

Closed form Maximum Likelihood Estimator for Generalized Linear Models in the case of categorical explanatory variables: Application to insurance loss modelling

Alexandre Brouste, Christophe Dutang, Tom Rohmer

▶ **To cite this version:**

Alexandre Brouste, Christophe Dutang, Tom Rohmer. Closed form Maximum Likelihood Estimator for Generalized Linear Models in the case of categorical explanatory variables: Application to insurance loss modelling. Computational Statistics, 2020, 10.1007/s00180-019-00918-7 . hal-01781504v3

HAL Id: hal-01781504
https://hal.science/hal-01781504v3

Submitted on 25 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Closed-form Maximum Likelihood Estimator for Generalized Linear Models in the case of categorical explanatory variables: Application to insurance loss modeling

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

Abstract

Generalized Linear Models with categorical explanatory variables are considered and parameters of the model are estimated by an exact maximum likelihood method. The existence of a sequence of maximum likelihood estimators is discussed and considerations on possible link functions are proposed. A focus is then given on two particular positive distributions: the Pareto 1 distribution and the shifted log-normal distributions. Finally, the approach is illustrated on an actuarial dataset to model insurance losses.

Keywords: Regression models, heavy-tailed distributions, explicit MLE, insurance claim modeling

[1]Institut du Risque de l'Assurance, Laboratoire Manceau de Mathématiques
Le Mans Université, Avenue Olivier Messiaen, 72085 LE MANS, France
[2]CEREMADE, CNRS, Univ. Paris-Dauphine, Univ. PSL
Place du Maréchal de Lattre de Tassigny, 75016 PARIS, France

1 Introduction

The assumption of identical distributions for random variables in an observation sample is relaxed for regression models by considering explanatory variables. Generalized Linear Models (GLMs) were introduced by Nelder and Wedderburn (1972) and popularized in McCullagh and Nelder (1989). GLMs rely on probability distribution functions of exponential type for the response variable which include most of the light and medium tailed distributions (such as normal, gamma or inverse Gaussian). Asymptotic properties of sequences of maximum likelihood estimators (MLE) for GLMs were studied by Fahrmeir and Kaufmann (1985).

The finite sample property of MLE of specific GLMs has been discussed in depth in statistical literature, see Fienberg (2007) and Haberman (1974). In addition, there are lots of literature studying the finite sample property of MLE for logistic regression models, see e.g. Albert and Anderson (1984) and Silvapulle (1981).

Regression models for heavy-tailed distributions have been mainly studied through the point-of-view of extreme value analysis, see Beirlant et al. (2004) for a review. A regression model for the generalized Pareto distribution (GPD) where the scale parameter depends on covariates are described in Davison and Smith (1990) with a least square estimation procedure and a model checking method. Beirlant et al. (1998) propose a Burr model by regressing the shape parameter with an exponential link on explanatory variables. In the aforementioned article, a simulation study with one explanatory variable is detailed as well as an application to fire insurance. Residual plots and asymptotic convergence towards the normal distribution are also discussed. Similarly, Ozkok et al. (2012) propose a regression model for Burr distribution where the scale parameter depends on covariates.

An estimation of the extremal tail index (used in generalized extreme value (GEV) distributions and GPD) by considering a class of distribution function with an exponential link on explanatory variables is also described in Beirlant and Goegebeur (2003). Using generalized residuals of explanatory variable makes possible the estimation of the tail index. Still by the extreme value theory approach, Chavez-Demoulin et al. (2015) and Hambuckers et al. (2016) both propose a semi-parametric regression model for GEV and GPD where the explanatory variables are time or known factor levels. They assume that all parameters depend on covariates and also use exponentially distributed residuals.

Outside the extreme value theory framework, there is also a literature studying the covariate modelling. For instance, the so-called double GLMs, where the dispersion parameter also depends on covariates, have been studied by McCullagh and Nelder (1989, Chap. 10) or Smyth and Verbyla (1999). Furthermore, Rigby and Stasinopoulos (2005) propose a general regression framework where all parameters are modeled by explanatory variable and the distribution is not restricted to exponential family. The only restriction that the authors impose is the twice differentiability of the density function w.r.t. parameters. However, there is no clear convergence result of the proposed estimators. Among the proposed distributions, Rigby and Stasinopoulos (2005) use 1-parameter Pareto, log-logistic (a special case of the Pareto 3 distribution) and GEV distributions.

In this paper, we propose closed-form estimators for GLMs in the case of categorical variables. The expression is valid for any distribution belonging to the one-parameter exponential family and any link function. To our knowledge, only Lipovetsky (2015) provides an explicit solution in the special case of a logit regression with categorical predictors. He assumes the response variable follows a Bernoulli distribution with a canonical link function and a particular set-up of intercept.

Then, the paper will continue by the application of such formulas not on classical distributions, but on distributions such as the log-transformed variable has a distribution in the exponential family. Therefore, the choice of probability distributions of this paper is led by

two aspects: distributions with positive values and distributions as the log-transformed variable belongs to the exponential family. The considered distributions have heavier tails than the exponential distribution. We choose to study two distributions: the Pareto 1 distribution and the shifted lognormal distribution with fixed threshold parameters. We could have considered log-logistic and GEV distributions being also appropriate in many situations but these distributions do not belong to the exponential family.

Applications of this distribution can be found in various disciplines such as finance, insurance, reliability theory, etc. Here, we are interested with an application to insurance loss modeling. Making insurance tariffs consists in appropriately selecting and transforming explanatory variables so that the prediction fairly estimates the mean of the response variable. When the relation between the transformed response variable and an explanatory variable is not affine, non-linearities is generally accommodated in three ways: binning the variable, adding polynomial terms or using piecewise linear functions in the predictor, see e.g. Goldburd et al. (2016). In this paper, we focus mainly on the first approach where continuous variables (typically the age of the policyholder) have been discretized so that explanatory variables are categorical.

More precisely, pricing non-life insurance relies on estimating the claim frequency and the claim severity distributions. The former is generally estimated by a regression model such as Poisson or zero-inflated models. However for modeling claim severity, we commonly split the claim dataset between attritional and atypical claims. A threshold μ is chosen either from the extreme value theory or by expert judgments. A classical GLM such as gamma or inverse-Gaussian is fitted on attritional claim amounts below μ, see e.g. Ohlsson and Johansson (2010). Atypical claim amounts above μ are not necessarily modeled at all. An empirical rule of the insurance pricing is used to mutualize atypical claims over the portfolio, i.e. the aggregate sum of atypical claims is shared equally among all policies. We aim at providing a regression model for those claims above μ in order to refine this empirical rule.

The threshold μ can also be interpreted in another insurance context. Generally in non-life insurance, contracts are underwritten with a deductible. This has two consequences: the policyholder will retain the risk of claims below the deductible; and the insurer will only know and be interested in claims above the deductible. In the numerical section, we consider only the example of large claim modeling.

The paper is organized as follows. In Section 2, we present the GLMs. Section 3 provide exact formulas for MLE in the case of categorical explanatory variables. Section 4 is dedicated to the Pareto 1 GLM, while Section 5 is dedicated to the shifted lognormal GLM. Finally, a simulation analysis is provided in Section 6 and an application to an actuarial dataset is carried out in Section 7, before Section 8 conludes.

2 Preliminaries on Generalized Linear Models

In this section, we consider the estimation problem in GLMs. 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 $i = 1, \dots, n$.

In the following, for the sake of clarity, bold notations are reserved for vector of $\mathbb{R}^p$ and bold notations with an underline are reserved for vector of $\mathbb{R}^n$. The index $i \in I = \{1, \dots, n\}$ is reserved for the observations, while the indexes j, k, l are used for the explanatory variables.

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

Precisely, the likelihood L associated to the statistical experiment generated by Y_i, $i \in I$

verifies

$$\log L(\boldsymbol{\vartheta} \,|\, y_i) = \frac{\lambda_i(\boldsymbol{\vartheta}) y_i - b\left(\lambda_i(\boldsymbol{\vartheta})\right)}{a(\phi)} + c(y_i, \phi), \quad y_i \in \mathbb{Y} \subset \mathbb{R}, \tag{1}$$

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

In Equation (1), the parameters $\lambda_1, \ldots, \lambda_n$ depend on a finite-dimensional parameter $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^p$. Direct computations lead to

$$b'(\lambda_i(\boldsymbol{\vartheta})) = \mathbf{E}_{\boldsymbol{\vartheta}} Y_i \quad \text{and} \quad b''(\lambda_i(\boldsymbol{\vartheta}))a(\phi) = \mathbf{Var}_{\boldsymbol{\vartheta}} Y_i. \tag{2}$$

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

$$g(b'(\lambda_i(\boldsymbol{\vartheta}))) = \langle \boldsymbol{x}_i, \boldsymbol{\vartheta} \rangle = \eta_i, \quad \text{for all } \boldsymbol{\vartheta} \in \Theta,$$

where η_i are the linear predictors and $\langle ., . \rangle$ denotes the scalar product. In other words, the bijective function $\ell = (b')^{-1} \circ g^{-1}$ is setted; then we have

$$\lambda_i(\boldsymbol{\vartheta}) = \ell(\eta_i). \tag{3}$$

We summarize with the following relations

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

where D is the space of linear predictor and X the possible set of value of $\boldsymbol{x}_i$ for $i \in I$. Here ℓ is chosen and, consecutively Θ, Λ and D must be set.

The parameter $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^p$ is to be estimated and g is called the link function in the regression framework. We talk of canonical link function, when ℓ is the identity function.

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

$$\log L(\boldsymbol{\vartheta} \,|\, \underline{\boldsymbol{y}}) = \sum_{i=1}^{n} \frac{y_i \ell(\eta_i) - b\left(\ell(\eta_i)\right)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi), \tag{4}$$

with b, h and ℓ being respectively defined in (1) and (3). Here, the vector of the parameters $\boldsymbol{\vartheta}$ is unknown. If the model is identifiable, it can be shown that the sequence of MLE $(\widehat{\boldsymbol{\vartheta}}_n)_{n \geq 1}$ defined by $\widehat{\boldsymbol{\vartheta}}_n = \arg\max_{\boldsymbol{\vartheta} \in \Theta} L(\boldsymbol{\vartheta} \,|\, \underline{\boldsymbol{y}})$ asymptotically exists and is consistent (for example Fahrmeir and Kaufmann, 1985, Theorem 2, 4).

The MLE $\widehat{\boldsymbol{\vartheta}}_n$, if it exists, is the solution of the non linear system

$$S_j(\boldsymbol{\vartheta}) = 0, \quad j = 1, \ldots, p, \tag{5}$$

with $S_j(\boldsymbol{\vartheta})$ are the component of the Score vector defined by

$$S_j(\boldsymbol{\vartheta}) = \frac{1}{a(\phi)} \sum_{i=1}^{n} x_i^{(j)} \ell'(\eta_i) \left(y_i - b'\left(\ell(\eta_i)\right)\right).$$

It is worth mentioning that for a small data set (small n) or large number of explanatory variables, the existence of the MLE is not guaranteed. Note that the MLE $\widehat{\boldsymbol{\vartheta}}_n$ does not depend on the value of the dispersion parameter ϕ. Indeed, the dispersion parameter is estimated in a second step using the sum of square residuals or the log-likelihood, see e.g. (McCullagh and Nelder, 1989, Chap. 9).

In a general setting, the system (5) does not have a closed-form solution and GLMs are generally fitted using a Newton-type method, such that an iteratively re-weighted least square (IWLS) algorithm also refereed to Fisher Scoring algorithm, see e.g. McCullagh and Nelder (1989).

In the case of categorical explanatory variables described later on, the non-asymptotic existence of the MLE depends on the conditional distribution and the chosen link function (see Examples 1, 2 and 3 on Section 4.2).

3 A closed-form MLE for categorical explanatory variables

In any regression model, categorical or nominal explanatory variables have to be coded since their value is a name or a category. When the possible values are unordered, it is common to use a binary incidence matrix or dummy variables where each row has a single unity in the column of the class to which it belongs. In the case of ordered values, a contrast matrix has to be used, see e.g. Venables and Ripley (2002).

3.1 A single explanatory variable

Let us first consider the case of a single categorical explanatory variable. That is $p = 2$ and for all $i \in I$, $x_i^{(1)} = 1$ is the intercept and $x_i^{(2)}$ takes values in a set of d modalities $\{v_1, \ldots, v_d\}$ with $d > 2$. We define the incidence matrix $(x_i^{(2),j})_{i,j}$ where $x_i^{(2),j} = \mathbf{1}_{x_i^{(2)}=v_j}$ is the binary dummy of the jth category for $i \in I$ and $j \in J = \{1, \ldots, d\}$. From this incidence matrix, we compute the number of appearance $m_j > 0$ of the jth category and $\overline{y}_n^{(j)}$ the mean value of $\underline{\boldsymbol{y}}$ taking over the jth category by

$$m_j = \sum_{i=1}^{n} x_i^{(2),j}, \quad j \in J \quad \text{and} \quad \overline{y}_n^{(j)} = \frac{1}{m_j} \sum_{i=1}^{n} y_i x_i^{(2),j}, \quad j \in J.$$

By construction, this incidence matrix has rows that sum to 1. Therefore if we use the combination of the incidence matrix with a 1-column for the intercept $(x_i^{(1)}, x_i^{(2),j})_{i,j}$: a redundancy appears. We must choose either to use *no intercept*, to *drop one column* for a particular modality of $x_i^{(2)}$, or to use a *zero-sum condition* on the parameters. We investigate below these three options in a single framework.

Consider the following GLM for the explanatory variables $x_i^{(1)}, x_i^{(2),1}, \ldots, x_i^{(2),d}$ assuming that

$$g(\mathbf{E}Y_i) = \vartheta_{(1)} + \sum_{j=1}^{d} x_i^{(2),j} \vartheta_{(2),j}, \quad i \in I, \tag{6}$$

where $\boldsymbol{\vartheta} = (\vartheta_{(1)}, \vartheta_{(2),1}, \ldots, \vartheta_{(2),d})$ is the unknown vector parameters. The model being not identifiable, we impose exactly one linear equation on $\boldsymbol{\vartheta}$

$$\langle \boldsymbol{R}, \boldsymbol{\vartheta} \rangle = 0, \tag{7}$$

with $\boldsymbol{R} = (r_0, r_1, \ldots, r_d)$ any real vector of size $d + 1$. A theorem and two corollaries are given below and corresponding proofs are postponed to Appendix A.1.

Theorem 3.1. *Suppose that for all $i \in I$, Y_i takes values in $b'(\Lambda)$. If the vector $\boldsymbol{R}$ is such that $\sum_{j=1}^{d} r_j - r_0 \neq 0$, then there exists a unique, consistent and explicit MLE $\widehat{\boldsymbol{\vartheta}}_n =$*

$(\widehat{\boldsymbol{\vartheta}}_{n,(1)}, \widehat{\boldsymbol{\vartheta}}_{n,(2),1}, \ldots, \widehat{\boldsymbol{\vartheta}}_{n,(2),d})$ *of* $\boldsymbol{\vartheta}$ *given by*

$$\widehat{\boldsymbol{\vartheta}}_{n,(1)} = \frac{\sum_{k=1}^{d} r_k g(\overline{Y}_n^{(k)})}{\sum_{k=1}^{d} r_k - r_0}, \widehat{\boldsymbol{\vartheta}}_{n,(2),j} = g(\overline{Y}_n^{(j)}) - \frac{\sum_{k=1}^{d} r_k g(\overline{Y}_n^{(k)})}{\sum_{k=1}^{d} r_k - r_0}, j \in J. \tag{8}$$

Note that if $\overline{Y}_n^{(j)}$ *does not belong to* $b'(\Lambda)$, $g(\overline{Y}_n^{(j)})$ *and hence* $\widehat{\boldsymbol{\vartheta}}_{n,(l),j}$ *are not defined.*

We give below the three most common examples of linear constraint, some details of these calculus are given in Appendix A.1.

Example 3.1. *No-intercept model*

The no-intercept model is obtained with $\boldsymbol{R} = (1, 0, \ldots, 0)$ *leading to* $\vartheta_{(1)} = 0$. *Therefore the unique, consistent and explicit MLE* $\widehat{\boldsymbol{\vartheta}}_n$ *of* $\boldsymbol{\vartheta}$ *is*

$$\widehat{\vartheta}_{n,(1)} = 0, \quad \widehat{\vartheta}_{n,(2),j} = g\left(\overline{Y}_n^{(j)}\right), j \in J. \tag{9}$$

Example 3.2. *Model without first modality*

The model without first modality is obtained with $\boldsymbol{R} = (0, 1, \ldots, 0)$ *leading to* $\vartheta_{(2),1} = 0$. *Therefore, the unique, consistent and explicit MLE* $\widehat{\boldsymbol{\vartheta}}_n$ *of* $\boldsymbol{\vartheta}$ *is*

$$\widehat{\vartheta}_{n,(1)} = g\left(\overline{Y}_n^{(1)}\right), \quad \widehat{\vartheta}_{n,(2),1} = 0, \quad \widehat{\vartheta}_{n,(2),j} = g\left(\overline{Y}_n^{(j)}\right) - \widehat{\vartheta}_{n,1}, j \in J \setminus \{1\}.$$

Example 3.3. *Zero-sum condition*

The zero-sum model is obtained with $\boldsymbol{R} = (0, 1, \ldots, 1)$ *leading to* $\sum_{j=1}^{d} \vartheta_{(2),j} = 0$. *Therefore, the unique, consistent and explicit MLE* $\widehat{\boldsymbol{\vartheta}}_n$ *of* $\boldsymbol{\vartheta}$ *is*

$$\widehat{\vartheta}_{n,(1)} = \frac{1}{d}\sum_{k=1}^{d} g\left(\overline{Y}_n^{(k)}\right), \ \widehat{\vartheta}_{n,(2),j} = g\left(\overline{Y}_n^{(j)}\right) - \widehat{\vartheta}_{n,1}, \quad j \in J.$$

Remark 3.1. *In Theorem 3.1, it is worth noting that the value of* $\widehat{\boldsymbol{\vartheta}}_n$ *does not depend on the distribution of the* Y_i. *This fact was stated in Goldburd et al. (2016) but without any proof.*

Remark 3.2. *The three different parametrizations (Examples 3.1, 3.2 and 3.3) depends on the type of application and on the modeler choice. In statistical software, there is a default choice: for instance in the statistical software R, the model without the first modality is the default parametrization (see functions* `lm()`, `glm()` *by R Core Team (2019)). The first option without intercept may be justified when no group can be chosen as the reference group.*

Remark 3.3. *When* g *is the identity function, the first and third options (Examples 3.1 and 3.3) can be interpreted as a generalized analysis of variance (ANOVA) for* Y_i *with respect to groups defined by the explanatory variable* $\boldsymbol{x}^{(2)}$. *Even for non-Gaussian random variables, some applications may justify these options.*

Remark 3.4. *The case when there is no explanatory variable, i.e.* Y_i *are identically distributed, cannot be obtained with Equation (8). But in that case, we get with similar arguments* $\widehat{\boldsymbol{\vartheta}}_{n,(1)} = g(\overline{Y}_n)$, *which is in line with Example 6.3.12 of Lehmann and Casella (1998) when* g *is the canonical link, see Appendix A.1.*

Remark 3.5. *Despite the distribution of* $\overline{Y}_n^{(j)}$ *still belongs to the exponential family, the bias of* $\widehat{\vartheta}_{n,(1)}$ *and* $\widehat{\vartheta}_{n,(2),j}$ *cannot be determined for a general link function* g*. However, we can show the consistency of the estimator and an asymptotic confidence interval, see Appendix A.1. In the following, we will investigate the bias and the distribution of the MLE for some special cases of distributions and link functions.*

Theorem 3.1 has two interesting corollaries which give some clues on the choice of the link function g. This corollary tempers the importance of the link function since it will not affect the predicted moments in the case of a single explanatory variable.

Corollary 3.1. *The value of the log-likelihood defined in (4) taken on the exact MLE* $\widehat{\boldsymbol{\vartheta}}_n$ *(if it exists) given by (8), under constraint (7), does not depend on the link function* g*. More precisely, we have* $\forall i \in I, \quad \ell(\hat{\eta}_i) = (b')^{-1}(\overline{y}_n^{(j)})$ *for* $j \in J$ *such that* $x_i^{(2),j} = 1$ *and*

$$\log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{y}}) = \frac{1}{a(\phi)} \sum_{j=1}^{d} \sum_{i, x_i^{(2),j}=1} \left(y_i \tilde{b}\left(\overline{y}_n^{(j)}\right) - b\left(\tilde{b}\left(\overline{y}_n^{(j)}\right)\right)\right) + \sum_{i=1}^{n} c(y_i, \phi),$$

with $\tilde{b} = (b')^{-1}$*. Therefore, the criteria AIC and BIC are also independent of the link function* g*. The estimator of* ϕ *is obtained by maximizing* $\log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{y}})$ *with respect to* ϕ *given* a, b, c *functions.*

Corollary 3.2. *The predicted mean and predicted variance for the* i*th individual is estimated by* $\widehat{\mathbf{E}Y_i} = b'(\ell(\widehat{\eta}_i))$ *and* $\widehat{\mathbf{Var}Y_i} = a(\widehat{\phi})b''(\ell(\widehat{\eta}_i))$ *respectively using (2). Both estimates do not depend on the link function* g *and the predicted mean does not depend on the function* b*. More precisely, when* v_j *is the modality of the* i*th individual (i.e.* $x_i^{(2),j} = 1$*), the predicted mean and predicated variance are given by*

$$\widehat{\mathbf{E}Y_i} = \overline{y}_n^{(j)}, \ \widehat{\mathbf{Var}Y_i} = a(\widehat{\phi})b'' \circ (b')^{-1}\left(\overline{y}_n^{(j)}\right).$$

Corollary 3.2 may be surprising because the predicted mean does not depend on the conditional distribution of $Y_1, \ldots, Y_n$. The predicted mean is just the mean of the response variable taken over the class j i.e. observations y_i such that the covariate $x_i^{(2)}$ takes the modality v_j.

The formula (8) of the MLE $\widehat{\boldsymbol{\vartheta}}_n$ can be reformulated as

$$\widehat{\boldsymbol{\vartheta}}_n = \begin{pmatrix} \boldsymbol{Q} \\ \boldsymbol{R} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{g}(\bar{\boldsymbol{Y}}) \\ 0 \end{pmatrix}$$

with $\boldsymbol{Q}$ is the $d \times (1+d)$ matrix defined by $\boldsymbol{Q} = (A_0, A_1)$, with A_0 is the ones vector of size d, A_1 the identity matrix of size d, and $\boldsymbol{g}(\bar{\boldsymbol{Y}})$ the vector $(g(\bar{Y}_n^{(1)}), \ldots, g(\bar{Y}_n^{(d)}))$. We have

$$\sum_{j=1}^{d} r_j - r_0 \neq 0 \Leftrightarrow \operatorname{rank} \begin{pmatrix} \boldsymbol{Q} \\ \boldsymbol{R} \end{pmatrix} = d + 1.$$

With this formulation, the estimator of $\boldsymbol{\vartheta}$ is $\widehat{\boldsymbol{\vartheta}}_n = (\boldsymbol{Q}'\boldsymbol{Q} + \boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{Q}'\boldsymbol{g}(\bar{\boldsymbol{Y}})$. This general form is particularly useful in the case of two categorical variables of the next subsection.

3.2 Two explanatory variables

Now, we consider the case of two explanatory categorical variables. That is $p = 3$ and for all $i = 1, \ldots, n$, $x_i^{(1)} = 1$ is the intercept and $x_i^{(2)}, x_i^{(3)}$ take values in $\{v_{j1}, \ldots, v_{jd_j}\}$ with d_2, d_3

modalities respectively. We define by $x_i^{(2),k}$ and $x_i^{(3),l}$, $k \in K = \{1, \ldots, d_2\}$ and $l \in L = \{1, \ldots, d_3\}$ the binary dummies of the kth and lth resp. categories, $m_k^{(j)} > 0$ the number of appearance of the kth modality of the jth variable, $j = 1, 2$, $m_{k,l}$ the number of appearance of the kth and lth category simultaneously and $\overline{y}_n^{(k,l)}$ the mean value of $\underline{\boldsymbol{y}}$ taking over the kth and lth categories. That is

Dummy	Frequency	Mean	Index
$x_i^{(2),k} = 1_{x_i^{(2)}=v_{2k}}$	$m_k^{(2)} = \sum_{i=1}^n x_i^{(2),k}$	$\overline{y}_n^{(2),k} = \frac{1}{m_k^{(2)}} \sum_{i=1}^n y_i x_i^{(2),k}$	$k \in K$
$x_i^{(3),l} = 1_{x_i^{(3)}=v_{3l}}$	$m_l^{(3)} = \sum_{i=1}^n x_i^{(3),l}$	$\overline{y}_n^{(3),l} = \frac{1}{m_l^{(3)}} \sum_{i=1}^n y_i x_i^{(3),l}$	$l \in L$
$x_i^{(k,l)} = x_i^{(2),k} x_i^{(3),l}$	$m_{k,l} = \sum_{i=1}^n x_i^{(k,l)}$	$\overline{y}_n^{(k,l)} = \frac{1}{m_{k,l}} \sum_{i=1}^n y_i x_i^{(k,l)}$	$(k,l) \in K \times L$

where $\overline{y}_n^{(k,l)}$ is computed over $KL^\star = (K \times L) \setminus \{(k,l) \in K \times L; m_{k,l} = 0\}$. Set $d_{2,3}^\star = \#KL^\star$, for $l \in L$, $K_l^\star = \{k \in K; m_{k,l} > 0\}$, $d_{(3),l}^\star = \#K_l^\star$ and for $k \in K$, $L_k^\star = \{l \in L; m_{k,l} > 0\}$, $d_{(2),k}^\star = \#L_k^\star$.

Note that $(m_{k,l})_{kl}$ are absolute frequencies of the contingency table resulting from cross-classifying factors and can be computed very easily. Be careful that $KL^\star$ is not equal to $K \times L$ but $\bigcup_{l \in L} K_l^\star = \bigcup_{k \in K} L_k^\star = KL^\star$, and $d_{2,3}^\star = d_2 d_3 - r$, where $r = \#\{(k,l) \in K \times L; m_{k,l} = 0\}$.

Let $\boldsymbol{Q}$ be the $d_{2,3}^\star \times (1 + d_2 + d_3 + d_{2,3}^\star)$ real matrix defined by $\boldsymbol{Q} = (A_0, A_1, A_2, A_{12})$ with $A_0 = \mathbf{1}_{d_{2,3}^\star}$ the $d_{2,3}^\star \times 1$ ones matrix; $A_1 = (\mathrm{diag}(\mathbf{1}_{d_{(2),k}^\star}))_{k \in K}$, the $d_{2,3}^\star \times K$ diagonal block matrix of ones vector of size $d_{(2),k}^\star$; $A_2 = (I_{d_3}^{\star,k})_{k \in K}$, the $d_{2,3}^\star \times L$ matrix where $I_{d_3}^{\star,k}$ is the identity matrix of size d_3 without rows l for which $m_{k,l} = 0$; $A_{12} = I_{d_{2,3}^\star}$ the $d_{2,3}^\star \times d_{2,3}^\star$ identity matrix.

Consider the following GLM for explanatory variables $x_i^{(1)}, x_i^{(2),j}, x_i^{(3),j}$

$$g(\mathbf{E}Y_i) = \vartheta_1 + \sum_{k=1}^{d_2} x_i^{(2),k} \vartheta_{(2),k} + \sum_{l=1}^{d_3} x_i^{(3),l} \vartheta_{(3),l} + \sum_{(k,l) \in KL^\star} x_i^{(k,l)} \vartheta_{kl}, \tag{10}$$

where $\vartheta_{(1)}, (\vartheta_{(2),k})_{k \in K}, (\vartheta_{(3),l})_{l \in L}, (\vartheta_{kl})_{(k,l) \in KL^\star}$ are the $d_2 + d_3 + d_{2,3}^\star + 1$ unknown parameters. Again at this stage, the model is not identifiable because of the redundancy on the vectors $(x_1^{(2),k}, \ldots, x_n^{(2),k})$, $k \in K$, the vectors $(x_1^{(3),l}, \ldots, x_n^{(3),l})$, $l \in L$ and the ones vector. As previously, we need to impose $q \geq 1 + d_2 + d_3$ linear constraints on the vector parameters $\boldsymbol{\vartheta}$

$$\boldsymbol{R}\boldsymbol{\vartheta} = \mathbf{0}_q, \tag{11}$$

where $\boldsymbol{R}$ is a $q \times (1 + d_2 + d_3 + d_{2,3}^\star)$ real matrix of linear contrasts, with $rank(\boldsymbol{R}) = 1 + d_2 + d_3$ and $\mathbf{0}_q$ the zeros vector of size q. Again, the proofs of the following theorem and corollaries are postponed to Appendix A.2.

Theorem 3.2. *Suppose that for all $i \in \{1, \ldots, n\}$, Y_i takes values in $b'(\Lambda)$. Under constraint (11) and if $\boldsymbol{R}$ such that $(\boldsymbol{Q}', \boldsymbol{R}')$ is of rank $d_{2,3}^\star$, there exists a unique, consistent and explicit MLE $\widehat{\boldsymbol{\vartheta}}_n$ of $\boldsymbol{\vartheta}$ given by*

$$\widehat{\boldsymbol{\vartheta}}_n = (\boldsymbol{Q}'\boldsymbol{Q} + \boldsymbol{R}'\boldsymbol{R})^{-1} \boldsymbol{Q}' \boldsymbol{g}(\bar{\boldsymbol{Y}}), \tag{12}$$

where the vector $\boldsymbol{g}(\bar{\boldsymbol{Y}})$ is $((g(\bar{Y}_n^{(k,l)}))_{l \in L_k^\star})_{k \in K}$.

Example 3.4. *No intercept and no single-variable dummy*

The model with no intercept and no single-variable dummy is $\vartheta_1 = 0$ and $\vartheta_{(2),k} = \vartheta_{(3),l} = 0$ $\forall k \in K \, \forall l \in L$. Therefore, the unique, consistent and explicit MLE $\widehat{\boldsymbol{\vartheta}}_n$ of $\boldsymbol{\vartheta}$ is

$$\widehat{\vartheta}_{n,kl} = g\left(\overline{Y}_n^{(k,l)}\right), \quad (k,l) \in KL^\star.$$

Example 3.5. *Zero-sum conditions The model with zero-sum conditions assumes*

$$\sum_{k\in K} m_k^{(2)}\vartheta_{(2),k} = \sum_{l\in L} m_l^{(3)}\vartheta_{(3),l} = 0,$$

$$\forall l \in L, \sum_{k\in K_l^\star} m_{k,l}\vartheta_{kl} = 0,\ \forall k \in K, \sum_{l\in L_k^\star} m_{k,l}\vartheta_{kl} = 0.$$

Therefore, the unique, consistent and explicit MLE $\widehat{\boldsymbol{\vartheta}}_n$ *of* $\boldsymbol{\vartheta}$ *is*

$$\begin{cases} \widehat{\vartheta}_{n,(1)} = \dfrac{1}{n}\displaystyle\sum_{(k,l)\in KL^\star} m_{k,l} g\left(\overline{Y}_n^{(k,l)}\right) \\ \widehat{\vartheta}_{n,(2),k} = \dfrac{1}{m_k^{(2)}}\displaystyle\sum_{l\in L_k^\star} m_{k,l} g\left(\overline{Y}_n^{(k,l)}\right) - \widehat{\vartheta}_{n,1}, \quad k \in K \\ \widehat{\vartheta}_{n,(3),l} = \dfrac{1}{m_l^{(3)}}\displaystyle\sum_{k\in K_l^\star} m_{k,l} g\left(\overline{Y}_n^{(k,l)}\right) - \widehat{\vartheta}_{n,1}, \quad l \in L \\ \widehat{\vartheta}_{n,kl} = g\left(\overline{Y}_n^{(k,l)}\right) - \widehat{\vartheta}_{n,(2),k} - \widehat{\vartheta}_{n,(3),l} - \widehat{\vartheta}_{n,1}, \quad (k,l) \in KL^\star. \end{cases}$$

Remark 3.6. *The MLE of the model with only main effects for two explanatory variables defined as* $g(\mathbf{E}Y_i) = \vartheta_1 + \sum_{k=1}^{d_2} x_i^{(2),k}\vartheta_{(2),k} + \sum_{l=1}^{d_3} x_i^{(3),l}\vartheta_{(3),l}$ *does not present such explicit formula whatever the matrix* $\boldsymbol{R}$ *of rank 2. In that case, the MLE does not solves a least square problem under a linear constraint, see Appendix A.2. In the special case of logit-regression, Lipovetsky (2015) also notice that least square estimation does not coincide with the MLE.*

For simplicity, we consider only the cases of one and two explanatory categorical variables. With a higher number of explanatory variables, we can perform a similar analysis to obtain an explicit solution of the MLE. As for one explanatory variable, Theorem 3.2 has two interesting corollaries on the value of the log-likelihood and the predicted moments.

Corollary 3.3. *The value of log-likelihood defined in (4) taken on the exact MLE* $\widehat{\boldsymbol{\vartheta}}_n$ *(if it exists) given by (12), under constraint (11), does not depend on the link function* g*. More precisely, we have* $\forall i \in I, \quad \ell(\hat{\eta}_i) = (b')^{-1}(\overline{y}_n^{(k,l)})$ *for* $l \in L$ *and* $k \in K$ *such that* $x_i^{(2),j} = 1$ *and* $x_i^{(3),k} = 1$ *and*

$$\log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{y}}) = \frac{1}{a(\phi)} \sum_{(k,l)\in KL^\star} \sum_{i\in\tilde{I}} \left(y_i \tilde{b}\left(\overline{y}_n^{(k,l)}\right) - b\left(\tilde{b}\left(\overline{y}_n^{(k,l)}\right)\right)\right) + \sum_{i=1}^{n} c(y_i,\phi),$$

with $\tilde{b} = (b')^{-1}$ *and* $\tilde{I} = \{i \in I, x_i^{(2),k} = x_i^{(3),l} = 1\}$*. The estimator of* ϕ *is obtained by maximizing* $\log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{y}})$ *with respect to* ϕ *given* a, b, c *functions.*

Corollary 3.4. *The predicted mean and predicted variance for the* i*th individual is estimated by* $\widehat{\mathbf{E}Y_i} = b'(\ell(\widehat{\eta}_i))$ *and* $\widehat{\mathbf{Var}Y_i} = a(\widehat{\phi})b''(\ell(\widehat{\eta}_i))$ *respectively using (2). Both estimates do not depend on the link function* g *and the predicted mean does not depend on the function* b*. Let* v_{2k} *and* v_{3l} *be the modalities of the* i*th individual of the two explanatory variables, i.e.* $x_i^{(2),k} = 1$ *and* $x_i^{(3),l} = 1$*. The predicted mean and variance are given by*

$$\widehat{\mathbf{E}Y_i} = \overline{y}_n^{(k,l)}, \ \widehat{\mathbf{Var}Y_i} = a(\widehat{\phi})b'' \circ (b')^{-1}(\overline{y}_n^{(k,l)}).$$

In the next two sections, we apply previous theorems and corollaries to two particular distributions: Pareto 1 and lognormal distribution. Our results do not only apply to continuous distributions but also for discrete distributions. But we choose these two distributions in order to model insurance losses.

4 GLM for Pareto I distribution with categorical explanatory variables

4.1 Characterization

Consider the sample $\underline{\boldsymbol{Y}} = (Y_1, \ldots, Y_n)$ composed of independent Pareto Type 1. Precisely, we assume that the independent random variables $Y_1, \ldots, Y_n$ are Pareto with known threshold parameter μ and respective shape parameter (depending on the unknown parameter $\boldsymbol{\vartheta}$) $\lambda_1(\boldsymbol{\vartheta}), \ldots, \lambda_n(\boldsymbol{\vartheta}) \in \Lambda = (0, \infty)$. The density f of Pareto distribution with threshold and shape parameter μ and $\lambda_i(\boldsymbol{\vartheta})$, $i \in I$ is

$$f(y) = \lambda_i(\boldsymbol{\vartheta}) \frac{\mu^{\lambda_i(\boldsymbol{\vartheta})}}{y^{\lambda_i(\boldsymbol{\vartheta})+1}}, \quad y \in \mathbb{Y} = [\mu, \infty), \tag{13}$$

and 0 if $y < \mu$.

We recall that for the Pareto Type 1 distribution

$$\mathbf{E}Y_i = \frac{\lambda_i(\boldsymbol{\vartheta})\mu}{\lambda_i(\boldsymbol{\vartheta}) - 1} < +\infty, \text{iff } \lambda_i(\boldsymbol{\vartheta}) > 1 \text{ and } \mathbf{E}Y_i^2 = \frac{\lambda_i(\boldsymbol{\vartheta})\mu^2}{\lambda_i(\boldsymbol{\vartheta}) - 2} < +\infty, \text{iff } \lambda_i(\boldsymbol{\vartheta}) > 2.$$

Unlike the known parameter μ, the parameter $\boldsymbol{\vartheta}$ is to be estimated. These closed-form formulas are particularly useful in an insurance context since the expectation and the variance are used in most premium computation. For instance, $\mathbf{E}Y$ is the pure premium and for $\gamma > 0$, $\mathbf{E}Y + \gamma \mathbf{Var}Y$ is the variance principle (see Bühlmann and Gisler, 2006, Section 1.2.2).

In the following, instead of $\underline{\boldsymbol{Y}}$ we consider the sample $\underline{\boldsymbol{Z}} = (T(Y_1), \ldots, T(Y_n))$. With the re-parametrization $z_i = T(y_i) = -\log(y_i/\mu)$, $i \in I$, this distribution belongs to the exponential family as defined in (1), with

$$a(\phi) = 1, b(\lambda) = -\log(\lambda), \text{ and } c(z, \phi) = 0, \quad z \in T(\mathbb{Y}) = \mathbb{R}^-, \lambda \in \Lambda. \tag{14}$$

In particular, for the Pareto I distribution, there is no dispersion parameter. It is also worth mentioning that $-Z_i$ is exponential with parameter $\lambda_i(\boldsymbol{\vartheta})$. Consecutively, all moments of Z_i exist and are given by $\mathbf{E}(Z_i)^m = (-1)^m m!/\lambda_i(\boldsymbol{\vartheta})^m$, $m \in \mathbb{N}^*$.

Consider the regression model with a link function g, a response variable Y_i Pareto I distributed where $Z_i = -\log(Y_i/\mu)$ and

$$g(\mathbf{E}Z_i) = \vartheta_{(1)} + x_i^{(2),1}\vartheta_{(2),1} + \ldots + x_i^{(2),d}\vartheta_{(2),d} = \langle \boldsymbol{x}_i, \boldsymbol{\vartheta} \rangle, \quad i \in I \tag{15}$$

with for $i \in I$, $\boldsymbol{x}_i = (1, x_i^{(2),1}, \ldots, x_i^{(2),d})^T$ are the covariate vectors and $\boldsymbol{\vartheta} = (\vartheta_{(1)}, \vartheta_{(2),1} \ldots, \vartheta_{(2),d})^T$ is the unknown parameter vector.

The choice of the link function g appearing in (15) is a crucial point. Let us start with the canonical link. That is, the chosen function g so that $\ell = (b')^{-1} \circ g^{-1}$ is the identity function. For our Pareto model, $g(t) = -\frac{1}{t}$ since $b'(\lambda) = -\frac{1}{\lambda}$. From (4), the choice of the canonical link function imposes constraints on the linear predictor space $D \subset \Lambda \subset (0, +\infty)$ in that case. Since D results from the scalar product of $\boldsymbol{\vartheta}$ parameters with explanatory variables $\boldsymbol{y}_i$, some negative values might be produced when the covariables take negative values. This make the choice of the canonical link inappropriate.

In order to remedy this issue, we can choose a link function such that the values of the function ℓ falls in $\Lambda \subset (0, \infty)$. A natural choice is $\ell(\eta) = \exp(\eta)$ which guarantees a positive parameter. The choice $\ell(\eta) = \exp(\eta) + 1$ guarantee a finite expectation for the random variables Y_i, $i = 1, \ldots, n$. We summarize in Table 1 the tested ℓ functions in our application in Section 7, and in Table 2, the spaces given a link function.

Table 1: Table of typical link functions for Pareto I

Names	$\ell(\eta_i)$	$g^{-1}(t)$	$g(t)$
canonical	η_i	$-\frac{1}{t}$	$-\frac{1}{t}$
log-inv	e^{η_i}	$-e^{-t}$	$\log(-\frac{1}{t})$
shifted log-inv	$e^{\eta_i}+1$	$-\frac{1}{e^t+1}$	$\log(-\frac{1}{t}-1)$

Table 2: Summary of spaces for Pareto I

Link name	Parameter–covariable space $\boldsymbol{\vartheta} \times X_i$		Linear predictor space D		Parameter space Λ	$b'(\Lambda)$
unspecified	$\boldsymbol{\vartheta} \times X_i \subset \mathbb{R}^p \times \mathbb{R}^p$	$\xrightarrow{\langle .,. \rangle}$	$D \subset \mathbb{R}$	$\underset{\ell}{\overset{\ell^{-1}}{\leftrightharpoons}}$	$\Lambda \subset (0,+\infty)$	
canonical	$\{(\boldsymbol{\vartheta}, \boldsymbol{x}_i) \in \mathbb{R}^p \times \mathbb{R}^p, \langle \boldsymbol{\vartheta}, \boldsymbol{x}_i \rangle > 0\}$	$\xrightarrow{\langle .,. \rangle}$	$(0,+\infty)$	$\underset{\mathrm{id}}{\overset{\mathrm{id}}{\leftrightharpoons}}$	$(0,+\infty)$	$(-\infty,0)$
log-inv	$\mathbb{R}^p \times \mathbb{R}^p$	$\xrightarrow{\langle .,. \rangle}$	$\mathbb{R}$	$\underset{\exp}{\overset{\log}{\leftrightharpoons}}$	$(0,+\infty)$	$(-\infty,0)$
shifted log-inv	$\mathbb{R}^p \times \mathbb{R}^p$	$\xrightarrow{\langle .,. \rangle}$	$\mathbb{R}$	$\underset{\exp(x)+1}{\overset{\log(x-1)}{\leftrightharpoons}}$	$(1,+\infty)$	$(-1,0)$

4.2 Estimation for categorical exogenous variables

Consider the case of one categorical exogenous variable. We expose the case of the re-parametrization without intercept, i.e. $\langle \boldsymbol{R}, \boldsymbol{\vartheta} \rangle = 0$ with $\boldsymbol{R} = (1, 0, \ldots, 0)$.

Let $\widehat{\boldsymbol{\vartheta}}_n$ the MLE defined in (9), if it exists, of $\boldsymbol{\vartheta}$. Using the following equality of Corollary 3.1

$$\sum_{j=1}^{d} \sum_{i, x_i^{(2),j}=1} z_i \tilde{b}\left(\overline{z}_n^{(j)}\right) = n, \text{ with } \tilde{b} = (b')^{-1},$$

the log likelihood evaluated on $\widehat{\boldsymbol{\vartheta}}_n$ for both the transformed sample $\underline{\boldsymbol{z}}$ and the original sample $\underline{\boldsymbol{y}}$ with one categorical exogenous variable is

$$\log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{z}}) = n - \sum_{j=1}^{d} m_j \log(-\overline{z}_n^{(j)}), \;\; \log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{y}}) = \log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{z}}) - \sum_{i=1}^{n} \log(y_i). \tag{16}$$

The log-likelihood computation is detailed on Appendix B.

The second remark is that $-Z_i$ are exponentially distributed $\mathcal{E}(\ell(\eta_i))$. Hence for $j \in J$, the estimators $\widehat{\vartheta}_{n,(2),j}$ of $\vartheta_{(2),j}$ are known transforms of a gamma random variable $\mathcal{G}a(m_j, m_j \ell(\vartheta_{2,j}))$. Below we analyze the choice of the link functions considered in Table 1 in Examples 4.1, 4.2, 4.3 and plotted in Figure 4a.

Example 4.1. *canonical link*

In the special case of canonical Pareto model, because $z_i < 0$ for all $i \in I$, we have $\overline{z}_n^{(j)} \in b'(\Lambda) = (-\infty, 0)$ for all $j \in J$ (g and Λ are respectively defined in Tables 1 and 2). With no-intercept using Equation (9), the MLE is

$$\widehat{\vartheta}_{n,(2),j} = -m_j \left(\sum_{i=1}^{n} x_i^{(2),j} Z_i \right)^{-1} = -\frac{1}{\overline{Z}_n^{(j)}}, \quad j \in J.$$

Hence, for $j \in J$, $\widehat{\vartheta}_{n,(2),j}$ follows an Inverse Gamma distribution with shape parameter m_j and rate parameter $m_j\vartheta_{(2),j}$, see e.g. (Johnson et al., 2000, Ch. 17). We can compute the moments of the Inverse Gamma distribution, for $m_j > 2$,

$$\mathbf{E}\widehat{\vartheta}_{n,(2),j} = \frac{m_j}{m_j - 1}\vartheta_{(2),j}, \text{ and } \mathbf{Var}\widehat{\vartheta}_{n,(2),j} = \frac{m_j^2}{(m_j-1)^2(m_j-2)}\vartheta_{(2),j}^2, j \in J.$$

An unbiased estimator of $\vartheta_{(2),j}$ is then $\widehat{\vartheta}^{\star}_{n,(2),j} = \frac{m_j-1}{m_j}\widehat{\vartheta}_{n,(2),j}$ which has a lower variance

$$\mathbf{Var}\widehat{\vartheta}^{\star}_{n,(2),j} = \frac{\vartheta_{(2),j}^2}{m_j - 2} \leq \mathbf{Var}\widehat{\vartheta}_{n,(2),j}, \quad j \in J.$$

A similar bias is also obtained by Bühlmann and Gisler (2006) in a credibility context. Of course this unbiased estimator is also applicable for two exogenous variables with the first parametrization of the Theorem 3.2. When some modalities (or couple of modalities) aren't much represented, it can be relevant to use this unbiased estimator.

Example 4.2. *log-inverse link*

In the special case of the log-inv Pareto model, we also have $\overline{z}_n^{(j)} \in b'(\Lambda) = (-\infty, 0)$ for all $j \in J$ (g and Λ are respectively defined in Tables 1 and 2). With no-intercept using Equation (9), the MLE is

$$\widehat{\vartheta}_{n,(2),j} = -\log\left(\frac{1}{-m_j}\sum_{i=1}^{n} x_i^{(2),j} Z_i\right) = -\log\left(-\overline{Z}_n^{(j)}\right), \quad j \in J.$$

Here, for $j \in J$, the distribution of $-\widehat{\vartheta}_{n,(2),j}$ is the distribution of the log of the gamma distribution with shape m_j and rate $m_j \exp(\vartheta_{(2),j})$. We can derive moments of this distribution which should not be confused with the log-gamma distribution studied e.g. in Hogg and Klugman (1984).

Let $L = \log(G)$ when G is gamma distributed with shape parameter $a > 0$ and rate parameter $\lambda > 0$. We have by elementary manipulations the moment generating function of L:

$$M_L(t) = \mathbf{E}e^{tL} = \frac{\Gamma(a+t)}{\Gamma(a)}\lambda^{-t}, \quad t > -a,$$

where Γ denotes the usual gamma function. Therefore by differentiating and evaluating at 0, we deduce that the expectation and the variance of L are $\mathbf{E}L = \psi(a) - \log\lambda$ and $\mathbf{Var}L = \psi'(a)$, where the functions ψ and ψ' are the digamma and trigamma function, see e.g. Olver et al. (2010). Getting back to our example, we deduce that

$$\mathbf{E}\widehat{\vartheta}_{n,(2),j} = \vartheta_{(2),j} + \log m_j - \psi(m_j) \quad \text{and} \quad \mathbf{Var}\widehat{\vartheta}_{n,(2),j} = \psi'(m_j), \qquad j \in J.$$

From Olver et al. (2010), we know that $\log(m_j) - \psi(m_j)$ tends to 0 as m_j tend to infinity. Hence $\widehat{\vartheta}_{n,(2),j}$ is asymptotically unbiased, and an unbiased estimator of $\vartheta_{(2),j}$ is

$$\widehat{\vartheta}^{\star}_{n,(2),j} = \widehat{\vartheta}_{n,(2),j} - (\log(m_j) - \psi(m_j)), \quad j \in J.$$

Example 4.3. *shifted log-inverse link*

In the special case of the shifted log-inv Pareto model, $\overline{z}_n^{(j)}$ is not necessarily in $b'(\Lambda) = (-1, 0)$ for all $j \in J$ (g and Λ are respectively defined in Tables 1 and 2). If there is an index j such as $\overline{z}_n^{(j)} \leq -1$, the MLE is not defined and we couldn't use the shifted log-inv link with the same incidence matrix.

Nevertheless, for sufficiently large n, for j such that $x_i^{(2)} = v_j$, $\overline{Z}_n^{(j)} \to \mathbf{E}Z_i$ almost surely, where $\mathbf{E}Z_i = -1/(\exp(\vartheta_{2,j}) + 1) > -1$. Hence for sufficiently large n, the conditions of Theorem 3.1 are satisfied. With no-intercept using Equation (9), the MLE (provided it exists) is

$$\widehat{\vartheta}_{n,(2),j} = \log\left(\frac{m_j}{-\sum_{i=1}^n x_i^{(2),j} Z_i} - 1\right) = \log\left(-1/\overline{Z}_n^{(j)} - 1\right), \quad j \in J.$$

The expectation of $\widehat{\vartheta}_{n,(2),j}$ is complex and should be done numerically. However by the strong law of large numbers and the continuity of the link function, $\widehat{\vartheta}_{n,(2),j} = -\log\left(-1/\overline{Z}_n^{(j)} - 1\right)$ converge almost surely to $\log((\exp(\vartheta_{(2),j}) + 1) - 1) = \vartheta_{(2),j}$.

Remark 4.1. *In Theorem 3.1, the condition Y_i takes values in $b'(\Lambda)$ might seem too restrictive. In fact the condition $\overline{y}_n^{(j)} \in b'(\Lambda)$ for all $j \in J$ is sufficient to define a vector value $\widehat{\boldsymbol{\vartheta}}_n$ which maximise the likelihood. But $\widehat{\boldsymbol{\vartheta}}_n$ fails to be a MLE estimator because the random variable $g(\overline{Y}_n^{(j)})$ can to be not defined. Nevertheless, when m_j tends to infinity for any $j \in J$, the random variables Y_i's defined by (6) such that $y_i^{(2),j} = 1$ are i.i.d. (not only independent) and the strong law of large numbers implies that $\overline{Y}_i^{(j)}$ converges almost surely to $\mathbf{E}Y_i = b'(\ell(\eta_i)) \in b'(\Lambda)$. Hence, the probability $P(\overline{Y}_n^{(j)} \notin b'(\Lambda))$ tends to zero which guarantees the asymptotically existence of the MLE estimator.*

4.3 Model diagnostic

In this paragraph, we propose residuals adapted at the case of Pareto problem. First note that Y_i is Pareto I with shape $\ell(\eta_i)$ and threshold μ and for the parametrization (14) $-Z_i = \log(Y_i/\mu) \sim \mathcal{E}(\ell(\eta_i))$. Let define the residuals

$$R_i = -\ell(\eta_i) Z_i, \quad i \in I.$$

Hence $R_1, \ldots, R_n$ are i.i.d. and have an exponential distribution $\mathcal{E}(1)$. The consistency of the MLE makes it possible to say that the estimated residuals $\widehat{R}_{n,i} = -\ell(\widehat{\eta}_i) Z_i$, $i \in I$, with $\widehat{\eta}_i = \langle \boldsymbol{x}_i, \widehat{\boldsymbol{\vartheta}}_n \rangle$ are asymptotically i.i.d..

It is also possible to verify the assumption of the Pareto distribution for Y_i conditionally to $\boldsymbol{y}_i$ with graphical diagnostic as an exponential Quantile-Quantile plot on the residuals $\widehat{R}_{n,i}$.

In the case of a single explanatory variable, for $i \in I$, the residuals $\widehat{R}_{n,i}$ do not depend on ℓ function. Their explicit forms are

$$\widehat{R}_{n,i} = \frac{Z_i}{\bar{Z}_n^{(j)}} \quad j \text{ such that } x_i^{(2)} = v_j, \quad i \in I. \tag{17}$$

Furthermore, the summation of $\widehat{R}_{n,1}, \ldots, \widehat{R}_{n,n}$ has the surprising property to be deterministic and is exactly equal to n.

5 GLM for shifted log-normal distribution with categorical explanatory variables

5.1 Characterization

Secondly, consider the sample $\underline{\boldsymbol{Y}} = (Y_1, \ldots, Y_n)$ composed of independent shifted log-normal variables respectively with mean $\lambda_1(\boldsymbol{\vartheta}), \ldots, \lambda_n(\boldsymbol{\vartheta})$, dispersion $\phi = \sigma^2$ and a known threshold

μ. The shifted log-normal is also known as the 3-parameter log-normal. Precisely, the density of Y_i is

$$f(y) = \frac{1}{(y-\mu)\sqrt{2\pi\phi}} \exp\left(-\frac{(\log(y-\mu) - \lambda_i(\boldsymbol{\vartheta}))^2}{2\phi}\right), \quad y \in \mathbb{Y} = (\mu, \infty), \tag{18}$$

and 0 for $y \leq \mu$. It is well known that the lognormal distribution has finite moment, see e.g. Johnson et al. (2000). In particular, the expectation and the variance are given by

$$\mathbf{E}Y_i = \mu + \exp(\lambda_i(\boldsymbol{\vartheta}) + \phi/2), \quad \mathbf{Var}Y_i = (\exp(\phi) - 1)\exp(2\lambda_i(\boldsymbol{\vartheta}) + \phi).$$

The transformed sample $\underline{\boldsymbol{Z}} = T(\underline{\boldsymbol{Y}}) = (T(Y_1), \ldots, T(Y_n))$ with $T(y) = \log(y-\mu)$ is belongs the exponential family with

$$a(\phi) = \phi, \quad b(\lambda) = \lambda^2/2, \quad c(z,\phi) = -\frac{1}{2}\left(\frac{z^2}{\phi} + \log(2\pi\phi)\right), z \in \mathbb{R}+, \lambda \in \mathbb{R}.$$

It is worth mentioning that Z_i are normally distributed with mean $\lambda_i(\theta)$ and variance ϕ. As a consequence, all moments of Z_i exists and $\mathbf{E}(Z_i - \lambda_i)^m = (2m)!\phi^m/(2^m m!)$ for m even and 0 for m odd. Consider the regression model with a link function g, a response variable Y_i lognormally distributed where $Z_i = \log(Y_i - \mu)$ and

$$g(\mathbf{E}Z_i) = \vartheta_{(1)} + \vartheta_{(2),1}x_i^{(2),1} + \ldots + \vartheta_{(2),d}x_i^{(2),d} = \langle \boldsymbol{x}_i, \boldsymbol{\vartheta} \rangle, \quad i \in I \tag{19}$$

with for $i \in I$, $\boldsymbol{x}_i = (1, x_i^{(2),1}, \ldots, x_i^{(2),d})^T$ are the covariate vectors and $\boldsymbol{\vartheta} = (\vartheta_{(1)}, \vartheta_{(2),1}, \ldots, \vartheta_{(2),d})^T \in \mathbb{R}^d$ is the unknown parameter vector.

The choice of the link function g for Equation (19) is less restrictive than for the Pareto case. Any differentiable invertible function from $\mathbb{R}$ to $\mathbb{R}$ will work. Since $b'(x) = x$, the canonical link function is obtained by choosing g such that $\ell = \mathrm{id} \circ g^{-1} = g^{-1}$ is the identity function. In other words, the canonical link function is the identity function.

Unlike the previous section, any moment of Y_i exist and there is no particular link needed to guarantee their existence. In the numerical application, we will also consider another link function: a real version of the logarithm.

5.2 Estimation for categorical exogenous variables

Again we consider the case of categorical variables and without intercept model, that is with a predictor $\eta_i = x_i^{(2),1}\vartheta_{(2),1} + \cdots + x_i^{(2),d}\vartheta_{(2),d}$. In the case of the lognormal dispersion, there is a dispersion to be estimated. The log-likelihood is given by

$$\log L(\boldsymbol{\vartheta} \,|\, \underline{\boldsymbol{z}}) = -\frac{1}{2\phi}\sum_{i=1}^{n}(z_i - \lambda_i(\boldsymbol{\vartheta}))^2 - \frac{n\log(2\pi\phi)}{2}.$$

The components of the MLE of $\boldsymbol{\vartheta}$ are given by $\widehat{\boldsymbol{\vartheta}}_{n,(2),j} = g(\bar{z}_n^{(j)}), j \in J$, and the estimated log likelihood for the transformed sample $\underline{\boldsymbol{z}}$ is

$$\log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{z}}) = -\frac{1}{2\phi}\sum_{j\in J}\sum_{i, x_i^{(2),j}=1}\left(z_i - \bar{z}_n^{(j)}\right)^2 - \frac{n\log(2\pi\phi)}{2}.$$

Maximizing over ϕ the log-likelihood leads to the empirical variance

$$\widehat{\phi} = \frac{1}{n}\sum_{j\in J}\sum_{i, x_i^{(2),j}=1}\left(z_i - \bar{z}_n^{(j)}\right)^2. \tag{20}$$

Hence the estimator of ϕ is the intra-class variance. This closed-form estimate $\widehat{\phi}$ differs from what classical statistical softwares carry out, where the dispersion parameter is estimated by a quasi-likelihood approach.

Using the previous equations, we compute the log likelihood evaluated on $\widehat{\phi}$ and on $\widehat{\boldsymbol{\vartheta}}_n$ for both the transformed sample $\underline{\boldsymbol{z}}$ and the original sample $\underline{\boldsymbol{y}}$ with one categorical exogenous variable (Corollary 3.1) is

$$\log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{z}}) = -\frac{n}{2}(1+\log(2\pi\widehat{\phi})),\ \log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{y}}) = -\frac{n}{2}(1+\log(2\pi\widehat{\phi})) - \sum_{i=1}^{n} z_i. \tag{21}$$

The log-likelihood computation is detailed on Appendix B. Below we analyze the choice of the link functions considered in Table 3 in Examples 5.1, 5.2 and plotted in Figure 4b and with parameter spaces given in Table 4.

Table 3: Table of typical link functions for lognormal

Names	$\ell(\eta_i)$	$g^{-1}(t)$	$g(t)$
canonical	η_i	t	t
sym. log	$e^{\eta_i}1_{\eta_i\geq 0} + (2-e^{-\eta_i})1_{\eta_i<0}$	$e^{t}1_{t\geq 0} + (2-e^{-t})1_{t<0}$	$\log(t)1_{t\geq 1} - \log(2-t)1_{t<1}$

Table 4: Summary of spaces for lognormal

Link name	Parameter–covariable space $\vartheta \times X_i$		Linear predictor space D		Parameter space Λ
canonical	$\mathbb{R}^p \times \mathbb{R}^p$	$\xrightarrow{\langle .,. \rangle}$	$\mathbb{R}$	$\underset{\mathrm{id}}{\overset{\mathrm{id}}{\rightleftarrows}}$	$\mathbb{R}$
sym. log	$\mathbb{R}^p \times \mathbb{R}^p$	$\xrightarrow{\langle .,. \rangle}$	$\mathbb{R}$	$\underset{l_g}{\overset{l_g^{-1}}{\rightleftarrows}}$	$\mathbb{R}$

Example 5.1. *canonical link*

With the canonical link function, there is no issue between the parameter space and the linear predictor space since $D = \Lambda = \mathbb{R}$. With no-intercept using Equation (9), the MLE is

$$\widehat{\vartheta}_{n,(2),j} = \frac{1}{m_j} \sum_{i=1}^{n} x_i^{(2,j)} Z_i = \overline{Z}_n^{(j)}, \quad j \in J.$$

Hence, the distribution $\widehat{\vartheta}_{n,(2),j}$ is simply a normal distribution with mean $\vartheta_{(2),j}$ and variance ϕ/m_j. Therefore, the MLE is unbiased and converges in almost surely to $\vartheta_{(2),j}$.

Example 5.2. *symmetrical log link*

We consider a central symmetry of the logarithm function given in Table 3 leading to $l_g(x) = e^x 1_{x\geq 0} + (2-e^{-x})1_{x<0}$. With this symmetrical log link function, there is no issue between the parameter space and the linear predictor space since again $D = \Lambda = \mathbb{R}$. With no-intercept using Equation (9), the MLE is

$$\widehat{\vartheta}_{n,(2),j} = \log\left(\overline{Z}_n^{(j)} 1_{\overline{Z}_n^{(j)}\geq 1}\right) - \log\left(2 - \overline{Z}_n^{(j)} 1_{\overline{Z}_n^{(j)}<1}\right), \quad j \in J.$$

The expectation of $\widehat{\vartheta}_{n,(2),j}$ is complex and should be done numerically. However by the strong law of large numbers and the continuity of the link function, $\widehat{\vartheta}_{n,(2),j} = l_g(\overline{Z}_n^{(j)})$ converge almost surely to $l_g(\mathbf{E}\overline{Z}_n^{(j)}) = \vartheta_{(2),j}$.

5.3 Model diagnostic

In this paragraph, we give some details about residuals in the lognormal case. As already said, the transformed variables $Z_i = \log(Y_i - \mu)$ is normally distributed with mean $\ell(\eta_i)$ and variance ϕ. Let define the residuals

$$R_i = \frac{Z_i - \ell(\eta_i)}{\sqrt{\phi}}, \quad i \in I.$$

Hence $R_1, \ldots, R_n$ are i.i.d. and have a normal distribution $\mathcal{N}(0,1)$. The consistency of the MLE makes it possible to say that the $\widehat{R}_{n,i} = \frac{Z_i - \ell(\widehat{\eta}_i)}{\sqrt{\widehat{\phi}}}$, $i \in I$, with $\widehat{\eta}_i = \langle \boldsymbol{x}_i, \widehat{\boldsymbol{\vartheta}}_n \rangle$ are asymptotically i.i.d. Furthermore, the summation of $\widehat{R}_{n,1}, \ldots, \widehat{R}_{n,n}$ is exactly equal to 0.

It is also possible to verify the assumption of the lognormal distribution for Y_i conditionally to $\boldsymbol{x}_i$ with graphical diagnostic as a normal Quantile-Quantile plot on the residuals $\widehat{R}_{n,i}$.

6 Simulations study

This section is devoted to the simulation study: all computations are carried out thanks to the R statistical software R Core Team (2019). The first part of the simulation analysis consists in assessing the uncertainty of the MLE with the two different approaches: either the explicit formula given in Example 4.1 or the IWLS algorithm described in McCullagh and Nelder (1989).

Therefore, the simulation process has the following steps. Firstly, given the parameter number p, we simulate n random variables which are Pareto 1 distributed in the case of no intercept and a canonical link (see Example 4.1). Secondly, we check that the fitted coefficients by both methods have identical values. Thirdly, we compute the exact confidence interval using the result of Example 4.1 and the asymptotic MLE confidence interval resulted from the IWLS algorithm.

Figure 1 shows the estimated parameters when $p = 3$ for different values of n. That is we plot $\widehat{\vartheta}_{n,(2),1}$, $\widehat{\vartheta}_{n,(2),2}$, $\widehat{\vartheta}_{n,(2),3}$ as solid lines for $n = 100$, 300, 500, 700, 900, 1000, 3000, 5000, 7000, 9000, 10000 against the true values $\boldsymbol{\vartheta} = (2, 3, 4)$ (horizontal dot-dashed lines) and $\mu = 150$. The confidence intervals are plotted in dashed lines (theoretical) and dotted lines (asymptotic). We observe that for any sample size n, explicit and IWLS produce the same value but the explicit confidence interval is much thinner in the explicit case.

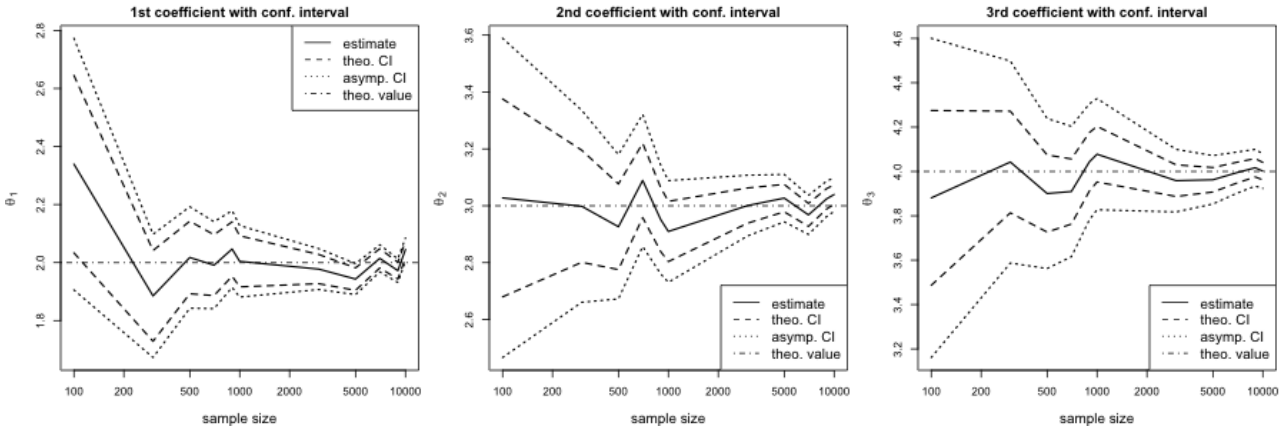


Figure 1: Estimated coefficient $\widehat{\vartheta}_{n,(2),i}$ (solid lines) with theoretical confidence intervals (dashed lines) and asymptotic confidence intervals (dotted lines)

The second part of the simulation analysis consists in assessing the gain in terms of computations between the two different approaches: either the explicit formula given in Example 4.1 or the IWLS algorithm described in McCullagh and Nelder (1989).

In Table 5, we provide a complexity analysis of the two procedure : the IWLS algorithm and the explicit solution. Again we simulate with Pareto 1 distributed random variables for a specific number of parameters p, and a known threshold $\mu = 150$ given a sample size n. Then we compute the floating point operation numbers given the size of the input dataset. Two different explanatory variables have been tested so that the parameter number is different: $p = 5$ or 7. We observe that the explicit solution is far less computer intensive (4000 times faster) than the IWLS algorithm which takes 5 or 6 iterations to reach the solution. Thus having an explicit solution can lead to substantial benefits for very large datasets.

IWLS algorithm				exact method				
n	p	iter. nb.	flop x1000	n	p	iter. nb.	flop x1000	rel. gain
200	5	6	1206.2	200	5	1	1.6	753.9
400	5	5	4010.4	400	5	1	3.2	1253.2
600	5	6	10818.6	600	5	1	4.8	2253.9
800	5	5	16020.8	800	5	1	6.4	2503.2
1000	5	5	25026	1000	5	1	8	3128.2
1200	5	5	36031.2	1200	5	1	9.6	3753.2
1400	5	5	49036.4	1400	5	1	11.2	4378.2
1600	5	5	64041.6	1600	5	1	12.8	5003.2
1800	5	5	81046.8	1800	5	1	14.4	5628.2
200	7	6	1686.2	200	7	1	2	843.1
400	7	6	6732.4	400	7	1	4	1683.1
600	7	6	15138.6	600	7	1	6	2523.1
800	7	6	26904.8	800	7	1	8	3363.1
1000	7	6	42031	1000	7	1	10	4203.1
1200	7	6	60517.2	1200	7	1	12	5043.1
1400	7	6	82363.4	1400	7	1	14	5883.1
1600	7	6	107569.6	1600	7	1	16	6723.1
1800	7	6	136135.8	1800	7	1	18	7563.1

Table 5: Floating point operation number given the size of the dataset

7 Application to large claim modeling

This section is devoted to the numerical illustration on a real dataset (again computations are carried out thanks to the R statistical software R Core Team (2019)). In our application, we focus on modeling non-life insurance losses (claim amount) of corporate business lines. Our data set comes from an anonymous private insurer: for privacy reason, amounts have been randomly scaled, dates randomly rearranged, variable modalities renamed. The data set consists of 211,739 claims which occurred between 2000 and 2010. In addition to the claim amount level, various explanatory variables are available.

We provide in Table 10 in Appendix C a short descriptive analysis of the two most important variables (risk class and guarantee type with respectively 5 and 7 modalities). Due to the very high value of skewness and kurtosis, we observe that claim amount is particularly heavy tailed.

In the sequel, we consider only large claims which are in our context claims above $\mu = 340,000$ (in euros). The threshold value has been chosen by expert opinion of practitioners. We refer to e.g. Reiss and Thomas (2007) for advanced selection methods based on extreme value theory.

7.1 A single explanatory variable

Firstly, we consider both Pareto 1 GLM and Shifted log-normal GLM with only one explanatory variable: the guarantee type. We choose Guarantee 1 as the reference level implying that $\vartheta_{(2),1} = 0$. So, $\vartheta_{(1)}$ representing the effect of the reference category and $(\vartheta_{(2),j})_j$ representing the differential effect of categories j relative to the reference category will be estimated through (15) and (19). Observations $y_1, \dots, y_n$ are observed claim amounts either from Pareto 1 (13) or shifted log-normal (18).

For theses two models, we have many possible choices for the link function g. Naturally, we choose link functions appearing in Tables 1 and 3 respectively. In accordance to Corollaries 3.1, the choice of g does not impact the values respectively given on (16) and (21) of the log-likelihoods applied on the MLE of $\boldsymbol{\vartheta}$.

Furthermore, for Pareto GLM, the choice of shifted log-inverse link function seems attractive because it guarantees the existence of $\mathbf{E}Y_i$. Nevertheless, alternative link functions (canonical or log-inv) allow to construct an unbiased estimator (see Section 4). For shifted log-normal model, the choice of canonical link function is more attractive because it leads to an unbiased and simpler MLE estimator (see Section 5).

Coefficients are estimated by explicit formulas given in Sections 4 and 5. In Table 6, the estimated coefficients are given in the five considered situations. Positive values of $\vartheta_{(2),j}$ in the Pareto GLM increase the shape parameter of the Pareto 1 distribution leading to a decrease in heavy-tailedness. Regarding the log-normal model, positive values of $\vartheta_{(2),j}$ increase the scale parameter of the log-normal distribution leading to a shrink of the distribution.

Irrespectively of the considered link function, the sign of the fitted coefficients are same except for intercept (Table 6) given a distribution. This convinces us that different model assumptions (i.e. link) do not lead to opposite conclusions on the claim severity. Furthermore from Table 10, we retrieve the fact that all guarantees except Guarantee 2 have heavier tails than the reference Guarantee 1.

Table 6: Coefficients for the guarantee variable

Model	Pareto 1			Shifted log normal	
Variable	canonical	loginv	shifted.loginv	canonical	symlog
Intercept	1.89	0.64	-0.11	11.75	2.46
Guarantee 2	0.04	0.02	0.04	0.10	0.01
Guarantee 3	-0.67	-0.43	-1.36	0.75	0.06
Guarantee 4	-0.86	-0.60	-3.13	1.04	0.08
Guarantee 5	-0.71	-0.47	-1.55	0.72	0.06
Guarantee 6	-0.42	-0.25	-0.63	0.42	0.04
Guarantee 7	-0.48	-0.29	-0.78	0.59	0.05
log likelihood	-14507.53	-14507.53	-14507.53	-14517.37	-14517.37

Whatever the considered link function g, the residuals defined in Section 4.3 by $\widehat{R}_{n,i} = -\ell(\widehat{\eta}_i)Z_i$, $i \in I$, do not depend on ℓ and are given by Equation (17). We show on Figure 3 (left) the quantile/quantile plots of residuals described on Section 4.3 against the standard exponential distribution. On Figure 2, we observe that the assumption of Pareto 1 for $Y_1, \dots, Y_n$ is better than the log-normal distribution at least for small quantiles. Moreover, comparing the value of the log-likelihood in Table 6, Pareto 1 distribution is also the best choice. In the following, we focus only on the Pareto 1 distribution.

For all coefficient, let us compute the p-values of statistical tests with the null hypothesis $\vartheta_{(1)} = 0$ (Interecept null), and for $j \in J \setminus \{1\}$, $\vartheta_{(2),j} = 0$ (no differential effect of the jth Guarantee). Table 7 reports the value of the coefficient, its standard error, the student statistics and the associated p-value. We observe that some modalities of the guarantee variable are

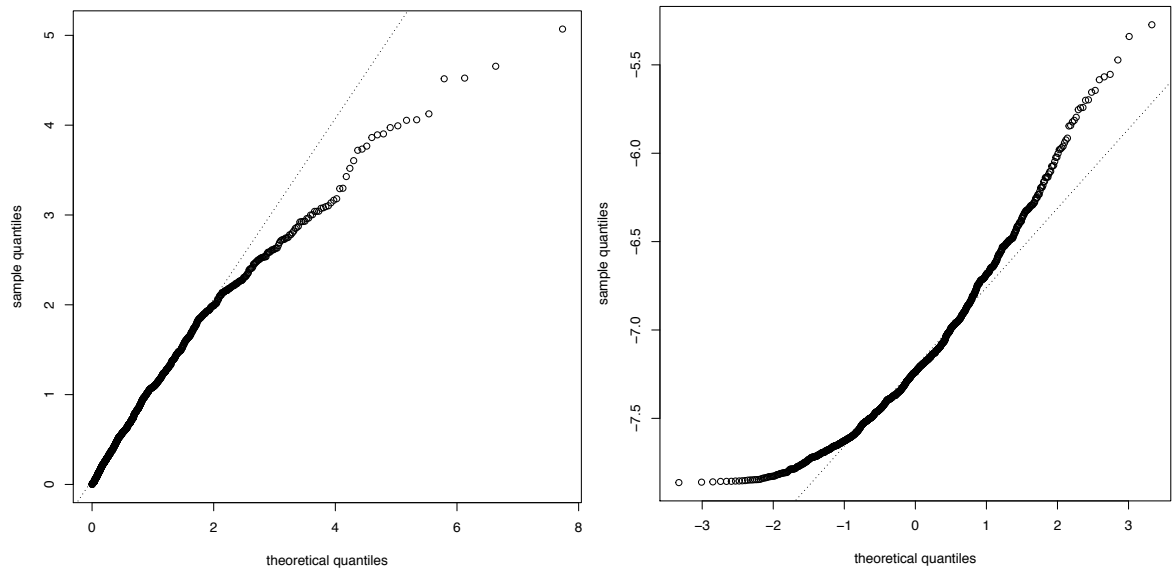


Figure 2: Quantile-quantile plots of the residuals defined in Section 4.3: (left) for Pareto 1, (right) for shifted log-normal distributions

statistically significant at the 5% level. Except for Guarantee 2 and Guarantee 6, other p-values are relatively small showing the Pareto 1 distribution with explanatory variables is relevant in this context.

Table 7: Statistics and p.values for the tests $\vartheta_{(1)} = 0$ and $\vartheta_j = 0$, $j \in J$ in the Pareto GLM model for the log-inverse link.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.6391	0.1644	3.8877	0.0001
Guarantee 2	0.0214	0.2219	0.0966	0.9230
Guarantee 3	-0.4332	0.1950	-2.2217	0.0263
Guarantee 4	-0.6009	0.1708	-3.5180	0.0004
Guarantee 5	-0.4660	0.1805	-2.5817	0.0098
Guarantee 6	-0.2485	0.2295	-1.0827	0.2789
Guarantee 7	-0.2949	0.1834	-1.6075	0.1080

7.2 Two explanatory variables

Secondly, we consider the Pareto GLM models and Shifted log-normal GLM models with the two explanatory variables (guarantee and risk class) without intercept nor single-variable (c.f. model (10) and example 3.4), that are

$$g(\mathbf{E}Z_i) = \sum_{(k,l)\in KL^\star} \vartheta_{kl} x_i^{(k,l)}, \quad i \in I \tag{22}$$

with $Z_i = -\log(Y_i/\mu)$ for the Pareto 1 modeling $Z_i = \log(Y_i - \mu)$ for the shifted log normal modeling and where for $(k,l) \in KL^\star$ the unknown parameters ϑ_{kl} represent the effect of the couple of the modalities k and l for the first and the second variable. In theses examples, as it describes in Table 8, we have $K = \{1,\dots,7\}$, $L = \{1,\dots,5\}$ but $KL^\star = \{1,\dots,7\} \times \{1,\dots,5\} \setminus \{(1,2),(6,5)\}$.

Consider the estimation procedures in (22). We compute the claim numbers according Guarantee and Risk in Table 8. This claim number per class might be too short to ensure the

existence of the MLE with the shifted log-inv link. We arbitrary choose the simple case of the canonical link and an unbiased estimator is relevant in this context.

The coefficients of the model are estimated using the exact method described in Section 3 and then unbiased in the same way of Example 1. The fitted coefficients are not shown but are available upon request to the authors. Furthermore, we compute the p-values of the statistical test $\vartheta_{kl} = 0$ in Table 9. We observe that most computed p-values are small: either less than 10^{-6} or less than 1%. Only 5 on the 33 p-values are above the usual 5% level, corresponding to the couples Guarantee/Risk class (1,5), (2,2), (2,3), (2,5) and (7,5) (claim number of 1,2 or 3). In the two variables setting, the Pareto 1 GLM is thus still relevant.

Table 8: Number of claim per Guarantee and per Risk class.

Claim number	Risk class 1	Risk class 2	Risk class 3	Risk class 4	Risk class 5
Guarantee 1	39	0	4	6	1
Guarantee 2	26	2	3	16	3
Guarantee 3	48	7	11	29	4
Guarantee 4	232	40	75	147	20
Guarantee 5	68	18	36	72	6
Guarantee 6	24	7	4	11	0
Guarantee 7	94	9	22	57	3

Table 9: p-values for the tests $\vartheta_{kl} = 0$, $(k, l) \in KL^\star$ in (22) for the canonical link.

p-values	Risk class 1	Risk class 2	Risk class 3	Risk class 4	Risk class 5
Guarantee 1	$< 10^{-6}$	-	0.04550	0.01431	0.31731
Guarantee 2	$< 10^{-6}$	0.15730	0.08301	0.00006	0.08326
Guarantee 3	$< 10^{-6}$	0.00815	0.00091	$< 10^{-6}$	0.04550
Guarantee 4	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	0.00001
Guarantee 5	$< 10^{-6}$	0.00002	$< 10^{-6}$	$< 10^{-6}$	0.01431
Guarantee 6	$< 10^{-6}$	0.00815	0.04550	0.00091	-
Guarantee 7	$< 10^{-6}$	0.00270	$< 10^{-6}$	$< 10^{-6}$	0.08326

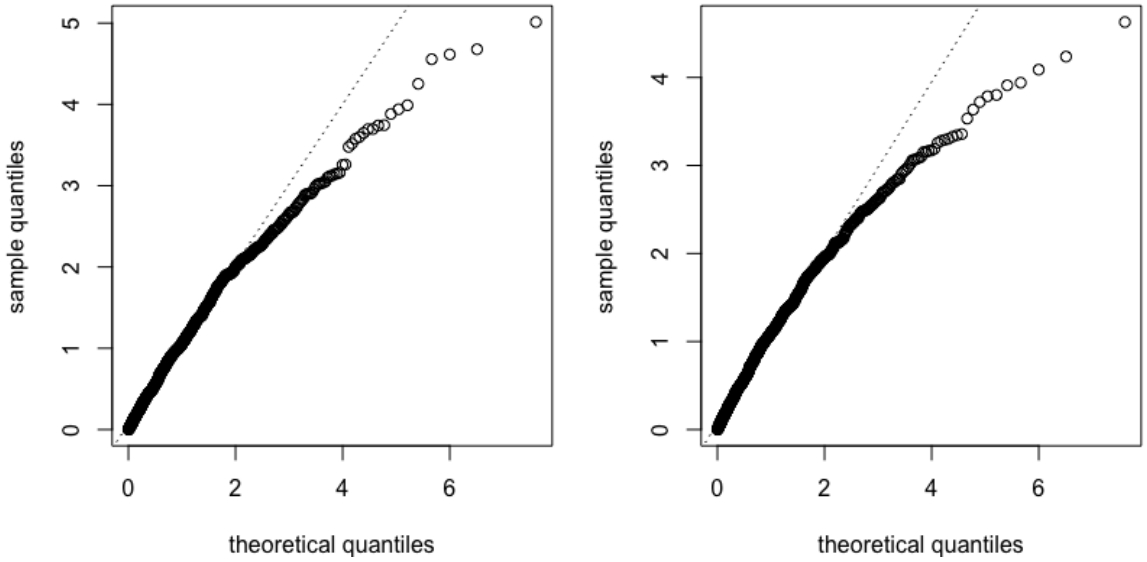


Figure 3: Quantile-quantile plots of the residuals defined in Section 4.3: (left) for one explanatory variable, (right) for two explanatory variables

8 Conclusion

In this paper, we deal with regression models where the response variable belongs to the general formulation of the exponential family, the so-called GLM. We focus on the estimation of parameters of GLMs and derive explicit formulas for MLE in the case of categorical explanatory variables. In this case, the closed-form estimators do not require any use of numerical algorithms, in particular the well-known IWLS algorithm. This is logical, because in the special setting of categorical variables, a regression model is equivalent to fitting the same distribution on subgroups defined with respect to explanatory variables. Hence, we get back to usual explicit solutions for the exponential family in the i.i.d. case.

Yet we work with one or two explanatory variables for the two derived theorems, the approach can be extended to d categorical variables as long as we consider interactions terms and a zero-sum condition. If we consider main effects only for d categorical variables, the MLE cannot be reformulated as a least-square problem. Nevertheless, having an explicit formula make a clear advantage compared to the IWLS algorithm, particularly for large scale datasets.

The explicit formulas are examplified on two particular positive distributions particularly useful in an insurance context: the Pareto 1 distribution and the shifted log-normal distribution. In both cases, we present typical link functions and derive in most cases the distribution of the MLE. In relevant cases, we also give an unbiased estimator. In the general setting, the exact standard error computation is not available, yet the Delta Method can be used to obtain an asymptotic standard error. Finally, we illustrate the estimation process for both distributions on simulated datasets and an actuarial data set.

For future research, a natural extension is to propose regression models for distribution outside the exponential family. Typically, we could consider the Pareto 1 distribution with unknown threshold and shape parameters. We could also consider generalized Pareto distribution based on the peak over thresholds approach. A natural extension could also be to jointly estimate the shape and the dispersion parameters of the distribution.

Acknowledgments

The authors thank Vanessa Desert for her active support during the writing of this paper. The authors are also very grateful for the useful suggestions of the two referees. This work is supported by the research project "PANORisk" and Région Pays de la Loire (France).

This is a post-peer-review, pre-copyedit version of an article published in Computational Statistics. The final authenticated version is available online at: http://dx.doi.org/10.1007/s00180-019-00918-7.

A Proofs of Section 3

A.1 Proof for the one-variable case

Proof of Theorem 3.1. We have to solve the system

$$\begin{cases} S(\boldsymbol{\vartheta}) = 0 \\ \boldsymbol{R\vartheta} = 0. \end{cases} \tag{23}$$

The system $S(\boldsymbol{\vartheta}) = 0$ is

$$\left\{ \begin{array}{l} \displaystyle\sum_{i=1}^{n} \ell'(\eta_i)\left(y_i - b' \circ \ell(\eta_i)\right) = 0 \\ \displaystyle\sum_{i=1}^{n} x_i^{(2),j} \ell'(\eta_i)\left(y_i - b' \circ \ell(\eta_i)\right) = 0, \quad \forall j \in J. \end{array} \right.$$

that is

$$\left\{ \begin{array}{l} \displaystyle\sum_{j\in J} \ell'(\vartheta_{(1)} + \vartheta_{(2),j}) \left(\sum_{i=1}^{n} x_i^{(2),j} y_i - m_j b' \circ \ell(\vartheta_{(1)} + \vartheta_{(2),j}) \right) = 0 \\ \displaystyle \ell'(\vartheta_{(1)} + \vartheta_{(2),j}) \left(\sum_{i=1}^{n} x_i^{(2),j} y_i - m_j b' \circ \ell(\vartheta_{(1)} + \vartheta_{(2),j}) \right) = 0, \quad \forall j \in J. \end{array} \right.$$

The first equation in the previous system is redundancy, and

$$S(\boldsymbol{\vartheta}) = 0 \Leftrightarrow \ell'(\vartheta_{(1)} + \vartheta_{(2),j}) \left(\sum_{i=1}^{n} x_i^{(2),j} y_i - m_j b' \circ \ell(\vartheta_{(1)} + \vartheta_{(2),j}) \right) = 0, \quad \forall j \in J.$$

Hence if Y_i takes values in $\mathbb{Y} \subset b'(\Lambda)$, and ℓ injective, we have

$$\vartheta_{(1)} + \vartheta_{(j)} = g(\overline{Y}_n^{(j)}) \quad \forall j \in J.$$

The system (23) is

$$\left\{ \begin{array}{l} \boldsymbol{Q}\boldsymbol{\vartheta} = \boldsymbol{g}(\bar{\boldsymbol{Y}}) \\ \boldsymbol{R}\boldsymbol{\vartheta} = 0. \end{array} \right. \Leftrightarrow \begin{pmatrix} \boldsymbol{Q} \\ \boldsymbol{R} \end{pmatrix} \boldsymbol{\vartheta} = \begin{pmatrix} \boldsymbol{g}(\bar{\boldsymbol{Y}}) \\ 0 \end{pmatrix}. \tag{24}$$

Let us compute the determinant of the matrix $M_d = \begin{pmatrix} \boldsymbol{Q} \\ \boldsymbol{R} \end{pmatrix}$. Consider $\boldsymbol{R} = (r_0, r_1, \ldots, r_d)$. We have

$$M_d = \begin{pmatrix} \mathbf{1}_d & I_d \\ r_0 & \boldsymbol{r} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \ldots & \\ 1 & 0 & 1 & 0 & \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 1 & 0 & \ldots & 0 & 1 \\ r_0 & r_1 & & \ldots & r_d \end{pmatrix}, \text{ with } \boldsymbol{r} = \begin{pmatrix} r_1 & \ldots & r_d \end{pmatrix}, \mathbf{1}_d = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

The determinant can be computed recursively

$$\det(M_d) = r_d \begin{vmatrix} 1 & 1 & 0 & \ldots \\ 1 & 0 & \ddots & 0 \\ \vdots & \vdots & \ddots & 1 \\ 1 & 0 & \ldots & 0 \end{vmatrix} - \begin{vmatrix} 1 & 1 & 0 & \ldots \\ 1 & 0 & \ddots & 0 \\ \vdots & \vdots & \ddots & 1 \\ r_0 & r_1 & \ldots & r_{d-1} \end{vmatrix} = (-1)^{d+1} r_d - \det(M_{d-1}).$$

Since $\det(M_1) = -r_0 + r_1$ and $\det(M_2) = -r_2 - (-r_0 + r_1) = r_0 - r_1 - r_2$, we get $\det(M_d) = (-1)^d r_0 + (-1)^{d+1}(r_1 + \cdots + r_d) = (-1)^d(r_0 - r_1 - \cdots - r_d)$. This determinant is non zero as long as $r_0 \neq \sum_{j=1}^{d} r_j$.

Now we compute the inverse of matrix M_d by a direct inversion.

$$\begin{pmatrix} \mathbf{1}_d & I_d \\ r_0 & \boldsymbol{r} \end{pmatrix} \begin{pmatrix} \boldsymbol{a}' & b \\ C & \boldsymbol{d} \end{pmatrix} = \begin{pmatrix} I_d & \mathbf{0} \\ \mathbf{0}' & 1 \end{pmatrix} \Leftrightarrow \left\{ \begin{array}{l} \mathbf{1}_d \boldsymbol{a}' + I_d C = I_d \\ b\mathbf{1}_d + I_d \boldsymbol{d} = \mathbf{0} \\ r_0 \boldsymbol{a}' + \boldsymbol{r} C = \mathbf{0}' \\ b r_0 + \boldsymbol{r}\boldsymbol{d} = 1 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} C = I_d - \frac{1}{-r_0 + \boldsymbol{r}\mathbf{1}_d} \mathbf{1}_d \boldsymbol{r} \\ \boldsymbol{d} = \frac{1}{-r_0 + \boldsymbol{r}\mathbf{1}_d} \mathbf{1}_d \\ \boldsymbol{a}' = \frac{\boldsymbol{r}}{-r_0 + \boldsymbol{r}\mathbf{1}_d} \\ b = \frac{-1}{-r_0 + \boldsymbol{r}\mathbf{1}_d} \end{array} \right.$$

Let us check the inverse of M_d

$$\begin{pmatrix} \mathbf{1}_d & I_d \\ r_0 & \boldsymbol{r} \end{pmatrix} \begin{pmatrix} \frac{\boldsymbol{r}}{-r_0+\boldsymbol{r}\mathbf{1}_d} & \frac{-1}{-r_0+\boldsymbol{r}\mathbf{1}_d} \\ I_d - \frac{\mathbf{1}_d\boldsymbol{r}}{-r_0+\boldsymbol{r}\mathbf{1}_d} & \frac{\mathbf{1}_d}{-r_0+\boldsymbol{r}\mathbf{1}_d} \end{pmatrix} = \begin{pmatrix} \frac{\mathbf{1}_d\boldsymbol{r}}{-r_0+\boldsymbol{r}\mathbf{1}_d} + I_d - \frac{\mathbf{1}_d\boldsymbol{r}}{-r_0+\boldsymbol{r}\mathbf{1}_d} & \frac{-\mathbf{1}_d}{-r_0+\boldsymbol{r}\mathbf{1}_d} + \frac{\mathbf{1}_d}{-r_0+\boldsymbol{r}\mathbf{1}_d} \\ r_0\frac{\boldsymbol{r}}{-r_0+\boldsymbol{r}\mathbf{1}_d} + \boldsymbol{r} - \frac{\boldsymbol{r}\mathbf{1}_d\boldsymbol{r}}{-r_0+\boldsymbol{r}\mathbf{1}_d} & \frac{-r_0}{-r_0+\boldsymbol{r}\mathbf{1}_d} + \frac{\boldsymbol{r}\mathbf{1}_d}{-r_0+\boldsymbol{r}\mathbf{1}_d} \end{pmatrix} = \begin{pmatrix} I_d & 0 \\ 0 & 1 \end{pmatrix}.$$

So as long as $r_0 \neq \sum_{j=1}^d r_j$

$$\widehat{\boldsymbol{\vartheta}}_n = \begin{pmatrix} \frac{\boldsymbol{r}}{-r_0+\boldsymbol{r}\mathbf{1}_d} & \frac{-1}{-r_0+\boldsymbol{r}\mathbf{1}_d} \\ I_d - \frac{\mathbf{1}_d\boldsymbol{r}}{-r_0+\boldsymbol{r}\mathbf{1}_d} & \frac{\mathbf{1}_d}{-r_0+\boldsymbol{r}\mathbf{1}_d} \end{pmatrix} \begin{pmatrix} \boldsymbol{g}(\bar{\boldsymbol{Y}}) \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{\boldsymbol{r}\boldsymbol{g}(\bar{\boldsymbol{Y}})}{-r_0+\boldsymbol{r}\mathbf{1}_d} \\ \boldsymbol{g}(\bar{\boldsymbol{Y}}) - \mathbf{1}_d\frac{\boldsymbol{r}\boldsymbol{g}(\bar{\boldsymbol{Y}})}{-r_0+\boldsymbol{r}\mathbf{1}_d} \end{pmatrix}.$$

In an other way, the system (24) is equivalent to

$$(\boldsymbol{Q}', \boldsymbol{R}') \begin{pmatrix} \boldsymbol{Q} \\ \boldsymbol{R} \end{pmatrix} \boldsymbol{\vartheta} = \boldsymbol{Q}'\boldsymbol{g}(\bar{\boldsymbol{Y}}),$$

and for $(\boldsymbol{Q}\,\boldsymbol{R})$ of full rank, the matrix $(\boldsymbol{Q}'\boldsymbol{Q}+\boldsymbol{R}'\boldsymbol{R})$ is invertible and $\boldsymbol{\vartheta} = (\boldsymbol{Q}'\boldsymbol{Q}+\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{Q}'\boldsymbol{g}(\bar{\boldsymbol{Y}})$. □

Examples - Choice of the contrast vector $\boldsymbol{R}$

1. Taking $r_0 = 1, \boldsymbol{r} = \mathbf{0}$ leads to $-r_0 + \boldsymbol{r}\mathbf{1}_d = -1 \Rightarrow \widehat{\boldsymbol{\vartheta}}_n = \begin{pmatrix} 0 \\ \boldsymbol{g}(\bar{\boldsymbol{Y}}) \end{pmatrix}$.

2. Taking $r_0 = 0, \boldsymbol{r} = (1, \mathbf{0})$ leads to

$$-r_0 + \boldsymbol{r}\mathbf{1}_d = 1 \Rightarrow \widehat{\boldsymbol{\vartheta}}_n = \begin{pmatrix} g(\bar{Y}_n^{(1)}) \\ 0 \\ g(\bar{Y}_n^{(2)}) - g(\bar{Y}_n^{(1)}) \\ \vdots \\ g(\bar{Y}_n^{(d)}) - g(\bar{Y}_n^{(1)})) \end{pmatrix}.$$

3. Taking $r_0 = 0, \boldsymbol{r} = \mathbf{1}$ leads to

$$-r_0 + \boldsymbol{r}\mathbf{1}_d = d \Rightarrow \widehat{\boldsymbol{\vartheta}}_n = \begin{pmatrix} \overline{\boldsymbol{g}(\bar{\boldsymbol{Y}})} \\ g(\bar{Y}_n^{(1)}) - \overline{\boldsymbol{g}(\bar{\boldsymbol{Y}})} \\ \cdots \\ g(\bar{Y}_n^{(d)}) - \overline{\boldsymbol{g}(\bar{\boldsymbol{Y}})} \end{pmatrix}, \text{ with } \overline{\boldsymbol{g}(\bar{\boldsymbol{Y}})} = \frac{1}{d}\sum_{j=1}^d g(\bar{Y}_n^{(j)}).$$

Proof of Remark 3.4. We have to solve the system

$$S(\vartheta) = 0 \Leftrightarrow \sum_{i=1}^n \ell'(\eta)\,(y_i - b' \circ \ell(\eta)) = 0.$$

If ℓ is injective, the system simplifies to

$$\sum_{i=1}^n y_i - nb' \circ (b')^{-1} \circ g^{-1}(\eta) = 0 \Leftrightarrow \eta = g\left(\overline{y}_n\right) \Leftrightarrow \theta = g\left(\overline{y}_n\right).$$

□

Proof of Remark 3.5. Let Y_i from the exponential family $F_{exp}(a, b, c, \lambda, \phi)$. It is well known, that the moment generating function of Y_i is

$$\mathbf{E}e^{tY_i} = \exp\left(\frac{b(\lambda + ta(\phi)) - b(\lambda)}{a(\phi)}\right).$$

Hence, the moment generating function of the average $\overline{Y}_m$ is

$$M_{\overline{Y}_m}(t) = \left(\exp\left(\frac{b(\lambda + \frac{t}{m}a(\phi)) - b(\lambda)}{a(\phi)}\right)\right)^m = \exp\left(\frac{b(\lambda + ta(\phi)/m) - b(\lambda)}{a(\phi)/m}\right).$$

So we get back to a known result that $\overline{Y}_m$ belongs to the exponential family $F_{exp}(x \mapsto a(x)/m, b, c, \lambda, \phi)$, e.g. McCullagh and Nelder (1989).

In our setting, random variables in the average $\overline{Y}_n^{(j)}$ are i.i.d. with functions a, b, c and parameters $\lambda = \ell(\vartheta_{(1)} + \vartheta_{(j)})$ and ϕ. And $\overline{Y}_n^{(j)}$ also belongs to the exponential family with the same parameter but with the function $\bar{a} : x \mapsto a(x)/m_j$. In particular,

$$\mathbf{E}\overline{Y}_n^{(j)} = b'(\ell(\vartheta_{(1)} + \vartheta_{(j)})) = g^{-1}(\vartheta_{(1)} + \vartheta_{(j)}), \ \mathbf{Var}\overline{Y}_n^{(j)} = \frac{a(\phi)}{m_j}b''(\ell(\vartheta_{(1)} + \vartheta_{(j)})).$$

But the computation of $\mathbf{E}g(\overline{Y}_n^{(j)})$ remains difficult unless g is a linear function. By the strong law of large numbers, as $m_j \to +\infty$, the estimator is consistent since

$$\overline{Y}_n^{(j)} \xrightarrow[n\to+\infty]{\text{a.s.}} g^{-1}(\vartheta_{(1)} + \vartheta_{(j)}) \Rightarrow g(\overline{Y}_n^{(j)}) \xrightarrow[n\to+\infty]{\text{a.s.}} g(g^{-1}(\vartheta_{(1)} + \vartheta_{(j)})) = \vartheta_{(1)} + \vartheta_{(j)}.$$

By the Central Limit Theorem (i.e. $\overline{Y}_n^{(j)}$ converges in distribution to a normal distribution) and using the Delta Method, we obtain that the following

$$\sqrt{m_j}\left(g(\overline{Y}_n^{(j)}) - \vartheta_{(1)} + \vartheta_{(j)}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, a(\phi)b''(\ell(\vartheta_{(1)} + \vartheta_{(j)}))g'(g^{-1}(\vartheta_{(1)} + \vartheta_{(j)}))^2\right).$$

□

Proof of Corollaries 3.1. The log likelihood of $\widehat{\boldsymbol{\vartheta}}_n$ is defined by

$$\log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{y}}) = \frac{1}{a(\phi)}\sum_{i=1}^{n}(y_i\ell(\widehat{\eta}_i) - b(\ell(\widehat{\eta}_i))) + \sum_{i=1}^{n}c(y_i, \phi).$$

In fact, we must be verified than $\ell(\widehat{\eta}_i)$ does not depend on g function. If we consider $\widehat{\boldsymbol{\vartheta}}_n$ defined by (8), we have $\boldsymbol{Q}\widehat{\boldsymbol{\vartheta}}_n = \boldsymbol{g}(\bar{\boldsymbol{y}})$, since $\widehat{\boldsymbol{\vartheta}}_n$ is solution of the system (23), i.e. $\boldsymbol{Q}(\boldsymbol{Q}'\boldsymbol{Q}+\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{Q}' = I$ Using $\widehat{\eta}_i = (\boldsymbol{Q}\widehat{\boldsymbol{\vartheta}}_n)_j$ for i such that $x_i^{(2),j} = 1$ we obtain

$$\ell(\widehat{\eta}_i) = \sum_{j=1}^{d}\ell \circ g(\bar{y}_n^{(j)})x_i^{(2),j} = \sum_{j=1}^{d}\ell \circ \ell^{-1} \circ (b')^{-1}(\bar{y}_n^{(j)})x_i^{(2),j} = \sum_{j=1}^{d}(b')^{-1}(\bar{y}_n^{(j)})x_i^{(2),j},$$

and

$$\log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{y}}) = \frac{1}{a(\phi)}\sum_{j=1}^{d}\sum_{i, x_i^{(2)}=v_j}\left(y_i(b')^{-1}\left(\overline{y}_n^{(j)}\right) - b\left((b')^{-1}\left(\overline{y}_n^{(j)}\right)\right)\right) + \sum_{i=1}^{n}c(y_i, \phi).$$

In the same way,

$$\widehat{\mathbf{E}Y_i} = b'(\ell(\widehat{\eta}_i)) = \sum_{j=1}^{d}\bar{y}_n^{(j)}x_i^{(2),j}, \quad \widehat{\mathbf{Var}Y_i} = a(\phi)b''(\ell(\widehat{\eta}_i)) = a(\phi)\sum_{j=1}^{d}b'' \circ (b')^{-1}(\bar{y}_n^{(j)})x_i^{(2),j}.$$

□

A.2 Proof for the two-variable case

Proof of Theorem 3.2. The system $S(\boldsymbol{\vartheta}) = 0$ is

$$\begin{cases} \sum_{i=1}^{n} \ell'(\eta_i)\left(y_i - b' \circ \ell(\eta_i)\right) = 0 \\ \sum_{i=1}^{n} x_i^{(3),l} \ell'(\eta_i)\left(y_i - b' \circ \ell(\eta_i)\right) = 0, \quad \forall l \in L \\ \sum_{i=1}^{n} x_i^{(2),k} \ell'(\eta_i)\left(y_i - b' \circ \ell(\eta_i)\right) = 0, \quad \forall k \in K \\ \sum_{i=1}^{n} x_i^{kl} \ell'(\eta_i)\left(y_i - b' \circ \ell(\eta_i)\right) = 0, \quad \forall (k,l) \in KL^\star. \end{cases}$$

that is

$$\begin{cases} \sum_{(k,l)\in KL^\star} \ell'(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl}) \left(\sum_{i=1}^{n} x_i^{(k,l)} y_i - m_{k,l} b' \circ \ell(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl}) \right) = 0 \\ \sum_{k\in K_l^\star} \ell'(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl}) \left(\sum_{i=1}^{n} x_i^{(k,l)} y_i - m_{k,l} b' \circ \ell(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl}) \right) = 0 \quad \forall l \in L \\ \sum_{l\in L_k^\star} \ell'(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl}) \left(\sum_{i=1}^{n} x_i^{(k,l)} y_i - m_{k,l} b' \circ \ell(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl}) \right) = 0 \quad \forall k \in K \\ \ell'(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl}) \left(\sum_{i=1}^{n} x_i^{(k,l)} y_i - m_{k,l} b' \circ \ell(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl}) \right) = 0 \quad \forall (k,l) \in KL^\star. \end{cases}$$

The system have exactly $1 + d_2 + d_3$ redundancies, and $S(\boldsymbol{\vartheta}) = 0$ reduces to

$$\ell'(\vartheta_{(1)}+\vartheta_{(2),k}+\vartheta_{(3),l}+\vartheta_{kl}) \left(\sum_{i=1}^{n} x_i^{(k,l)} y_i - m_{k,l} b' \circ \ell(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl}) \right) = 0 \quad \forall (k,l) \in KL^\star.$$

Hence the system has rank $KL^\star$ and if Y_i takes values in $\mathbb{Y} \subset b'(\Lambda)$, and ℓ injective, we have

$$\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl} = g(\bar{Y}_n^{(k,l)}) \quad \forall (k,l) \in KL^\star.$$

In the same way of proof of Theorem 3.1, we have to solve

$$\begin{cases} \boldsymbol{Q\vartheta} = \boldsymbol{g}(\bar{\boldsymbol{Y}}) \\ \boldsymbol{R\vartheta} = \boldsymbol{0}. \end{cases} \tag{25}$$

that is, because $\boldsymbol{QQ}' + \boldsymbol{RR}'$ is full rank, in the same way of proof of Theorem 3.1

$$\boldsymbol{\vartheta} = (\boldsymbol{Q}'\boldsymbol{Q} + \boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{Q}'\boldsymbol{g}(\bar{\boldsymbol{Y}}).$$

In that case, the MLE solves a least square problem with response variable $\boldsymbol{g}(\bar{\boldsymbol{Y}})$, explanatory variable $\boldsymbol{Q}$ under a linear constraint $\boldsymbol{R}$.

1. Under linear contrasts $(\tilde{C}_0)$, the model (10) is equivalent to model (6) with $J = KL^\star$ modalities. Hence the solution is evident.

2. Under linear contrasts ($\tilde{C}_\Sigma$), the system

$$\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl} = g(\bar{Y}_n^{(k,l)}) \quad \forall (k,l) \in KL^\star$$

implies that

$$\sum_{(k,l)\in KL^\star} m_{k,l}(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l} + \vartheta_{kl}) = \sum_{(k,l)\in KL^\star} m_{k,l} g(\bar{Y}_n^{(k,l)}).$$

Using

$$\sum_{(k,l)\in KL^\star} m_{k,l} = n, \quad \sum_{(k,l)\in KL^\star} m_{k,l}\vartheta_{(2),k} = \sum_{k\in K}\sum_{l\in L_k^\star} m_{k,l}\vartheta_{(2),k} = \sum_{k\in K} m_k^{(2)}\vartheta_{(2),k} = 0,$$

$$\sum_{(k,l)\in KL^\star} m_{k,l}\vartheta_{(3),l} = \sum_{l\in L}\sum_{k\in K_l^\star} m_{k,l}\vartheta_{(3),l} = \sum_{l\in L} m_l^{(3)}\vartheta_{(3),l} = 0, \quad \sum_{(k,l)\in KL^\star} m_{k,l}\vartheta_{kl} = 0,$$

we get $\vartheta_{(1)} = \dfrac{1}{n}\displaystyle\sum_{(k,l)\in KL^\star} m_{k,l} g(\bar{Y}_n^{(k,l)})$. In the same way, taking summation over $K_l^\star$ for $l \in L$ and over $L_k^\star$ for $k \in K$, we found $\vartheta_{(2),k}$ and $\vartheta_{(3),l}$, and then ϑ_{kl}.

With main effect only, the system $S(\boldsymbol{\vartheta}) = 0$ is

$$\begin{cases} \displaystyle\sum_{i=1}^n \ell'(\eta_i) y_i = \sum_{i=1}^n g^{-1}(\eta_i)\ell'(\eta_i) \\ \displaystyle\sum_{i=1}^n x_i^{(3),l}\ell'(\eta_i) y_i = \sum_{i=1}^n x_i^{(3),l} g^{-1}(\eta_i)\ell'(\eta_i) \quad \forall l \in L \\ \displaystyle\sum_{i=1}^n x_i^{(2),k}\ell'(\eta_i) y_i = \sum_{i=1}^n x_i^{(2),k} g^{-1}(\eta_i)\ell'(\eta_i), \quad \forall k \in K \end{cases}$$

There are $1 + d_2 + d_3$ equations for $1 + d_2 + d_3$ parameters, but each explanatory variable are colinear. So, the two additional constraints $\boldsymbol{R\vartheta} = 0$ ensures that a solution exist for the remaining $d_2 + d_3 - 1$ parameters. Using $\sum_k x_i^{(2),k} = 1$, the second set of equations becomes $\forall l \in L$

$$\sum_{k\in K} \ell'(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l})\bar{y}_n^{(k,l)} m_{k,l} = \sum_{k\in K} g^{-1}(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l})\ell'(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l}) m_{k,l}$$

Similarly, the third set of equations becomes $\forall k \in K$

$$\sum_{l\in L} \ell'(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l})\bar{y}_n^{(k,l)} m_{k,l} = \sum_{l\in L} g^{-1}(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l})\ell'(\vartheta_{(1)} + \vartheta_{(2),k} + \vartheta_{(3),l}) m_{k,l}$$

Even with a canonical link $\ell(x) = x$ so that $\ell'(x) = 1$, this system is not a least-square problem for a nonlinear g function.

□

B Calculus of the Log-likelihoods appearing in Sections 4 and 5

Consider the Pareto GLM described on (13) and (15). The b function is $b(\lambda) = -\log(\lambda)$, using corollary 3.1 we have $\ell(\hat{\eta}_i) = (b')^{-1}(\overline{z}_n^{(j)}) = -(\overline{z}_n^{(j)})^{-1}$ for j such that $x_i^{(2),j} = 1$ and

$$\log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{z}}) = \sum_{j=1}^d \sum_{i, x_i^{(2),j}=1} \left(z_i/\overline{z}_n^{(j)} - \log\left(-\overline{z}_n^{(j)}\right)\right) = n - \sum_{j=1}^d m_j \log\left(-\overline{z}_n^{(j)}\right).$$

Compute the original log likelihood of Pareto 1 distribution:

$$\log L(\boldsymbol{\vartheta} \,|\, \underline{\boldsymbol{y}}) = \sum_{i=1}^{n} \big(\log \ell(\eta_i) + \ell(\eta_i) \log \mu - (\ell(\eta_i) + 1) \log y_i \big).$$

Hence with $z_i = -\log(y_i/\mu)$,

$$\begin{aligned} \log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{y}}) &= \sum_{j=1}^{d} \sum_{i, x_i^{(2),j}=1} \left(-\log(-\overline{z}_n^{(j)}) - \frac{\log \mu}{\overline{z}_n^{(j)}} + \frac{\log(y_i)}{\overline{z}_n^{(j)}} - \log y_i \right) \\ &= n - \sum_{j=1}^{d} m_j \log(-\overline{z}_n^{(j)}) - \sum_{i=1}^{n} \log y_i = \log L(\widehat{\boldsymbol{\vartheta}}_n \,|\, \underline{\boldsymbol{z}}) - \sum_{i=1}^{n} \log y_i. \end{aligned}$$

Now consider the shifted log-normal GLM described on (18) and (19). Here, the b function is $b(\lambda) = \lambda^2/2$, hence using Corollary 3.1, we have $\ell(\hat{\eta}_i) = (b')^{-1}(\overline{z}_n^{(j)}) = \overline{z}_n^{(j)}$ for j such that $x_i^{(2),j} = 1$ and equation (21) holds.

Let us compute the original log likelihood of the shifted log normal distribution:

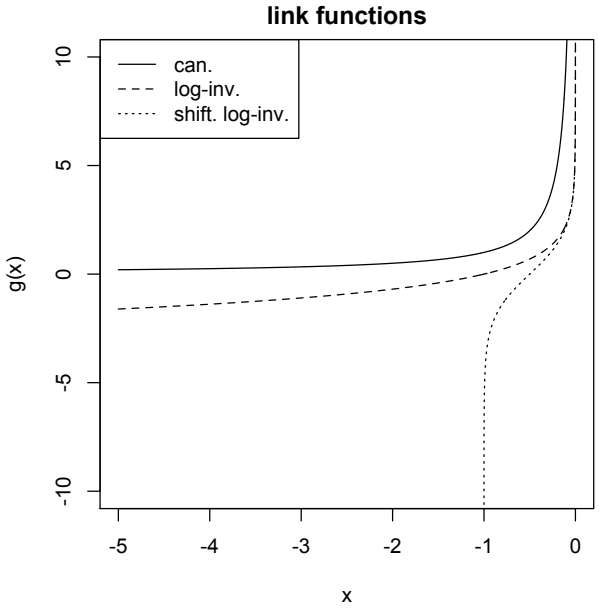
$$\begin{aligned} \log L(\boldsymbol{\vartheta} \,|\, \underline{\boldsymbol{y}}) &= \sum_{i=1}^{n} \left(-\log(x_i - \mu) - \log(\sqrt{2\pi\phi}) - \frac{(\log(x_i - \mu) - \ell(\eta_i))^2}{2\phi} \right) \\ &= -\sum_{i=1}^{n} z_i - n \log(\sqrt{2\pi\phi}) - \sum_{i=1}^{n} \frac{(z_i - \ell(\eta_i))^2}{2\phi}, \end{aligned}$$

with $z_i = \log(y_i - \mu)$. Hence

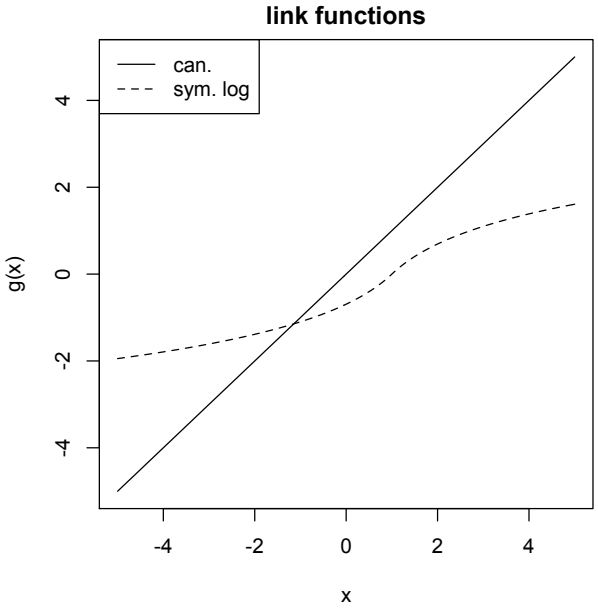
$$\log L(\widehat{\boldsymbol{\vartheta}} \,|\, \underline{\boldsymbol{y}}) = -\sum_{i=1}^{n} z_i - n \log(\sqrt{2\pi\phi}) - \frac{1}{2\phi} \sum_{j=1}^{d} \sum_{i, x_i^{(2),j}=1} (z_i - \overline{z}_n^{(j)})^2.$$

Using $\widehat{\phi} = \frac{1}{n} \sum_{j \in J} \sum_{i, x_i^{(2),j}=1} \left(z_i - \overline{z}_n^{(j)} \right)^2$ leads to the desired result.

C Link functions and descriptive statistics



(a) Link functions for Pareto 1



(b) Link functions for shifted log-normal

Figure 4: Graphs of link functions

Table 10: Empirical quantiles and moments (in euros)

	Amount	Risk class 1	Risk class 2	Risk class 3	Risk class 4	Risk class 5	
Min.	0	0	0	0	0	0	
1st Qu.	51	140	39	130	150	27	
Median	761	1120	652	1073	1015	737	
3rd Qu.	3,003	4,169	2,474	4,486	4,155	3,113	
Max.	15,688,300	15,315,173	15,688,300	11,916,121	6,078,593	10,833,825	
Mean	10,745	14,508	7,265	28,082	18,193	11,179	
Std dev.	128,146	148,380	98,141	275,175	140,607	125,004	
Skewness	54	48	96	24	24	38	
Kurtosis	4,473	3,933	12,753	751	796	2,124	
	Guarantee 1	Guarantee 2	Guarantee 3	Guarantee 4	Guarantee 5	Guarantee 6	Guarantee 7
Min.	0	0	0	0	0	0	0
1st Qu.	123	155	235	128	2	1	2
Median	1,253	814	1,955	893	2,977	2	564
3rd Qu.	4,994	2,664	8,246	3,726	39,647	1,560	2,097
Max.	3,882,524	4,529,249	15,315,173	14,272,522	15,688,300	4,888,656	4,670,686
Mean	7,022	4,055	28,429	32,328	110,056	8,388	7,157
Std dev.	39,581	24,620	280,500	273,958	534,337	74,969	60,916
Skewness	49	85	38	22	16	42	35
Kurtosis	3,955	13,620	1,833	738	366	2,399	1,927

References

Albert, A. and Anderson, J.A. (1984), 'On the existence of maximum likelihood estimates in logistic regression models', Biometrika **71**(1), 1–10.

Beirlant, J. and Goegebeur, Y. (2003), 'Regression with response distributions of pareto-type', Computational statistics & data analysis **42**(4), 595–619.

Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004), Statistics of extremes: Theory and applications, Wiley & Sons.

Beirlant, J., Goegebeur, Y., Verlaak, R. and Vynckier, P. (1998), 'Burr regression and portfolio segmentation', Insurance: Mathematics and Economics **23**(3), 231–250.

Bühlmann, H. and Gisler, A. (2006), A course in credibility theory and its applications, Springer Science & Business Media.

Chavez-Demoulin, V., Embrechts, P. and Hofert, M. (2015), 'An extreme value approach for modeling operational risk losses depending on covariates', Journal of Risk and Insurance.

Davison, A. and Smith, R. (1990), 'Models for exceedances over high thresholds', Journal of the Royal Statistical Society. Series B **52**(3), 393–442.

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

Fienberg, S.E. (2007), The Analysis of Cross-Classified Categorical Data, 2nd Ed, Springer.

Goldburd, M., Khare, A. and Tevet, D. (2016), Generalized linear models for insurance rating, CAS Monograph Series Number 5, Casualty Actuarial Society.

Haberman, S.J. (1974), 'Log-linear models for frequency tables with ordered classifications', Biometrics **30**(4), 589–600.

Hambuckers, J., Heuchenne, C. and Lopez, O. (2016), A semiparametric model for generalized pareto regression based on a dimension reduction assumption. HAL.
URL: *https://hal.archives-ouvertes.fr/hal-01362314/*

Hogg, R. V. and Klugman, S. A. (1984), Loss distributions, John Wiley & Sons.

Johnson, N., Kotz, S. and Balakrishnan, N. (2000), Continuous Univariate Distributions, Vol. 1, 2nd edn, Wiley.

Lehmann, E.L. and Casella, G. (1998), Theory of Point Estimation, 2nd Ed, Springer.

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

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

Nelder, J. A. and Wedderburn, R. W. M. (1972), 'Generalized linear models', Journal of the Royal Statistical Society. Series A **135**(3), 370–384.

Ohlsson, E. and Johansson, B. (2010), Non-Life Insurance Pricing with Generalized Linear Models, Springer.

Olver, F. W. J., Lozier, D. W., Boisvert, R. F. and Clark, C. W., eds (2010), NIST Handbook of Mathematical Functions, Cambridge University Press.
URL: *http://dlmf.nist.gov/*

Ozkok, E., Streftaris, G., Waters, H. R. and Wilkie, A. D. (2012), 'Bayesian modelling of the time delay between diagnosis and settlement for critical illness insurance using a burr generalised-linear-type model', Insurance: Mathematics and Economics **50**(2), 266 – 279.
URL: *http://www.sciencedirect.com/science/article/pii/S0167668711001326*

R Core Team (2019), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Reiss, R. and Thomas, M. (2007), Statistical Analysis of Extreme Values, 3rd edn, Basel: Birkhauser.

Rigby, R. and Stasinopoulos, D. (2005), 'Generalized additive models for location, scale and shape', Applied Statistics **54**(3), 507–554.

Smyth, G.K. and Verbyla, A.P. (1999), 'Adjusted likelihood methods for modelling dispersion in generalized linear models', Environmetrics **10**(6), 696–709.

Silvapulle, M.J. (1981), 'On the existence of maximum likelihood estimators for the binomial response models', Journal of the Royal Statistical Society. Series B (Methodological) **43**(3), 310–313.

Venables, W. and Ripley, B. (2002), Modern Applied Statistics with S, Springer.