J COMP SCI APPL MATH
Journal of Computer Science and Applied Mathematics
Vol. 5, no.2, (2023), 103–116
ISSN: 1857-9582 https://doi.org/10.37418/jcsam.5.2.5

STATISTICAL INFERENCE IN THE GENERALIZED EXTREME VALUE REGRESSION MODEL BASED ON SIMULATION STUDY

LO Fatimata, Ba Demba Bocar[1], and Diop Aba

ABSTRACT. Generalized extreme value (GEV) regression model is widely used when the dependent variable Y represents a rare event. In this case the logistic regression model shows relevant drawbacks. The quantile function of the GEV distribution is used as link function to investigate the relationship between the binary outcome Y and a set of potential predictors $\mathbf{X}$. Maximum likelihood estimators in this model has been proposed, and their asymptotic properties recently established. We conduct a detailed simulation study of its numerical properties. We evaluate its accuracy and the quality of the normal approximation of its asymptotic distribution. We study the quality of the approximation for constructing asymptotic Wald-type tests of hypothesis. Several others aspects of this model, such as the event probabilities still deserve attention. We also propose estimator of this quantity and we investigate its properties both theoretically and via simulations. Based on these results, we provide recommendations about the range of minimum sample size under which a reliable statistical inference on the event probabilities can be obtained in a GEV regression model. A real-data example illustrates the proposed estimators.

[1]*corresponding author*

Key words and phrases. Generalized extreme value, Regression model, Maximum likelihood estimation, Simulation study, Stroke.

Submitted: 17.08.2023; *Accepted:* 02.09.2023; *Published:* 19.10.2023.

1. INTRODUCTION

Generalized extreme value regression model has become a standard tool to investigate the relationship between a binary response Y which is often present in medical studies and a set of potential predictors [12]. The binary response Y may represent the infection status with respect to some disease ($Y = 1$ if the individual is infected and $Y = 0$ otherwise). If $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})'$ denotes the corresponding (p-dimensional, say) predictor (or covariate) for the i-th individual, generalized extreme value regression models the conditional probability of infection $\pi(\mathbf{X}_i) = \mathbb{P}(Y_i = 1|\mathbf{X}_i)$ as

$$\frac{1 - [\log(1 - \pi(\mathbf{X}_i))]^{-\tau}}{\tau} = \beta'\mathbf{X}_i, \tag{1.1}$$

where

$$\pi(\mathbf{X}_i) = 1 - \exp\left(-\left[(1 - \tau(\beta'\mathbf{X}_i))_+\right]^{-1/\tau}\right) = 1 - \text{GEV}(-\beta'\mathbf{X}_i; \tau), \tag{1.2}$$

and $\beta = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ is an unknown regression parameter measuring the association between potential predictors and the risk of infection (for a susceptible individual) and $GEV(x; \tau)$ represents the cumulative probability at $\mathbf{X}$ for the GEV distribution with a location parameter $\mu = 0$, a scale parameter $\sigma = 1$, an unknown shape parameter $\tau \in \mathbb{R}$. A more detailed discussion on the extreme value distributions can be found in [6] and [4]. Estimation and testing procedures in model (1.1) are well established (see, e.g., [7], [3]). These procedures are usually based on the maximum likelihood estimator (MLE) of β, which is consistent and approximately normally distributed in large samples.

Interpreting the results after estimating and testing the vector parameter β usually requires the additional estimation of: (i) the relative risk $R_j = \frac{\pi(x_j)}{\pi(\bar{x}_j)}$ between the response Y and the predictors $x_j (j = 2, \ldots, p)$, where $\bar{x}$ is the complementary of x (ii) the probabilities $\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$ of a positive outcome at $\mathbf{x}$. On the contrary and to the best of our knowledge, estimators of the odds and event probabilities in the GEV regression model have never been investigated. The present work aims at filling this gap. We construct estimators of these quantities and we derive their asymptotic properties and we evaluate their finite-sample behaviors via simulations. Finally, we illustrate the proposed estimators on a dataset about stroke.

The rest of the paper is organized as follows. In Section 2, we recall the properties of the MLE of the regression parameter in the GEV regression model. Then, we construct estimators of the odds and probability $\pi(\mathbf{x})$ of observing the outcome on a susceptible individual at $\mathbf{x}$, and we obtain their asymptotic distributions. Section 3 describes our simulation study. We examine the finite-sample behaviors of the proposed estimators. The real data example is treated in Section 4. Discussion and some recommendations about the use of GEV regression model are summarized in Section 5, along with some perspectives for future work.

2. ODDS AND EVENT PROBABILITY ESTIMATION IN THE GEV REGRESSION MODEL

Let $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$ be independent and identically distributed copies of the random vector $(Y, \mathbf{X})$ defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For every individual $i = 1, \ldots, n$, Y_i is a binary response variable indicating say, the infection status with respect to some disease (that is, $Y_i = 1$ if the i-th individual is infected, and $Y_i = 0$ otherwise). Let $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})'$ be random vectors of predictors or covariates.

The likelihood function for the unknown p-dimensional parameter β from the independent sample $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ is as follows:

$$L_n(\beta) = \prod_{i=1}^{n} [1 - \text{GEV}(-\beta'\mathbf{x}_i; \tau)]^{y_i} \times [\text{GEV}(-\beta'\mathbf{x}_i; \tau)]^{1-y_i}\,. \tag{2.1}$$

We define the maximum likelihood estimator $\widehat{\beta}_n$ as the solution of the p-dimensional score equation

$$\frac{\partial \log L_n(\beta)}{\partial \beta} = 0. \tag{2.2}$$

The consistency and asymptotic normality of $\hat{\beta}_n$ of β have been established by [7]. They have proved that the asymptotic covariance matrix

$$\mathcal{I}_{\beta}^{-1} = \left(-\mathbb{E}\left[\frac{\partial^2 \log L_n(\beta)}{\partial \beta \partial \beta'} \right] \right)^{-1}$$

of $\hat{\beta}_n$ can be consistently estimated by $\hat{\mathcal{I}}_{\hat{\beta}_n}^{-1}$ where $\hat{\mathcal{I}}_{\hat{\beta}_n} = -\frac{1}{n}\left[\frac{\partial^2 \log L_n(\beta)}{\partial \beta \partial \beta'} \right] |_{\beta = \hat{\beta}_n}$.

Theorem 2.1 (Existence and consistency ([7])). *The maximum likelihood estimator $\hat{\beta}_n$ exists almost surely as $n \to \infty$ and converges almost surely to β_0, if and only if λ_n tends to infinity as $n \to \infty$.*

Theorem 2.2 (Asymptotic normality ([7])). *Assume that $\hat{\beta}_n$ converges almost surely to β_0. Let $\widehat{\Sigma} = \mathbb{X}\mathbf{D}(\hat{\beta}_n)\mathbb{X}'$ and I_p denote the identity matrix of order p. Then $\widehat{\Sigma}_n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0)$ converges in distribution to the Gaussian vector $\mathcal{N}(0, I_p)$.*

These results shown by [7] allow us to construct estimators of the probability $\pi(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$ of event occurrence for a susceptible individual with covariate $\mathbf{x}$ and of the odds in the GEV regression models (1.1) and (1.2).

The conditional probability and the relative risk are naturaly estimated by

$$\hat{\pi}_n(\mathbf{x}) = 1 - \exp\left(\left[(1 - \hat{\tau}(\hat{\beta}_n'\mathbf{x}))_+\right]^{-1/\hat{\tau}}\right) \text{ and } \hat{R}_{j,n} = \frac{\hat{\pi}_n(x_j)}{\hat{\pi}_n(\bar{x}_j)}. \tag{2.3}$$

The asymptotic properties of $\hat{\pi}_n(\mathbf{x})$ are are summarized in the following theorem.

Theorem 2.3. *As n tends to infinity, $\sqrt{n}(\hat{\pi}_n(\mathbf{x}) - \pi(\mathbf{x}))$ converges in distribution to a zero mean Gaussian with variance*

$$\sigma_{\mathbf{x}}^2 = (-\hat{\tau}x')^{(-1/\tau-1)} \exp\left(\left[(1 - \hat{\tau}(\hat{\beta}_n'\mathbf{x}))_+\right]^{-1/\hat{\tau}}\right) \mathbf{x}'\mathcal{I}_{\hat{\beta}_n}^{-1}\mathbf{x}/\hat{\tau}.$$

Moreover $(-\hat{\tau}x')^{(-1/\tau-1)} \exp\left(\left[(1 - \hat{\tau}(\hat{\beta}_n'\mathbf{x}))_+\right]^{-1/\hat{\tau}}\right) \mathbf{x}'\hat{\mathcal{I}}_{\hat{\beta}_n}^{-1}\mathbf{x}/\hat{\tau}$ converge in probability to $\sigma_{\mathbf{x}}^2$.

Proof. The result follows by applying the delta-method ([11]) to the transformation

$$\hat{\beta}_n \mapsto 1 - \exp\left(\left[(1 - \hat{\tau}(\hat{\beta}_n'\mathbf{x}))_+\right]^{-1/\hat{\tau}}\right).$$

The consistency of the variance estimator follows from the consistency of $\hat{\beta}_n'$ and $\hat{\mathcal{I}}_{\hat{\beta}_n}^{-1}$ and the continuous mapping theorem. □

In the next section, we investigate via simulations the asymptotic properties of the estimators $\hat{\beta}_n'$ and $\hat{\pi}_n(\mathbf{x})$ by considering various measures of the accuracy of these estimators.

3. SIMULATION STUDY

In this section, we investigate the numerical properties of the maximum likelihood estimators $\hat{\beta}_n$ and $\hat{\pi}_n(\mathbf{x})$, under various conditions. We compare via simulations, the performance of three links functions (the real model GEV and the others models misspecifies the susceptibility probability $\pi(x_i)$: Logit and Truncated normal distribution values in interval (0,1)).

3.1. **Simulation-based study of $\hat{\beta}_n$.** The simulation setting is as follows. We first consider the following models

- **GEV link function**

$$\mathbb{P}(Y_i = 1|X_i) = 1 - GEV(\beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5}; \tau) \tag{3.1}$$

- **Logit link function**

$$\log\left(\frac{\mathbb{P}(Y_i = 1|X_i)}{1 - \mathbb{P}(Y_i = 1|X_i)}\right) = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} \tag{3.2}$$

- **Truncated normal distribution values in interval (0,1) link function**

$$\mathbb{P}(Y_i = 1|X_i) = F(\beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5}; \mu, \sigma) \tag{3.3}$$

where $\mu = 0$, $\sigma = 1$ and $F(x; \mu, \sigma) = 1 - \frac{1-\Phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{\mu}{\sigma})}, x > 0$ and Φ is the cumulative distribution function of the standard normal distribution.

Here $X_{i1} = 1$ for each individual i $(i = 1, \ldots, n)$. The covariates X_{i2}, X_{i3}, X_{i4} and X_{i5} are independently drawn from normal $\mathcal{N}(0, 1)$, exponential $\mathcal{E}(1)$, $\mathcal{P}(2)$ and $\mathcal{N}(1, 1)$ respectively. The true parameter β is set such that the proportion of 1's in the simulated data sets is around 15% (considered as Model $\mathcal{M}_1$: $\beta = (1.5, -1.2, 0, -2.5, -0.3)'$) and 30% (considered as Model $\mathcal{M}_2$, $\beta = (-1.3, 0, 2.5)'$).

An i.i.d. sample of size $n \geq 1$ of the vector $(Y, \mathbf{X})$ is generated from the model (1.1-1.2), and for each individual i, we get a realization $(y_i, \mathbf{x}_i)$. A maximum likelihood estimator $\hat{\beta}_n$ of $\beta = (\beta_1, \beta_2, \beta_3)'$ is obtained from this dataset by solving the score equation (2.2), using the `optim` function of the software `R`. An estimate is also obtained for τ, but it is not the primary parameter of interest hence we only focus on the simulation results for $\hat{\beta}_n$. The finite-sample behavior of the maximum likelihood estimator $\hat{\beta}_n$ was assessed for several sample sizes $(n = 200, 500, 1000)$.

For each configuration (sample size, proportion of 1's) of the design parameters, $N = 1000$ samples were obtained. Based on these $N = 1000$ replicates, we obtain averaged values for the estimates of the parameters β_j, $j = 1, \ldots, 3$, which are calculated as $N^{-1}\sum_{k=1}^{N}\hat{\beta}_{j,n}^{(k)}$, where $\hat{\beta}_{j,n}^{(k)}$ is the estimate obtained from the k-th simulated sample. The quality of estimates is evaluated by using the Bias and the Root Mean Square Error (RMSE) defined as, for $j = 1, 2, 3$:

$$\text{Bias}(\hat{\beta}_{n,j}) = \mathbb{E}(\hat{\beta}_{n,j} - \beta) \approx \frac{1}{N}\sum_{k=1}^{N}\left(\hat{\beta}_{j,n}^{(k)} - \beta\right),$$

$$\text{RMSE}(\hat{\beta}_{n,j}) = \sqrt{\mathbb{E}\left[(\hat{\beta}_{n,j} - \beta)^2\right]} \approx \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\hat{\beta}_{j,n}^{(k)} - \beta\right)^2}.$$

The results from the models (3.1)-(3.2)-(3.3) are summarized in Tables 1 and 2.

TABLE 1. Simulation results for the estimator Model $\mathcal{M}_1$

		n = 200			n = 500			n = 1000		
Paramater	**Model**	**MLE**	**BIAS**	**RMSE**	**MLE**	**BIAS**	**RMSE**	**MLE**	**BIAS**	**RMSE**
$\widehat{\beta}_{1,n}$	Logit	2.417	0.917	0.982	2.414	0.914	0.932	2.419	0.919	0.919
	Truncnorm	2.005	0.506	0.677	1.995	0.495	0.593	1.884	0.384	0.420
	Gev	1.702	0.202	0.485	1.646	0.146	0.421	1.570	0.070	0.268
$\widehat{\beta}_{2,n}$	Logit	-1.924	-0.724	0.756	-1.864	-0.664	0.672	-1.784	-0.584	0.587
	Truncnorm	-1.585	-0.385	0.667	-1.401	-0.201	0.296	-1.375	-0.175	0.225
	Gev	-1.291	-0.091	0.355	-1.251	-0.051	0.256	-1.228	-0.028	0.168
$\widehat{\beta}_{3,n}$	Logit	-0.038	-0.038	0.449	-0.012	-0.012	0.235	0.001	0.001	0.167
	Truncnorm	-0.012	-0.012	0.276	-0.007	-0.007	0.130	0.001	0.001	0.092
	Gev	-0.006	-0.006	0.244	-0.006	-0.006	0.115	-0.001	-0.001	0.081
$\widehat{\beta}_{4,n}$	Logit	-4.002	-1.502	1.530	-3.995	-1.495	1.499	-3.939	-1.439	1.434
	Truncnorm	-3.293	-0.793	1.347	-2.914	-0.414	0.565	-2.854	-0.354	0.434
	Gev	-2.708	-0.208	0.650	-2.617	-0.117	0.480	-2.563	-0.063	0.318
$\widehat{\beta}_{5,n}$	Logit	-0.783	-0.483	0.693	-0.620	-0.320	0.412	-0.612	-0.312	0.349
	Trunnorm	-0.411	-0.111	0.314	-0.345	-0.045	0.150	-0.336	-0.036	0.088
	Gev	-0.370	-0.070	0.302	-0.310	-0.010	0.139	-0.309	-0.009	0.085

TABLE 2. Simulation results for the estimator Model $\mathcal{M}_2$

		n = 200			n = 500			n = 1000		
Paramater	**Model**	**MLE**	**BIAS**	**RMSE**	**MLE**	**BIAS**	**RMSE**	**MLE**	**BIAS**	**RMSE**
$\widehat{\beta}_{1,n}$	Logit	-0.551	-0.051	1.034	-0.485	0.015	0.508	-0.513	-0.013	0.262
	Truncnorm	-0.548	-0.048	0.392	-0.458	0.042	0.181	-0.461	0.039	0.110
	Gev	-0.536	-0.036	0.192	-0.512	-0.012	0.177	-0.510	-0.010	0.096
$\widehat{\beta}_{2,n}$	Logit	-0.014	-0.014	0.463	0.014	0.014	0.236	0.007	0.007	0.157
	Truncnorm	-0.007	-0.007	0.248	0.007	0.007	0.130	0.004	0.004	0.086
	Gev	-0.006	-0.006	0.192	0.004	0.004	0.115	0.003	0.003	0.076
$\widehat{\beta}_{3,n}$	Logit	-2.987	-0.987	1.006	-2.868	-0.868	0.872	-2.717	-0.717	0.743
	Truncnorm	-2.408	-0.408	0.615	-2.326	-0.326	0.474	-2.244	-0.244	0.329
	Gev	-2.178	-0.178	0.441	-2.124	-0.124	0.349	-2.056	-0.056	0.234
$\widehat{\beta}_{4,n}$	Logit	2.475	0.975	0.991	2.383	0.883	0.889	2.315	0.815	0.821
	Truncnorm	1.967	0.467	0.701	1.728	0.228	0.315	1.683	0.183	0.234
	Gev	1.677	0.177	0.360	1.595	0.095	0.264	1.540	0.040	0.167
$\widehat{\beta}_{5,n}$	Logit	-4.686	-1.686	1.702	-4.383	-1.383	1.383	-4.382	-1.382	1.381
	Truncnorm	-3.623	-0.623	0.865	-3.364	-0.364	0.483	-3.355	-0.355	0.438
	Gev	-3.157	-0.157	0.462	-3.124	-0.124	0.407	-3.075	-0.075	0.303

From the Tables 1 and 2, it appears that the proposed maximum likelihood estimator $\widehat{\beta}_n$ given by the model (3.1) with GEV link function provides a reasonable approximation of the true parameter value better than those provided by the others link functions (3.2-3.3), even when the sample size is less than 200 with a poor percentage of 1's in the model (that is, when the GEV link function is used to generate the data, the estimates in the GEV model are of good quality). On the other hand, the quality of the estimates may be quite poor when the sample size is less than 100. Finally, these results indicate that a reliable statistical inference on the regression effects and event probabilities in the regression model for binary data with a GEV link function should be based on a sample having, at least, a moderately large size ($n \geq 500$, say).

Now, we investigate the quality of the normal approximation of the asymptotic distribution of $\widehat{\beta}_{j,n}$. For each configuration of the design parameters, we obtain the histograms of the $\hat{\beta}_{j,n}^{(k)}, k = 1, \ldots, N$ and the corresponding density plots. Figures 1 display the graphs for $j = 1, \ldots, 5$ (the graphs for the model $\mathcal{M}_2$ are similar and are thus not given).

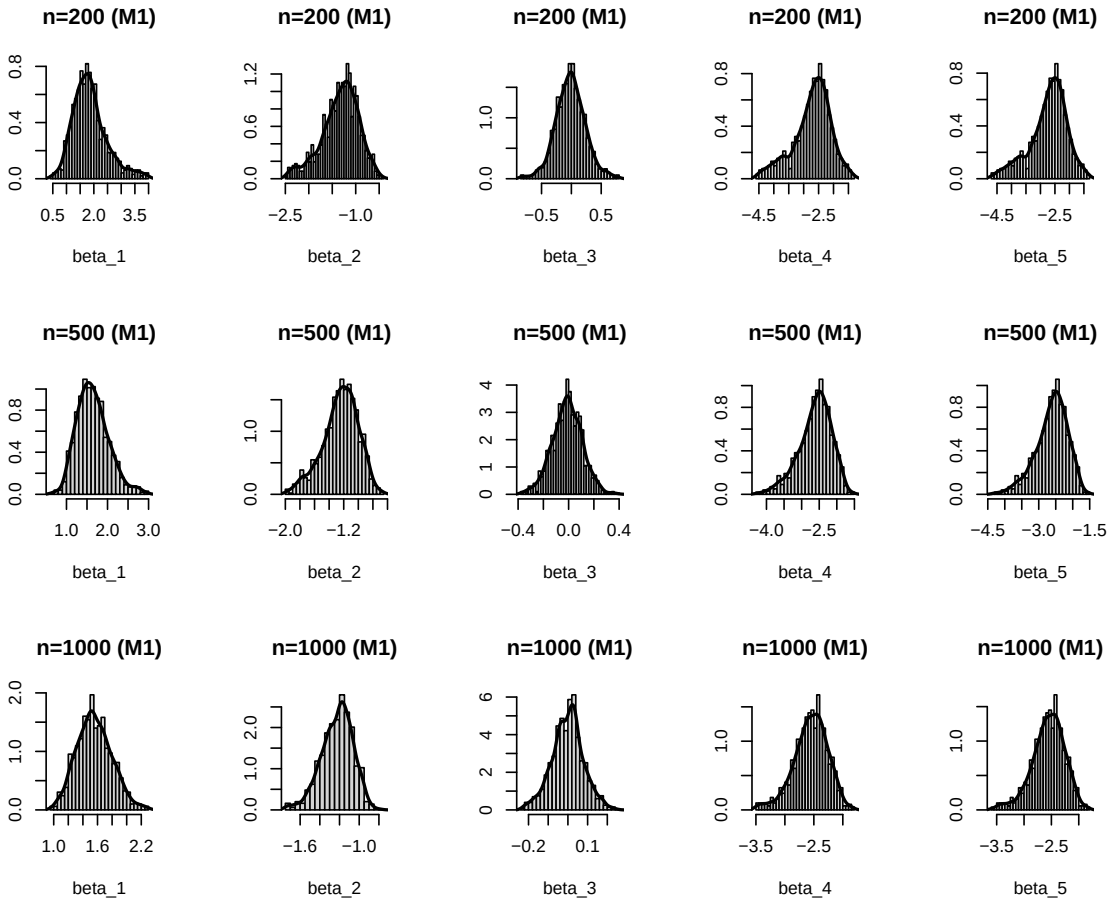

FIGURE 1. Histograms and Density plots for $\hat{\beta}_{j,n}$, in model $\mathcal{M}_1$.

It appears from these graphs that the normal approximation is reasonably satisfied when the sample size is large enough ($n \geq 500$ say). The quality of the approximation is poor when the sample size is less than 100. All these graphs corroborate the conclusions we drew from Table 1. We now investigate the properties of the estimator $\hat{\pi}_n(\mathbf{x})$.

3.2. **Simulation-based study of $\hat{\pi}_n(\mathbf{x})$.** The simulation setting is as follows. We consider the same following model

$$\frac{1-[\log(1-\pi(\mathbf{X}_i))]^{-\tau}}{\tau} = \beta_1\beta_2 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5}, \tag{3.4}$$

where the covariates X_{i2}, X_{i3}, X_{i4} and X_{i5} are independently drawn from normal $\mathcal{N}(0,1)$, exponential $\mathcal{E}(1)$, $\mathcal{P}(2)$ and $\mathcal{N}(1,1)$ respectively. The true parameter β is set such that the proportion of 1's in the simulated data sets is around 15% (considered as Model $\mathcal{M}_1$: $\beta = (1.5, -1.2, 0, -2.5, -.3)'$) and 30% (considered as Model $\mathcal{M}_2$, $\beta = (-.5, 0, -2, 1.5, -3)'$).

We consider the problem of estimating one value of $\pi(\mathbf{x})$ for each model $\mathcal{M}_1$ and $\mathcal{M}_2$, for a given $\mathbf{x}$. As an example, we wish to estimate the event probability for an individual (labeled h) having $x_{h,1} = 1$, $x_{h,2} = 0.5$, $x_{h,3} = 1$, $x_{h,4} = 0.5$, $x_{h,5} = -2$ and $\tau = -0.45$. Using these values, the probability $\pi(\mathbf{x}_h)$ to be estimated is:

- in the model $\mathcal{M}_1$: $\pi(\mathbf{x}_h) = 1 - \exp\big(-\big[(1 + 0.45(1.5 - 1.2 \times 0.5 + 0 \times 1 - 2.5 \times 0.5 + 0.3 \times 2))_+\big]^{1/0.45}\big) \approx 0.718$.
- in the model $\mathcal{M}_2$: $\pi(\mathbf{x}_h) = 1 - \exp\big(-\big[(1 + 0.45(-.5 + 0 \times 0.5 - 2 \times 1 + 1.5 \times .5 + 3 \times 2))_+\big]^{1/0.45}\big) \approx 0.999$.

The results from the model (3.4)are summarized in Table 3.

TABLE 3. Simulation results for the predicted probability $\pi(\mathbf{x}_h)$

	Model $\mathcal{M}_1$			Model $\mathcal{M}_2$		
n	MLE	BIAS	RMSE	MLE	BIAS	RMSE
200	0.722	0.004	0.208	0.998	0.001	0.001
500	0.720	0.002	0.123	0.998	0.001	0.001
1000	0.717	-0.001	0.014	0.999	0.000	0.001

From 3, it appears that $\hat{\pi}_n(\mathbf{x}_h)$ provides a reasonable estimate of $\pi(\mathbf{x}_h)$ when the proportion of 1's is moderate (namely, 15%) or large (namely, 30%), even for $n = 200$.

Now, we investigate the quality of the normal approximation of the asymptotic distribution of $\hat{\pi}_n(\mathbf{x}_h)$. For each configuration of the design parameters, we obtain the histograms of the $\hat{\pi}_n^{(k)}(\mathbf{x}_h), k = 1, \ldots, N$ and the corresponding density plots (for the model $\mathcal{M}_1$ only; the results for the model $\mathcal{M}_2$ yield similar observations and are thus omitted). The graphs are displayed in Figure 2. From these figures, the normal approximation is reasonably when the sample size is is sufficiently large ($n \geq 500$, say). The distribution of $\hat{\pi}_n(\mathbf{x}_h)$ can be highly skewed otherwise. These findings are coherent with our previous observations for $\hat{\beta}_{j,n}$ especially when the sample size is small (n around 500, say).

Overall, our results indicate that a reliable statistical inference on the parameters and event probabilities in the GEV regression model can be obtained in samples having at least a moderately large size ($n \geq 500$, say), when the proportion of 1's is low ($\leq 15\%$) to moderately large (30% say). When the sample size is small (n around 200, say), the results should be considered very carefully, considering the increase in the variability of the estimators and the skewness of their distributions.

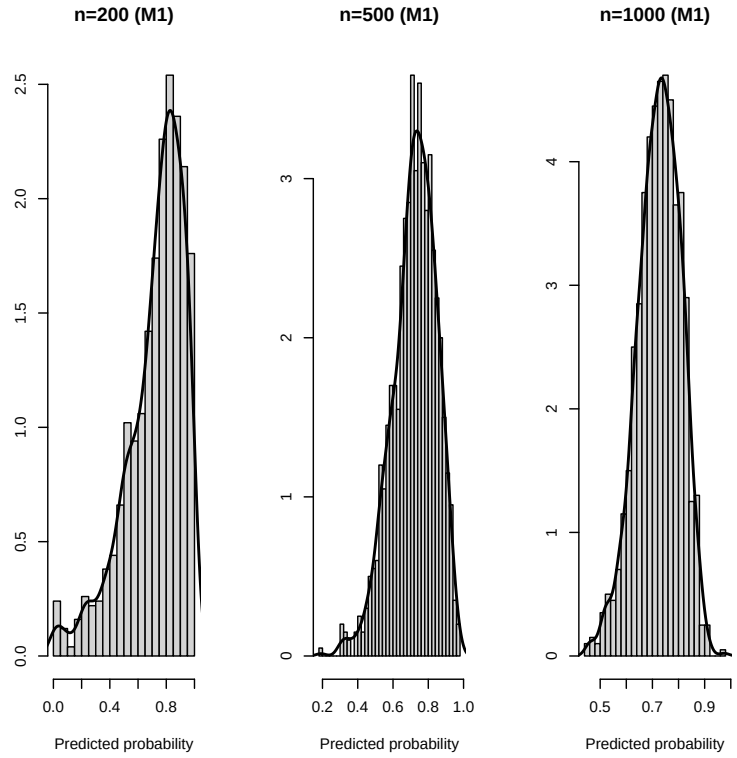

FIGURE 2. Histograms and Density plots for the predicted probabilitiy, in model $\mathcal{M}_1$.

4. REAL DATA APPLICATION

In this setion, we consider an application on Stroke data in central Senegal. Stroke is a sudden neurological deficit of vascular origin caused by an infarct or haemorrhage in the brain (see [2], [1] for more details). We consider here a database of size $n = 162$. The data was collected in the context of a prospective

and analytical study, carried out on a period of 8 months from april 5 to november 30, 2016 at Medical Imagery Service of both Matlaboul Fawzeini Hospital in Touba and Elhadj Ibrahima Niass regional hospital in Kaolack located in central Senegal. In Senegal, stroke is the most frequent neurological disease. Known for their high mortality and morbidity rates, they account for more than 30% of hospital admissions and nearly two-thirds of the loss of human life (see [9], [10]).

Patients with CT confirmation of stroke were included in the study. We aim at investigating, based on this dataset, the factors which may explain an unfavourable evolution of their health status. In this study, the dependent variable is the evolution of the health status of stroke patients (vital prognosis). We denote Y the binary variable defined as follows:

$$Y_i = \begin{cases} 1, & \text{if the vital prognosis evolves favourably,} \\ 0, & \text{if the vital prognosis evolves unfavourably.} \end{cases}$$

We consider the following covariates: *age*, *delay* (delay between the first symptoms and admission to hospital), *severity cerebral commitment* (displacement of parts of the nervous structure contained in the cranium through an orifice) and *intraventricular haemorrhages* (bleeding into the ventricles of the brain).

We ran a generalized extreme value regression analysis of the model defined as follows:

$$\begin{aligned} \mathbb{P}(Y_i = 1|\mathbf{x}) = 1 - GEV(-(&\beta_1 + \beta_2 \times \textit{age} + \beta_3 \times \textit{delay} \\ &+ \beta_4 \times \textit{severity} + \beta_5 \times \textit{intraventricular}; \tau)). \end{aligned} \tag{4.1}$$

The final results of these fitting procedure are given in Table 4. We compare these results with those obtained from the logistic regression model.

TABLE 4. Stroke data analysis

	Model GEV		Model LOGIT	
Variable	Estimate	Stand. error	Estimate	Stand. error
Intercept	1.1766	0.1375	-0.9621	0.9452
Age	-0.6065	0.0691	-0.6476	0.3277
Delay	0.1469	0.0691	0.9202	0.1801
Severity	-0.1521	0.0698	-0.9632	0.5080
Intraventricular	-0.1424	0.0687	-2.1483	0.6596
τ	-2.0127	0.0002		

All these variables are significant at level 5%. The adjusted probability of vital prognosis for the i-th individual given by the GEV regression model is defined as follows:

$$\mathbb{P}(Y_i = 1|\mathbf{x}) = 1 - GEV(-(1.1766 - 0.6065 \times \textit{Age}0.1469 \times \textit{Delay} - 0.1521 \times \textit{Severity} - 0.1424 \times \textit{Intraventricular}; -2.0127)).$$

We observe that the model concludes to the significance of the covariates age, delay, severity cerebral commitment and intraventricular haemorrhages. We also find that the estimators of the GEV regression model are more accurate than those of the logistic regression model. Indeed they have much smaller standard errors.

Age has a significant influence on the vital prognosis of stroke patients. Older patients are more at risk than other patents. Indeed, age favours the degradation of functioning of blood vessels. This imbalance is also highlighted by the chronicity of cardiovascular risk factors, which increases the vulnerability to stroke ([13]). The scanner must be performed urgently, at best within the first six hours after the onset of symptoms. We were also interested in the time between the first signs of stroke and admission to the health facility and the time between admission and CT scan. The first was more significant for unfavourable evolution of the vital prognosis of stroke patients. This explains why this delay increases the vital prognosis because it reduces the possibilities of functional rehabilitation. For severity signs, their appearance seriously engages the vital prognosis.

5. Discussion and perspectives

This article reports the results of a detailed simulation study of maximum likelihood based estimators of the event probabilities in a GEV regression model. From our results, these estimators perform quite well under reasonable conditions regarding the sample size and proportion of 1's. More precisely, reliable statistical inferences on the event probabilities and (point estimation, normal approximation) should be obtained when the sample size is low ($n \leq 100$) to moderately large ($n = 500$, say). These findings may serve as practical guidelines for the analysis of binary datasets for rare events in a variety of settings: medicine especially.

Another issue of interest deals with the inference in the GEV regression model in a high-dimensional setting, when the predictor dimension is much larger than the sample size (this problem arises, for example, in genetic studies where high-dimensional data are generated using microarray technologies). Some articles ([5], [8]) have addressed the estimation in the logistic model for binary data. Extending these methods to the GEV regression model constitutes another non-trivial topic for future work.

ACKNOWLEDGMENTS

The authors are grateful to the associate editor and one referee for their detailed comments that helped us to improve a previous version of this article.

DECLARATIONS

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no conflicts of interest to declare that are relevant to the content of this article.

REFERENCES

[1] V. BIOUSSE: *Etiologie et mécanisme des accidents vasculaires cérébraux*, Ann. Radiol., **37** (1994), 11-16.

[2] J. BOGOUSSLAVSKY, M.G. BOUSSER, J.L. MAS: *Les accidents vasculaires cérébraux*, Radiologie ed Doinville, **145** (1993).

[3] R. CALABRESE, S. OSMETTI: *Modelling SME Loan Defaults asRare Events: an Application to Credit Defaults*, Journal of Applied Statistics **40**(6) (2013), 1172-1188.

[4] S.G. COLES: *An Introduction to Statistical Modeling of Extreme Values*, Springer, New York, 2001.

[5] J. HUAN, S. MA, C.H. ZHANG: *The Iterated Lasso forHigh-Dimensional Logistic Regression*, Technical Report, **392**, 2008.

[6] S. KOTZ, S. NADARAJAH: *Extreme Value Distributions*, Theory and Applications, Imperial Colleg Press, London, 2000.

[7] F. LO, D.B. BA, A. DIOP: *Maximum likelihood estimation in the generalized extreme value regression model for binary data*, Gulf Journal of Mathematics. **12**(2) (2022), 49-56.

[8] L. MEIER, S. VAN DE GEER, P. BUHLMANN: *The group Lasso for logistic regression*, Journal of the Royal Statistical Society, Series B **70** (2008), 53–71.

[9] F. SENE-DIOUF: *The management of cerebrovascular events in Senegal*, Revue Neurologique, **163**(8–9) (2007), 823-827.
[10] K. TOURÉ *et al.*: *Mortalité des patients hospitalisés pour AVC ischémique en neurologie au CHU de Fann à Dakar*, Neurologie-Psychiatrie-Gériatrie, **17**(100) (2016), 230-234.
[11] A.W. VAN DER VAART: *Asymptotic Statistics*, Cambridge: Cambridge University Press, 1998.
[12] X. WANG: *Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption*, Ann. Appl. Stat. **4** (2010), 2000-2023.
[13] M. ZUBER, J.L. MAS: *Epidémiologie des infarctus cérébraux*, Ann. Radiol., **37** (1994), 7-10

DEPARTMENT OF MATHEMATICS
UNIVERSITY IBA DER THIAM OF THIES
THIES,
SENEGAL.
Email address: fatymalo@hotmail.fr

DEPARTMENT OF MATHEMATICS
UNIVERSITY IBA DER THIAM OF THIES
THIES,
SENEGAL.
Email address: dbba@univ-thies.sn

DEPARTMENT OF MATHEMATICS
UNIVERSITY ALIOUNE DIOP OF BAMBEY
BAMBEY,
SENEGAL.
Email address: aba.diop@uadb.edu.sn