Review of: "Variable selection in generalized extreme value regression model using Bootstrap method"

Minghui Zhang[1]

1 Tongji University

Potential competing interests: No potential competing interests to declare.

After reviewing the paper titled “Variable selection in generalized extreme value regression model using Bootstrap method” I wrote some notes that could be useful for the researchers:

1. The abstract needs to concisely describe the background and object, methods, results, and conclusions of the research.
2. In the introduction part, the author needs to sort out the research background, rather than a simple list (such as “Becker et al. (2021) has developped an approch of variable selection in regression models using global sensitivity analysis.”). The introduction needs to analyze the previous literature, simply and clearly introduce their research content, the problems that have been solved, what are the current problems, and what methods are used in this study to solve the remaining issues. In my humble opinion, this study does not clearly express the significance and value of the study.
3. The original sample data or the proportion of each variable to the original sample should be provided, and compared with the proportion of each variable to the original sample after resampling by the bootstrap method.
4. Why choose 1000 bootstrap samples? There are only 12 variables in the original sample, but the overall sample data should be no less than three hundred. Is it too small to choose only 1000 bootstrap samples for such a large sample data? Moreover, a small number of bootstrap samples significant uncertainty, and statistical analyses should be performed on multiple selections of 1000 bootstrap samples to determine which variables can be selected as independent predictors of stroke patients’s vital prognosis.
5. Please explain why ROC shows a trend of increasing and then decreasing in the model from M1 to M4, but AIC and BIC have been showing a decreasing trend.
6. From Table 3 in the text, it can be seen that the ROC area of the M2 model is 0.024 larger than that of the M1 model, with a percentage increase of 3.448%, and the authors do not believe that there is a significant increase in area. But does the value in Table 4 reflect the specific value or percentage of the area lower than other model areas? If this value is a specific area difference, with a maximum difference of 0.001, how does it compare to 0.024? If it is a percentage value, how does it compare to 3.448%? Why is the conclusion in the article that " The Table 4 shows that the area under the ROC curve of the model M1 is significantly lower than that for each of the other models."?
7. Is there a contradiction between " The area under the ROC curve of M1**is not lower than that** for each of the other models are summarized in Table 4." and " The Table 4 shows that the area under the ROC curve of the **model M1 is significantly lower than** that for each of the other models." mentioned in the text?

8. Please reanalyze and explain the results of Tables 3 and 4.
9. If an M1-M4 model is constructed based on the proportion of each variable in the original data, what is the difference between the results of this model and those obtained using the Delong and Bootstrap methods?
10. The researchers did not mention in the Discussion the final result of the research.
11. In this PDF version of the manuscript, the spacing between the front and back paragraphs is not uniform, and the indentation of the first line is not uniform. Additionally, the table should be a three-line table.
12. This manuscript contains a large number of English spelling and grammar errors. For example, “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)).”, “Becker et al. (2021) has developped an approach 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. ”, “such as Lasso proposed by Tibshirani (1996) and many improved ~~lasso~~ **Lasso** methods. ”, “Outliers have high infl uence on the regression parameters in that removing them would radically change the estimates.” , “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~~ **non-Gaussian**, the bootstrap provides a way to get such estimates regardless of the method for estimating the parameters.”, “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.”, “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). ”, “Austin and Tu (2004) proposed model selection method based upon drawing repeated bootstrap samples from the original dataset. ”, “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.”.

The above is my personal humble opinion, perhaps it is my personal judgment mistake. But I still hope it will be helpful to you.

Ph.D candidate Minghui Zhang

School of Aerospace Engineering and Applied Mechanics, Tongji University, Shanghai, China