Computational Statistics (2024) 39:181–213
https://doi.org/10.1007/s00180-023-01330-y

ORIGINAL PAPER

Bootstrapping binary GEV regressions for imbalanced datasets

Michele La Rocca[1] · **Marcella Niglio**[1] · **Marialuisa Restaino**[1]

Received: 2 May 2022 / Accepted: 26 January 2023 / Published online: 4 February 2023

Abstract

This paper proposes and discusses a bootstrap scheme to make inferences when an imbalance in one of the levels of a binary variable affects both the dependent variable and some of the features. Specifically, the imbalance in the binary dependent variable is managed by adopting an asymmetric link function based on the quantile of the generalized extreme value (GEV) distribution, leading to a class of models called *GEV regression*. Within this framework, we propose using the fractional-random-weighted (FRW) bootstrap to obtain confidence intervals and implement a multiple testing procedure to identifying the set of relevant features. The main advantages of FRW bootstrap are as follows: (1) all observations belonging to the imbalanced class are always present in every bootstrap resample; (2) the bootstrap can be applied even when the complexity of the link function does not allow to easily compute second-order derivatives for the Hessian; (3) the bootstrap resampling scheme does not change whatever the link function is, and can be applied beyond the GEV link function used in this study. The performance of the FRW bootstrap in GEV regression modelling is evaluated using a detailed Monte Carlo simulation study, where the imbalance is present in the dependent variable and features. An application of the proposed methodology to a real dataset to analyze student churn in an Italian university is also discussed.

Keywords GEV regression · FRW bootstrap · Imbalanced data · Variable selection

✉ Marialuisa Restaino
mlrestaino@unisa.it

Michele La Rocca
larocca@unisa.it

Marcella Niglio
mniglio@unisa.it

[1] Department of Economics and Statistics, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, Salerno, Italy

1 Introduction

In recent years, imbalanced data has attracted researchers' attention, given the exponential growth of data and rise of the phenomenon of big data, for two main reasons. First, the class imbalance problem is pervasive and intrinsic in many real situations and domains (for a review of the main applications, see Krawczyk 2001; Haixiang et al. 2017; Sun et al. 2009). Second, some statistical models and methods may be inadequate when encountering this imbalance and rareness because they are based on the assumption of equal class distribution for data (King and Zeng 2001; Wang and Dey 2010).

The imbalanced binary variables are characterized as having more instances of certain classes than others. Particularly, one class is represented by a large number of units (that is, the *majority class*, corresponding to the non-events class), while another class has only a few samples (that is, the *minority class*, related to the events class).

When the degree of imbalance is extreme, and the data are characterized by the number of ones being hundreds to thousands of times smaller than the number of zeros, the events become rare (King and Zeng 2001; Wang and Dey 2010; Bergtold et al. 2018).

Given that rare instances occur frequently and the minority class is usually the group of interest, statistical models should consider this imbalance and avoid producing biased estimates (McCullagh and Nelder 1989).

Indeed, when this rareness affects the response variable, using logistic regression based on the symmetric logit link function, could be inappropriate because the probability of rare events may be underestimated. Therefore, the units should be allocated into the majority class (the non-events) so that the bias of the maximum likelihood estimators increases (among the others see Agresti 2002).

Since the 90s, many methods have been developed for dealing with imbalanced data. The following two main groups of methods have been developed in the literature: (1) balancing the class distribution and making it suitable for the statistical models using preprocessing and/or sampling techniques, and then applying traditional models, and (2) modifying the existing classifiers to improve the bias toward majority classes to obtain better results from imbalanced data.

Sampling techniques re-balance the sample for an imbalanced dataset and mitigate the effect of skewed class distribution. Among various sampling methods developed to address this problem and eliminate the issue of skewed distribution, *oversampling* and *undersampling* are the two most used. The first creates new minority class samples, while the second removes the samples from the majority class. The two widely used oversampling methods are randomly duplicating the minority samples and SMOTE (Synthetic Minority Over-Sampling technique), which show good results across various applications (Chawla et al. 2002). The simplest undersampling method is the Random Under-Sampling (RUS), which involves the random elimination of majority class examples (Tahir et al. 2012).

However, these resampling schemes have some disadvantages. First, they change the data structure, because with undersampling, the balanced classes

loss a lot of majority class data. Moreover, oversampling creates multiple samples within the minority class, resulting in overfitting of the model. Furthermore, both procedures focus on the rareness and imbalance in binary response variable, neglecting if these characteristics are also present in the categorical features. Consequently, they might be able to re-balance the response variable, but simultaneously increase the imbalance and rareness in the covariates. For a review of the main characteristics of sampling techniques, see among the others (Japkowicz and Stephen 2002; Estabrooks et al. 2004.

Apart from sampling methods, another common technique to handle imbalanced data is reweighting the likelihood function, which consists of directly passing the weights to each observation to the likelihood function. Seiffert et al. (2008) investigate the difference between the reweighting and sampling methods for imbalanced data.

Olmus et al. (2022) present an approach to parameter estimation bias using inverse conditional distributions, and compare their approach with different penalized LR methods.

Moreover, the use of asymmetric link function has been proposed because the probability of binary response approaches zero at a significantly different rate than it approaches one (Chen et al. 1999; Kim et al. 2007). Furthermore, to appropriately model the large skewness caused by the rareness, Wang and Dey (2010) and Calabrese and Osmetti (2013) propose the use of an asymmetric link function based on the quantile of the Generalized Extreme Value (GEV) distribution, introducing the *GEV regression*.

The appropriate skewed link function remains an open problem, motivating some authors to investigate more flexible models to accommodate such imbalances.

Using the GEV regression framework, here we investigate the effect of imbalanced data on dependent and independent variables. According to the results in Calabrese and Osmetti (2013), the main novelty of our method is the use of a specific bootstrap scheme to make inferences about GEV regression models. Particularly, we implement the Fractional-Random-Weighted (FRW) bootstrap, presented in Xu et al. (2020), for GEV regression, to build confidence intervals and implement multiple testing for identifying the set of relevant features of the model. The main advantage of using the FWR bootstrap is that it offers an alternative resampling method that never fails to capture every single class, regardless of the underlying probability distribution of the classes. Given that the observations remain across all bootstrap samples, it prevents the rare events from not being in the bootstrap resample. Thus, it makes be easier to deal with the imbalance and rareness. It also avoids the estimation procedure failures and accelerates the optimization algorithm, avoiding poorly behaved likelihoods that require extra time to converge.

Thus, our proposal has some advantages. It is flexible because it can be used for all link functions, including standard link functions (i.e. logit and probit) and others beyond the GEV (i.e. cauchit and skewed probit). Moreover, it can be easily applied, especially when the link function is difficult to analytically manage. Moreover, it overcomes the disadvantage of other sampling techniques (i.e. oversampling and undersampling), which might change the data structure. Finally, it considers the rareness and imbalance across response variable and features.

The performance in the finite samples of the FRW bootstrap in GEV regression modelling, is evaluated using a detailed Monte Carlo simulation study, where the imbalance and rareness are present across the dependent variable and features. Finally, the proposed methodology is applied to a real dataset to analyze student churn in an Italian university.

The paper is organized as follows. In Sect. 2 we introduce the notation and recall the generalized linear and generalized extreme value models. In Sect. 3 for GEV regression, we introduce the weighted bootstrap resampling scheme (Sect. 3.1), along with a short review of the bootstrap confidence intervals used throughout the study (Sect. 3.2) and a bootstrap testing procedure for variable selection that controls for the Familywise Error Rate (Sect. 3.3). Next, we evaluate the performance of the proposed procedure in finite samples for rare events regression, using a simulation study (Sect. 4.1) and apply it to a real dataset to study student churn (Sect. 4.2). Some concluding remarks conclude the paper (Sect. 5).

2 GEV binary regression models

The generalization of linear regression models allows the management of cases where the response variable is not continuous but dichotomous, polytomous, et cetera.

Let Y be the response variable whose distribution belongs to the exponential family, with expectation $E[Y_i] = \mu_i$, for $i = 1, 2, \dots, n$, and n the sample size. Let $g(\cdot)$ be a monotone and differentiable function such that:

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ is the $(k \times 1)$ vector of parameters, with $k = p + 1$ and $\boldsymbol{\beta} \in \mathbb{R}^k$, $\mathbf{x}_i' = (1, x_{i1}, x_{i2}, \dots, x_{ip})'$ is the vector of explanatory variables (covariates or dummy variables for factor levels) of unit i. The function $g(\cdot)$, called link function, relates $\mathbf{x}_i'\boldsymbol{\beta}$ to μ_i and has to be chosen to properly deal with the set of values assumed by μ_i, for $i = 1, 2, \dots, n$ (e.g. with a dichotomous variable $0 < \mu_i < 1$). The Eq. (1) defines the generalized linear model (GLM) and, different from the linear regression model (where $g(\mu_i) = \mu_i$), has a link function that is an increasing or decreasing function of μ_i.

Assume that Y is a binary response variable that takes the value of 1 if the event of interest occurs, and 0 otherwise; $\mathbf{X}$ is the design $(n \times k)$ matrix; the probability associated with $Y_i = 1$ is π_i, and the corresponding probability of $Y_i = 0$ is $1 - \pi_i$. Therefore, the event of interest for the i-th unit can be modelled using a Bernoulli random variable Y_i, with $E[Y_i] = \pi_i$ and $P(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$, for $y_i = 0, 1$ and $i = 1, 2, \dots, n$.

Furthermore,

$$E[Y_i] = \pi_i = P(Y_i = 1) = F(\mathbf{x}_i'\boldsymbol{\beta}), \tag{2}$$

where $F(\cdot)$ is the proper chosen cumulative distribution function. Using the GLM notation (1):

$$\pi_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}) = F(\mathbf{x}_i'\boldsymbol{\beta}).$$

This implies that $F^{-1}(\cdot)$ is the link function and if $F^{-1}(\pi_i)$ is the logit link function, $\text{logit}(\pi_i) = \ln[\pi_i/(1-\pi_i)]$, the distribution function $F(\mathbf{x}_i'\boldsymbol{\beta})$ becomes

$$\pi_i = \frac{\exp\left\{\mathbf{x}_i'\beta\right\}}{1+\exp\left\{\mathbf{x}_i'\beta\right\}} \quad i = 1, 2, \ldots, n,$$

(see Dobson and Barnett 2008, among the others).

When Y_i $(i = 1, \ldots, n)$ is imbalanced, the logit link function has several drawbacks (see McCullagh and Nelder 1989, shortly listed in Sect. 1. The response curve is symmetric and approaches zero at the same rate that it approaches one; the logistic regression model could underestimate the probability π_i; the bias of the maximum likelihood estimators of the parameters $\boldsymbol{\beta}$ further increases in the presence of finite samples.

Calabrese and Giudici (2015), Calabrese and Osmetti (2013) and Wang and Dey (2010) have largely discussed these points and proposed the use of the GEV distribution function to estimate the probability π_i.

Accordingly, let W be a random variable. It belongs to the GEV family, if its distribution function is as follows:

$$F_W(w) = \exp\left\{-\left[1+\xi\left(\frac{w-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right\}, \tag{3}$$

where $\{w : 1+\xi\left(\frac{w-\mu}{\sigma}\right) > 0\}$ and $(\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+, \xi \in \mathbb{R})$ are the location, scale and shape parameters, respectively.

This class of random variables, widely presented in Kotz and Nadarajah (2000) and Coles (2001), includes the following three types of extreme values distributions: if $\xi > 0$, the Fréchet family is obtained; if $\xi < 0$, the Weibull family is achieved, if $\xi \to 0$, the Gumbel family is attained.

Furthermore, the advantages of using the GEV distribution function include the definition and application of skewed link functions and flexibility of the GEV family because the parameter ξ controls the shape and size of the tails of the distribution. This characteristic is particularly important in the presence of rare and imbalanced data, because different proportions of zeroes and ones are required for the selection of a link function that approaches one at a different rate than it approaches zero.

To provide the corresponding empirical evidence, consider the standardized GEV distribution with $\xi < 0$ (Weibull distribution function) and $\xi > 0$ (Fréchet distribution function). In Fig. 1 it can be noted that the distributions in both plots become more asymmetric as $|\xi|$ increases and then tails change. Particularly, the Weibull distribution on the left side of Fig. 1 approaches 1 sharply, and 0 slowly. Conversely, the Fréchet distribution approaches 1 more slowly as ξ increases on the right side of Fig. 1.

Following Calabrese and Osmetti (2013), for the GEV distribution function (3), π_i is given by

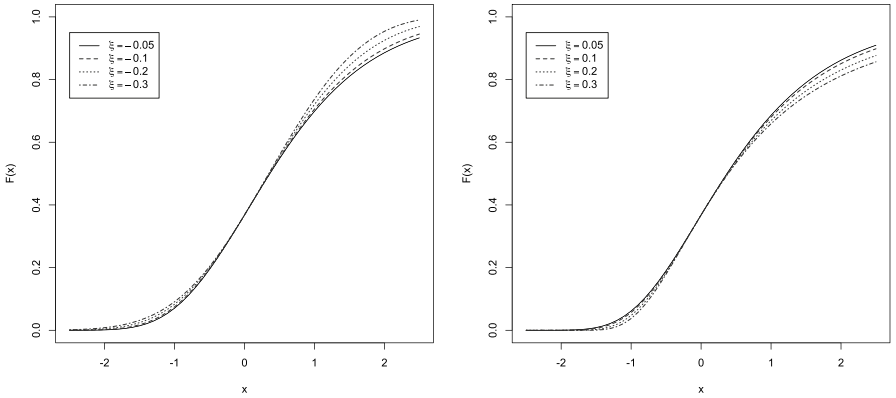


Fig. 1 GEV distribution function with different values of the shape parameter ξ

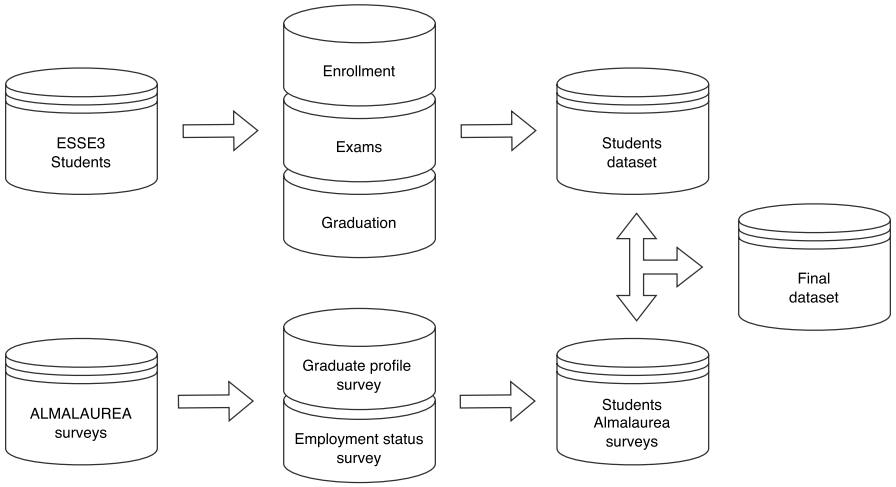


Fig. 2 The scheme of the merging procedure between ESSE3 system and Almalaurea surveys

$$\pi_i = \exp\{-[1 + \xi \mathbf{x}_i' \boldsymbol{\beta}]^{-\frac{1}{\xi}}\} \tag{4}$$

where $(1 + \xi \mathbf{x}_i' \boldsymbol{\beta}) > 0$, with a non-canonical link function

$$\frac{[-\ln(\pi_i)]^{-\xi} - 1}{\xi} = \mathbf{x}_i' \boldsymbol{\beta}, \tag{5}$$

for $i = 1, 2, \ldots, n$, and correspondingly the log-likelihood function becomes:

$$\begin{aligned}\ell(\boldsymbol{\beta}, \xi; \mathbf{X}, \mathbf{y}) &= \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}, \xi; \mathbf{x}_i, y_i) \\ &= \sum_{i=1}^{n} \{-y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)\},\end{aligned} \tag{6}$$

where π_i is given by (4).

Calabrese and Osmetti (2013) discussed some computational issues related to the maximization of (6) and clarified that, because the Fisher information matrix is not diagonal, the parameters $\boldsymbol{\beta}$ and ξ need to be jointly estimated.

In the Appendix, they report that the gradient and Hessian of the log-likelihood function allow the attainment of the asymptotic variance of the parameters but simultaneously provide evidence of the analytical burden faced during the computation of the first and second-order derivatives. Moreover, given the asymptotic normality of the $\boldsymbol{\beta}$ and ξ maximum likelihood estimators, confidence intervals and proper tests can be built to evaluate the accuracy of the estimates and their significance, respectively.

These results further reveal that, as the complexity of the link function increases, the analytical computations related to the second-order derivatives, can be challenging to calculate.

Resampling techniques can significantly help eliminate analytical complexities that make it difficult for practitioners to build confidence intervals and testing procedures. Here, we introduce inferential results based on the bootstrap. To the best of our knowledge, this has not been previously addressed in the GEV regression domain. Thus, we reach two main objectives. First, no analytical computation of the Hessian matrix is needed, to overcome the aforementioned analytical issues. Second, the bootstrap procedure can be made almost automatic and used along with other link functions beyond the GEV function used in this study.

In Sect. 3, we consider the FRW bootstrap, largely discussed in Xu et al. (2020), to build confidence intervals for the parameters included in the vector $\boldsymbol{\beta}$ of the GEV regression model. Next, the bootstrap distribution is used to make variable selection. This takes place in the multiple testing setting, which will be discussed in Sect. 3.3.

An important advantage of the FRW bootstrap is that it can be properly used when the number of successes in the binary dependent variable, is related to rare events or where there is insufficient mixing of successes and failures across the features. Moreover, as will be discussed in Sect. 3, under mild conditions, the FRW version of the maximum likelihood (ML) estimator is consistent and asymptotically follows the normal distribution.

3 Fractional-random-weighted bootstraping for GEV regression

The weighted bootstrap has had a long-established role in the bootstrap literature since the seminal paper of Efron (1982), where the standard bootstrap was shown to be equivalent to a weighting resampling scheme with random integer weights,

and the weights were given by the number of times each observation is drawn in the resampling.

In the presence of binary imbalanced data, resampling bootstrap approaches do not work well because of the very low number of ones that could entail the selection of bootstrap samples with only zeroes. This leads to a preference for different bootstrap schemes where the resampling is replaced by a proper random weighting of the observations.

In this domain, we describe the FRW bootstrap (presented in Xu et al. 2020) which will be applied to gain inference regarding the parameters of the GEV regression, which is mainly used in the presence of imbalanced and rare events datasets.

3.1 Weighted bootstrap for GEV regression

Let $\ell(\boldsymbol{\beta}, \xi; \mathbf{X}, \mathbf{y})$ be the log-likelihood function (6), with $\ell_i(\boldsymbol{\beta}, \xi; \mathbf{x}_i, y_i)$, for $i = 1, \ldots, n$, the contribution for the observation $(\mathbf{x}_i', y_i)$.

The random weighted log-likelihood is given as follows

$$\ell^*(\boldsymbol{\beta}, \xi; \mathbf{X}, \mathbf{y}, \mathbf{w}^*) = \sum_{i=1}^{n} w_i^* \ell_i(\boldsymbol{\beta}, \xi; \mathbf{x}_i, y_i), \tag{7}$$

where the weight vector $\mathbf{w}^* = (w_1^*, w_2^*, \ldots, w_n^*)'$ is generated using a uniform Dirichlet distribution, multiplied by n. Therefore, $\sum_{i=1}^{n} w_i^* = n$.

The probability density function of the Dirichlet distribution of order $n \geq 2$ with parameters $\alpha_1, \ldots, \alpha_n$ is given by:

$$f(w_1, \ldots, w_n; \alpha_1, \ldots, \alpha_n) = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} w_i^{\alpha_i - 1},$$

with $\Gamma(\cdot)$ denoting the Gamma function, $\alpha_i > 0$, $\sum_{i=1}^{n} w_i = 1$ and $w_i \geq 0$. When $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 1$, we get the uniform Dirichlet distribution.

Operationally, generating fractional weights using uniform Dirichlet distribution is equivalent to generating random weights using normalized exponential distribution with mean one. More precisely, the random vector $\mathbf{w}^* = (w_1^*, w_2^*, \ldots, w_n^*)'$ is generated as follows:

$$w_i^* = n \cdot \frac{Z_i}{\sum_{i=1}^{n} Z_i}, \quad i = 1, \ldots, n, \tag{8}$$

where $Z_1, \ldots, Z_n$ are iid exponential distributions with mean one.

Generating the weights according to the previous scheme delivers FRW bootstrap estimators with good asymptotic properties, as long as the weights are positive, independent, and identically distributed from continuous random variables with equal mean and variance, as for the uniform Dirichlet case (see Jin et al. 2001 and empirical results in Xu et al. 2020).

The fractional random weight counterpart of the likelihood estimate $\hat{\boldsymbol{\beta}}$ is obtained by maximizing (7):

$$\hat{\boldsymbol{\beta}}^{*} = \arg\max_{\boldsymbol{\beta}} \ell^{*}(\boldsymbol{\beta}, \xi; \mathbf{X}, \mathbf{y}, \mathbf{w}^{*}).$$

The probability law of $\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{*} - \hat{\boldsymbol{\beta}}\right)|\mathbf{X}$ delivers the bootstrap approximation for the unknown sampling distribution $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$.

There is a clear advantage in using fractional random weights in our framework. Integer weighting schemes, such as those based on the Multinomial distribution, have a random number of weights equal to zero. Consequently, some observations from the log-likelihood function (7) are excluded. This might cause serious estimation problems in contexts where the dependent variable or some predictors have rare levels, making the bootstrap procedure fail altogether or deliver poor results. In such cases, integer weights or sampling schemes (as in the under-sampling case) may lead to the selection of samples with binary variables, originally affected by rarity, having all values equal to zero because, for example, the small number of ones has not been sampled from the procedure. In a fractional weighted bootstrap scheme, the weights are never zero and all observations remain in the bootstrap samples. Therefore, the estimation difficulties associated with the resampling process do not arise.

The fractional-weighted bootstrap scheme for GEV regression delivers consistent results. Therefore, for $n \to \infty$,

$$\hat{\boldsymbol{\beta}}^{*} \xrightarrow{p} \boldsymbol{\beta} \tag{9}$$

and

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{*} - \hat{\boldsymbol{\beta}}\right)|\mathbf{X} \xrightarrow{d} N\left(\mathbf{0}, I(\boldsymbol{\beta})^{-1}\right). \tag{10}$$

where $I(\boldsymbol{\beta})$ is the Fisher information matrix for $\boldsymbol{\beta}$, as computed in the Appendix by Calabrese and Osmetti (2013).

The proof is straightforward. It starts with the results in Smith (1985) and Calabrese and Osmetti (2013), showing the regularity of the GEV maximum likelihood estimators when $\xi > -0.5$. Particularly, under this condition, the maximum likelihood estimators have the usual asymptotic properties. Thus, for $n \to \infty$

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$$

and

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, I(\boldsymbol{\beta})^{-1}\right).$$

Accordingly to Result 1 of Xu et al. (2020), the consistency of the fractional random weight $\hat{\boldsymbol{\beta}}^{*}$ estimators follows (see Eq. 9). The asymptotic normality (see Eq. 10) is instead obtained from Result 2 and the corresponding proof details given in the Supplement of Xu et al. (2020).

Based on the previous results, we can state that the probability laws of $\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{*} - \hat{\boldsymbol{\beta}}\right)|\mathbf{X}$ and $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ are asymptotically equivalent. Thus, for large n, the

unknown sampling distribution of $\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ can be approximated with the distribution of $(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})|\mathbf{X}$.

The bootstrap distribution is difficult to analytically derive, and, as usual, it will be approximated using Monte Carlo, according to Algorithm 1. The approximated bootstrap distribution can be used to construct confidence intervals and tests for the parameters $\boldsymbol{\beta}$. The adaptation of our framework is straightforward and shown in Sects. 3.2 and 3.3.

Algorithm 1 MC approximation of the bootstrap distribution $(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})|\mathbf{X}$

Require: $B \geq 1000$, number of bootstrap runs
1: $b = 1$
2: **while** $b \leq B$ **do**
3: Draw the bootstrap weights $\mathbf{w}^b$ from a uniform Dirichlet distribution
4: Compute

$$\hat{\boldsymbol{\beta}}_b^* = \arg\max_{\boldsymbol{\beta}} \ell^*(\boldsymbol{\beta}, \xi; \mathbf{X}, \mathbf{y}, \mathbf{w}^b)$$

5: Compute $\boldsymbol{\delta}_b^* = \hat{\boldsymbol{\beta}}_b^* - \hat{\boldsymbol{\beta}}$
6: $b = b + 1$
7: **end while**
8: Approximate the bootstrap distribution with the Empirical Cumulative Distribution Function (ECDF), namely $\hat{G}(\mathbf{z})$ of

$$\left\{\boldsymbol{\delta}_b^* = \hat{\boldsymbol{\beta}}_b^* - \hat{\boldsymbol{\beta}}, \quad b = 1, 2, \ldots, B\right\}.$$

3.2 Bootstrap confidence intervals

Given the distribution derived in Algorithm 1, several alternative confidence intervals can be constructed (see Shao and Tu 1995 for a general introduction).

Let β_j be the parameter of interest and $\hat{\beta}_j$ its maximum likelihood estimator (MLE), with $j = 0, 1, 2, \ldots, p$. The distribution derived from Algorithm 1 can be used to construct approximate confidence intervals using the *hybrid (or basic) bootstrap method*. The approximate $100(1-\alpha)\%$ bootstrap interval for β_j is given by:

$$\left[\hat{\beta}_{j,\text{lo}}^*, \hat{\beta}_{j,\text{up}}^*\right] = \left[\hat{\beta}_j - q_{1-\alpha/2}, \hat{\beta}_j - q_{\alpha/2}\right],$$

where q_α denotes the percentile of order α of the ECDF of $\{\delta_{j,1}^*, \delta_{j,2}^*, \ldots, \delta_{j,B}^*\}$ with $\delta_{j,b}^* = \hat{\beta}_{j,b}^* - \hat{\beta}_j$.

This method delivers a valid asymptotic approximation, but the interval limits are neither range-preserving nor transformation invariant. This is an essential and crucial property of our framework, where some quantities of interest in the applications need to be written as smooth functions of the model parameters vector.

An alternative approach is the bootstrap confidence interval obtained using the *percentile method*. The approximate $100(1-\alpha)\%$ bootstrap percentile interval for β_j is given as follows:

$$\left[\hat{\beta}^*_{j,\text{lo}}, \hat{\beta}^*_{j,\text{up}}\right] = \left[\hat{\beta}^*_{j,(\alpha/2)}, \hat{\beta}^*_{j,(1-\alpha/2)}\right],$$

where $\hat{\beta}^*_{j,(\alpha)}$ denotes the percentile of order α of the empirical distribution of the bootstrap replicates $\hat{\beta}^*_{j,1}, \hat{\beta}^*_{j,2}, \ldots, \hat{\beta}^*_{j,B}$. This is the simplest way to derive a bootstrap confidence interval because it is easy to compute, range-preserving, and transformation invariant. However, it tends to be too narrow for small n (Hesterberg 2015).

A better alternative is given by the *bias corrected* (BC) *percentile method*. The basic idea of the BC method is to replace the percentiles $\alpha/2$ and $1-\alpha/2$ used in the simple percentile method with the adjusted percentiles α_1 and α_2. Particularly, the confidence interval using the BC percentile method is given as follows:

$$\left[\hat{\beta}^*_{j,\text{lo}}, \hat{\beta}^*_{j,\text{up}}\right] = \left[\hat{\beta}^*_{j,(\alpha_1)}, \hat{\beta}^*_{j,(1-\alpha_2)}\right],$$

where

$$\alpha_1 = \Phi\left[2z_{\hat{b}} - z_{1-\alpha/2}\right]$$
$$\alpha_2 = \Phi\left[2z_{\hat{b}} + z_{1-\alpha/2}\right],$$

with $\Phi(\cdot)$ denoting the CDF of the Standard Gaussian distribution, z_α denoting the percentile of order α of the Standard Gaussian distribution, and $\hat{b}$ denoting the fraction of the values $\left\{\hat{\beta}^*_{j,b}, b = 1, 2, \ldots, B\right\}$ that are less than $\hat{\beta}_j$. The value $z_{\hat{b}}$ is the bias-correction, that is, the value that corrects for the median bias in the distribution of $\hat{\beta}^*_j$, on the Standard Gaussian scale.

The BC percentile method is less intuitive than the other two methods and requires the estimation of a bias-correction term. However, it delivers confidence intervals that are range-preserving and transformation invariant, works well for a variety of parameters, and is second-order accurate (see DiCiccio and Efron 1996).

3.3 Bootstrap variable selection

The variable selection problem can be seen as a multiple testing problem and is implemented using the bootstrap distribution derived in Algorithm 1. Let

$$H_j : \beta_j = 0 \quad vs \quad H'_j : \beta_j \neq 0, \qquad j = 1, 2, \ldots, p$$

and consider the following test statistics

$$T_j = |\hat{\beta}_j|, \qquad j = 1, 2, \ldots, p.$$

Clearly "large" values of the T_j are indicative of the alternative.

Here, the problem is how to perform the test given the multitude of tests. Accordingly, we refer to a bootstrap procedure suggested by Romano and Wolf (2005a, 2005b) to control *Familywise Error Rate* (FWE), which indicates the probability of having at least one false rejection. In our case, it is the probability of having at least one wrongly labeled variable as relevant to the model.

The procedure runs as follows. Relabel all hypotheses in descending order of the observed test statistics, $T_{r_1} \geq T_{r_2} \geq \cdots \geq T_{r_p}$. Accordingly to the labels $\{r_1, r_2, \ldots, r_p\}$, H_{r_1} denotes the "most significant" variable and H_{r_p}, the "least significant" variable. Now, consider the absolute values of the bootstrap replicates obtained in Algorithm (1)

$$\left\{ |\delta^*_{j,b}| = |\hat{\beta}^*_{j,b} - \hat{\beta}_j|, \quad j = 1, 2, \ldots, p; b = 1, 2, \ldots, B \right\},$$

and compute

$$\max^{*,b}_{T,s} = \max \left\{ |\delta^*_{r_s,b}|, \ldots, |\delta^*_{r_p,b}| \right\},$$

for $s = 1, 2, \ldots, p$ and $b = 1, 2, \ldots, B$.

Finally, let $\hat{c}(1-\alpha, s)$ be the $1-\alpha$ percentile of the set $\left\{ \max^{*,b}_{T,s}, b = 1, 2, \ldots, B \right\}$.

Now, we can apply Algorithm 3.1 described in Romano and Wolf (2016), which we report in Algorithm 2 adapted to our variable selection testing problem.

Algorithm 2 Multiple testing algorithm for controlling FWE at level α

Require: Fix the level of the test α
for $j = 1$ to p **do**
  Reject H_{r_j} if, and only if, $T_{r_j} > \hat{c}(1-\alpha, 1)$
end for
Let R_1 be the number of hypotheses rejected.
if $R_1 = 0$ **then**
  Stop
else
  $s = 2$
end if
for $j = R_{s-1} + 1$ to p **do**
  Reject H_{r_j} if, and only if, $T_{r_j} > \hat{c}(1-\alpha, R_{s-1}+1)$
end for
if No further hypotheses are rejected **then**
  Stop
else
  Denote by R_s the number of hypotheses rejected so far
  $s = s + 1$
  Return to step 10
end if

When the number of the hypothesis is in the hundreds or thousands, controlling the FWE might lead to test procedures that are too conservative. Therefore, variables may be wrongly labeled as irrelevant to the model, when they actually are relevant. Alternatively, the k-FWE, defined as the probability of rejecting at least k of the true null hypotheses, can be used to construct more powerful tests. Algorithm 2 can be easily extended for controlling the k-FWE along the lines described in Algorithm 4.2 in Romano et al. (2008). Finally, when $k = 1$, controlling the k-FWE reduces to controlling the FWE.

4 Numerical study

4.1 Monte Carlo simulation

This section discusses the results of a Monte Carlo simulation study used to assess the performance in finite samples of the proposed fractional-weighted bootstrap scheme for GEV regression models. These performances are compared with those obtained from the maximum likelihood method. Particularly, the primary purpose is to investigate the effect on the accuracy of alternative inference procedures when the data are affected by unbalanced and rare events in the dependent and independent variables. The latter scenario has received less attention in the literature. However, it appears to be very frequent in real applications, especially when considering one-hot-encoding transformations used to deal with categorical predictors. Such transformations often generate very unbalanced binary variables when some of the levels are associated with rare events.

For this purpose, we have considered a GEV regression model with p features, where $p = \{2, 4, 10\}$, including numeric and binary variables. The first $p/2$ predictors are numeric variables and the last $p/2$ are binary variables. Let p_{num} be the number of numeric variables and p_{bin} the number of binary variables, such that $p = p_{\text{num}} + p_{\text{bin}}$. The p_{bin} binary variables are generated using independent Bernoulli distributions, where the probability of success assumes the values p_X given in the set $\{0.05, 0.10, 0.20, 0.50\}$, which includes balanced and imbalanced classes. The p_{num} numeric variables are generated from a p_{num}-variate normal random variable where all variables have mean zero, unit variance, and $\text{corr}(X_i, X_j) = \rho$, for $i \neq j$ and $i = j = 1, \ldots, p_{\text{num}}$, where $\rho \in \{0.0, 0.5\}$. Given the p features, the dependent variable Y is generated using the binary GEV regression model with $P[Y|\mathbf{x}]$ given in (4). Three different values are considered for the shape parameter $\xi = \{0.10, -0.10, -0.20\}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ with $\beta_j = 0$, for $j = 1, \ldots, p$, whereas β_0 is set at different levels to guarantee that $P(Y|\mathbf{x}) \in \{0.05, 0.10, 0.20, 0.50\}$.

Furthermore, we have assumed $B = 1999$ bootstrap runs, 1000 Monte Carlo replicates, and sample size n varying in the set $\{250, 500, 1000, 2000\}$. Finally, the fractional weights of the FRW bootstrap are generated using a uniform Dirichlet distribution (with all parameters equal to one). The overall simulation design comprises a total of 1152 design points.

To better clarify the simulation design, in Algorithm 3, all settings and the structure of the Monte Carlo study, are shortly described. All procedures are implemented in R (R Core Team 2022) and the code is included as Supplementary file.

Algorithm 3 Monte Carlo simulation

Simulation design: settings
$n = \{250, 500, 1000, 2000\}$; # sample size
$p = \{2, 4, 10\}$; # number of covariates ($p_{bin} + p_{num}$)
$p_X = \{0.05, 0.10, 0.20, 0.50\}$; # proportion of ones in binary covariates
$p_Y = \{0.05, 0.10, 0.20, 0.50\}$; # $P(Y|\mathbf{X})$
$\rho = \{0.0, 0.5\}$; # correlation among numeric variables
$\boldsymbol{\beta}^0 = (\beta_0, 0, \ldots, 0)$;
$\xi = \{-0.20, -0.10, 0.10\}$; # shape parameter
$nmc = 1000$ # Monte Carlo replicates
$B = 1999$ # Bootstrap replicates

procedure GENERATE.DATA(n, p, p_X, p_Y, ρ, $\boldsymbol{\beta}^0$, ξ)
    $\mathbf{X}_{bin}$: generate n random numbers from $p/2$ independent Bernoulli r.v.'s, $Be(p_X)$;
    $\mathbf{X}_{num}$: generate n random numbers from a $p/2$-variate Normal r.v., $N(\mathbf{0}; \Sigma)$, with $\Sigma = [\sigma_{ij}]$, where $\sigma_{ij} = 1$ if $i = j$ and $\sigma_{ij} = \rho$ otherwise;
    $\mathbf{y}$: draw n binary variables from (4) with $\mathbf{X} = [\mathbf{X}_{bin} : \mathbf{X}_{num}]$, $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ and ξ;
    return (data=[$\mathbf{y} : \mathbf{X}$])
end procedure

procedure MONTE CARLO(nmc, B, n, p, p_X, p_Y, ρ, $\boldsymbol{\beta}^0$, ξ)
    for i=1 to nmc **do**
        generate.data(n, p, p_X, p_Y, ρ, $\boldsymbol{\beta}^0$, ξ)
        $\hat{\boldsymbol{\beta}}$: estimate $\boldsymbol{\beta}$ maximizing (6);
        for b=1 to B **do**
            draw n random weights $w_1^*, \ldots, w_n^*$, as in (8)
            $\hat{\boldsymbol{\beta}}_b^*$: estimate $\boldsymbol{\beta}$ maximizing (7)
        end for
    end for
    return(matrix($\hat{\boldsymbol{\beta}}$); array($\hat{\boldsymbol{\beta}}^*$))
        ▷ [matrix($\hat{\boldsymbol{\beta}}$) is a $(nmc \times k)$ matrix; array($\hat{\boldsymbol{\beta}}^*$) includes B $(nmc \times k)$ matrices.]
end procedure

Given the large simulation study, we only discuss the cases with $\xi = -0.10$ and $\rho = 0.5$, because the overall results are significantly similar when $\xi = \{-0.20, 0.10\}$ and $\rho = 0$. Moreover, the moderately correlated features scenario appears to be more realistic than the scenario where all numeric features are uncorrelated.

Given the copious number of plots, we include only the plots where the number of predictors is $p = 4$. This is because the results obtained by fixing $p = 2$ and $p = 10$, are very similar. In any case, the complete set of results is available from the authors as supplementary material.

Given the aim of the simulation study, the accuracy of the bootstrap estimators has been evaluated by comparing the variance of $\hat{\beta}_j^*$, for $j = 0, 1, \ldots, 4$, with the "true" variance (based on further 20000 Monte Carlo replicates where the corresponding ML estimates are obtained for each of them). To compare the bootstrap results with that of a competing method, the ratios between the variance of the maximum likelihood estimator $\hat{\beta}_j$, obtained from the Monte Carlo study described in Algorithm 3, and the "true" variance are further computed. In Figs. 3 and 4, the empirical distributions of the bootstrap and maximum likelihood (ML)

variance ratios are shown for all β_j, with $j \in \{0, 1, 2, 3, 4\}$, for different levels of imbalance in the dependent variable Y and across the binary covariates, X_3 and X_4. It can be noted that the variability of the methods increases as the rarity of Y, X_3, and X_4 grows, and no remarkable differences are seen among them. This latter point could be seen as an advantage for the bootstrap method. This is because when the complexity of the link function does not allow to easily obtain the second derivatives for the Hessian, the bootstrap approach can be considered as a valid alternative to maximize the likelihood and effectively gain inference about the unknown parameters of the model. Moreover, the bootstrap resampling scheme does not change whatever the complexity of the chosen link function is. Therefore, it can be applied beyond the GEV regression considered in this study.

the FRW bootstrap distributions of the GEV regression parameters are further used to build bootstrap confidence intervals (CIs). As clarified in Sect. 3.2, we consider the following three methods: the percentile, bias-corrected, and hybrid methods.

Given the nominal confidence level $1 - \alpha = 0.90$, in Figs. 5 and 6 the lengths of the bootstrap confidence intervals are compared with those obtained from the maximum likelihood approach. The similar behavior of the lengths, for all values of p_X and p, further provide evidence of what was previously noted.

Finally, the empirical coverage of the three different methods in the FRW bootstrap domain with imbalanced data is shown and compared with the empirical coverage of the confidence intervals based on the likelihood approach. To simplify the presentation of results and obtain clearer plots, in Figs. 7, 8, 9 and 10, the empirical coverage of the confidence intervals is evaluated by separately considering the lower and upper confidence bounds. In Figs. 7 and 8, we consider the empirical percentage error obtained by comparing the true β_j value, for $j \in \{0, 1, 2, 3, 4\}$ with the lower confidence bound. In Figs. 9 and 10, the corresponding empirical percentage error for the upper bound is considered.

If we evaluate the results of the lower bounds, the four methods are almost equivalent for all β_j, $j \in \{0, 1, 2, 3, 4\}$, for all rates of imbalance in the data, and for small values of n. In all cases, the empirical error rate is close to the nominal level $\alpha/2 = 0.05$.

A different evaluation arises from Figs. 9 and 10, panels (c) and (d). In these cases, when n is not sufficiently large, the small number of ones in Y, X_3, and X_4 comprise the upper empirical percentage errors of the bootstrap CIs for β_3 and β_4, far from the nominal level $\alpha/2 = 0.05$ (the dark gray line). In any case, the BC and Hybrid methods outperform the Percentile method as expected. However, all bootstrap methods are almost equivalent to the likelihood method, as n grows.

Finally, as presented in Sect. 3.3, the FWR bootstrap distributions are used in a multiple testing setting, for variable selection. To the best of our knowledge, the FWR bootstrap has not been previously used in this domain.

Following the steps of Algorithm 2, the test is built by controlling the probability of having at least one false rejection (FWE), which, in practice, corresponds to the case where at least one variable is wrongly labeled as relevant.

Given the nominal FWE level 0.10, Fig. 11 shows the corresponding empirical values obtained using the FWR bootstrap distribution of the GEV regression parameters.

Note that what mainly affects the testing performance is the imbalance of Y and the predictors, which could lead to the inclusion of irrelevant variables in the model. In all other cases, the performance of multiple testing procedure is satisfactory.

It confirms the following results of the construction of the CIs: the FRW bootstrap delivers reasonable good results in all cases, except when the dependent and independent variables are highly imbalanced across small sample sizes.

4.2 Empirical data analysis

Our proposal for imbalanced binary data is evaluated by analyzing a complex dataset related to university students' careers. Particularly, we are interested in investigating students' churn, defined as students' choice to enroll for a master course in the same university they graduated from with a bachelor's degree. Thus, the event of interest can be modeled by a binary response variable Y, which takes the value of 1 if students stay in the same university, and 0 otherwise.

The analysis aims to i) identify which students' characteristics influence their choice and ii) sketch a profile of the students who are willing to enroll in a master program from the same university they received their bachelor degrees from.

For these purposes and given the complexity of the factors contributing to the university churn under analysis, we collected information from two main sources. The first is the Student Information System (ESSE3), a student management system used by most Italian universities, which manages the entire career of students from enrollment to graduation. It contains information about students' high school diplomas, personal characteristics, exams, abroad experience, internship, and degrees. Therefore, given the large amount of available data, from this source, we collected and merged information on students' enrollment, exams, and graduation for all years under analysis (Fig. 2). The second source is the AlmaLaurea Consortium (https://www.almalaurea.it/en). It is an inter-university consortium that collects information and assessments of partner universities and their activities every year, for statistical and research purposes. It also facilitates the entry of young graduates into the labour market using its innovative online platform. Accordingly, Almalaurea also carries out the following two statistical surveys:

1. graduates' profiles, which provide a portrait of the characteristics of graduates, their university achievements, experiences they have gained during their studies, and the evaluation of the studies completed;
2. graduates' employment status, which provides, at certain time points (one/three/five year(s) after graduation), a portrait of the graduates' job placement in the labor market, characteristics of the jobs found, and skills acquired during university studies.

The micro-data collected by both surveys are delivered to all affiliated universities. The datasets of the two AlmaLaurea surveys are characterized by a large number of variables (i.e. 145 and 159 variables in the surveys on the graduates' profiles and employment status, respectively) (Fig. 2). The data from ESSE3 and Almalaurea are not freely available and can be acquired from the university staff for research purposes only.

Given the complexity and large size of the datasets, we only focused on the University of Salerno, established in 1968 in Southern Italy. It has 17 Departments and about 90 bachelor and master programs. Additionally, among all departments and programs, we opted for analyzing the bachelor courses in Business Administration (BA), Economics (E), and Administration and Organization (A &O) at the Department of Economics and Statistics for eight academic years (2013–2020).

Thus, the analysis covers 1543 students (BA = 697; E = 654; A &O = 192) that have started a master program at the University of Salerno $Y = 1$ (1036 students) or elsewhere $Y = 0$ (507 students). Next, we merged the datasets from ESSE3 and Almalaurea using students' identification numbers, to finalize the data matrix, which contains information about students' high school diplomas, university career, evaluation and satisfaction of their experience in university, first job experience, and family background (Fig. 2).

Given the large number of features resulting from the merging procedure, we reduce the number of risk factors by taking into account the aim of the study. A first screening of the covariates was performed by testing their statistical significance on the response variable Y. The final set of the variables identified as potential characteristics, affecting the students' churn, is shown in Table 1. They are classified into the following four groups: high school, bachelor degree, socio-demographic information, and job position.

The analyzed dataset consists of some binary variables with different levels of imbalance, which have to be managed along with the imbalance in the response variable. This is because they might affect the estimates, as discussed in Sect. 1 and shown in the simulation study (Sect. 4.1).

Thus, for sketching the profiles of students who enrolled for a master program in the same university they received their bachelor degrees from, we estimate the GEV regression model and make inferences on the estimated parameters using the FRW bootstrap distribution. Particularly, because we aim to estimate students' churn and identify the main students' characteristics that might affect their choice of continuing their career in the same university they received their bachelor degree from, we compared the variables selected by multiple testing based on controlling FWE, with those obtained by estimating a generalized linear model with *c-log-log* link function and elastic-net regression.

Given that the FRW bootstrap procedure is based on the numerical maximization of the log-likelihood, it requires the specification of starting values for the model parameters. The starting values for the $\boldsymbol{\beta}$s are fixed at the values estimated using *c-log-log* link function because, for ξ close to zero, the GEV distribution becomes a Gumbel distribution, which corresponds to *c-log-log* link. Furthermore, for the shape parameter ξ, two approaches are suggested in Calabrese and Osmetti (2013), Calabrese and Giudici (2015); Calabrese et al. (2016). The first consists of fixing a

Table 1 Table of covariates chosen as potential factors for students' churn

Group	Variable	Short description	Type
High school	Type of diploma	High school type (classical studies, technical, scientific, ...)	Nominal
	Final exam mark	Total marks (from 60 to 100)	Integer
	Geographic area where the diploma has been obtained	Province and geographic area (the same province where the University is located, North, South or Center of Italy, abroad ...)	Nominal
Bachelor degree	Course of study	BA, E, SAO	Nominal
	Enrollment age	University enrollment after the diploma: the next year of after one year, two or more years later	Nominal
	Enrollment motivation	Reasons behind the course of study choice	Nominal
	Graduation years	Number of years from the enrollment to the graduation	Integer
	Residence	Place of residence (in the same province as the University, different province but in the same region, different region, abroad)	Nominal
	Final mark	First degree final mark (from 66 to 110 cum laude; 110 + laude = 113)	Integer
	Degree age	First degree age	Integer
	International experience	Participation to international exchange projects	Binary
	Satisfaction	Satisfaction of the University experience	Binary
	Organization	Satisfaction of the teaching and administrative organization (exams and lectures timetable, information...)	Binary
	Teachers relationship	Satisfaction of the relationship with teachers	Binary
	Exams	Correspondence between the exam marks and the knowledge	Binary
	Library	Evaluation of the library services	Binary
	Equipments	Evaluation of the equipments used during the course of study (i.e. laboratories)	Binary
	Master wish	University where the student wishes to enroll for the master degree	Nominal
	Back to the start	The student is asked if, going back to its enrollment, he would have chosen the same University	Nominal

Table 1 (continued)

Group	Variable	Short description	Type
Socio-demo	Gender	Gender of the graduated student	Nominal
	Parents title	Title of study of the student's parents	Ordinal
	Father title	Title of study of the student's father	Ordinal
	Mother title	Title of study of the student's mother	Ordinal
	Social status	Social status as coded by the AlmaLaurea survey	Nominal
Job	Job position	Current job position	Nominal
	Job province	The student is asked if they are willing to work in the same province they reside in	Binary
	Job Region	The student is asked if they are willing to work in the same region they reside in	Binary
	Job North	The student is asked if they are willing to work in the North of Italy	Binary
	Job Center	The student is asked if they are willing to work in the Center of Italy	Binary
	Job South	The student is asked if they are willing to work in the South of Italy	Binary
	Job Europe	The student is asked if they are willing to work in Europe	Binary
	Job no-Europe	The student is asked if they are willing to work in a non-European country	Binary

Table 2 The relevant covariates by maximum likelihood, elastic net, and FWE

Variable	ML	Elastic net	FWE
Course of study: economics	✓	✓	
Course of study: administration and organization	✓	✓	
Social status: high	✓	✓	
Social status: white-collars middle class			
Social status: autonomous middle class			
High school diploma: scientific	✓		
High school diploma: psychopedagogical			
High school diploma : technical-professional	✓	✓	
High school diploma: other			
Degree age			
Enrollment age: the same year of diploma	✓	✓	
Residence: same region of the study	✓	✓	
Residence: different region of the study/abroad			
final degree mark			
Gender: female			
International experience: yes	✓	✓	
Satisfaction: no satisfied	✓	✓	
Back to the start: yes, but same course and University		✓	
Back to the start: yes, but different course, same University			
Back to the start: yes, but same course, different University	✓	✓	✓
Back to the start: yes, but different course & University			
Teachers relationship: no satisfied		✓	
Job province: yes		✓	
Job Europe: yes			
Job no-Europe: yes		✓	
Job region: yes		✓	
Exams: $\leq 50\%$	✓		
Exams: $> 50\%$	✓		
Exams: always or almost always	✓		

grid of values for ξ, and choosing the value that maximizes the likelihood or gives the best empirical predictive performance. The second approach suggests to jointly estimate the shape parameter ξ and coefficients $\boldsymbol{\beta}$ by maximizing the likelihood. In this analysis, we adopt both proposals. Given that the estimates for ξ are substantially equal for both approaches ($\xi = -0.25$ for the first approach and $\xi = -0.26$ for the second approach), the initial value for the shape parameter is set at $\xi = -0.25$.

Table 2 shows the variables that are significant by observing the p-values of the maximum likelihood estimates in *c-log-log* regression and using the elastic net procedure. When controlling FWE, only one variable is selected as relevant. Thus, all other previously identified features might be wrongly labeled as relevant.

For the significant variable, *Back to the start: Yes, but same course, different University*, we plot the bootstrap distribution of the corresponding estimates based on the GEV maximum likelihood, with its BC bootstrap confidence interval (Fig. 12). For sake of comparison, we also report the *c-log-log* maximum likelihood estimate and corresponding confidence interval (Fig. 12). It is evident that the bootstrap distribution is slightly negative skewed. Moreover, the likelihood-based confidence interval is wider than the BC. Finally, the negative value of the estimate denotes a decrease in the probability of starting a master program at the University of Salerno for those students that, going back to the first-level enrolment, would choose the same course but different university. This might appear quite obvious but provide clear implications in terms of university policies that need to focus their attention on the overall satisfaction of students. In fact, it emerges that students are not unsatisfied with the selection of the type of course of study, but other factors might compromise their positive experience at the University of Salerno (Table 2).

5 Concluding remarks

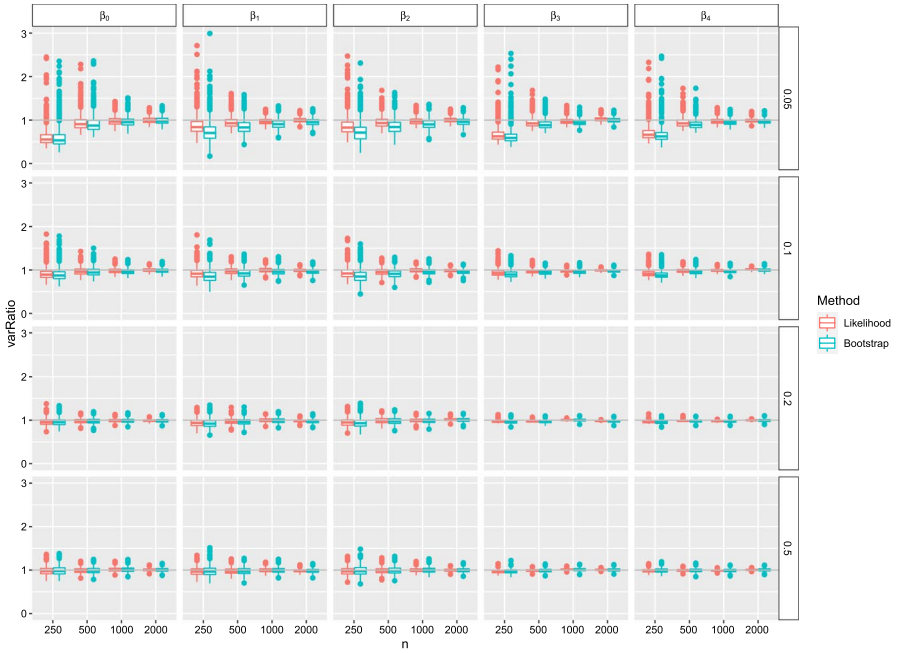
We addressed the problem of imbalance and rareness in binary dependent and independent variables, which may produce inaccurate inferences. To model these data, we employed GEV regression models from Calabrese and Osmetti (2013) and Wang and Dey (2010), which use an asymmetric link function based on the quantile function of the Generalized Extreme Value (GEV) distribution.

Furthermore, instead of using the inferential results presented in Wang and Dey (2010) and Calabrese and Osmetti (2013), we propose to implement the Fractional-Random-Weighted (FRW) bootstrap, proposed by Xu et al. (2020), to construct both bootstrap confidence intervals and a multiple testing procedure for selecting the set of relevant variables.

The following advantages can be obtained from the use of the FRW bootstrap: i) it is flexible because the same algorithm can be used for all link functions; ii) it can be easily applied when the link function is challenging to analytically manage; iii) it overcomes the disadvantage of other sampling techniques (i.e. oversampling and undersampling), which might change the data structure; and iv) it considers the rareness and imbalance in both response variable and features, while other sampling techniques usually focus only on the dependent variable.

The simulation study shows that the imbalance in the binary independent variables seems to have a higher impact on the variability of the estimates, compared to the binary imbalanced response variable. In fact, when the probability of having one in the features is less than 0.10, the estimates have larger variability, especially for small sample sizes. Moreover, the effect of rare events in the response variable is mitigated by the asymmetric distribution of GEV. These results are achieved in the presence of ML and FRW bootstrap estimators. However, the advantages of the latter approach enable the use of FRW for practitioners.

The FRW bootstrap distribution is then used to construct a variable selection procedure which considers the multiple testing structure of the problem. The simulation

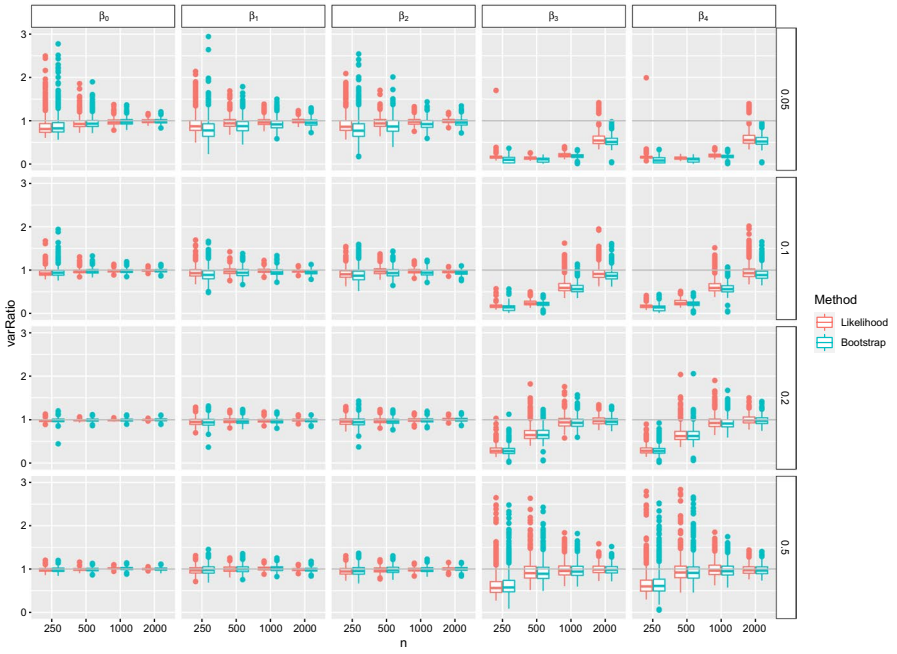


(a) $p_X = 0.50$

(b) $p_X = 0.20$

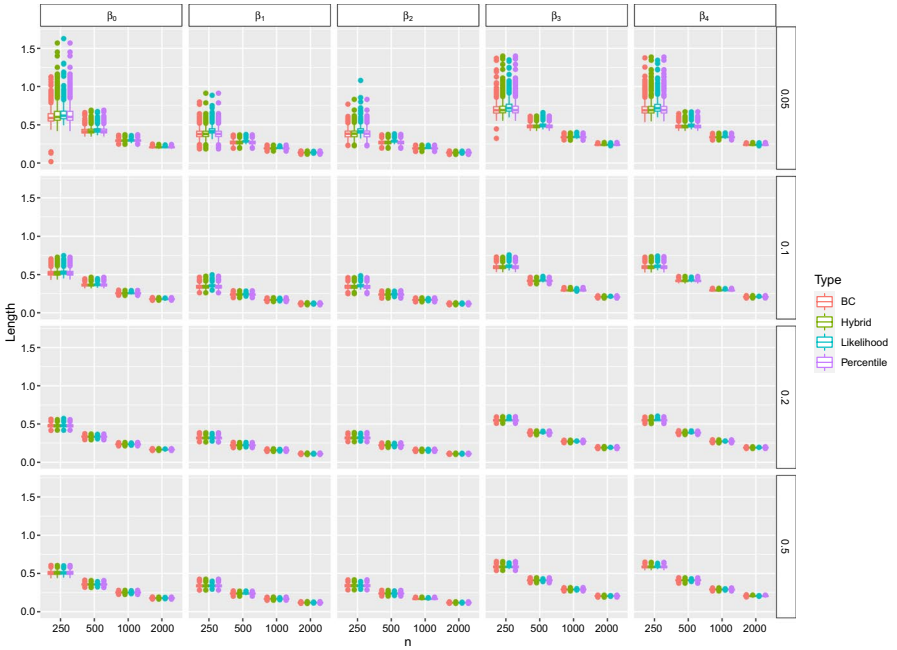
Fig. 3 Variance ratios of the bootstrap estimators and of the maximum likelihood estimators with the "true" variance of β_j, $j = 0, 1, 2, 3, 4$, for $p_X = \{0.20, 0.50\}$ and $p = \{0.05, 0.10, 0.20, 0.50\}$

(a) $p_X = 0.10$

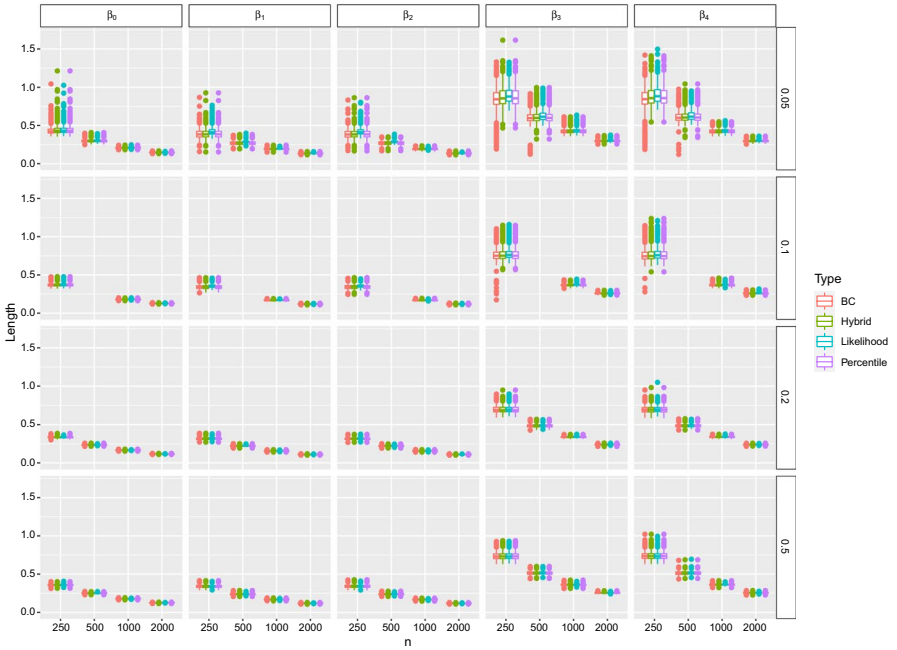


(b) $p_X = 0.05$

Fig. 4 Variance ratios of the bootstrap estimators and of the maximum likelihood estimators with the "true" variance of β_j, $j = 0, 1, 2, 3, 4$, for $p_X = \{0.05, 0.10\}$ and $p = \{0.05, 0.10, 0.20, 0.50\}$

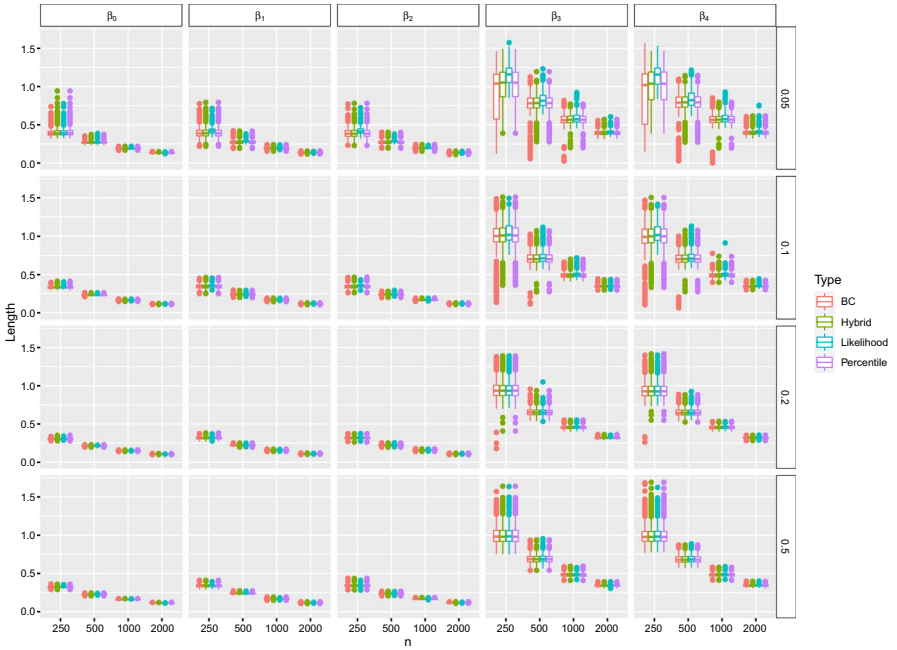


(a) $p_X = 0.50$

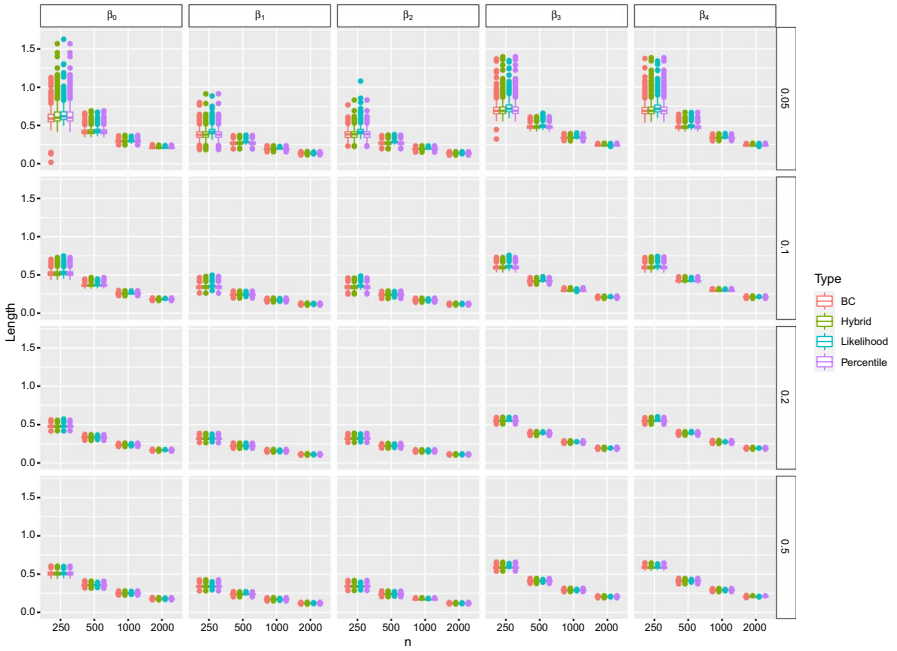


(b) $p_X = 0.20$

Fig. 5 Confidence intervals length of the percentile, bias corrected and hybrid bootstrap method and of the confidence intervals based on likelihood, for different values of $p_X = \{0.20, 0.50\}$ and $p = \{0.05, 0.10, 0.20, 0.50\}$

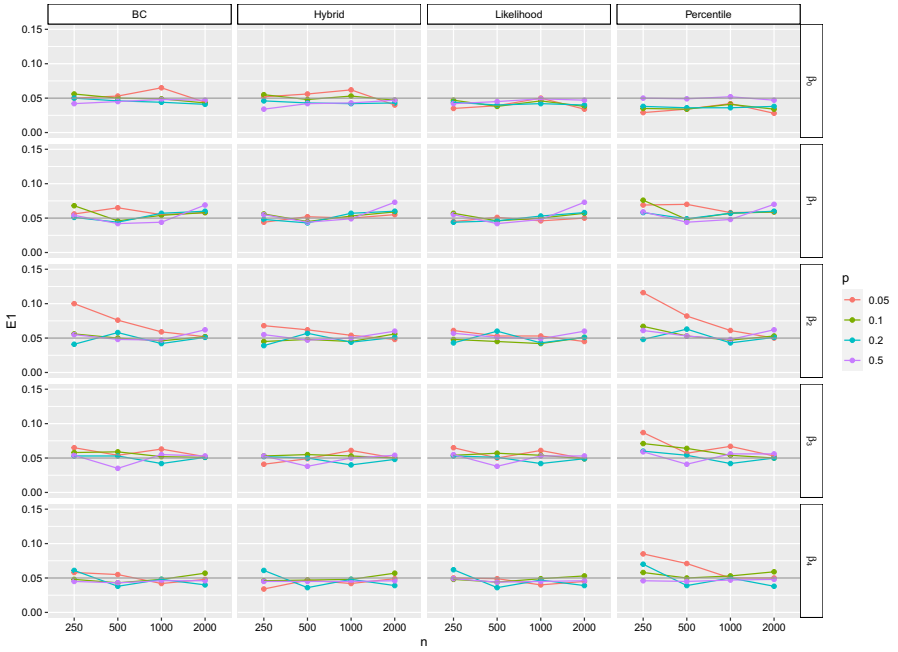


(a) $p_X = 0.10$

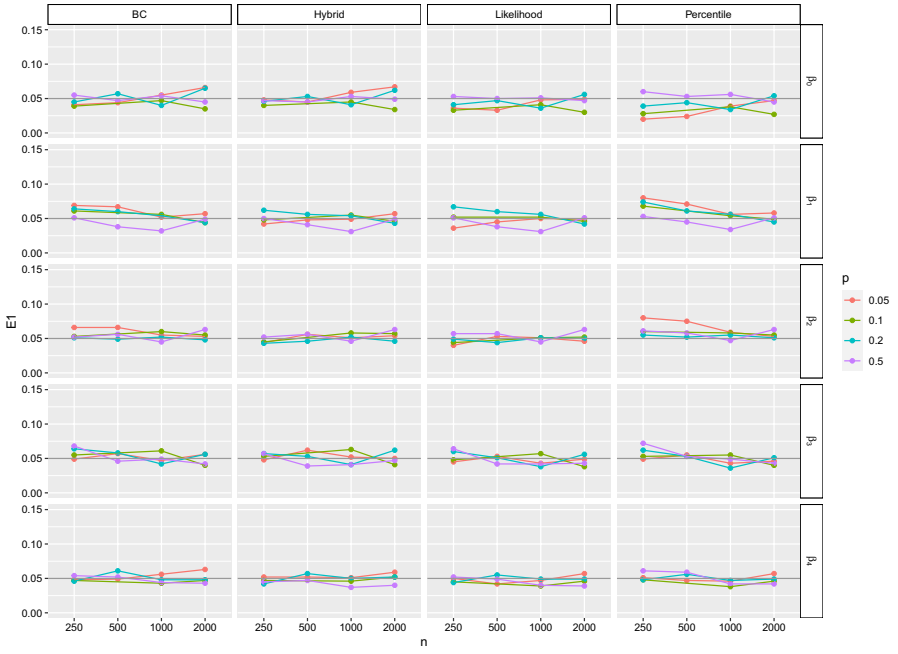


(b) $p_X = 0.05$

Fig. 6 Confidence intervals length of the percentile, bias corrected and hybrid bootstrap method and of the confidence intervals based on likelihood, for different values of $p_X = \{0.05, 0.10\}$ and $p = \{0.05, 0.10, 0.20, 0.50\}$

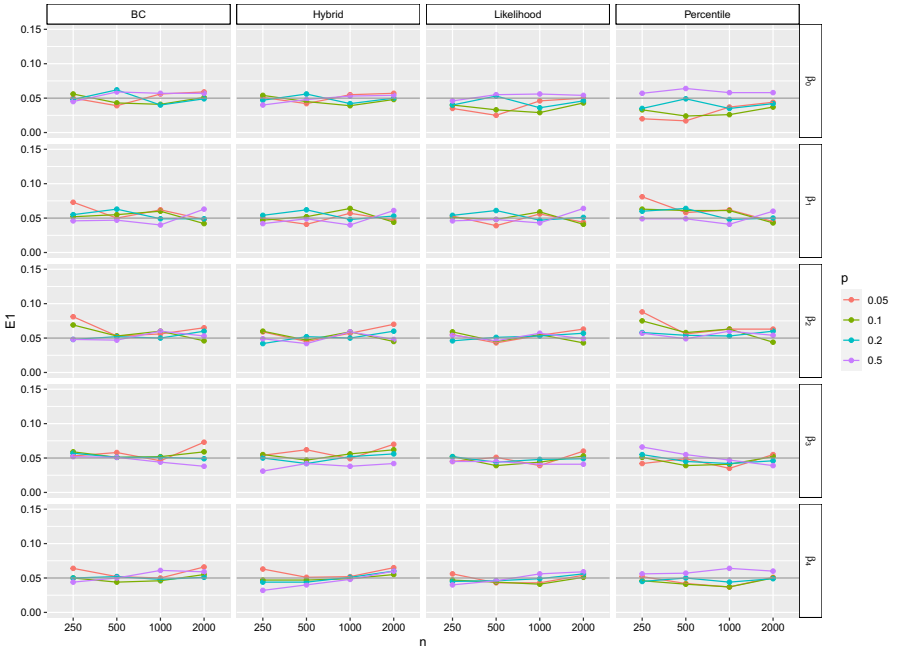


(a) $p_X = 0.50$

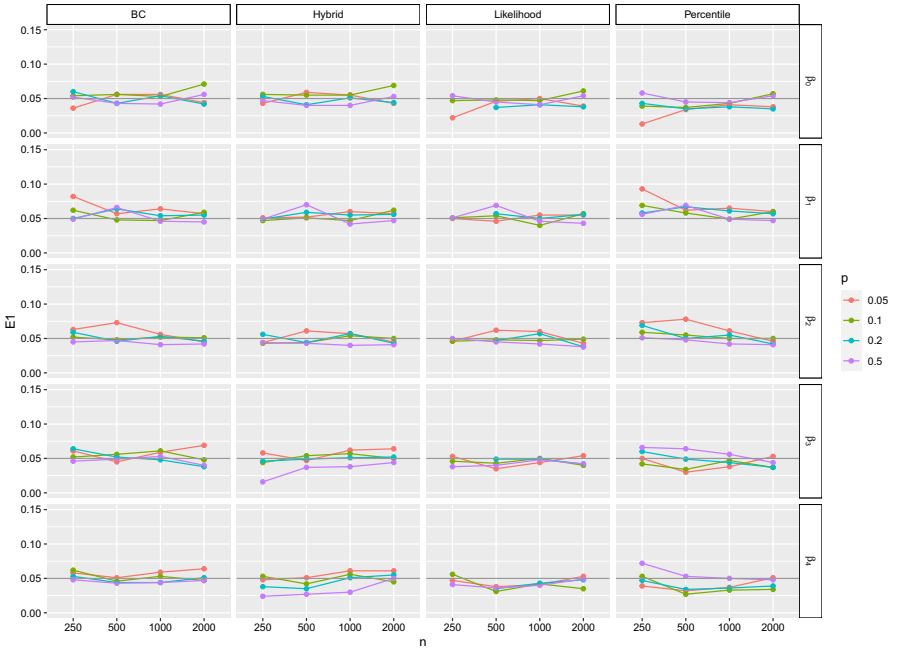


(b) $p_X = 0.20$

Fig. 7 Empirical percentage error of the lower FRW bootstrap confidence bound, with nominal level $\alpha/2 = 0.05$, $p_X = \{0.20, 0.50\}$ and $p = \{0.05, 0.10, 0.20, 0.50\}$. The gray line is the nominal level

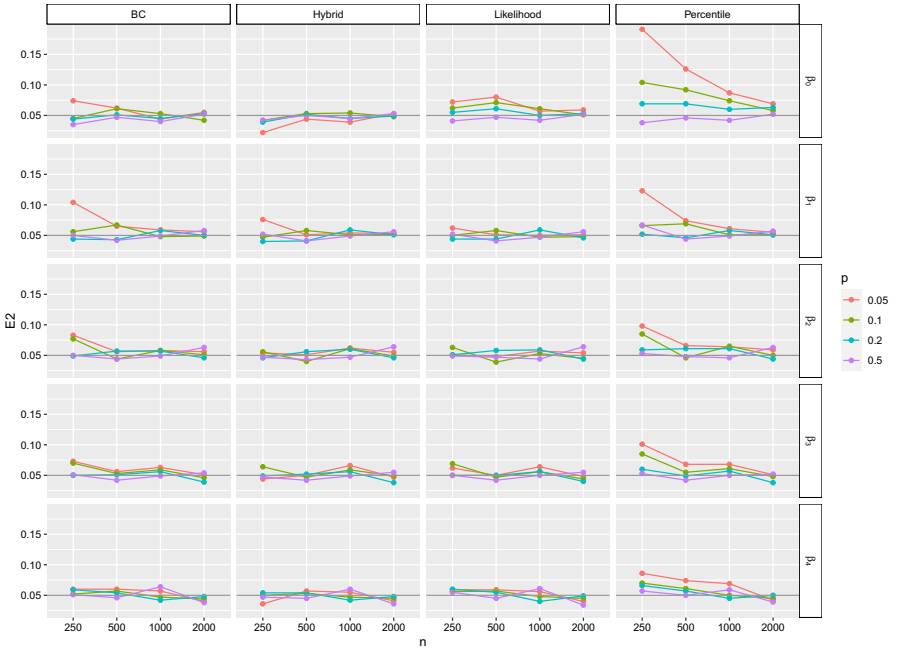


(a) $p_X = 0.10$

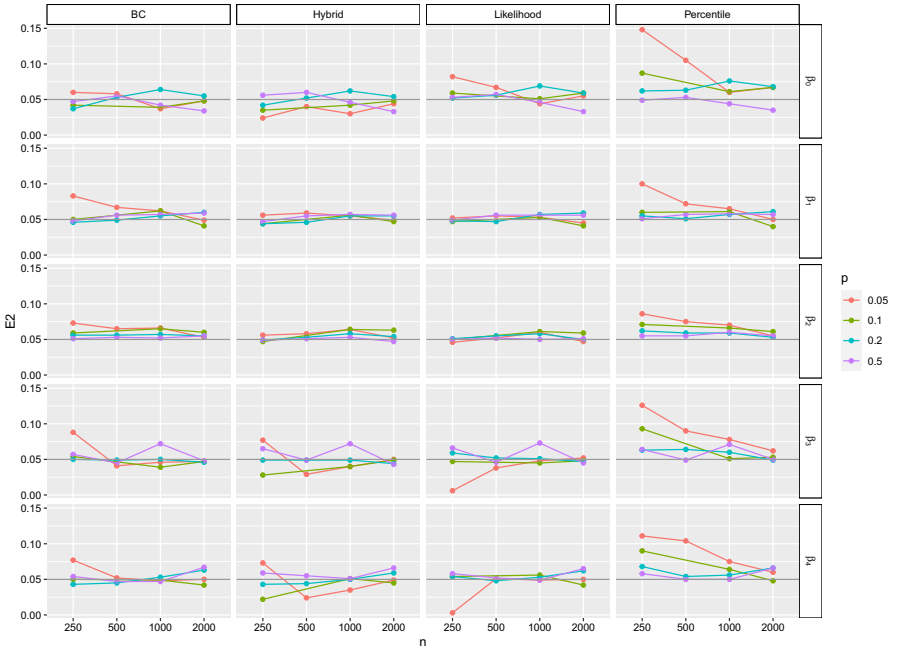


(b) $p_X = 0.05$

Fig. 8 Empirical percentage error of the lower FRW bootstrap confidence bound, with nominal level $\alpha/2 = 0.05$, $p_X = \{0.05, 0.10\}$ and $p = \{0.05, 0.10, 0.20, 0.50\}$. The gray line is the nominal level

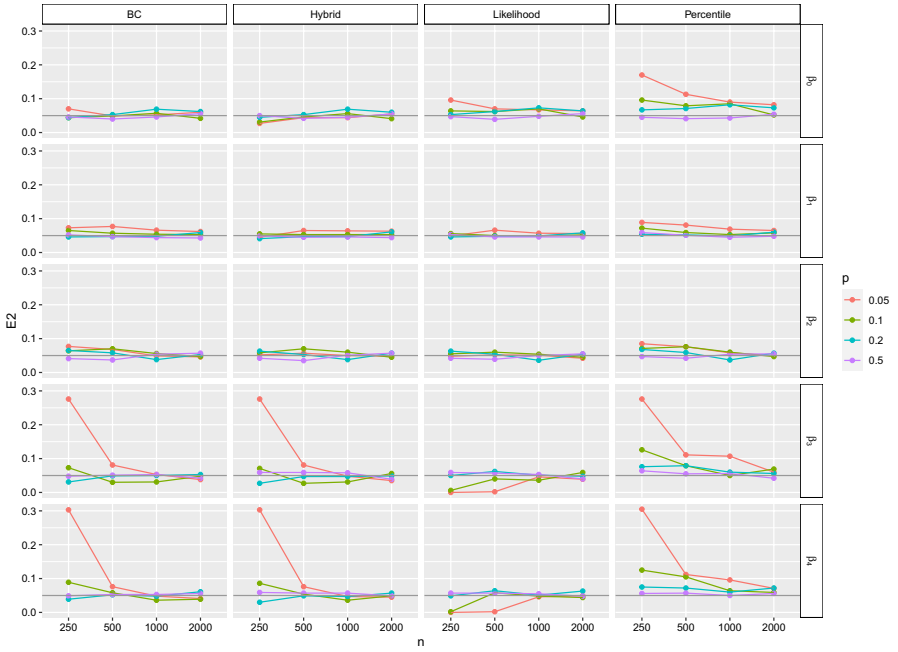


(a) $p_X = 0.50$

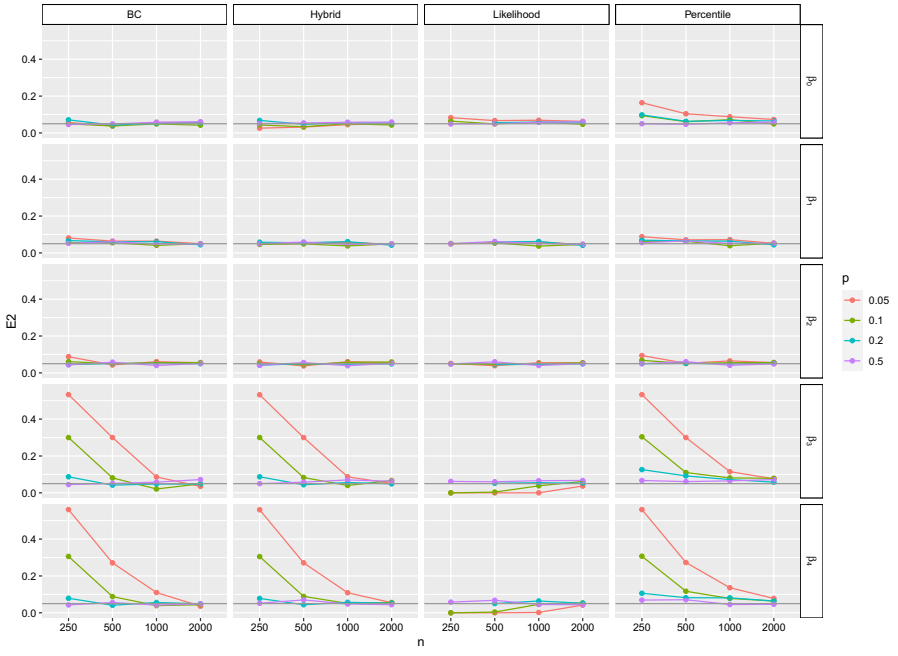


(b) $p_X = 0.20$

Fig. 9 Empirical percentage error of the upper FRW bootstrap confidence bound, with nominal level $\alpha/2 = 0.05$, $p_X = \{0.20, 0.50\}$ and $p = \{0.05, 0.10, 0.20, 0.50\}$. The gray line is the nominal level

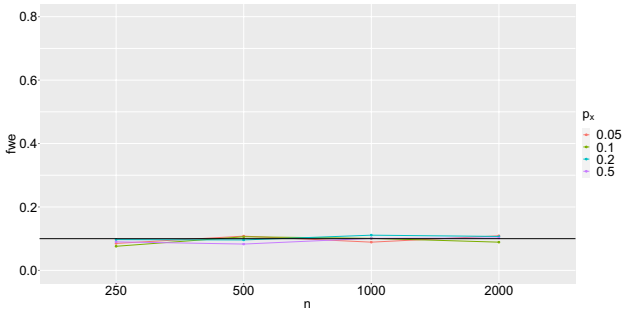


(a) $p_X = 0.10$

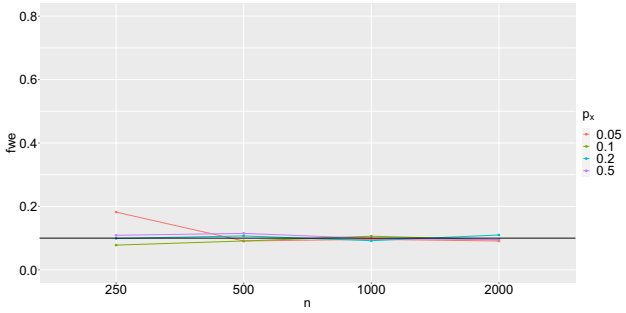


(b) $p_X = 0.05$

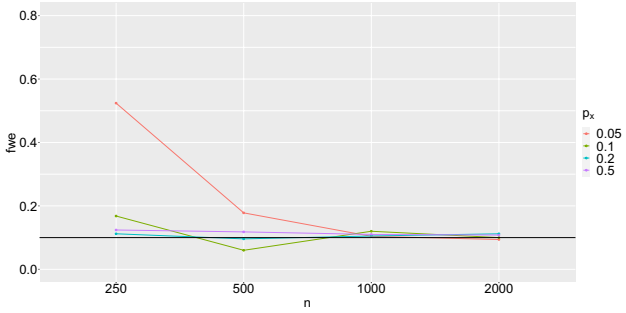
Fig. 10 Empirical percentage error of the upper FRW bootstrap confidence bound, with nominal level $\alpha/2 = 0.05$, $p_X = \{0.05, 0.10\}$ and $p = \{0.05, 0.10, 0.20, 0.50\}$. The gray line is the nominal level



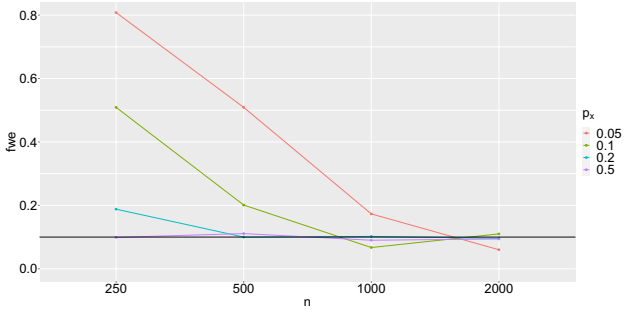
(a) $P(Y = 1|\mathbf{x}) = 0.50$



(b) $P(Y = 1|\mathbf{x}) = 0.20$



(c) $P(Y = 1|\mathbf{x}) = 0.10$



(d) $P(Y = 1|\mathbf{x}) = 0.05$

Fig. 11 Empirical FWE obtained from the FRW bootstrap distributions

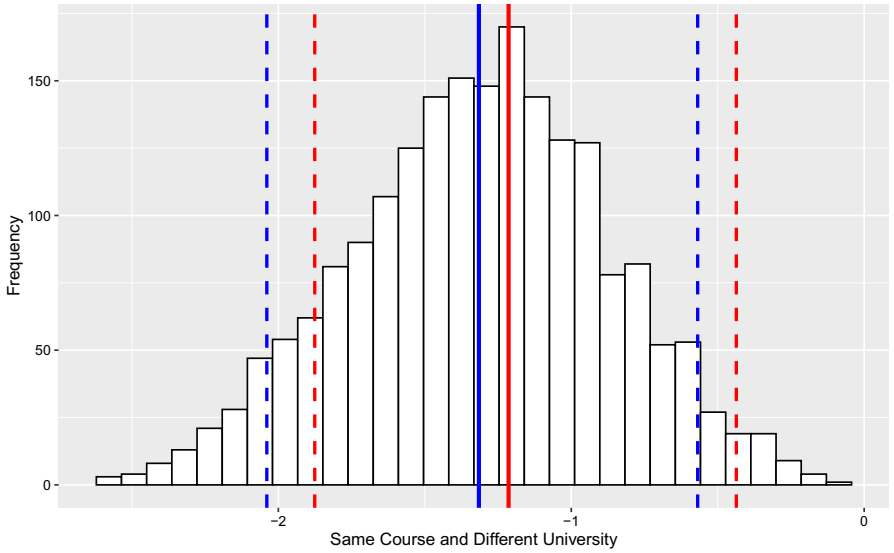


Fig. 12 Histogram of the FRW bootstrap distribution along with the GEV regression estimates (solid red line) and the BC confidence interval (dashed red lines). The blue solid line is the estimate of β by *c-log-log* regression with its confidence interval (dashed blue lines)

results show that the imbalance of Y and the predictors, combined with that of the small values of n, leads to the inclusion of irrelevant variables in the model. However, satisfactory results are obtained even with these imbalanced data, as n grows.

As an application to a real dataset, we analyzed university students' churn, defined as their choice to opt for continuing their studies in other universities after earning their first-level graduation at a specific university. We identified the main factors that might contribute to this choice using different variable selection approaches. The multiple testing procedure based on the FRW bootstrap distributions of the GEV regression parameters with fixed FWE, reduced the number of false positives.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00180-023-01330-y.

Acknowledgements The authors would like to thank the anonymous reviewers for their careful inspection, expert opinions, and detailed comments, which were crucial to the revision of the manuscript. The authors would like to express their gratitude to the University of Salerno for providing the dataset on which the empirical analysis was based.

Funding Open access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York

Bergtold JS, Yeager EA, Featherstone AM (2018) Inferences from logistic regression models in the presence of small samples, rare events, nonlinearity, and multicollinearity with observational data. J Appl Stat 45(33):528–546

Calabrese R, Giudici P (2015) Estimating bank default with generalised extreme value regression models. J Oper Res Soc 66(11):1783–1792

Calabrese R, Osmetti S (2013) Modelling SME loan defaults as rare events: an application to credit defaults. J Appl Stat 40(6):1172–1188

Calabrese R, Marra G, Osmetti SA (2016) Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. J Oper Res Soc 67:604–615

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Chen M-H, Dey DK, Shao Q-M (1999) A new skewed link model for dichotomous quantal response data. J Am Stat Assoc 94(448):1172–1186

Coles S (2001) An introduction to statistical modeling of extreme values. Springer, Berlin

DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals. Stat Sci 11(3):189–228

Dobson AJ, Barnett AG (2008) An introduction to generalized linear models, 3rd edn. CRC Press, New York

Efron B (1982) The Jackknife, the bootstrap, and other resampling plans. CBMS-NF n038, S.I.A.M., Philadelphia

Estabrooks A, Taeho J, Japkovicz N (2004) A multiple resampling method for learning form imbalanced data sets. Comput Intell 20(1):18–36

Jin Z, Ying Z, Wei L (2001) A simple resampling method by perturbing the minimand. Biometrika 88(2):381–390

Kim S, Chen M-H, Dey DK (2007) Flexible generalized t-link models for binary response data. Biometrika 95(1):93–106

Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 73(1):220–239

Hesterberg TC (2015) What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum. Am Stat 61:371–386

Japkowicz N, Stephen S (2002) The class imbalance problem: a systematic study. Intell Data An Js 6(5):429–449

King G, Zeng L (2001) Logistic regression in rare events data. Polit Anal 9(2):137–163

Kotz S, Nadarajah S (2000) Extreme values distributions. Theory and methods. Imperial College Press, London

Krawczyk B (2001) Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 5:221–232

McCullagh P, Nelder JA (1989) Generalized linear models. Chapmann Hall, New York

Olmus H, Nazman E, Erbaş S (2022) Comparison of penalized logistic regression models for rare event case. Commun Stat Simul Comput 51(1578–1590):1578–1590

R Core Team (2022) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Romano JP, Wolf M (2005a) Exact and approximate stepdown methods for multiple hypothesis testing. J Am Stat Assoc 100(469):94–108

Romano JP, Wolf M (2005b) Stepwise multiple testing as formalized data snooping. Econometrica 73(4):1237–1282

Romano JP, Wolf M (2016) Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. Stat Prob Lett 113:38–40

Romano JP, Shaikh AM, Wolf M (2008) Formalized data snooping based on generalized error rates. Econom Theory 24(2):404–447

Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A (2008) Resampling or reweighting: a comparison of boosting implementations. In: 20th IEEE international conference tools with artificial intelligence, vol 1. IEEE, pp 445–451

Shao J, Tu D (1995) The Jackknife and bootstrap. Springer, New York

Smith RL (1985) Maximum likelihood estimation in a class of non-regular cases. Biometrika 72:67–90

Sun Y, Wong AC, Kamel MS (2009) Classification of imbalanced data: a review. Int J Pattern Recognit Artif Intell 23(4):687–719

Tahir MA, Kittler J, Mikolajczyk K, Yan F (2012) A multiple expert approach to the class imbalance problem using inverse random under sampling. In: Multiple classifier systems. Springer, pp 82–91

Wang X, Dey DK (2010) Generalised extreme value regression for binary response data: an application to b2b electronic payments system adoption. Ann Appl Stat 4(4):2000–2023

Xu L, Gotwalt C, Hong Y, King CB, Meeker WQ (2020) Applications of the fractional-random-weight bootstrap. Am Stat 74(4):345–358