Comparing Variable Selection Methods: A Simulation Study

Brenton Graham

12/16/2022

Introduction

Variable selection describes the process of selecting the "best" subset of predictors for modeling a given outcome. Variable selection is important in statistics since data sets rarely (i.e., realistically never) come with a full set of useful predictor variables. That is, not all predictors are helpful for explaining an outcome; redundant, non-informative, or collinear predictors should be handled before statistical modeling (e.g., through removal or feature engineering). Parsimonious models, or models that are fit using a simple set of predictors (that still fit the data well), are often preferred for statistical inference. Compared to more complex models, parsimonious models are less prone to overfitting the data and are characterized by increased generalizabilty.

While the importance of variable selection is evident, variable selection can be quite tedious and automated methods for variable selection are flawed. A simulation framework is used in this study to quantitatively compare the performance of five common variable selection methods. Further, we aim to test the effects of sample size and between-predictor correlation on the performance of each method. Variable selection methods of interest include three backward selection approaches (i.e., stepwise selection using p-values, AIC, and BIC for selection criteria) and two regularization approaches (i.e., LASSO and elastic net).

Methods

Data Set Configurations

Data sets were simulated using the `genData` package in R. Six separate parameter configurations were used to simulate these data, representing different combinations of sample size ($n = 250, 500$) and induced between-predictor correlation ($\rho = 0.00, 0.40, 0.80$, representing no correlation, moderate correlation and high correlation, respectively) (Table 1). Each simulated data set includes 20 predictors and a normally distributed outcome variable. Five predictors were set to non-zero, using equally spaced β coefficient values of $\frac{0.5}{3}$, $\frac{1}{3}$, $\frac{1.5}{3}$, $\frac{2}{3}$ and $\frac{2.5}{3}$. Non-zero predictors represent the variables that are associated with the outcome; these are the predictors that ought to remain in the model after variable selection. The other 15 predictor β coefficients were set to 0. In theory, these predictors do not hold prediction value with respect to the outcome and should be removed through variable selection.

Table 1: Data Set Configurations for Simulation Study

Configuration	Sample Size (n)	Correlation	ρ
1	250	None	0.00
2	250	Moderate	0.40
3	250	High	0.80
4	500	None	0.00
5	500	Moderate	0.40
6	500	High	0.80

Variable Selection Methods

The variable selection methods tested in this study include three backward selection approaches (i.e., stepwise selection using p-values, AIC, and BIC for selection criteria) and two regularization approaches (i.e., LASSO and elastic net). Additionally, we explore two methods of selecting the λ penalization hyperparameter in LASSO and elastic net. This hyperparameter specifies the amount of regularization applied to the model; larger λ values result in an increasingly stringent variable selection process. Each λ selection method tested used 5-fold cross-validation and mean-squared error (MSE) as the loss to determine the value of λ. The first method used the λ that corresponds to the minimum mean cross-validated error (λ.min). The second method used the largest value of λ such that the error is within one standard error of the of the cross-validated errors for λ.min (λ.1se). In principal, λ.1se is larger than λ.min and should result in more parsimonious models (i.e., less variables should be selected when setting λ to λ.1se). Overall, seven variable selection models were fit for each simulated data set configuration.

Simulations & Quantitative Assessment

100 data sets were simulated per data set configuration. Each variable selection method was tested with each data set and β estimates from all iterations and model types were stored. Quantitative measurements were deducted following the simulation. We determined true positive rates (TPRs), false positive rates (FPRs), β estimate bias and 95% confidence interval (CI) coverage from the variables (and corresponding β estimates) retained by each model. The TPR in this report reflects the rate at which non-zero predictors remain after variable selection. The FPR reflects the rate at which non-informative predictors are falsely retained. Estimate bias measures the mean difference between model-based point estimates and true population parameters. 95% CI coverage estimates the proportion of 95% CIs that capture true β parameter values (i.e., how often the true β parameter is contained within the 95% CI of a model's β estimate). Variable-specific selection rates were also tracked to evaluate the effect of true β parameters on variable selection.

Results

Variable-Specific Selection Rates

Variable-specific selection rates are shown for each model type and multiple data set configurations in Figure 1. The effect of β on variable retention is quite clear and confirms what we might expect. Non-zero predictors in Figure 1 include variables 1-5 while non-informative predictors include variables 6-20. Non-zero predictors are associated with increased selection rates when compared to non-informative predictors. Further, we would generally expect non-zero predictors with larger β coefficients to be retained during variable selection at higher rates than non-zero predictors with smaller β coefficients. This trend is observed in Figure 1. β coefficients increase from variable 1 to variable 5, with the smallest non-zero beta coefficient belonging to variable 1 and the largest beta coefficient belonging to variable 5. Of the non-zero predictors, variable 1 is associated with the lowest selection rate. Additionally, an increase in β is generally associated with an increase in selection rate. These trends are present across all variable selection methods.

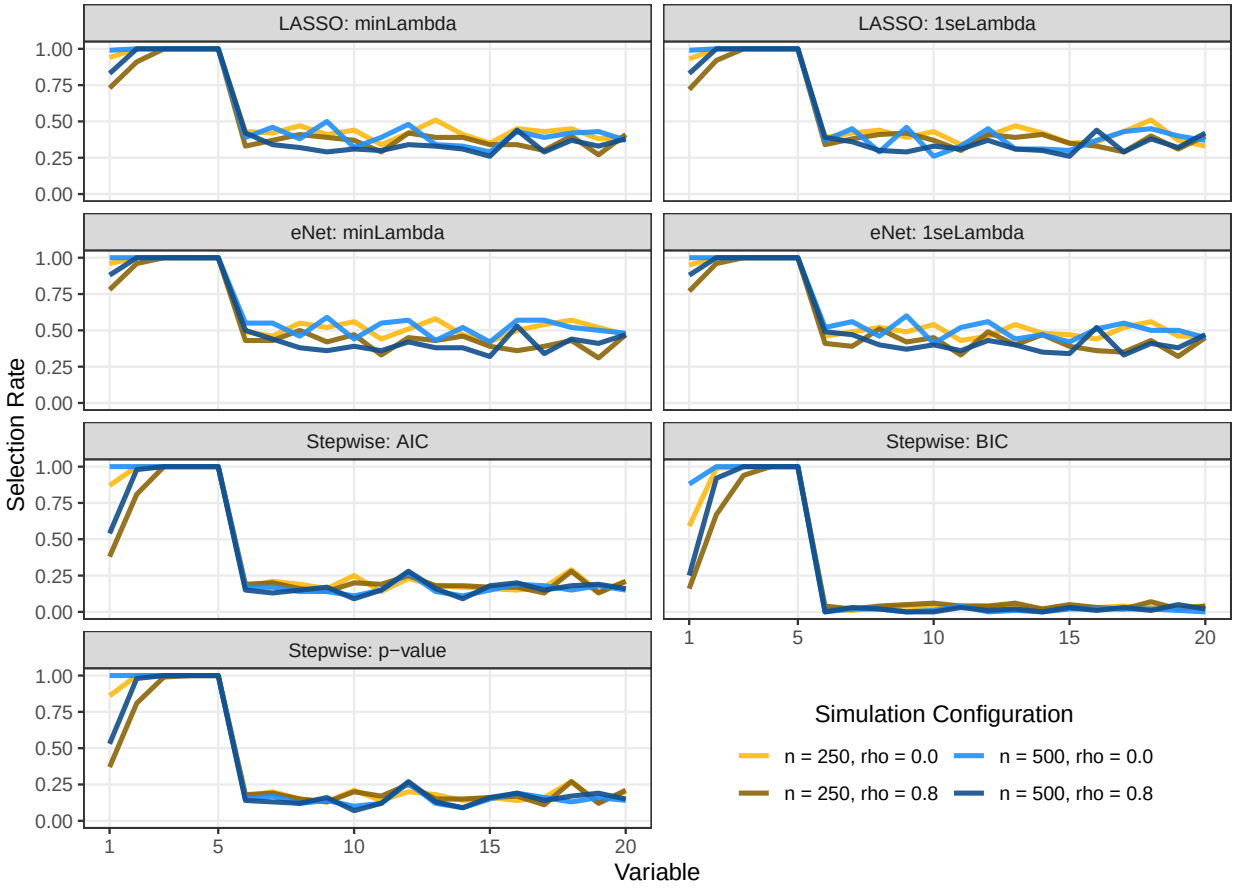

Figure 1: Selection rates for predictors in simulated data sets. Variables 1-5 are non-zero predictors while variables 6-20 are non-informative with respect to the outcome. Moderate between-predictor correlation results ($\rho = 0.4$) are not shown to improve interpretability.

The Effect of Sample Size

Evaluation metrics are shown for all variable selection methods and simulated data configurations in Figure 2. Reducing the sample size appears to have adverse effects on all performance metrics, including TPR, FPR, estimate bias, and 95% CI coverage. While all variable selection methods are adversely impacted by reduced sample size, the stepwise selection methods appear to be more sensitive to sample size than the regularization methods. That is, the effect of sample size seems to be greater when using stepwise selection.

The Effect of Multicollinearity

As shown in Figure 2, between-predictor correlation has a strong influence over variable selection performance. Interestingly, however, the effects of multicollinearity are different when comparing stepwise selection methods to regularization methods. In general, TPRs decrease (worsen) as between-predictor correlation increases. This trend is much more apparent for the stepwise selection methods, however. In fact, inducing moderate correlation ($\rho = 0.40$) yields no adverse effects when looking at TPR for the regularization methods; TPR only appears to decrease when strong between-predictor correlation is present ($\rho = 0.40$). Next, multicollinearity has no apparent effect on FPR when using stepwise selection methods. An unexpected result is observed for the regularization methods, however. Increased multicollinearity appears to lower (or improve) FPRs in the LASSO and elastic net models. Finally, increased between-predictor correlation appears to have adverse effects on estimate bias and 95% CI coverage, although this trend is observed for stepwise selection methods.

Variable Selection Method Comparison

Stepwise Selection Methods

Variable selection performance was nearly identical when using p-values and AIC as the selection criterion during stepwise selection. Stepwise selection using BIC as the selection criterion produced vastly different results, however. FPRs were markedly low when using BIC as the selection criterion for stepwise selection. Low FPRs paired with relatively diminished TPRs (Figure 1 and Figure 2) suggest that the BIC stepwise selection method was more stringent than all other variable selection methods. This method is associated with more severe estimate bias and improved 95% CI coverage when compared to the other stepwise selection methods.

Regularization Methods

Regularization methods, characterized by high TPRs and high FPRs, were less stringent than stepwise selection methods with respect to variable removal. Results show that TPRs were similar between the LASSO and elastic net methods, but that the LASSO yielded lower FPRs. No detectable differences resulted from the λ hyperparameter selection technique.

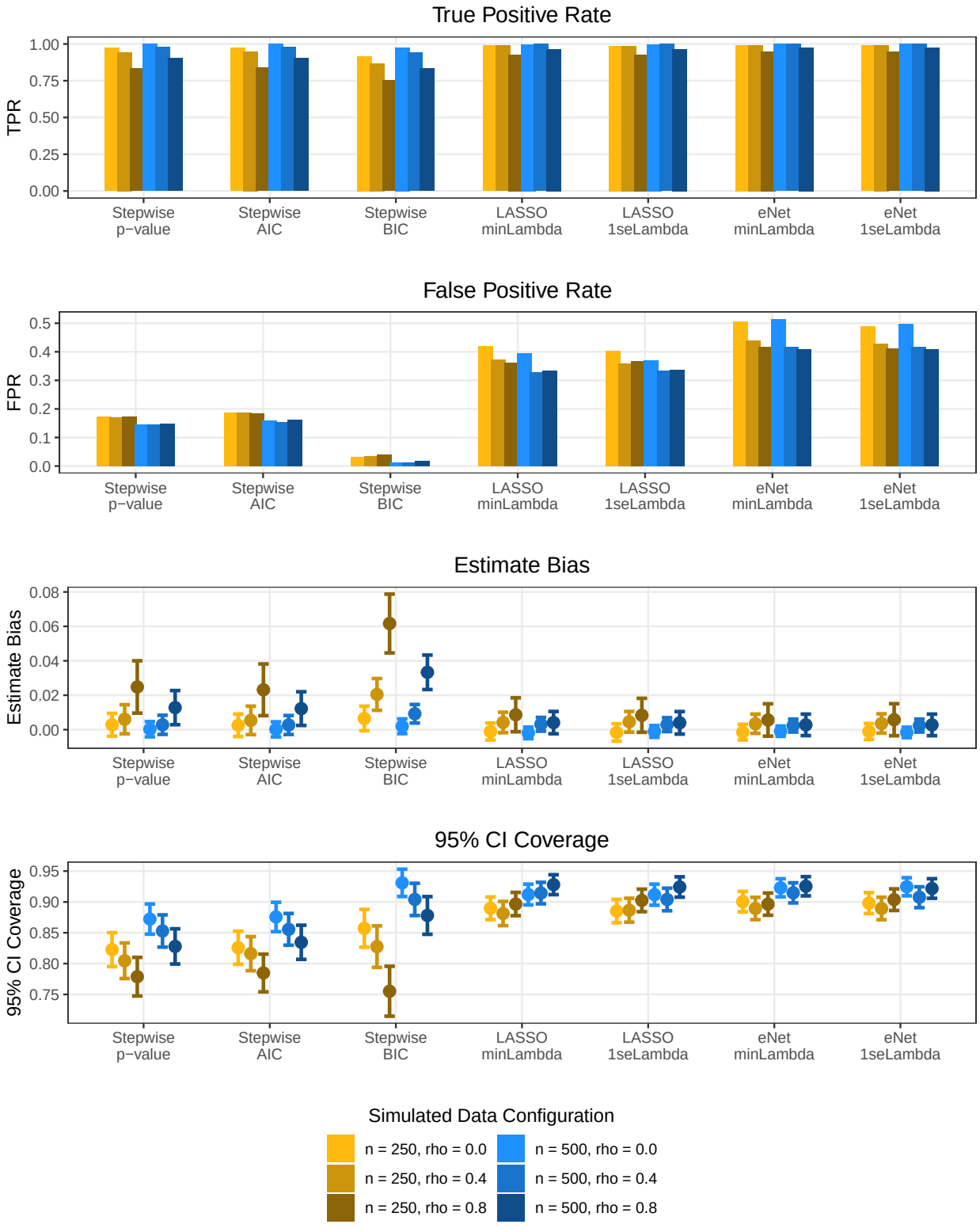

Figure 2: Model performance by variable selection method and simulated data configuration. Performance metrics include TPR, FPR, estimate bias and coverage of the 95% CI.

Discussion & Limitations

Results from this report echo an important theme of variable selection. In general, no variable selection method stands out as the clear-cut *best*. All methods are associated with some characteristics that can be considered features or flaws in specific circumstances. Nonetheless, results here can serve as a guide when deciding which variable selection method might work best in a particular use-case. For example, regularization methods should be preferred if researchers are primarily concerned with all important predictors making it into the model (optimal TPR) and are willing to deal with some non-important features being selected (worsened FPR). On the other hand, backwards selection with BIC should be preferred if researchers are aiming to fit the most parsimonious model (i.e., they are okay with some important predictors not making it into the model as long non-important predictors are rarely selected). Depending on which variable selection method is chosen, researchers can use the results here to understand limitations that are likely associated with their results. For example, in the case of using backwards selection with BIC, model point estimates are likely biased.

In light of the results here, limitations in this report are prevalent. First, computational resources were limited which caused us to only use 100 simulations per data set configuration. The reliability of our results would be improved by increasing the number of simulations (say to 1000 iterations). Second, estimate bias and 95% CI coverage estimates are certainly biased for the regularization methods. In our approach, we used regularization methods to first select the variables to keep. We then refit linear regression models using the selected variables to obtain point estimates and confidence intervals. In general, this approach is associated with drawbacks; the reason we used this approach is because confidence intervals cannot be confidently determined when using LASSO or elastic net. We should be careful when making estimate bias and coverage comparisons between the regularization methods and the stepwise selection methods.