Variable selection in generalized extreme value regression model using Bootstrap method

DIOP Aba*
E-mail: aba.diop@uadb.edu.sn
Dpartement de Mathmatiques
Universit Alioune Diop, Bambey, Sngal
NIANG Mouhamed Amine
E-mail: amine.niang@uadb.edu.sn
Dpartement de Mathmatiques
Universit Alioune Diop, Bambey, Sngal

Abstract

In applied research in general, analysts frequently use variable selection methods in order to identify independent predictors of an outcome. The bootstrap method replaces complex analytical procedures by computer intensive empirical analysis. It relies heavily on Monte Carlo Method where several random resamples are drawn from a given original sample. The bootstrap method has been shown to be an effective technique in situations where it is necessary to determine the sampling distribution of (usually) a complex statistic with an unknown probability distribution using these data in a single sample. This work investigates the use of bootstrap tools in the context of variable selection in the generalized extreme value regression model. The treatment is based specifically upon drawing repeated bootstrap samples from the original dataset by founding the proportion of bootstrap samples in which each variable was identified as an independent predictor of the outcome. We performed a real data application and compared this approch with traditional model selection methods.

Keywords: Regression model, Model selection, Bootstrap, Multivariate analysis, Information criterion.

*Corresponding author

https://doi.org/10.32388/WC481X

1 Introduction

In public health and in applied research in general, analysts frequently use variable selection methods such as backward elimination or forward selection in order to identify independent predictors of an outcome or for developing parsimonious regression models (Miller (2002)). Model choice is also of primary concern in many areas of applied studies. Becker *et al.* (2021) has developped an approch of variable selection in regression models using global sensitivity analysis. Zhang *et al.* (2014) proposed propose a new variable selection method called logistic elastic net for the logistic regression model in pattern recognition. In linear regression model, there are many methods can solve the problem of variable selection, such as Lasso proposed by Tibshirani (1996) and many improved lasso methods. It shrinks some coefficients and sets others to zero, and hence makes it retain the good features of both subset selection and ridge regression. The lasso has injected great vitality for the area of variable selection, especially when least angle regression (LARS) algorithm was proposed by Efron *et al.* (2004).

In regression model, if the error distribution is heavy tailed or there are a few outliers in the data, the least squares estimates will give too much weight to the outliers, trying to fi t the outliers well at the expense of the rest of the data. Outliers have high infl uence on the regression parameters in that removing them would radically change the estimates. When the error distribution has heavy tails, robust procedures such as least absolute deviations, M-estimation, and repeated medians may be better approaches even though they are analytically more complex. Regardless of the procedure used to estimate the regression parameters if we are interested in confi dence regions for the parameters or want prediction intervals for future cases, we need to know more about the error distribution. However, when the error distribution is unknown and non - Gaussian, the bootstrap provides a way to get such estimates regardless of the method for estimating the parameters. Other compli-

cations in regression modeling can also be handled by bootstrapping. These include model selection, heteroscedasticity of variances, nonlinearity in the model parameters, and bias due to transformation (see Chernick and Labudde (2011), Zoubir and Iskander (2004)).

Bootstrap methods introduced by Efron (1979), have been used for assessing the performance of regression models. Both Efron and Tibshirani (1994) and Davison and Hinkley (1997) described bootstrap methods to assess the prediction error of a specific regression model. In choosing between competing candidate models, one can select the model with the lowest bootstrap-corrected prediction error. Similarly, bootstrap sampling allows one to assess the statistical significance of individual regressors (Efron and Tibshirani (1994)).

The bootstrap method has been applied effectively in a variety of situations. Many studies have shown that the bootstrap resampling technique provides a more accurate estimate of a parameter than the analysis of any one of the samples (see forexample Diop and Deme (2021), Carpenter and Bithell (2000), Zoubir and Iskander (2004), Austin (2008)). Harrll (2015) and Austin and Tu (2004) had exhibited how bootstrap methods can be used for variable selection. This included simple bootstrapping and bootstrapping incorporating automated methods.
In this work we propose a procedure variable selestion in linear regression model using bootstrap method accordind to the approch proposed by Austin and Tu (2004). The rest of this paper is organized as follows. In Section 2, we describe the problem of generalized extreme value regression model and bootstrap method procedure for variable selection. Section 3 present the obtained results on real data application. A discussion and some perspectives are given in Section 4.

2 Method

2.1 Generalized extreme value regression model

Generalized extreme value is widely used to model rare and extreme event (see Coles (2001). In the case where the dependent variable Y represents a rare event, the logistic regression model (obviously used for this category of data) shows relevant drawbacks. We suggest the quantile function of the GEV distribution as link function to investigate the relationship between the binary response variable Y and the potential predictors $\mathbf{X}$ (see Wang and Dey (2010) and Calabrese and Osmetti (2013) for more details). We use Bootstrapping method as a tool to implement a variables selection procedure and bootstrapping residuals in the generalized extreme value regression model in order to develop a parsimonious predictive model. For a binary response variable Y_i and the vector of explanatory variables x_i, let $\pi(x_i) = \mathbb{P}(Y_i = 1|\mathbf{X}_i = x_i)$ the conditional probability of infection. Since we consider the class of Generalized Linear Models, we suggest the GEV cumulative distribution function proposed by Calabrese and Osmetti (2013) as the response curve given by

$$\begin{aligned} \pi(x_i) &= 1 - \exp\{[(1 - \tau(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}))_+]^{-1/\tau}\} \\ &= 1 - GEV(-x_i'\beta; \tau) \end{aligned} \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ is an unknown regression parameter measuring the association between potential predictors and the response variable Y and $GEV(x; \tau)$ represents the cumulative probability at x for the GEV distribution with a location parameter $\mu = 0$, a scale parameter $\sigma = 1$, an unknown shape parameter τ.

For $\tau \to 0$, the previous model (1) becomes the response curve of the log-log model, for $\tau > 0$ and $\tau < 0$ it becomes the Frechet and Weibull response curve respectively, a particular case of the GEV one.

The link function of the GEV model is given by

$$\frac{1 - [\log(1 - \pi(x_i))]^{-\tau}}{\tau} = x_i'\beta = \eta(x_i) \tag{2}$$

The unknown vector parameter β can be estimated with $(1-\alpha)\%$ confidence intervals ($\alpha \in [0;1]$) and a test of hypothesis H_0: $\beta_j = 0$ by both classical approch of GEV regression model and bootstrap methods (see Diop and Deme (2021)).

2.2 Bootstrapping variable selection

The bootstrap is a well-known statistical method used to assess the variability of test statistics (Efron and Tibshirani (1994), Davison and Hinkley (1997)). The nonparametric bootstrap allows one to estimate an empirical distribution function by repeated sampling from the observed data. The use of bootstrap methods allows one to approximate the distribution of test statistics in settings in which analytic calculations are intractable or in small samples in which large-scale asymptotic results may not hold.

Austin and Tu (2004) proposed model selection method based upon drawing repeated bootstrap samples from the original dataset. Within each bootstrap sample, backwards elimination (by taking a threshold of $\alpha = 0.05$ for eliminating a variable from the model) is used to develop a parsimonious predictive model. For each candidate variable, the proportion of bootstrap samples in which that variable was identified as an independent predictor of the outcome is determined. Candidate variables are then ranked according to the proportion of bootstrap samples in which they were identified as independent predictors of the outcome. The algorithm is summarize in the following box:

1. *Draw B bootstrap samples $\{(y_i^{(b)}, \mathbf{X}_i^{(b)}), i = 1, \ldots, n\}$ $(b = 1, \ldots, B)$ from the original data sample, and for each bootstrap sample, compute the last squares estimate $\hat{\beta}_j^{(b)}$ of β in the model 1 and its estimate standard error $\hat{\sigma}_{\hat{\beta}_j^{(b)}}$ for $j = 1, \ldots, p$.*

2. *For each bootstrap sample, estimate observed statistic test under the null hypothesis $H_0 : \beta_j = 0$, $t^{obs}_{\hat{\beta}_j^{(b)}} = \hat{\beta}_j^{(b)} / \hat{\sigma}_{\hat{\beta}_j^{(b)}}$, for $j = 1, \ldots, p$ and calculate the p-value for each variable $\mathbf{X}_j$ by: $p - value_j^{(b)} = \mathbb{P}(|t_{n-p}| > |t^{obs}_{\hat{\beta}_j^{(b)}}| \ / \ H_0)$.*

3. *Taking a threshold of $\alpha = 0.05$, for $b = 1, \ldots, B$ and for each candidate variable X_j, calculate the proportion p_j of bootstrap samples in which that variable was identified as an independent predictor of the outcome is determined by*

$$p_j = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\left\{p-value_j^{(b)} < \alpha\right\}}, \quad j = 1, \ldots, p.$$

A preliminary selection model would consist of those variables that were identified as significant predictors in all bootstrap samples. Variables could then be sequentially added to this preliminary model according to the proportion of bootstrap samples in which they were selected as significant predictors. Each candidate model can then be assessed for its predictive accuracy and a final model identified.

3 Real data application

3.1 Data Source

Data for the case study consisted of 162 patients discharged with a diagnosis of stroke (which is a sudden neurological deficit of vascular origin caused by an infarct or haemorrhage in the

brain. See Biousse (1994) for more details) from Medical Imagery Service of both Matlaboul Fawzeini Hospital in Touba and Elhadj Ibrahima Niass regional hospital in Kaolack located in central Senegal between April 5, 2016, and November 30, 2016. All variables used in the current study were dichotomous. They denoted the present or absence of a specific condition or risk factor. The data was collected in the context of a prospective and analytical study. Patients with stroke confirmation were included in the study and the dependant variable Y is the patient's vital prognosis (engaged : yes or no).
We consider the following covariates in the dataset:

3.2 Results

We used the bootstrap model selection method describe in Section 2 to develop parsimonious models for predicting stroke patient's vital prognosis. For the backwards model selection process, we used a threshold of $\alpha = 0.05$ for eliminating a variable from the model. We used 1000 bootstrap samples. The number of times that each variable was identified as a significant predictor is summarized in Table 2. Four variables (Delay, Intraventricular haemorrhages, Stroke type and Severity Cerebral commitment) were selected as independant predictors of stroke patients's vital prognosis in at least 70% of the bootstrap samples. An additional variable (Motor deficiency) was selected in at least 50% of the bootstrap samples. The remaining seven candidate variables were selected as independent predictors of stroke patient's vital prognosis in a minority of the bootstrap samples. A further three variables (Disturbance of consciousness, Sex and Hemiplegia) were selected as independant predictors in at least 30% of the bootstrap samples. Four (Age, Cardiopathy, Hypertension and Diabetes) of the variables were identified as independent predictors of stroke patients's vital prognosis in fewer than 15% of the bootstrap,samples using backwards elimination. Finnaly aach candidate variable was identified as a significant predictor of stroke patient's vital prognosis in at least 16.2% of the bootstrap samples.

Table 1: Explain variables description

Variable	Description	Abbreviation
Age	age of stroke patients	Age
Sex	sex of stroke patients	Sex
Stroke type	Ischemic stroke or Hemorrhagic stroke	Stroke-Type
Cardiopathy	coronary insufficiency	Cardiopathy
Diabetes	insufficient insulin production by the pancreas	Diabetes
Hypertension	abnormal increase of blood pressure on artery walls	Hypertension
Hemiplegia	total or partial paralysis of one half (left or right) of the body	Hemiplegia
Disturbance of consciousness	any disturbance of vigilance and conscious thinking	Disturb-cons
Motor deficiency	affected mobility of the upper and/or lower limbs	Motor-def
Severity Cerebral commitment	displacement of parts of the nervous structure contained in the cranium through an orifice	SC-commitment
Intraventricular haemorrhages	bleeding into the ventricles of the brain	Ivh
Hospital Admission Delay	delay between the first symptoms and admission to hospital	Delay
Vital Prognosis	vital prognosis	Prognosis

Table 2: Frequency selected variables

Variable	Frequency selected
Delay	0.995
Ivh	0.894
Stroke-Type	0.886
SC-commitment	0.730
Motor-def	0.660
Disturb-cons	0.445
Sex	0.372
Hemiplegia	0.367
Age	0.288
Cardiopathy	0.264
Hypertension	0.184
Diabetes	0.162

Using the results of the bootstrap sampling, we created a series of candidate models for predicting stroke patient's vital prognosis. They contained the variables that were selected in at least 70% (model $\mathcal{M}_1$), 50% (model $\mathcal{M}_2$), 30% (model $\mathcal{M}_3$) and 15% (model $\mathcal{M}_4$) of the bootstrap samples using backwards elimination. Note that the model with selected candidate variables in at least 15% contain all 12 variables. For each model we assessed its predictive ability using the probability of concordance between predicted probability and outcome, the Aikaike's information criterion (AIC) and Schwartz's Bayesian information criterion (BIC). This predictive ability correspond to the area under a receiver operating characteristic (ROC) curve, a widely used measure of diagnostic discrimination. Results for each predictive model are summarized in Table 3. The first model (containing variables

Table 3: Goodness of fit and discrimination of each predictive model

Model	Area under the ROC curve	AIC	BIC
$\mathcal{M}_1$	0.696	-126.7167	-117.4539
$\mathcal{M}_2$	0.720	-113.6718	-101.3215
$\mathcal{M}_3$	0.619	-99.6414	-74.9407
$\mathcal{M}_4$	0.427	-89.2917	-52.2405

Delay, Intraventricular haemorrhages, Stroke type and Severity Cerebral commitment) had an area under the ROC curve of 0.696 and the lowest values of AIC and BIC. Once we included those variables that were identified as significant predictors of mortality in at least 50% of the bootstrap samples (model $\mathcal{M}_2$), the area under the ROC curve increased to 0.720. Adding additional variables did not result in a substantial increase in the area under the ROC curve. We also compared the area under the ROC curve for each predictive model using an algorithm proposed by Delong *et al.* (1988). The p-values of the null-hypothesis testing H_0: *The area under the ROC curve of* $\mathcal{M}_1$ *is not lower than that for each of the*

other models are summarized in Table 4. We used both Delong and Bootstrap methods. The Table 4 shows that the area under the ROC curve of the model $\mathcal{M}_1$ is significantly

Table 4: Comparaison of Area under the ROC curve

Model	Delong-Method	Bootstrap-Method
$\mathcal{M}_1$ vs $\mathcal{M}_2$	0.000512	0.000536
$\mathcal{M}_1$ vs $\mathcal{M}_3$	0.001048	0.001077
$\mathcal{M}_1$ vs $\mathcal{M}_4$	0.000824	0.000882

lower than that for each of the other models. The use of either AIC or BIC resulted in the selection of the model $\mathcal{M}_1$ identically to the model selected by the bootstrap model selection procedure.

4 Discussion and perspectives

The model selection procedure describeb in this work using bootstrap resampling and applied to clinical dataset in order to develop a predictive models presents good performance. By combining bootstrap sampling with automated variable selection methods, we were able to determine the empirical distribution of a variable's likelihood of being identified as an independent predictor of stroke patient's vital prognosis. Several question can be asked about for example the appropriate value of bootstrap samples for model choice.

References

Ariffin S.B., Midi H., 2012. Robust Bootstrap Methods in Logistic Regression Model. 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE), Langkawi, 10-12, 1-6.

Austin P.C., Tu J.V., 2004. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. Journal of Clinical Epidemiology 57, pp.11381146.

Austin P.C., 2008. Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backward variable elimination: a simulation study. Journal of Clinical Epidemiology 61, pp.1009-1017.

Adjei1 I.A., Karim R., 2016. An Application of Bootstrapping in Logistic Regression Model. Open Access Library Journal, 3: e3049.

Becker W., Paruolo P., Saltelli A., 2021. variable selection in regression models using global sensitivity analysis J. Time Ser. Econom. pp 147.

Biousse V., 1994. Etiologie et mcanisme des accidents vasculaires crbraux. AnnRadiol 37: 11-16.

Brunea, F. 2008. Consistent Selection via the Lasso for High Dimensional Approcimating Regression Models. In Pushing the Limits of Contemporary Statistics: Essays in Honor of J. K. Gosh, edited by B. Clarke, and S. Ghosal, 12237. Dordrecht: IMS.

Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Multimodel Inference A Practical Information-Theoretic Approach, 2nd ed. New York: Springer.

Chernick M.R., 1999. Bootstrap methods: a practitioner's guide. New York: Wiley.

Chernick M.R., Labudde R.A., 2011. An introduction to bootstrap methods with applications to R. John Wiley & Sons, Inc., Hoboken, New Jersey.

Claeskens, G., and N. L. Hjort. 2003. The Focused Information Criterion, with Discussion. Journal of the American Statistical Association 98: 900845.

Carpenter J., Bithell J., 2000. Bootstrap Confidence Intervals: When, Which, What ? A Practical Guide for Medical Statisticians. Statistics in Medicine, 19, 1141-1164.

Coles S.G., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer, New York.

Calabrese R., Osmetti S., 2013. Modelling SME Loan Defaults as Rare Events: an Application to Credit Defaults. Journal of Applied Statistics 40 (6), 1172-1188.

Davison A.C., Hinkley D.V., Young, G.A., 2003. Recent Developments in Bootstrap Methodology. Statistical Science , 18, 141-157.

Davison A.C., Hinkley D.V, 1997. Bootstrap Methods and Their Application, Cambridge University Press.

Diop A., Deme E., 2021. Parametric bootstrapping method in generalized extreme value regression model for binary response. Preprint.

DeLong E.R., DeLong D.M., Clarke-Pearson D.L., 1988. Compar ing the Areas Under Two or More Correlated Receiver Operating Curves: A Nonparametric Approach. Biometrics, 44, 837-845.

Efron B., 1979. Bootstrap Methods: Another Look at the Jacknife. Annals of Statistics, vol. 7, N. 1; pp: 1-26.

Efron B., Hastie T., Johnstoneand I., Tibshirani R., 2004. Least angle regression. The Annals of Statistics, Vol.32, No.2, pp.407-499.

Efron B., Tibshirani R.J., 1994. An Introduction to the Bootstrap. Chapman and Hall/CRC, UK.

Hall P., 1992. The bootstrap and Edgeworth expansion. New York: Springer

Harrell F.E. Jr., 2015. Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Ordinal Regression, and Survival Analysis. Second edition. Springer Series in Statistics ISBN 0-387-95232-2.

Hjort N.L., Claeskens G., 2003. Frequentist Model Average Estimators. Journal of the American Statistical Association 98: 87999.

Lger C., Politis D.H.N., Romano J.P., 1992. Bootstrap technology and applications. Technometrics 34, pp.378398.

Leeb H., Poetscher B.M., 2006. Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators ? Annals of Statistics 34: 255491.

Liu W., Yang Y., 2011. Parametric or Nonparametric ? A Parametricness Index for Model Selection. The Annals of Statistics 39 (4): 2074102.

Miller, A.J. 2002. Subset Selection in Regression, 2nd ed. Boca Raton, USA: Chapman and Hall, CRC Press.

Reynolds J.H., Templin W.D., 2004. Comparing Mixture Estimates by Parametric Bootstrapping Likelihood Ratios. Journal of Agricultural, Biological, and Environmental Statistics , 9, 57-74.

Romano J.P., Wolf M., 2005. Stepwise Multiple Testing as Formalized Data Snooping. Econometrica 73: 123782.

Shao X., 2010. The Dependent Wild Bootstrap, Journal of the American Statistical Association Volume 105, 2010 - Issue 489 , Pages 218-235.

Shao J., 1996. Bootstrap model selection. Journal of the American Statistical Association Vol. 91, No. 434, pp. 655-665.

Tibshirani R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, Vol.58, No.1, pp.267-288.

Wang X., Dey D.K., 2010. Generalized extreme value regression for binary response data:An application to B2B electronic payments system adoption. Ann. Appl. Stat. 4, 2000-2023.

Zhang S., Zhang L., Qiu K., Lu Y., Cai B., 2014. Variable selection in logistic regression model. Chinese Journal of Electronics, Vol.23, No.4.

Zoubir A.M., Iskander D.R., 2004. Bootstrap Techniques for Signal Processing. Cambridge University Press, Cambridge.