ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII^th SISMEC Congress - Vol. 20 Suppl. 1

Evaluating Variable Selection Methods in a Classification Framework: A Simulation Study

Samuele Minari[1,2], Dario Pescini[1], Antonella Zambon[1,2], Davide Soranna[2]

(1) Laboratory of Quantitative methods for Life, Health and Society (QmLHS-Lab), Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy
(2) Biostatistics Unit, IRCCS Istituto Auxologico Italiano, Milano, Italy

CORRESPONDING AUTHOR: *Minari Samuele, samuele.minari@unimib.it*

INTRODUCTION

Variable selection is a common step in clinical research, where large datasets often include many, potentially highly correlated, variables. The main objective is to identify the most relevant predictors for an outcome, thereby enhancing model interpretability, simplicity, and predictive performance [1]. However, data-driven variable selection also carries several underappreciated risks. These include the potential exclusion of important predictors, inclusion of irrelevant ones, biased coefficient estimates, underestimated standard errors, invalid confidence intervals, and overall model instability [2].

Simulation studies are a valuable approach for evaluating statistical methods, provided they are carefully designed. Yet, many such studies exhibit bias in favor of the newly proposed methods [3]. To address this, we developed a neutral comparison simulation study to fairly evaluate the performance of several variable selection techniques.

OBJECTIVE

To systematically evaluate and compare different variable selection methods across multiple simulated scenarios.

METHODS

To improve the design and reporting of our simulation study, we followed the ADEMP structure [4], this involves specifying the aim (A), the data-generating process (D), the estimand or target of inference (E), the analytical methods (M), and the criteria used to evaluate performance (P).

We designed different simulation scenarios by varying the number of observations, total variables, and number of true predictors. Predictor correlations were modeled to decay exponentially with increasing distance between variables, and effect sizes for true predictors were varied [5, 6]. Noise was introduced into the correlation structures to better mimic real-world data.

We focused on a binary classification setting, evaluating each method on two key outcomes: model selection accuracy (i.e. whether the true model is selected) and predictive performance. Five methods for selecting variables were compared: stepwise logistic regression, LASSO logistic regression, Elastic net logistic regression, Random Forest Classifier with OOB error based backward elimination [7] and Genetic Algorithm (GA) [8, 9]. Performance metrics included the Area Under the Curve (AUC), number of variables selected, and True Positive Rate (TPR). All the analyses were performed using Python 3.12.

RESULTS

We ran 1,000 Monte Carlo simulations per scenario, varying key factors such as sample size, number of predictors, true signal strength, and correlation strength. Elastic Net consistently achieved the highest mean AUC and TPR, particularly in high-dimensional or strong-signal settings (e.g., Scenarios 5–8), showing robust performance across conditions. Random Forest and Genetic Algorithm performed comparably in some scenarios but incurred substantially higher computational costs. LASSO achieved competitive AUC with significantly lower runtime, though it tended to underselect in weaker signal scenarios. Stepwise selection, while the fastest method, had the lowest overall predictive performance and true positive rates (Table 1).

DOI: 10.54103/2282-0930/29431

Table 1. Mean AUC, TRP and number of variable selected (s) over 1000 Monte Carlo simulations by each variable selection methods. The execution time (t) of the 1000 simulations is also reported

	Stepwise				LASSO				Elastic Net				RF				GA			
Scenario*	t	s	TPR	AUC	t	s	TPR	AUC	t	s	TPR	AUC	t	s	TPR	AUC	t	s	TPR	AUC
n=200/m=20/p=3/e=0.15	20s	2	0.259	0.538	4m 4s	1	0.194	0.523	31m 44s	13	0.747	0.558	94m 59s	11	0.683	0.545	122m 14s	6	0.602	0.64
n=200/m=20/p=3/e=0.40	18s	2	0.644	0.753	2m 51s	1	0.391	0.753	32m 20s	5	0.610	0.742	95m 31s	13	0.875	0.679	122m 15s	6	0.686	0.731
n=200/m=50/p=5/e=0.40	1m 1s	4	0.389	0.784	3m 16s	2	0.279	0.794	37m 43d	6	0.371	0.788	188m 49s	12	0.469	0.721	136m 45s	19	0.545	0.767
n=500/m=50/p=5/e=0.45	3m 18s	3	0.103	0.591	9m 2s	6	0.207	0.605	53m 47s	39	0.819	0.616	269m 24s	13	0.258	0.583	164m 12s	19	0.421	0.641
n=500/m=100/p=10/e=0.75	7m 43s	11	0.623	0.852	5m 27s	14	0.626	0.861	77m 16s	26	0.706	0.859	393m 7s	16	0.526	0.821	178m 18s	44	0.643	0.809
n=500/m=50/p=5/e=1.00	1m 13s	4	0.522	0.809	3m 30s	7	0.575	0.809	55m 19s	22	0.752	0.803	252m	16	0.558	0.789	157m 59s	19	0.618	0.760
n=1000/m=100/p=10/e=0.5	8m 36s	11	0.593	0.837	7m 17s	17	0.64	0.838	168m 32s	35	0.760	0.837	553m 16s	14	0.557	0.806	311m 15s	44	0.662	0.776
n=1000/m=50/p=5/e=1.50	2m 7s	5	0.764	0.938	12m 3s	7	0.779	0.938	81m 42s	9	0.783	0.937	321m 21s	13	0.702	0.909	187m 17s	20	0.790	0.865

Parameters of scenarios: n = observations, m = features, p = relevant features, ρ = correlation, e = effect size

CONCLUSION

Among the five evaluated methods, Elastic Net provided the best trade-off between predictive performance and model stability, particularly in realistic, high-dimensional settings. Our results reinforce the importance of carefully considering the variable selection method in the context of the data structure and research goals. This neutral comparison contributes to evidence-based guidance for method selection in clinical research and similar applied settings.

REFERENCES

1. Chan, J. Y. L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022). Mitigating the multicollinearity problem and its machine learning approach: a review. Mathematics, 10(8), 1283.
2. Ullmann T, Heinze G, Hafermann L, Schilhart-Wallisch C, Dunkler D, for TG2 of the STRATOS initiative (2024) Evaluating variable selection methods for multivariable regression models: A simulation study protocol. PLoS ONE 19(8): e0308543.
3. Kipruto E, Sauerbrei W (2022) Comparison of variable selection procedures and investigation of the role of shrinkage in linear regression-protocol of a simulation study in low dimensional data. PLoS ONE 17(10): e0271240
4. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019; 38: 2074–2102
5. Bag, S., Gupta, K., & Deb, S. (2022). A review and recommendations on variable selection methods in regression models for binary data. arXiv preprint arXiv:2201.06063.
6. Hardin, J., Garcia, S. R., & Golan, D. (2013). A method for generating realistic correlation matrices. The Annals of Applied Statistics, 1733-1762
7. Díaz-Uriarte, R., Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics *7*, 3 (2006).
8. M. Mitchell, An Introduction to Genetic Algorithms, Cambridge, MA: MIT Press, 1998.
9. Zhang Z, Trevino V, Hoseini SS, et al. Variable selection in Logistic regression model with genetic algorithm. Ann Transl Med. 2018;6(3):45.

DOI: 10.54103/2282-0930/29431