Fondation du Risque, a joint initiative by Le Mans University 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 University and Groupama Centre-Manche company, member of Groupama group.

References

Brouste, A., Dutang, C. & Rohmer, T. (2020), '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.

Denuit, M., Hainaut, D. & Trufin, J. (2020), *Effective Statistical Learning Methods for Actuaries I: GLMs and extensions*, Springer Nature.

Dutang, C. & Charpentier, A. (2020), *CASdatasets: Insurance datasets*, Univ. Paris-Dauphine and Univ. du Québec à Montreal. R package version 1.0-11.

Dutang, C. & Guibert, Q. (2021), 'An explicit split point procedure in model-based trees allowing for a quick fitting of glm trees and glm forests', *Statistics and Computing* **32**(1).

Fahrmeir, L. & Kaufmann, H. (1985), 'Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models', *The Annals of Statistics* pp. 342–368.

Frischknecht, B., Eckert, C., Louviere, J. & Ribeiro, T. (2014), Simple ways to estimate choice models for single consumers, *in* S. Hess & A. Daly, eds, 'Handbook of Choice Modelling', Chapters, Edward Elgar Publishing, chapter 21, pp. 498–518.

Kadarmideen, H., Thompson, R. & Simm, G. (2000), 'Linear and threshold model genetic parameters for disease, fertility and milk production in dairy cattle', *Animal Science* **71**, 411–419.

Lindsey, J. (1997), *Applying Generalized Linear Models*, Springer Texts in Statistics.

Lipovetsky, S. (2015), 'Analytical closed-form solution for binary logit regression by categorical predictors', *Journal of applied statistics* **42**(1), 37–49.

Lipovetsky, S. & Conklin, M. (2014), 'Best-worst scaling in analytical closed-form solution', *Journal of choice modelling* **10**, 60–68.

Lipovetsky, S., Liakhovitski, D. & Conklin, M. (2015), What's the right sample size for my maxdiff study, *in* 'Sawtooth Software Conference'.

Lu, T.-T. & Shiou, S.-H. (2002), 'Inverses of 2× 2 block matrices', *Computers & Mathematics with Applications* **43**(1-2), 119–129.

Marley, A., Islam, T. & Hawkins, G. (2016), 'A formal and empirical comparison of two score measures for best–worst scaling', *Journal of choice modelling* **21**, 15–24.

McCullagh, P. & Nelder, J. A. (1989), *Generalized linear models*, Vol. 37, CRC press.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Sunwoo, H. (1996), 'Simple algorithms about kronecker products in the linear model', *Linear algebra and its applications* **237**, 351–358.

Van der Vaart, A. W. (2000), *Asymptotic statistics*, Vol. 3, Cambridge university press.

Wuethrich, M. & Merz, M. (2021), Statistical foundations of actuarial learning and its applications. SSRN papers.
URL: *https://ssrn.com/abstract=3822407*

A Proof of Section 3

A.1 Identifiability

Consider the subspace $\Theta = \{\boldsymbol{x} \in \mathbb{R}^p, R\boldsymbol{x} = 0\}$. The parameter $\boldsymbol{\vartheta}$ is identifiable if $\boldsymbol{x} \mapsto Q\boldsymbol{x}$ is injective on Θ. That is

$$\forall \boldsymbol{x} \in \mathbb{R}^p, R\boldsymbol{x} \neq 0 \text{ or } \begin{cases} R\boldsymbol{x} = 0 \\ Q\boldsymbol{x} \neq 0 \end{cases} \Leftrightarrow \forall \boldsymbol{x} \in \mathbb{R}^p, \|R\boldsymbol{x}\|^2 + \|Q\boldsymbol{x}\|^2 > 0$$

$$\Leftrightarrow \forall \boldsymbol{x} \in \mathbb{R}^p, \boldsymbol{x}^T(R^TR + Q^TQ)\boldsymbol{x} > 0.$$

Hence the identifiability condition is equivalent to $R^TR + Q^TQ$ being positive definite.

A.2 Proof of Theorem 1

Consider the empirical average $\overline{Y}_n^{(k_2,\ldots,k_{m+1})}$ of Equation (11). Because random responses Y_i are i.i.d. on the set of observations $\{i; \eta_{x_i} = \eta_{k_2,\ldots,k_{m+1}}\}$, $\overline{Y}_n^{(k_2,\ldots,k_{m+1})}$ converges almost surely to $\mu_{k_2,\ldots,k_{m+1}}$. By the continuous mapping theorem $g(\overline{Y}_n^{k_2,\ldots,k_{m+1}})$ converges almost surely to $g(\mu_{k_2,\ldots,k_{m+1}})$. Hence we obtain directly the strong consistency of $\widetilde{\boldsymbol{\vartheta}}_n$.

Using the Delta method (Van der Vaart 2000) and the central limit theorem, we have

$$\sqrt{m_{k_2,\ldots,k_{m+1}}}\left(g(\overline{Y}_n^{k_2,\ldots,k_{m+1}}) - g(\mu_{k_2,\ldots,k_{m+1}})\right) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\left(0, a(\phi)b''((b')^{-1}(\mu_{k_2,\ldots,k_{m+1}}))(g'(\mu_{k_2,\ldots,k_{m+1}}))^2\right),$$

where $m_{k_2,\ldots,k_{m+1}}$ are empirical frequencies (10) Using the theoretical relative frequencies (10), we obtain

$$\sqrt{n}\left(g(\overline{Y}_n^{k_2,\ldots,k_{m+1}}) - g(\mu_{k_2,\ldots,k_{m+1}})\right) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\left(0, \sigma^2_{k_2,\ldots,k_{m+1}}\right),$$

Therefore we obtain the distribution of the vector of transformed empirical averages $g(\overline{\boldsymbol{Y}})$

$$\sqrt{n}\left(g(\overline{\boldsymbol{Y}}) - g(\boldsymbol{\mu})\right) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}(\mathbf{0}, \Sigma), \ \Sigma = \text{diag}(\sigma^2_{k_2,\ldots,k_{m+1}}).$$

Multiplying by Q_R and using $g(\boldsymbol{\mu}) = Q\boldsymbol{\vartheta}$ leads to

$$\sqrt{n}\left(Q_R g(\overline{\boldsymbol{Y}}) - Q_R Q\boldsymbol{\vartheta}\right) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, Q_R \Sigma Q_R^T\right).$$

Using $R\boldsymbol{\vartheta} = 0$, we have

$$R^TR\boldsymbol{\vartheta} = \mathbf{0}_{d_2+1} \Leftrightarrow Q^TQ\boldsymbol{\vartheta} = (Q^TQ + R^TR)\boldsymbol{\vartheta} \Leftrightarrow Q_RQ\boldsymbol{\vartheta} = \boldsymbol{\vartheta}. \tag{20}$$

Replacing $Q_RQ\boldsymbol{\vartheta} = \boldsymbol{\vartheta}$ leads to the desired result.

A.3 Proof of Theorem 2

In the general case of m categorical explanatory variables, the log-likelihood (3) becomes

$$J(\boldsymbol{\vartheta}) = \log \mathcal{L}(\boldsymbol{\vartheta} \,|\, \boldsymbol{y}) = \sum_{i=1}^{n} \frac{y_i \ell(\eta_{\boldsymbol{x}_i}) - b\left(\ell(\eta_{\boldsymbol{x}_i})\right)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi).$$

where $\eta_{\boldsymbol{x}_i} = \boldsymbol{x}_i^T \boldsymbol{\vartheta}$. Let us compute partial derivative of the linear predictor $\eta_{\boldsymbol{x}_i}$

$$\frac{\partial \eta_{\boldsymbol{x}_i}}{\partial \vartheta^{(1)}} = 1, \; \frac{\partial \eta_{\boldsymbol{x}_i}}{\partial \vartheta_k^{(j)}} = x_i^{(j),k}, \frac{\partial \eta_{\boldsymbol{x}_i}}{\partial \vartheta_{k_2,k_3}^{(j_2,j_3)}} = x_i^{(j_2),k_2} x_i^{(j_3),k_3}, \ldots, \frac{\partial \eta_{\boldsymbol{x}_i}}{\partial \vartheta_{k_2,\ldots,k_{m+1}}^{(2,\ldots,m+1)}} = x_i^{(2),k_2} \ldots x_i^{(m+1),k_{m+1}}.$$

In order to deal with partial derivatives of J, we introduce the derivative w.r.t. intercept

$$S^{(1)} = \frac{\partial J(\boldsymbol{\vartheta})}{\partial \vartheta^{(1)}} = \sum_{i=1}^{n} \frac{y_i \ell'(\eta_{\boldsymbol{x}_i}) - \ell'(\eta_{\boldsymbol{x}_i}) b'\left(\ell(\eta_{\boldsymbol{x}_i})\right)}{a(\phi)} \frac{\partial \eta_{\boldsymbol{x}_i}}{\partial \vartheta^{(1)}} = \sum_{i=1}^{n} \ell'(\eta_{\boldsymbol{x}_i}) \frac{y_i - b'\left(\ell(\eta_{\boldsymbol{x}_i})\right)}{a(\phi)},$$

the derivative w.r.t. single-effect coefficients $j = 2, \ldots, m+1$, $k = 1, \ldots, d_j$,

$$S_k^{(j)} = \frac{\partial J(\boldsymbol{\vartheta})}{\partial \vartheta_k^{(j)}} = \sum_{i=1}^{n} \frac{y_i \ell'(\eta_{\boldsymbol{x}_i}) - \ell'(\eta_{\boldsymbol{x}_i}) b'\left(\ell(\eta_{\boldsymbol{x}_i})\right)}{a(\phi)} \frac{\partial \eta_{\boldsymbol{x}_i}}{\partial \vartheta_k^{(j)}} = \sum_{i=1}^{n} \ell'(\eta_{\boldsymbol{x}_i}) x_i^{(j),k} \frac{y_i - b'\left(\ell(\eta_{\boldsymbol{x}_i})\right)}{a(\phi)},$$

the derivative w.r.t. double-effect coefficients $j_i = 2, \ldots, m+1$, $k_i = 1, \ldots, d_{j_i}$, $i = 2, 3$,

$$S_{k_2,k_3}^{(j_2,j_3)} = \frac{\partial J(\boldsymbol{\vartheta})}{\partial \vartheta_{k_2,k_3}^{(j_2,j_3)}} = \sum_{i=1}^{n} \frac{y_i \ell'(\eta_{\boldsymbol{x}_i}) - \ell'(\eta_{\boldsymbol{x}_i}) b'\left(\ell(\eta_{\boldsymbol{x}_i})\right)}{a(\phi)} \frac{\partial \eta_{\boldsymbol{x}_i}}{\partial \vartheta_{k_2,k_3}^{(j_2,j_3)}} = \sum_{i=1}^{n} \ell'(\eta_{\boldsymbol{x}_i}) x_i^{(j_2),k_2} x_i^{(j_3),k_3} \frac{y_i - b'\left(\ell(\eta_{\boldsymbol{x}_i})\right)}{a(\phi)},$$

the derivative w.r.t. full-effect coefficients $j_i = 2, \ldots, m+1$, $k_i = 1, \ldots, d_{j_i}$, $i = 2, \ldots, m+1$,

$$\begin{aligned} S_{k_2,\ldots,k_{m+1}}^{(2,\ldots,m+1)} &= \frac{\partial J(\boldsymbol{\vartheta})}{\partial \vartheta_{k_2,\ldots,k_{m+1}}^{(j_2,\ldots,j_{m+1})}} = \sum_{i=1}^{n} \frac{y_i \ell'(\eta_{\boldsymbol{x}_i}) - \ell'(\eta_{\boldsymbol{x}_i}) b'\left(\ell(\eta_{\boldsymbol{x}_i})\right)}{a(\phi)} \frac{\partial \eta_{\boldsymbol{x}_i}}{\partial \vartheta_{k_2,\ldots,k_{m+1}}^{(j_2,\ldots,j_{m+1})}} \\ &= \sum_{i=1}^{n} \ell'(\eta_{\boldsymbol{x}_i}) x_i^{(2),k_2} \ldots x_i^{(m+1),k_{m+1}} \frac{y_i - b'\left(\ell(\eta_{\boldsymbol{x}_i})\right)}{a(\phi)}. \end{aligned}$$

The maximum likelihood optimization problem is

$$\begin{cases} \nabla_{\boldsymbol{\vartheta}} J(\boldsymbol{\vartheta}) - R^T \boldsymbol{u} = \boldsymbol{0} \\ R\boldsymbol{\vartheta} = \boldsymbol{0}, \end{cases} \tag{21}$$

where $\boldsymbol{u}$ is the Lagrange multiplier and the log-likelihood gradient and contrasts are

$$\nabla_{\boldsymbol{\vartheta}} J(\boldsymbol{\vartheta}) = \begin{pmatrix} S^{(1)} \\ S_k^{(j)} \\ \vdots \\ S_{k_2,k_3}^{(j_2,j_3)} \\ \vdots \\ S_{k_2,\ldots,k_{m+1}}^{(j_2,\ldots,j_{m+1})} \end{pmatrix}, R = \left(R^{(1)}, R^{(j)}, \ldots, R^{(j_2,j_3)}, \ldots, R^{(2,\ldots,m+1)}\right) = (R_{-m}, R^{(2,\ldots,m+1)}). \tag{22}$$

Using the binary structures of dummies, see Equation 9, the linear predictor η_{x_i} simplifies. Hence introducing multiple sums over $k_2, \ldots, k_{m+1}$ makes η_{x_i} no longer depend on i but on other sum indexes. Using empirical average (11), empirical frequency (10), and $M_m^{()}$ matrices (7) Equation (22) simplifies to

$$\begin{aligned} S^{(1)} &= \sum_{k_2,\ldots,k_{m+1}} \sum_{i=1;\eta_{x_i}=\eta_{k_2,\ldots,k_{m+1}}}^{n} \ell'(\eta_{k_2,\ldots,k_{m+1}}) \frac{y_i - b'\left(\ell(\eta_{k_2,\ldots,k_{m+1}})\right)}{a(\phi)} \\ &= \sum_{k_2,\ldots,k_{m+1}} \ell'(\eta_{k_2,\ldots,k_{m+1}}) \left(\sum_{i=1;\eta_{x_i}=\eta_{k_2,\ldots,k_{m+1}}}^{n} \frac{y_i}{a(\phi)} - \sum_{i=1;\eta_{x_i}=\eta_{k_2,\ldots,k_{m+1}}}^{n} \frac{b'\left(\ell(\eta_{k_2,\ldots,k_{m+1}})\right)}{a(\phi)} \right) \\ &= \sum_{k_2,\ldots,k_{m+1}} \xi_{k_2,\ldots,k_{m+1}} = \left(M_m^{()}\right)^T \Xi, \end{aligned}$$

with $\Xi = \left(\xi_{k_2,\ldots,k_{m+1}}\right)_{k_2,\ldots,k_{m+1}}$ and

$$\xi_{k_2,\ldots,k_{m+1}} = \ell'(\eta_{k_2,\ldots,k_{m+1}}) m_{k_2,\ldots,k_{m+1}} \left(\frac{\overline{y}_n^{k_2,\ldots,k_{m+1}} - b'\left(\ell(\eta_{k_2,\ldots,k_{m+1}})\right)}{a(\phi)} \right).$$

Now for $j = 2, \ldots, m+1$ consider the single-effect score vector $(S_{k^\star}^{(j)})_{k^\star=1,\ldots d_j}$.

$$\left(S_{k^\star}^{(j)}\right)_{k^\star=1,\ldots d_j} = \left(\sum_{k_2,\ldots,k_{j-1},k_{j+1},\ldots,k_{m+1}} \xi_{k_2,\ldots,k_{j-1},k^\star,k_{j+1},\ldots,k_{m+1}} \right)_{k^\star=1,\ldots d_j} = \left(M_m^{(j)}\right)^T \Xi.$$

Then for $j_2 < j_3 = 2, \ldots, m+1$ consider the double-effect score vector $\left(S_{k_2^\star,k_3^\star}^{(j_2,j_3)}\right)_{k_2^\star=1,\ldots,d_{j_2},k_3^\star=1,\ldots,d_{j_3}}$.

$$\begin{aligned} \left(S_{k_2^\star,k_3^\star}^{(j_2,j_3)}\right)_{k_2^\star k_3^\star} &= \left(\sum_{\ldots,k_{j_2-1},k_{j_2+1},\ldots,k_{j_3-1},k_{j_3+1},\ldots} \xi_{k_2,\ldots,k_{j_2-1},k_2^\star,k_{j_2+1},\ldots,k_{j_3-1},k_3^\star,k_{j_3+1},\ldots,k_{m+1}} \right)_{k_2^\star k_3^\star} \\ &= \left(M_m^{(j_2,j_3)}\right)^T \Xi. \end{aligned}$$

For all-cross-effect score vector $\left(S_{k_2,\ldots,k_{m+1}}^{(2,\ldots,m+1)}\right)_{k_2,\ldots,k_{m+1}}$, we have

$$\left(S_{k_2^\star,\ldots,k_{m+1}^\star}^{(2,\ldots,m+1)}\right)_{k_2^\star,\ldots,k_{m+1}^\star} = \left(\xi_{k_2^\star,\ldots,k_{m+1}^\star}\right)_{k_2^\star,\ldots,k_{m+1}^\star} = M_m^{(2,\ldots,m+1)}\Xi.$$

Note that $S^{(1)}$ (resp. other terms) can be written as a sum of all (resp. some) $S_{k_2,\ldots,k_{m+1}}^{(2,\ldots,m+1)}$. Using the Q matrix (8), see also examples in Table 8, we have

$$\left(b'(\ell(\eta_{k_2,\ldots,k_{m+1}})\right)_{k_2,\ldots,k_{m+1}} = g^{-1}(Q\boldsymbol{\vartheta}).$$

Using the preceding equations and $M_m^{()}$ matrices (7), the gradient of the log-likelihood can be written as

$$\nabla J(\boldsymbol{\vartheta}) = \begin{pmatrix} \left(M_m^{()}\right)^T \Xi \\ \vdots \\ \left(\left(M_m^{(j)}\right)^T \Xi\right)_{j=1,\dots,m+1} \\ \vdots \\ \left(\left(M_m^{(j_2,j_3)}\right)^T \Xi\right)_{j_2<j_3=1,\dots,m+1} \\ \vdots \\ (M_m^{(2,\dots,m+1)})^T \Xi \end{pmatrix} = Q^T \Xi.$$

The first rows of the score equation (21) become $Q^T\Xi = R^T\boldsymbol{u}, \quad R = (R^{(1)}$. Then

$$\begin{cases} \left((R^{(1)})^T - M_0 (R^{(2,\dots,m+1)})^T\right)\boldsymbol{u} = 0 \\ \left((R^{(j)})_j^T - M_1 (R^{(2,\dots,m+1)})^T\right)\boldsymbol{u} = 0 \\ \left((R^{(j_2,j_3)})_{j_2,j_3}^T - M_2 (R^{(2,\dots,m+1)})^T\right)\boldsymbol{u} = 0 \\ \vdots \end{cases} \Leftrightarrow \left(R_{-m}^T - Q_{-m}^T (R^{(2,\dots,m+1)})^T\right)\boldsymbol{u} = 0.$$

Hence the system for $\boldsymbol{u}$ is

$$R_\Delta^T \boldsymbol{u} = \mathbf{0}, \text{ with } R_\Delta = R_{-m} - R^{(2,\dots,m+1)} Q_{-m}. \tag{23}$$

This system admit an unique solution that is $\boldsymbol{u} = 0$ if $det(R_\Delta) \neq 0$. In this case, because $\Xi = G(\boldsymbol{\vartheta})(\overline{\boldsymbol{Y}} - g^{-1}(Q\boldsymbol{\vartheta}))$, with $G(\boldsymbol{\vartheta})$ the diagonal matrix

$$G(\boldsymbol{\vartheta}) = \begin{pmatrix} \frac{\ell'(\eta_{1,\dots,1}) m_{1,\dots,1}}{a(\phi)} & 0 & \cdots \\ & \ddots & \\ \cdots & 0 & \frac{\ell'(\eta_{d_2,\dots,d_{m+1}}) m_{d_2,\dots,d_{m+1}}}{a(\phi)} \end{pmatrix},$$

because $\ell' \neq 0$, the system (21) rewrites

$$\begin{cases} \boldsymbol{u} = 0 \\ Q\boldsymbol{\vartheta} = g(\overline{\boldsymbol{Y}}) \\ R\boldsymbol{\vartheta} = \mathbf{0} \end{cases} \Leftrightarrow \begin{cases} \boldsymbol{u} = 0 \\ \boldsymbol{\vartheta} = (Q^TQ + R^TR)^{-1} Q^T g(\overline{\boldsymbol{Y}}). \end{cases}$$

A.4 Examples of Section 3.2

A.4.1 $q = 1$

The proposed estimator is correctly defined and

$$Q^TQ + R^TR = \begin{pmatrix} d_2 & \mathbf{1}_{d_2}^T \\ \mathbf{1}_{d_2} & I_{d_2} \end{pmatrix} + \begin{pmatrix} r_{0,1}^2 & r_{0,1}\boldsymbol{r} \\ r_{0,1}\boldsymbol{r}^T & \boldsymbol{r}^T\boldsymbol{r} \end{pmatrix} = \begin{pmatrix} d_2 + r_{0,1}^2 & \mathbf{1}_{d_2}^T + r_{0,1}\boldsymbol{r} \\ \mathbf{1}_{d_2} + r_{0,1}\boldsymbol{r}^T & I_{d_2} + \boldsymbol{r}^T\boldsymbol{r} \end{pmatrix}. \tag{24}$$

In the case of $q = 1$, using

$$\det(Q^TQ + R^TR) = \left(\det \begin{pmatrix} Q \\ R \end{pmatrix}\right)^2,$$

and Appendix A of Brouste et al. (2020), we get

$$\det(Q^TQ + R^TR) = \left((-1)^{d_2}(r_{0,1} - \boldsymbol{r}\mathbf{1}_{d_2})\right)^2 = (r_{0,1} - \boldsymbol{r}\mathbf{1}_{d_2})^2. \tag{25}$$

Table 8: Notations of Q and R matrices for $m = 2, 3, 4$

m	$Q =$	notation	corresp. contrast	effect
4	$(\mathbf{1}_{d_5 d_4 d_3 d_2},$	$M_4^{()}$	$R^{(1)}$	intercept
	$\mathbf{1}_{d_5 d_4 d_3} \otimes I_{d_2}, \mathbf{1}_{d_5 d_4} \otimes I_{d_3} \otimes \mathbf{1}_{d_2}, \mathbf{1}_{d_5} \otimes I_{d_4} \otimes \mathbf{1}_{d_3 d_2}, I_{d_5} \otimes \mathbf{1}_{d_4 d_3 d_2},$	$M_4^{(j)}$	$R^{(j)}$	single effect
	$\mathbf{1}_{d_5 d_4} \otimes I_{d_3 d_2}, \mathbf{1}_{d_5} \otimes I_{d_4} \otimes \mathbf{1}_{d_3} \otimes I_{d_2}, I_{d_5} \otimes \mathbf{1}_{d_4 d_3} \otimes I_{d_2},$			
	$\mathbf{1}_{d_5} \otimes I_{d_4 d_3} \otimes \mathbf{1}_{d_2}, I_{d_5} \otimes \mathbf{1}_{d_4} \otimes I_{d_3} \otimes \mathbf{1}_{d_2}, I_{d_5 d_4} \otimes \mathbf{1}_{d_3 d_2},$	$M_4^{(j_2,j_3)}$	$R^{(j_2,j_3)}$	double effect
	$\mathbf{1}_{d_5} \otimes I_{d_4 d_3 d_2}, I_{d_5} \otimes \mathbf{1}_{d_4} \otimes I_{d_3 d_2}, I_{d_5 d_4} \otimes \mathbf{1}_{d_3} \otimes I_{d_2}, I_{d_5 d_4 d_3} \otimes \mathbf{1}_{d_2},$	$M_4^{(j_2,j_3,j_4)}$	$R^{(j_2,j_3,j_4)}$	triple effect
	$I_{d_5 d_4 d_3 d_2}),$	$M_4^{(2,3,4,5)}$	$R^{(2,3,4,5)}$	all effect
3	$(\mathbf{1}_{d_4 d_3 d_2},$	$M_0^{(0),3}$	$R^{(1)}$	intercept
	$\mathbf{1}_{d_4 d_3} \otimes I_{d_2}, \mathbf{1}_{d_4} \otimes I_{d_3} \otimes \mathbf{1}_{d_2}, I_{d_4} \otimes \mathbf{1}_{d_3 d_2},$	$M_k^{(1),3}$	$R^{(j)}$	single effect
	$\mathbf{1}_{d_4} \otimes I_{d_3 d_2}, I_{d_4} \otimes \mathbf{1}_{d_3} \otimes I_{d_2}, I_{d_4 d_3} \otimes \mathbf{1}_{d_2},$	$M_k^{(2),3}$	$R^{(j,j+1)}$	double effect
	$I_{d_4 d_3 d_2}),$	$M_0^{(3),3}$	$R^{(2,3,4)}$	all effect
2	$(\mathbf{1}_{d_3 d_2},$	$M_0^{(0),2}$	$R^{(1)}$	intercept
	$\mathbf{1}_{d_3} \otimes I_{d_2}, I_{d_3} \otimes \mathbf{1}_{d_2},$	$M_k^{(1),2}$	$R^{(j)}$	single effect
	$I_{d_3 d_2})$	$M_0^{(2),2}$	$R^{(2,3)}$	all effect

The case of no intercept Consider $R = (1, 0, \ldots, 0)$, i.e. $r_0 = 1$ and $\boldsymbol{r} = \mathbf{0}^T$.

Using (24), we have

$$Q^T Q + R^T R = \begin{pmatrix} d_2 + 1 & \mathbf{1}_{d_2}^T \\ \mathbf{1}_{d_2} & I_{d_2} \end{pmatrix}.$$

Using (25), the determinant is non null $\det(Q^T Q + R^T R) = 1 \neq 0$.

The case of no first-level Consider $R = (0, 1, 0, \ldots, 0)$, i.e. $r_0 = 0$ and $\boldsymbol{r} = (1, \mathbf{0}^T)$.

Using (24), we have

$$Q^T Q + R^T R = \begin{pmatrix} d_2 & 1 & \mathbf{1}_{d_2-1}^T \\ 1 & 2 & \mathbf{0}_{d_2-1}^T \\ \mathbf{1}_{d_2-1} & \mathbf{0}_{d_2-1} & I_{d_2-1} \end{pmatrix}.$$

Using (25), the determinant is non null $\det(Q^T Q + R^T R) = (0 - 1 - 0)^2 = 1 \neq 0$.

Zero-sum condition Consider $R = (0, \mathbf{1}^T)$, i.e. $r_0 = 0$ and $\boldsymbol{r} = \mathbf{1}^T$.

Using (24), we have

$$Q^T Q + R^T R = \begin{pmatrix} d_2 & \mathbf{1}_{d_2}^T \\ \mathbf{1}_{d_2} & I_{d_2} + \mathbf{1}_{d_2 \times d_2} \end{pmatrix}.$$

Using (25), the determinant is non null $\det(Q^T Q + R^T R) = (0 - \mathbf{1}^T \mathbf{1})^2 = (d_2)^2 \neq 0$.

B Linear algebra on Q and R

B.1 Linear algebra on Q, R in (17) without interaction

Consider the matrices Q_1, Q and R_1, R, defined in (18) and (19).

Using Theorem 2.1 of Lu & Shiou (2002), $Q^T Q + R^T R$ is invertible iff $Q_1^T Q_1 + 2R_1^T R_1$ invertible. In this case we have

$$(Q^T Q + R^T R)^{-1} = \begin{pmatrix} (\frac{1}{2} Q_1^T Q_1 + R_1^T R_1)^{-1} & -\frac{1}{2}(\frac{1}{2} Q_1^T Q_1 + R_1^T R_1)^{-1} Q_1^T \\ -\frac{1}{2} Q_1 (\frac{1}{2} Q_1^T Q_1 + R_1^T R_1)^{-1} & \frac{1}{2} I + \frac{1}{4} Q_1 (\frac{1}{2} Q_1^T Q_1 + R_1^T R_1)^{-1} Q_1^T \end{pmatrix}$$

$$= 2 \begin{pmatrix} (Q_1^T Q_1 + 2R_1^T R_1)^{-1} & -\frac{1}{2}(Q_1^T Q_1 + 2R_1^T R_1)^{-1} Q_1^T \\ -\frac{1}{2} Q_1 (Q_1^T Q_1 + 2R_1^T R_1)^{-1} & \frac{1}{4} I + \frac{1}{4} Q_1 (Q_1^T Q_1 + 2R_1^T R_1)^{-1} Q_1^T \end{pmatrix}.$$

It can be verified that $(Q_1^TQ_1 + 2R_1^TR_1)^{-1}Q_1^T = (Q_1^TQ_1 + R_1^TR_1)^{-1}Q_1^T$. So because $Q = (Q_1, I)$, we have

$$\begin{aligned} Q_R &= 2\begin{pmatrix} (Q_1^TQ_1 + R_1^TR_1)^{-1}Q_1^T - \frac{1}{2}(Q_1^TQ_1 + R_1^TR_1)^{-1}Q_1^T \\ -\frac{1}{2}Q_1(Q_1^TQ_1 + R_1^TR_1)^{-1}Q_1^T + \frac{1}{4}I + \frac{1}{4}Q_1(Q_1^TQ_1 + R_1^TR_1)^{-1}Q_1^T \end{pmatrix} \\ &= \begin{pmatrix} (Q_1^TQ_1 + R_1^TR_1)^{-1}Q_1^T \\ \frac{1}{2}(I - Q_1(Q_1^TQ_1 + R_1^TR_1)^{-1}Q_1^T) \end{pmatrix}. \end{aligned}$$