Statistical learning vs Machine learning:

LAMOUR Samanta, GASPARIN-GRANGER Lia, AANKOUD Nora
under the supervision of Mr. Philippe DE PERETTI

February 2024

Abstract

In the dynamic landscape of data science, the choice between machine learning (ML) and statistical learning (SL) methods is of crucial importance, especially when tackling the challenges posed by big data. This research delves into the complex task of variable selection, shedding light on its inherent difficulties within the vast domain of large-scale data analysis, to construct models that not only capture the complexity of voluminous datasets but also facilitate informed decision-making while avoiding overfitting for robust generalization to new data.

In this paper, we explore the challenges related to variable selection that statisticians and data scientists may encounter in the context of big data. The primary focus is on variable selection, with an evaluation of its performance on different types of data generated through Monte Carlo simulation, under the assumption H0 of independence and the H1 assumption of linear dependence(independence, correlation or outliers)

Contents

1 Introduction

Variable selection is a crucial step in econometrics and machine learning to identify the most relevant predictors in a large dataset. Two main approaches distinguish themselves in this selection: procedures with inference (Statistical learning) and those without inference (Machine Learning).

Inference-based methods (statistical variable selection) rely on statistical significance and evaluation criteria to measure the importance of variables. For instance, stepwise procedures use tests such as the F-test, and the T-test, or criteria like AIC to selectively include or eliminate variables step by step based on their contribution to improving the model.

Conversely, inference-free methods, often associated with machine learning techniques, diverge from traditional statistical tests and utilize procedures such as LASSO, LARS, or ELASTICNET, for example. Instead of relying on statistical tests, these methods prioritize regularization approaches to reduce the coefficients of less significant variables towards zero, thereby eliminating them from the model.

In summary, variable selection is divided between methods based on classical statistical criteria and those more focused on predictive performance, without relying on these tests. This distinction between approaches with and without inference provides diverse choices for selecting optimal predictors tailored to each problem, as a statistically significant coefficient does not necessarily imply significant economic impact.

2 Litterature overview: statistical learning and machine learning methods

2.1 Statistical learning

The Statistical Learning Theory (SLT) is a framework within machine learning. Originating in the 1960s, initially focused on the theoretical understanding of how we can estimate functions from data, its major evolution occurred in the 1990s with the theoretical contributions of Vladimir Vapnik, transforming this theory into a practical tool for creating algorithms for estimating multidimensional functions using statistical inference. Its significance lies in its ability to formalize key concepts such as learning, generalization, and overfitting, providing a formal framework for designing and evaluating learning algorithms. It has also inspired new algorithmic approaches to solving function estimation problems, offering perspectives for the design of more robust and accurate models.
To achieve this, we will first explore variable selection, including Backward, Forward and Stepwise selection, which is a crucial step in the field of statistical learning. A proper variable selection primarily aims to reduce the model's dimension (adding or removing non-significant variables) and avoid overfitting. Indeed, a large number of variables can lead to overfitting, where the model fits too closely to the training data, resulting in poor generalization to new data. Thoughtful variable selection can mitigate this problem by favoring simpler and more generalizable models.
In the literature, authors have distinguished three major classes of data selection methods. In the following two sections, we will focus on the subset selection class. This approach involves choosing a specific group of predictors believed to be associated with the expected response. Several methods fall within this class, such as the best subset selection method and the stepwise selection method (both forward and backward).

2.1.1 Stepwise selection

The stepwise selection method aims to create a predictive model by iteratively choosing explanatory variables, and combining forward and backward selection techniques. Each step of the process involves both forward selection, which tests the variables (using the F-statistic) to assess their significance in the model and determine whether they should be added or not, and backward selection, which tests the variables present in the model to determine their significance; if not significant, they are removed. The stepwise approach is not rigid, allowing the addition and removal of variables during the process (possibility to reconsider decisions). The process stops when it is no longer possible to add or remove variables. Therefore, stepwise selection combines two selection methods: forward and backward selection.

2.1.2 Forward Stepwise selection

The forward procedure contrasts with the backward method. This approach begins with a null model (M0) with no explanatory variables, meaning it initializes by incorporating only a constant:

$$Y = \beta_0$$

Then, variables are sequentially added based on their ability to improve the model. This ability is determined by conducting tests of statistical significance. The process of the forward method typically unfolds as follows:

1. Null Model: Start with a model that contains no explanatory variables.
2. Addition Criteria: Use a criterion to assess the impact of adding each variable. This criterion may be based on measures such as improvement in information criteria (such as AIC or BIC) or threshold of the p-value of a T-stat or an F-test.
3. Addition of the Best Variable: Add the variable that most significantly improves the defined criterion, choosing the one with the lowest p-value.
4. Iteration: Repeat steps 2 and 3 until a certain stopping criterion is met. In other words, continue until the model no longer improves with the addition of new variables.

2.1.3 Backward Stepwise selection

In contrast to forward stepwise selection, backward stepwise selection begins with a full model containing all potential explanatory variables. The process then iteratively removes variables based on their significance. This method aims to simplify the model by eliminating variables that do not contribute significantly to explaining the response variable.
The process of backward stepwise selection typically unfolds as follows:

1. Full Model: Start with a model that includes all potential explanatory variables.
2. Removal Criteria: Use a criterion to assess the impact of removing each variable. Similar to the addition criteria in forward stepwise selection, the removal criterion may be based on measures such as a decrease in information criteria (AIC or BIC) or exceeding a threshold for the p-value of a T-stat or an F-test.
3. Removal of the Least Significant Variable: Remove the variable that least significantly contributes to the model, choosing the one with the highest p-value.
4. Iteration: Repeat steps 2 and 3 until a certain stopping criterion is met. The process continues until the model no longer improves with the removal of variables.

Both forward and backward stepwise selection methods provide a balance between including relevant variables and avoiding overfitting by iteratively adjusting the model's complexity. The decision to add or remove variables is guided by statistical criteria, ensuring that the resulting model is both interpretable and generalizable to new data.

2.2 Machine learning

Machine learning is a scientific field, particularly a subfield of Artificial Intelligence, that enables computers to learn without being explicitly programmed. It focuses on creating systems that learn or improve their performance based on the data they process and involves training algorithms from a learning dataset to enable them to make predictions or automate tasks.
Machine learning is commonly applied in various fields such as image recognition, trend prediction, fraud detection, and many others. Therefore, there are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning.
In machine learning, feature or variable selection is the process of choosing a subset of relevant variables to use in model construction. It is a crucial step that enhances the model's accuracy by eliminating irrelevant variables. There are several variable selection methods, but we will focus on Forward Stagewise, Elasticnet, LARS, and LASSO.

2.2.1 Forward Stagewise

The Forward Stagewise method aims to select the most relevant variables for building a machine learning model. This technique is a variant of forward stepwise selection, as instead of gradually adding the most relevant variables until the stopping criterion is met, Forward Stagewise selection incrementally adds variables, adjusting the coefficients of existing variables at each iteration. In other words, we have a set of potential explanatory variables X1, ..., Xp, and we seek the explanatory variables from this set that best explain the target variable y.
The coefficients' weights are initialized to zero, and the algorithm examines which coefficient is most correlated with the residual r. Once identified, it updates the coefficient by adding an increment j and includes it (along with the associated variable) in the set of variables used for regression. The residuals are then also updated, and the process iterates until no variable in the set is correlated with the residual r.

Algorithm : Incremental Forward Stagewise Regression

1. Start with $r = y - \bar{y}, \beta_1, \beta_2, ..., \beta_p = 0$
2. Find the predictor x_j most correlated with r
3. Update $\beta_j \leftarrow \beta_j + \delta_j$ where $\delta_j = \epsilon \cdot \text{sign}\left[\text{corr}(r, x_j)\right]$
4. Update $r \leftarrow r - \epsilon x_j$
5. Repeat steps 2 and 3 until no predictor has any correlation with r

2.2.2 Least Angle Regression (LAR)

Introduced in 2004 by Efron et al., the Least Angle Regression (LAR) algorithm is a tool used in regression for high-dimensional data (data with numerous attributes) to select a reduced set of important variables. It is a "Forward Stagewise" method as it adds and removes variables from the model based on their correlation with the current residuals. However, LAR

still possesses qualities that the classical Forward Stagewise method does not have, such as the fact that the direction of movement ensures the continuity of the common correlation between the two predictors most correlated with the residual until a third predictor has the same correlation with the residual as the first two already integrated into the model.
The LAR method is also effective when the number of predictors is high compared to the data size, especially when there are correlations between these predictors. This makes the algorithm particularly well-suited for handling datasets with numerous predictive variables.

Algorithm : Least Angle Regression

1. Standardize the predictors to have a mean zero and unit norm. Start with $r = y - \bar{y}, \beta_1, \beta_2, ..., \beta_p = 0$.
2. Find the predictor x_j most correlated with r.
3. Move β_j from 0 towards its least-squares coefficient $\langle x_j, r \rangle$, until some other competitor x_k has as much correlation with the current residual as does x_j.
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on (x_j , x_k), until some other competitor x_l has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. After min(N - 1, p) steps, we arrive at the full least-squares solution.

2.2.3 Least Absolute Shrinkage and Selection Operator

The Least Absolute Shrinkage and Selection Operator (LASSO), introduced in 1996 by Robert Tibshirani, is a method designed to address the limitations of linear regression in a high-dimensional context. To mitigate the instability issues of linear regression predictions, regularization techniques are employed. The need for this method arises from overfitting or underfitting problems in the data. Using Ordinary Least Squares (OLS) when dealing with numerous regressors (i.e., in high-dimensional models) poses a significant risk of overfitting. Reducing the variance of estimates through Lasso regression estimators can help avoid such problems.
Unlike classical linear regression, the goal of this regularized regression is to eliminate unnecessary variables and retain only relevant ones. Lasso regression aims to balance the bias-variance trade-off by reducing some coefficients to zero, making it a predictor selection optimizer.
The Lasso method involves estimating the parameter vector by minimizing the quadratic least squares criterion while adding a penalty on the sum of absolute values of coefficients or, equivalently, by minimizing the quadratic criterion penalized by the L1 norm of the coefficients. The introduction of a penalty reduces estimation variability, thereby improving prediction accuracy. The optimization problem is the following :

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad \text{where} \quad \sum_{j=1}^{p} |\beta_j| \leq t$$

where $\lambda \geq 0$ is the parameter that controls the strength of the penalty, the larger the value of, the greater the amount of shrinkage.

The choice of the penalty parameter is crucial. Indeed, when is equal to zero, we obtain the Ordinary Least Squares (OLS) estimation. Conversely, for very high values, all Lasso estimates are zero.
This choice involves a trade-off between model fitting and its generalization capacity. The model experiences underfitting when high bias and low variance occur, and overfitting when there is low bias and high variance. We need to find a balance between bias and variance to achieve the perfect combination and minimize errors in our predictions. To choose this parameter , various methods can be used, but cross-validation is often employed.

2.2.4 Ridge Regression and Elastic net

Similar to LASSO regression, Ridge regression is also a regularized version of the least squares method for linear regression, but in the case of Ridge, L2 regularization is used. Unlike LASSO regression, L2 regularization reduces coefficients but does not allow setting coefficients to zero. Indeed, Ridge regression is constructed by adding to the sum of squared residuals to minimize, not the sum of absolute values of coefficients, but rather the sum of the coefficients squared.

The elastic net is a regularized regression method that combines both L1 penalty (used in LASSO regression) and L2 penalty (used in Ridge regression). It is a technique aimed at overcoming certain limitations of individual methods, particularly in the presence of high correlations between explanatory variables.
The elastic net introduces two penalty parameters, typically denoted as and , where controls the mix between L1 and L2 regularization terms. Specifically, when is equal to zero, the elastic net is equivalent to Ridge regression, and when is equal to one, it is equivalent to LASSO regression. So, the elastic net provides flexibility by allowing one to benefit from the variable selection advantages of LASSO regression while preserving the stabilization properties of Ridge regression. This method can be particularly useful in situations where there are groups of highly correlated variables, as it has the ability to select entire groups of variables.

2.3 Criteria

In this section, we present the criteria used throughout our study, grouped into two distinct categories: the stopping criterion and the choose criterion. The stopping criterion specifies the conditions to end the variable selection process, such as AIC or SBC (BIC). Finally, the selection criterion determines the optimal model at each step of the selection based on measures like AIC or SBC (BIC). Each of these categories plays a crucial role in constructing a robust and data-adaptive model, ensuring to avoid overfitting while promoting generalization. Our analysis will focus on comparing the model performances based on different

combinations of these criteria.

2.3.1 Statistical Learning criteria

AIC and AICc The Akaike Information Criterion (AIC) is a measure of the quality of a model proposed by Hirotugu Akaike in 1973. The AIC is a method used to choose the best statistical model by striking a balance between the goodness of fit and the model's complexity. The fundamental idea is that adding parameters to a model can improve its fit to the data, but it could also make it too complex (overfitting). AIC penalizes models with a large number of parameters. The model with the lowest Akaike Information Criterion is then chosen. The Akaike information criterion is defined as :

$$AIC = -2\log(L) + 2k$$

where
n : number of observations
k : the number of parameters to be estimated in the model
L : the likelihood function of the model
SSE : Residuals Sum of Squares

The AICc is a correction of the AIC. It is recommended to use it when the number of parameters is large compared to the number of observations. It penalizes additional variables more than the AIC. As the sample size increases, AICc and AIC converge. AICc is defined by :

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

where
n: the number of observations k : the number of parameters to be estimated in the model

BIC and SBC The Bayesian Information Criterion, BIC, is also used for model comparison. BIC imposes a stronger penalty than AIC for the number of parameters, favoring simpler models. While AIC was introduced to retain relevant variables during forecasting, BIC aims at selecting statistically significant variables in the model. It is defined as follows:

$$BIC = -2log(L) + klog(n)$$

where
k : the number of parameters to be estimated in the model
L : the likelihood function of the model
n : the number of observations

Moreover, it should be noted that the BIC is called SBC by the SAS language. There is also a criterion named BIC by SAS but which corresponds to another criterion which is calculated in the following way :

$$BIC(SAS) = nlog(\frac{SSE}{n}) + 2(p+2)q - 2q^2$$
$$\text{where } q = \frac{n\hat{\sigma}^2}{SSE}$$

where
n = the number of observations
p = the number of parameters including the intercept $\hat{\sigma}^2$ = Estimate of pure error variance from fitting the full model SSE = Error sum of squares

2.3.2 Machine Learning criteria

Cross-validation The general idea of cross-validation is to assess the performance of a predictive model in generalizing to out-of-sample data by retaining a portion of the training dataset for testing the model. There are different cross-validation (CV) methods, such as K-folds, which involves dividing the dataset into K subsets, using K-1 subsets for training and the remaining one for validation. This approach maximizes data utilization and reduces the variance associated with a simple train/test split. The choice of k often depends on the dataset size, with common values like 5 or 10. By averaging performance over the k iterations, k-fold cross-validation provides a more reliable overall estimation of the model's performance, making it a valuable technique for model comparison or hyperparameter selection, such as the in Lasso and Ridge regressions.

Leave-one-out cross-validation Another cross-validation method is leave-one-out cross-validation (LOOCV), a technique used almost exclusively in small-sized datasets due to its computational cost. With LOOCV, for each data point in the training set, the model is trained on all points except one and then tested on the excluded point. This process is repeated for each unique data point, hence the term "leave-one-out." This method is also known as "PRESS."

2.3.3 Criteria Overview

The criteria play a crucial role in model construction, stopping the selection process, and choosing the most appropriate model. Below, we summarize the key criteria, including those from statistical learning methods (Section 2.3.1) and machine learning techniques (Section 2.3.2).

Criterion	Stop=criterion	Choose=criterion
AIC	The variable selection process stops when adding an additional variable no longer minimizes the AIC of the model.	The chosen model is the one with the smallest AIC.
AICc	The variable selection process stops when adding an additional variable no longer minimizes the AICc of the model.	The chosen model is the one with the smallest AICc.
BIC	The variable selection process stops when the model with the smallest BIC is reached.	The chosen model is the one with the smallest BIC.
CV	The variable selection process stops when the model achieves optimal performance in terms of cross-validation. It stops when the predictive performance of the model on data not used for training is maximized.	The chosen model is the one with the best average performance when evaluated on distinct subsets of the training data.
PRESS	The variable selection process stops when the model has the smallest PRESS.	The chosen model is the one with the smallest predicted residual sum of squares.

Table 1: Criteria Summary

3 Selection method performance comparison

3.1 Methodology

The objective of this project is to assess the performance of various variable selection methods when applying the "GLM SELECT" procedure. To achieve this, we created four empty datasets, each specific to a data type, in which three variables "Methods, Criteria, Results" were defined with well-defined attributes. These datasets will be subsequently used to store the results of the simulations.

```
DATAout.PERFORMANCE_GDP1 out.PERFORMANCE_GDP2 out.PERFORMANCE_GDP3
out.PERFORMANCE_GDP4;
ATTRIB
    METHODE length=$15 format=$15. label="Selection Method"
    CRITERE length=$15 format=$15. label="Stop Criterion"
    RESULTAT length=$15 format=$15. label="Result";
STOP;
RUN;
```

Each table will consist of [(6*6)-1]*1000= 35000 observations resulting from different combinations of possible criteria and methods iterated 1000 times (excluding the ELASTICNET and PRESS combination, which is not feasible). The criteria and methods used are: AIC, AICc, BIC, SBC, CV, PRESS; Forward, Backward, Stepwise, LASSO, LAR, ELASTICNET.

We subsequently define the number of observations (n=250) and potential variables (p=50), including the true variables that we have specified in the "true" model:

$$Y = \text{Intercept} + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

We will use the GLMSELECT procedure to select the model for each method, selection criterion among different models, and stopping criterion within a given model. These criteria are respectively defined by the SAS options CHOOSE= and STOP=. This procedure is iterated m = 1000 times.

The different metrics that we will calculate are:

1. Perfect fitting: the probability that the model selects exactly the true variables.
2. Overfitting: the probability that the model selects the variables pre-defined as true and additional variables (not true).
3. Underfitting: the probability that the model does not select all pre-defined true variables.
4. Fail: the probability that the model does not select all pre-defined true variables and adds additional variables (not true).

To properly compare the performance of the selection methods, we assume that the data is always comparable for a given structure and number of simulations (sufficiently large number of observations and iterations).

Data types

We will use four different types of data to observe the performance of our variable selection methods under two assumptions: the null hypothesis of data independence and the alternative hypothesis of linear data dependence.

- **Gaussian and independent data** :These data are standardized and follow a multivariate normal distribution as they are generated under the null hypothesis.
- **Gaussian and correlated data** : Here, correlated data is generated under the alternative hypothesis. The generation process is as follows: We select a subset of 5 variables and include a predefined correlation among them using the Toeplitz form. This form has two advantages, the first being that it allows us to incorporate correlation into our variables, and the second is that it ensures the correlation matrix is semi-positive definite. The other 45 variables are independent and identically distributed in a Gaussian manner. This table effectively shows the correlation among the chosen 5 variables.

Coefficients de corrélation de Pearson, N = 250 Proba > |r| sous H0: Rho=0					
	X1	X2	X3	X4	X5
X1	1.00000	0.82205 <.0001	0.73655 <.0001	0.48749 <.0001	0.39733 <.0001
X2	0.82205 <.0001	1.00000	0.82294 <.0001	0.73695 <.0001	0.45094 <.0001
X3	0.73655 <.0001	0.82294 <.0001	1.00000	0.82448 <.0001	0.71032 <.0001
X4	0.48749 <.0001	0.73695 <.0001	0.82448 <.0001	1.00000	0.79864 <.0001
X5	0.39733 <.0001	0.45094 <.0001	0.71032 <.0001	0.79864 <.0001	1.00000

Figure 1: Correlation between true model variables

- **Independent data with extreme values** : To generate these data, we first generate independent data and then add random outliers to all variables in the model. In 90% of the cases, it randomly generates data following a uniform distribution, and in the remaining 10%, it generates outliers. These extreme values correspond to normal data to which we add 5. To get an overview of the distributions of the extreme values, we displayed the following histogram.

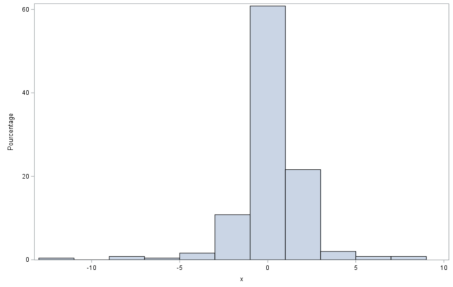

Figure 2: Histogram of outlier distribution

- **Correlated data with extreme values**: In this section, dependent data with outliers are generated. The dataset consists of 250 observations and 50 explanatory variables, including the constant. These variables are generated from a multivariate normal distribution and are then subjected to a multivariate transformation using the Iman-ConoverTransform module. This transformation aims to introduce some correlation between the explanatory variables. The procedure for adding measurement errors is integrated through the creation of the dependent variable Y. This is obtained by linearly combining the explanatory variables with predefined coefficients and adding a normally distributed error component (EPS). Thus, each observation of the dependent variable is influenced by the explanatory variables as well as random measurement errors. The procedure for adding extreme values (outliers) is carried out randomly within a loop. For each observation, a random test is performed, and if the condition is satisfied ($u < 0.9$), the values of the explanatory variables remain unchanged. However, if the condition is not satisfied, outlier values are added to each explanatory variable.

Once the nature of the data is defined, we will be able to proceed with the interpretation of the performances obtained for each method, based on the criteria and types of data, through the histograms.

3.2 Comparison of model selection performances based on identical stopping and choosing criteria.

In this first part, we have decided to choose a single predefined criterion for both the stopping and selection criteria. Therefore, we will have (STOP=criterion CHOOSE=criterion)

3.2.1 Normal and independent data

SBC : Statistical Learning VS Machine Learning

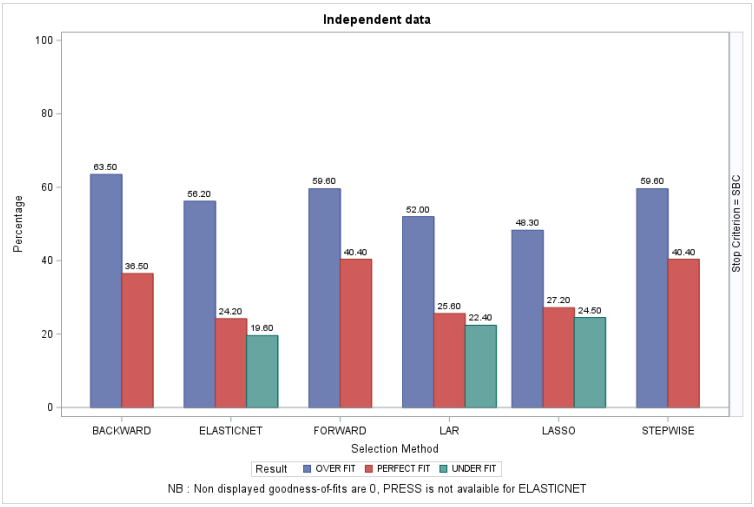

Figure 3: Variable selection performance with STOP= SBC and CHOOSE= SBC

In the case where SBC (Schwarz Bayesian Criterion) is applied as a stopping criterion and as a selection criterion, we find that the statistical learning models (Backward, Forward, and Stepwise) perform similarly. Indeed, the probability of overfitting with Forward and Stepwise is 59.60% and that of perfect fitting is 40.40%, putting them in the first place. They are closely followed by the Backward selection method, which has a 36.50% probability of selecting the same variables as the true model. As for the machine learning models, their performance is just as similar. However, their chance of obtaining the true model is lower than that of the statistical learning models, with 27.20% for LASSO, 25.60% for LARS, and 24.20% for ELASTICNET. In contrast to the statistical learning methods, underfitting is observed for each machine learning model, with a probability of not identifying all the true variables of 24.50% for LASSO, 22.40% for LARS, and 19.60% for ELASTICNET. Thus, in the context of normal, independent data, the best-performing models with the SBC stopping and selection criterion are Stepwise and Forward.

BIC : Statistical Learning VS Machine Learning

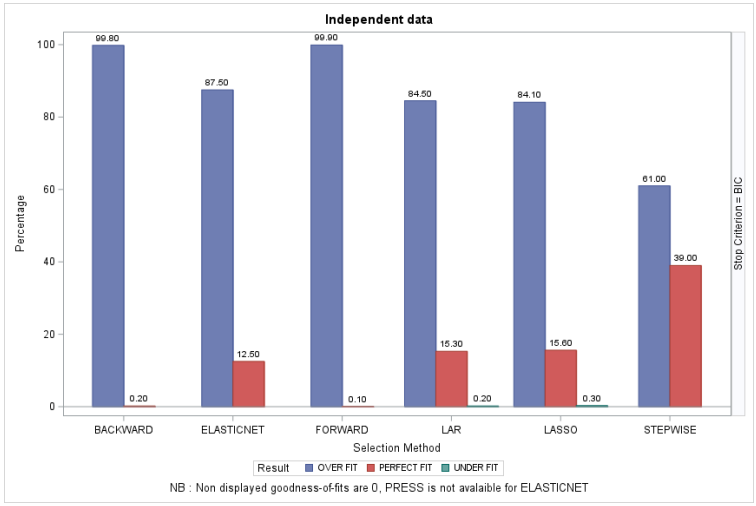

Figure 4: Variable selection performance with STOP= BIC and CHOOSE= BIC

Compared with the SBC criterion, the use of BIC led to a significant drop in the probability of obtaining the true model, whatever the method, and increased the probability of over-fitting significantly. This translates into figures of up to 99.80% for Backward, 99.90% for Forward, 84.50% for LARS, 84.10% for LASSO, and 61% for Stepwise. It's important to note that the Stepwise method is the only one whose performance does not vary significantly, and it's also the only one with such a high probability of obtaining the true model, making it the best performer according to the BIC criterion.

AIC and AICc : Statistical Learning VS Machine Learning

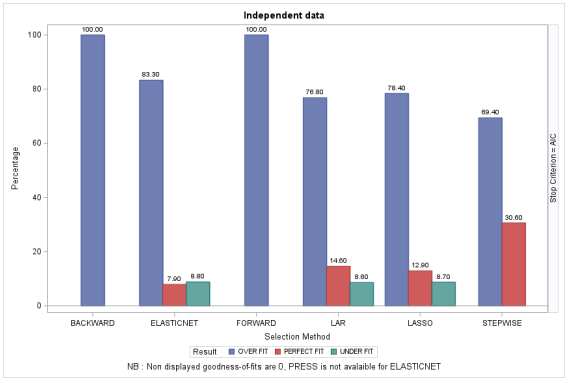

Figure 5: Variable selection performance with STOP= AIC and CHOOSE= AIC

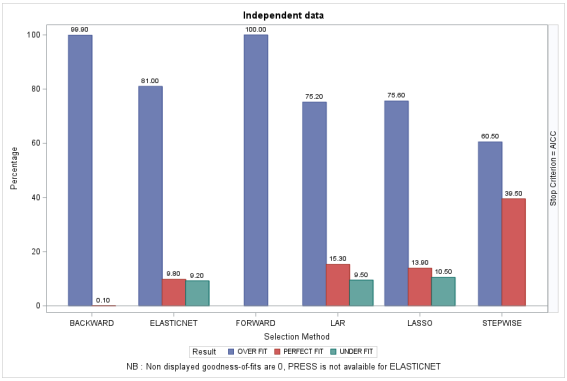

Figure 6: Variable selection performance with STOP= AICc and CHOOSE= AICc

The AICc criterion, being a correction of the AIC criterion designed for smaller samples, was compared simultaneously with the latter, as well as with other criteria. In the end, the performances were virtually identical. Overfitting remains predominant regardless of the model. The results obtained with LARS and LASSO show a similarity of around 2%,

whether with AIC or AICc. In light of these observations, we can conclude that the AIC and AICc criteria tend to improve the Stepwise selection method, with respective probabilities of obtaining exactly the right model of 30.60% and 39.50%.

CV : Statistical Learning VS Machine Learning

Figure 7: Variable selection performance with STOP= CV and CHOOSE= CV

In this context, the stopping and choosing criterion used is the CV, which means that the selection process ends when the CV reaches a minimum, thus characterizing an empirical approach and where the CVPRESS score is calculated at each stage, and the model selected is the one with the lowest CVPRESS score at the first stage. Similar to the AIC and AICc criteria, the Backward model has a 100% probability of selecting more variables than necessary in its model. However, as with AIC/AICc, the CV criterion shows a preference for the STEPWISE method, with a 38.90% probability of obtaining exactly the right model. This indicates that, according to the CV criterion, the STEPWISE method is more likely to produce a model that faithfully represents the data without including excessive variables.

PRESS : Statistical Learning VS Machine Learning

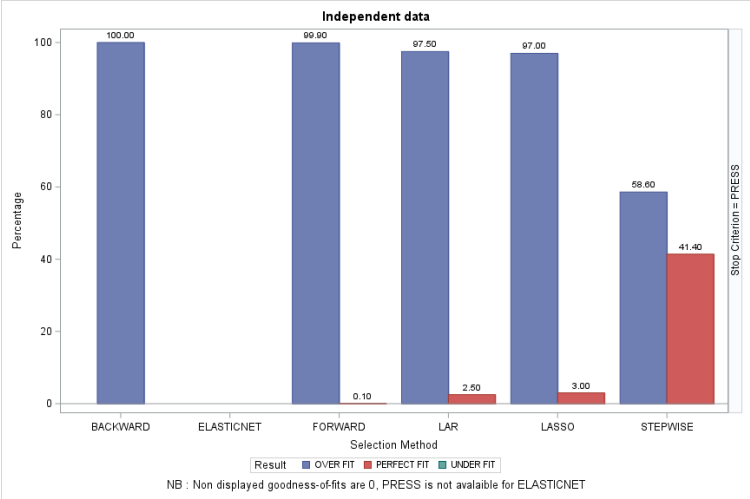

Figure 8: Variable selection performance with STOP= PRESS and CHOOSE= PRESS

In this scenario, each observation is used successively as a test set, making this criterion stricter than the standard K-fold CV. This rigor explains the decrease in the percentage chance of selecting the correct model from all possibilities. Except for LAR and LASSO, which respectively have a 2.50% and 3% probability of obtaining the true model, the Stepwise method remains the best performing of all the methods evaluated, even in the context of the PRESS criterion. The probability of obtaining the true model with the Stepwise method reaches 41.40%, indicating superior performance over the other methods in this particular configuration of the stopping criterion.

3.2.2 Independent data with outliers

SBC : Statistical Learning VS Machine Learning

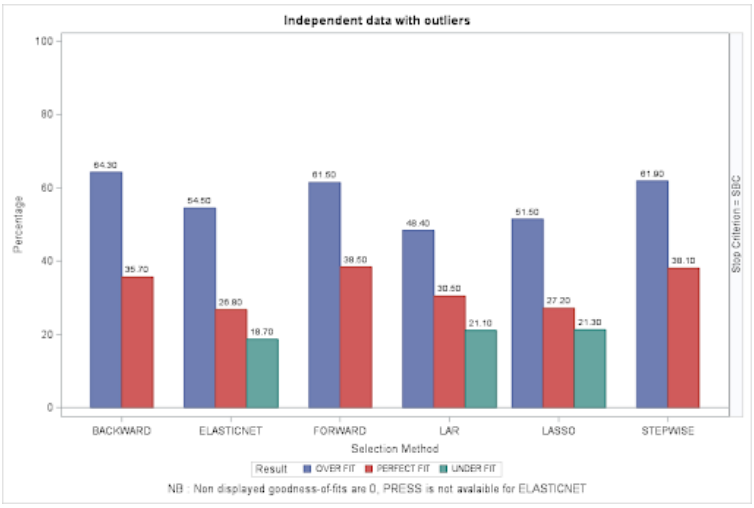

Figure 9: Variable selection performance with STOP= SBC and CHOOSE= SBC

We observe that compared to independent normal data, machine learning models here have no chance of encountering models with failure. Overall, the "best model" with independent data containing outliers and the SBC criterion is Backward (identically to the case of independent normal data). Conversely, the model with the lowest percentage chance of obtaining the true model is ELASTICNET with 26.80% overfitting. The percentages of overfitting remain nearly the same as without outliers, and machine learning models remain the only ones capable of finding models that do not include all the true variables.

BIC : Statistical Learning VS Machine Learning

Figure 10: Variable selection performance with STOP= BIC and CHOOSE= BIC

It can be observed that with the BIC criterion, the selection performances do not improve. Compared to SBC, BIC increases overfitting for all models except for Stepwise, which decreases from 61.90% chance of overfitting to 59.40%. The statistical learning models Backward and Forward reach up to 99% chance of obtaining models with additional variables.

The best method when SBC is used as both the stopping and selection criterion is therefore Stepwise, which has the highest percentage chance of obtaining the true model (40.60% perfect fitting).

AIC and AICc : Statistical Learning VS Machine Learning

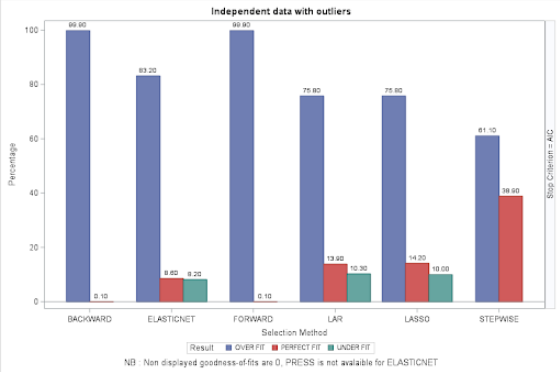

Figure 11: Variable selection performance with STOP= AIC and CHOOSE= AIC

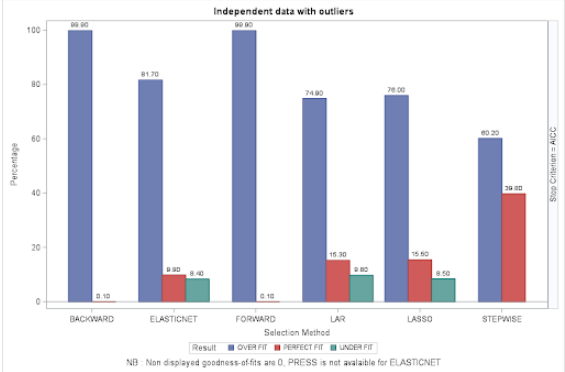

Figure 12: Variable selection performance with STOP= AICc and CHOOSE= AICc

When comparing the performance associated with statistical learning models with that of normal and independent data for the same criterion, it is observed that they are similar within a few percentage points. The Backward method shows identical performance to that with normal and independent data, while the percentage chance of overfitting with the Forward method in normal independent data only reduces very slightly (from 100% to 99.90%). For machine learning models, unlike in the case of normal independent data, here there is no chance of model failure (i.e., not finding all the true variables and adding extra variables). When considering independent data with outliers, it is unsurprising that model performances are similar within a few percentage points, as the AICc is a correction of the AIC for small samples, and here the sample size is large compared to the number of parameters, so the AIC and AICc tend to yield similar results.

CV : Statistical Learning VS Machine Learning

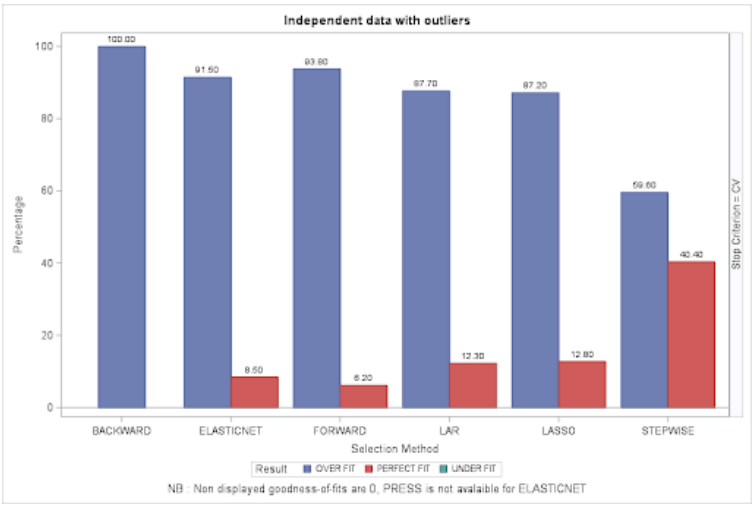

Figure 13: Variable selection performance with STOP= CV and CHOOSE= CV

When the CV criterion is used as both the stopping and selection criterion, the method giving the highest chance of finding the correct model is the Stepwise method (identically to the case of normal and independent data) with a 40.40% chance of finding the correct model. The Backward method, on the other hand, seems to be the least suitable for this criterion as it has a 100% chance of not selecting all the true variables (X1 to X5). No method appears to be prone to overfitting or underfitting situations with the addition of extra non-true variables (failure). In numerical terms, the machine learning methods LARS, LASSO, and ELASTICNET have respective chances of finding the correct model of 12.30%, 12.80%, and 8.50%, compared to 6.20% for Forward.

PRESS : Statistical Learning VS Machine Learning

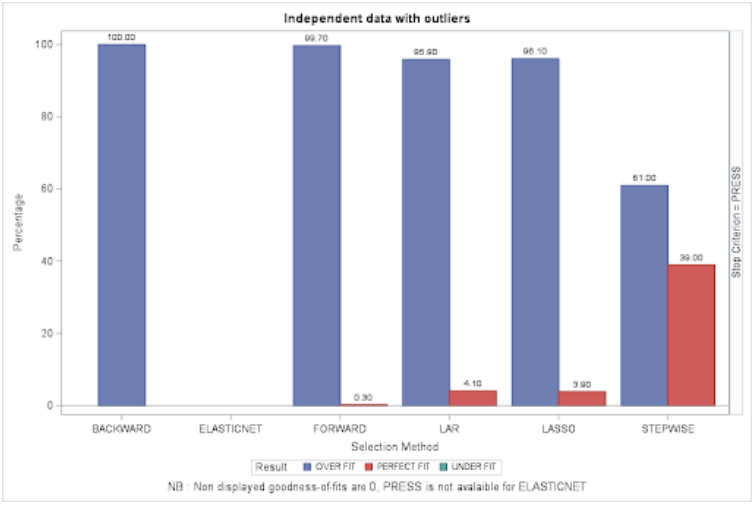

Figure 14: Variable selection performance with STOP= PRESS and CHOOSE= PRESS

As seen in the first part, the PRESS criterion is not applicable to the ELASTICNET machine learning method. With this criterion as the stopping and selection criterion, the chances of obtaining the correct model are reduced for all possible methods compared to using CV. This phenomenon is mainly due to the restrictive nature of the PRESS criterion, which is an extreme case of the CV criterion. For Forward, the perfect fitting performance has decreased from 6.20% (with the CV criterion) to 0.30% (with PRESS), for LAR it decreased from 12.30% with CV to 4.10% with PRESS, from 12.80% to 3.90% for LASSO, and from 40.40% to 39% for Stepwise. The percentages associated with perfect fitting lost when using PRESS have been added to the chance of obtaining a model with additional variables (overfitting), which has increased for all methods except for Backward.

3.2.3 Correlated Data

SBC : Statistical Learning VS Machine Learning

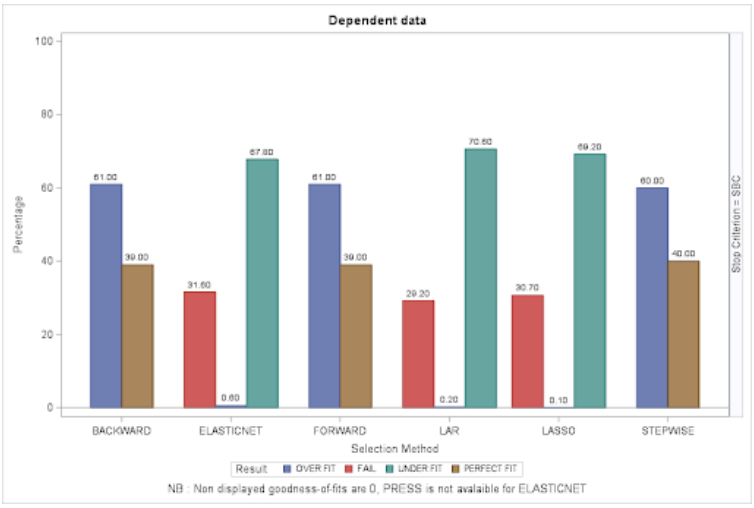

Figure 15: Variable selection performance with STOP= SBC and CHOOSE= SBC

In this subpart, we consider a scenario with correlated data and observe the performance of each statistical and machine learning method when the SBC criterion is used as both the stopping and selection criterion. The presence of correlation significantly alters the performance of each method in selecting the correct variables.
When comparing these performances to those associated with using SBC with normal and independent data, we notice that in the case of correlated data, machine learning methods no longer have any chance of obtaining the correct model.
For instance, to provide some numbers, the ELASTICNET model saw its percentage chance of underfitting increase from 16% with normal independent data to 67.80% with correlated data. As for statistical learning methods, we observe very little change in performance between normal independent data and correlated data.

BIC: Statistical Learning VS Machine Learning

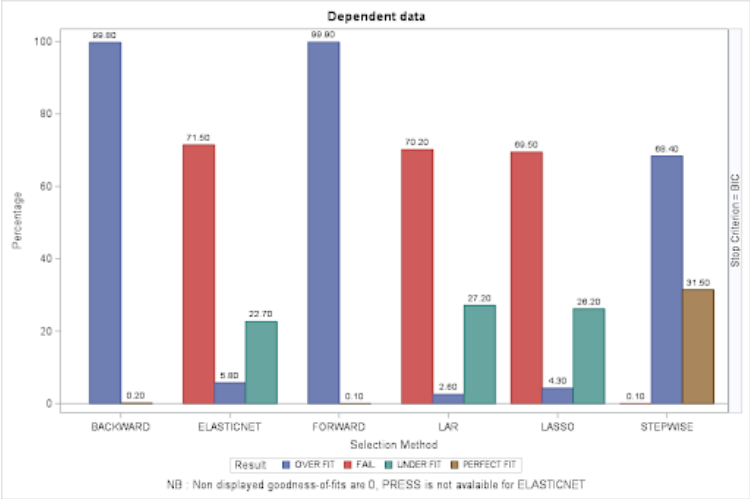

Figure 16: Variable selection performance with STOP= BIC and CHOOSE= BIC

Similarly to the SBC criterion, the use of the BIC criterion as both the stopping and selection criterion, in the context of correlated data, significantly reduces the probability of obtaining the correct model for the various existing methods.
Concerning machine learning methods, the probability that the model does not contain all the true variables and adds additional variables (failure) is quite high for all machine learning methods, with 71.50% for ELASTICNET, 70.20% for LARS, and 69.50% for LASSO.
As for statistical learning methods, overfitting occurs with a high probability, with 99.80%, 99.90%, and 68.40% chance, respectively, that models generated with the Backward, Forward, and Stepwise methods do not include all the true variables.

AIC and AICc: Statistical Learning VS Machine Learning

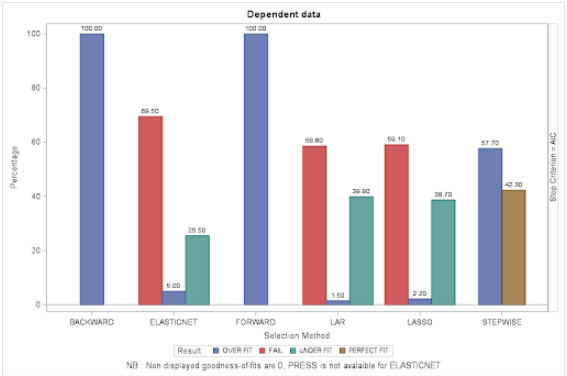

Figure 17: Variable selection performance with STOP= AIC and CHOOSE= AIC

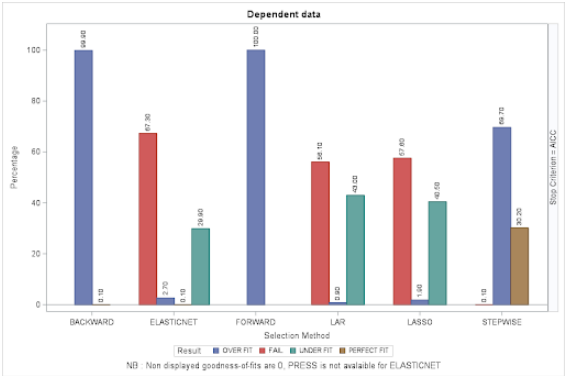

Figure 18: Variable selection performance with STOP= AICc and CHOOSE= AICc

Identically to the case of normal and independent data, in this section, we notice that with AIC as both the stopping and selection criterion, the percentage chance that the Backward

and Forward methods find models that do not include all the true variables (underfitting) is almost 100%. The performance of the Stepwise method is better than with normal and independent data, increasing from 38.30% chance of obtaining a perfect fit to 42.30% with correlated data.
As for machine learning methods, we observe that the probability of underfitting with additional variables and simple underfitting is extremely high. The most optimal method to obtain the true model with AIC will be Stepwise with a 42.30% chance of finding the correct model. The same holds true for the AICc criterion, with performance trends remaining generally similar for all methods except Stepwise, whose probability of underfitting increases and that of perfect fitting decreases (due to the more restrictive nature of the AICc criterion).

CV: Statistical Learning VS Machine Learning

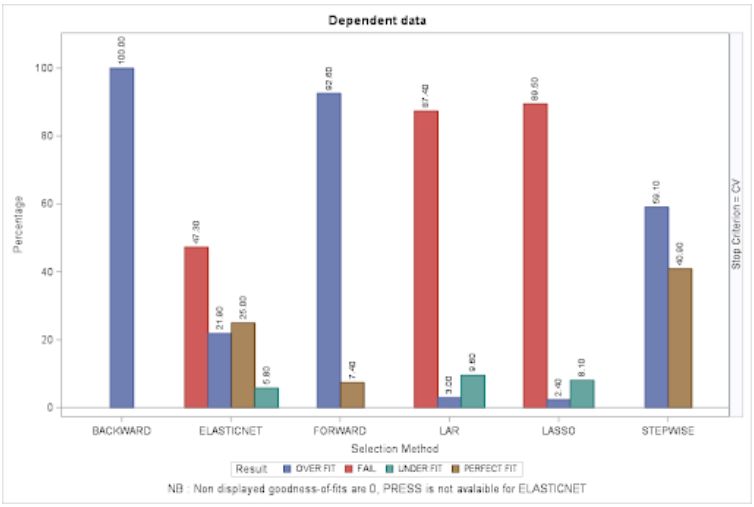

Figure 19: Variable selection performance with STOP= CV and CHOOSE= CV

When CV is used as the STOP and CHOOSE criterion, two main distinctions are identified. Statistical learning methods are more prone to overfitting, with percentages of chance to obtain a model with additional variables of 100% for Backward, 92.60% for Forward, and 59.10% for Stepwise. In contrast, machine learning methods are more likely to obtain models of underfitting with additional variables (failure): 47.30% failure for ELASTICNET, 87.40% for LARS, and 89.50% for LASSO.In this configuration, the method most likely to find the correct model is Stepwise, with 40.90% of perfect fit models found.

PRESS: Statistical Learning VS Machine Learning

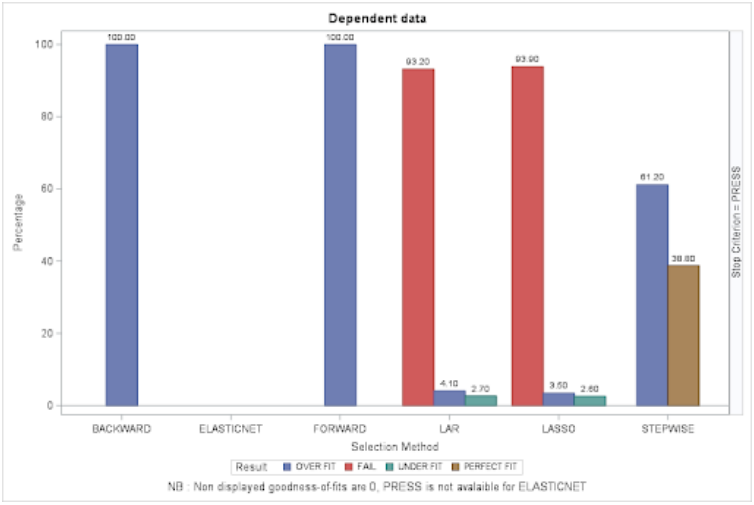

Figure 20: Variable selection performance with STOP= PRESS and CHOOSE= PRESS

Similarly to normal and independent data, overfitting is very prevalent in statistical learning models with 100 % chance for both Backward and Forward, and 61.20% for Stepwise. However, with correlated data, machine learning models no longer tend to find models with additional variables (overfit), but rather models with missing true variables and with additional variables (failure). Overfitting is still present but to a lesser extent, with only 4.10% chance for LARS and 3.50% for LASSO. The best model in this case is, once again, Stepwise with a 38.80% chance of finding the correct model.

3.2.4 Correlated Data with outliers

SBC : Statistical Learning VS Machine Learning

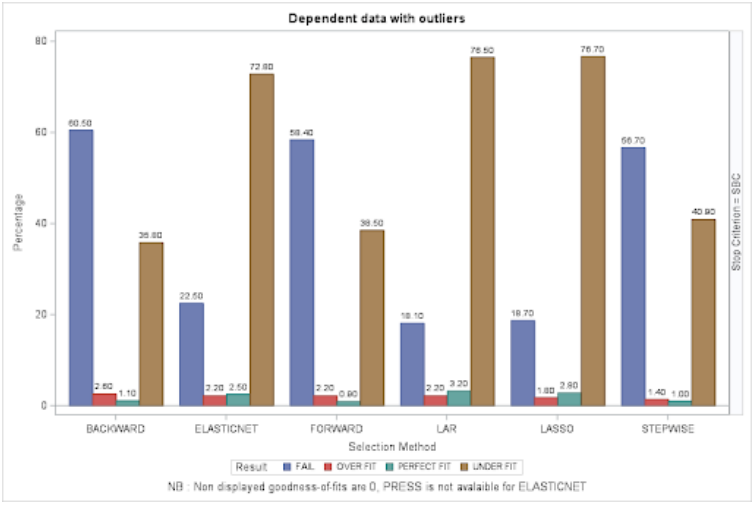

Figure 21: Variable selection performance with STOP= SBC and CHOOSE= SBC

The first overarching observation from this graph is the over-representation of underfitting

and failure (underfitting with additional variables) across all model selection performances. In other words, regardless of the method's nature (statistical or machine learning), underfitting and failure are predominant. No method provides a significant chance of obtaining the correct model, with the method maximizing perfect fitting being LAR with 3.20%.

BIC : Statistical Learning VS Machine Learning

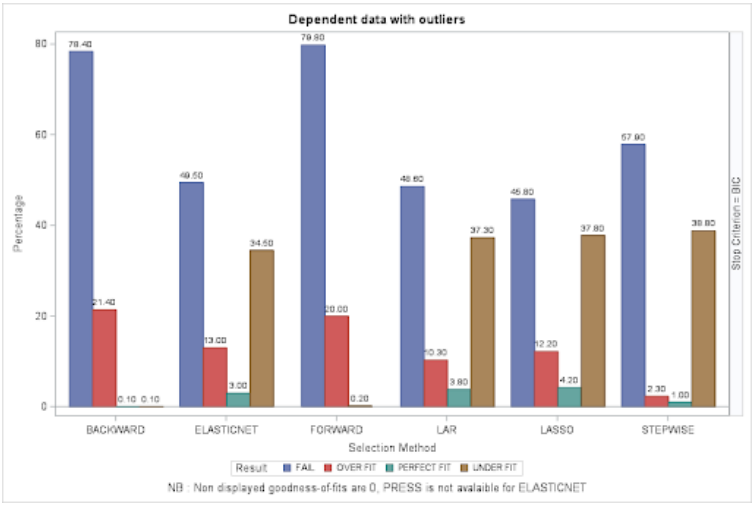

Figure 22: Variable selection performance with STOP= BIC and CHOOSE= BIC

In comparison to the previous section, the BIC criterion increases overfitting from 2.20% for Forward under the SBC criterion to 20% under the BIC criterion. However, when comparing this overfitting to the case of simple correlated data, it has significantly reduced. Furthermore, we observe that the methods most likely to select insufficient true variables are statistical learning models. Thus, for the SBC criterion as STOP and CHOOSE with correlated data with outliers, the model most likely to find the correct model is LASSO with a 4.20% chance.

AIC and AICc : Statistical Learning VS Machine Learning

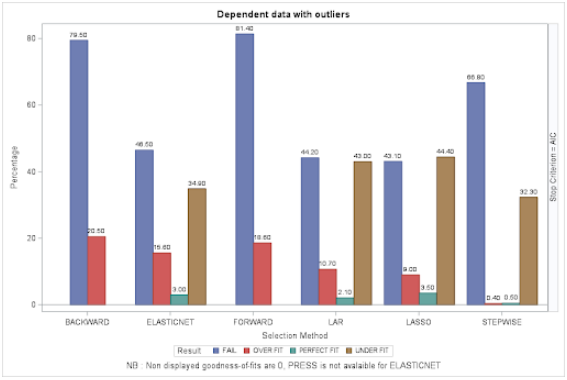

Figure 23: Variable selection performance with STOP= AIC and CHOOSE= AIC

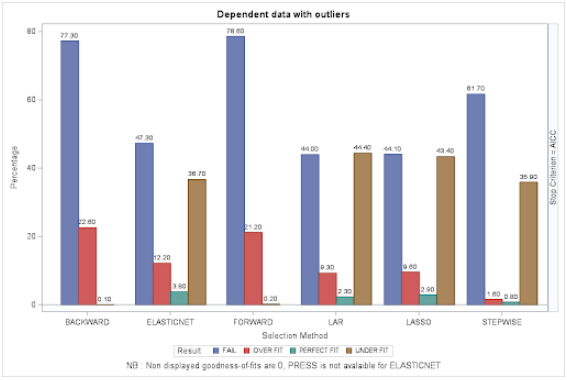

Figure 24: Variable selection performance with STOP= AICc and CHOOSE= AICc

In contrast to correlated data without outliers, in this case, even the Stepwise method does not yield good results in terms of model selection. Here, for both AIC and AICC, the perfect fitting percentage for Stepwise approaches zero, and the failure rate (underfitting with additional variables) is very high, at 66.80% and 61.70% respectively. Other statistical learning models generally follow the same trend, albeit with almost no chance of having failing models. As for machine learning methods, they balance the proportion of underfitting and failure more evenly.

CV: Statistical Learning VS Machine Learning

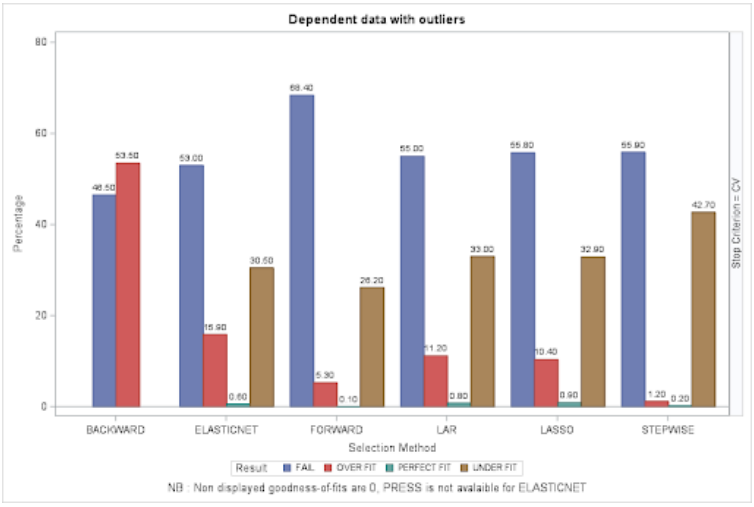

Figure 25: Variable selection performance with STOP= CV and CHOOSE= CV

When the CV criterion is defined as both the stopping and choosing criterion, we observe that the Forward and Stepwise methods have similar performances, while Backward is much more likely to have overfit models with a 53.50% chance compared to 5.30% for Forward and 1.20% for Stepwise. Machine learning methods, on the other hand, show relatively similar behavior, with a high chance that none of the true variables are selected but additional variables are (fail).

PRESS: Statistical Learning VS Machine Learning

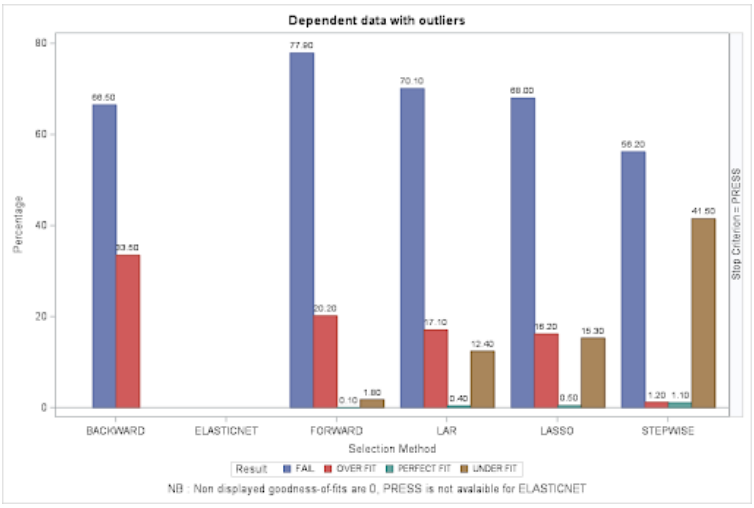

Figure 26: Variable selection performance with STOP= PRESS and CHOOSE= PRESS

The PRESS criterion is a restrictive special case of the CV criterion, which may explain the change in performances. The Backward method will now tend to find more fail-type models than models with additional variables (overfit). Machine learning methods LARS and LASSO also undergo changes with a higher chance of having overfit models (20.20% and 17.10% respectively) and fewer cases of pure underfitting (12.40% and 15.30% respectively).

3.3 Model selection performance according to combinations of criteria

Studying cases where the criterion is identical in both the STOP and CHOOSE options is not optimal, and it has allowed us to observe that combining criteria would yield models with a better fit. In this second section, we will compare the performance of various model selection methods and then examine the sensitivity of our performance.

In this section, we will observe from a global perspective the best combinations of criteria for each type of data. To do this, we will first impose a stopping criterion and determine the criterion to associate with it in the CHOOSE option in order to have the highest chance of obtaining the correct model. Among all these best combinations, we will choose the one that maximizes the percentage of perfect fitting and define this combination as the best for the given type of data.

3.3.1 Normal and independent data

The table above allows us to identify all the best combinations for a given stopping criterion. We can see that if the stopping criterion is SBC, the best criterion to set as CHOOSE would be PRESS, as once these criteria are combined, the LASSO method would have the highest chance of obtaining the correct model (50.10%). Furthermore, this combination is the best among all, as it has the highest perfect fit percentage. Beyond that, we observe that Stepwise is predominantly the best method to use within the best combinations, except when the

Given stopping criterion (stop=)	Best selection criterion (choose=)	Best method	Perfect fit percentage (%)
SBC	PRESS	LASSO	50.10
BIC	CV	Stepwise	40.40
AIC	BIC	Stepwise	40.80
AICc	BIC	Stepwise	40.00
CV	PRESS/BIC	Stepwise	42.10
PRESS	SBC	LARS	42.60

Table 2: Best combinations for a given criterion (Independent normal data)

stopping or choosing criterion is PRESS. Below is the graph allowing us to observe in more detail the results associated with the use of the combination STOP = SBC and CHOOSE = PRESS.

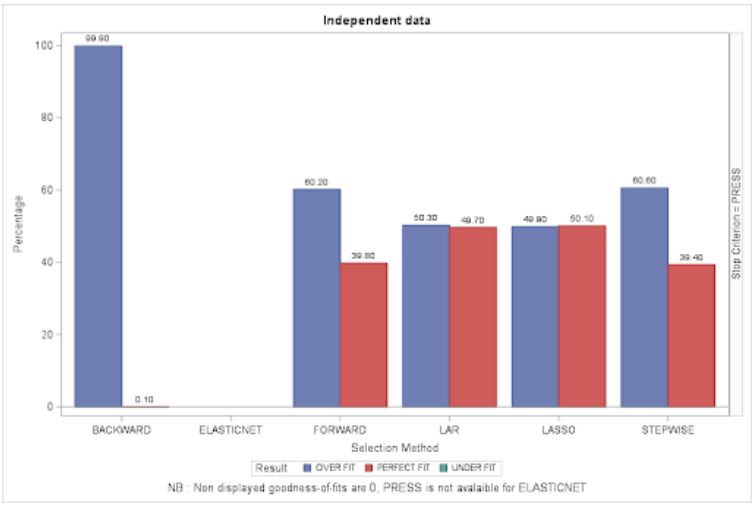

Figure 27: Details of selection performance associated with the best combination of criteria (Independent normal data).

3.3.2 Independent data with outliers

Once all combinations have been tested for independent data with outliers, we can determine the best combinations of criteria as listed in the table above. The optimal combination for use with independent data containing outliers is to employ SBC as the stopping criterion and PRESS as the selection criterion (STOP = SBC, CHOOSE = PRESS). When utilizing this combination, the LASSO method is most likely to yield the correct model (49% perfect fit). Additionally, it's noteworthy that the best combinations are similar to those for normal independent data, except that the BIC criterion no longer appears as an option for CHOOSE. Below is the graph providing a more detailed view of the results associated with using the STOP = SBC and CHOOSE = PRESS combination.

Given stopping criterion (stop=)	Best selection criterion (choose=)	Best method	Perfect fit percentage (%)
SBC	PRESS	LASSO	49.00
BIC	CV	Stepwise	40.70
AIC	CV	Stepwise	40.60
AICc	CV	Stepwise	41.50
CV	PRESS	Stepwise	41.90
PRESS	SBC	LARS	45.00

Table 3: Best combinations for a given criterion (Independent data with outliers)

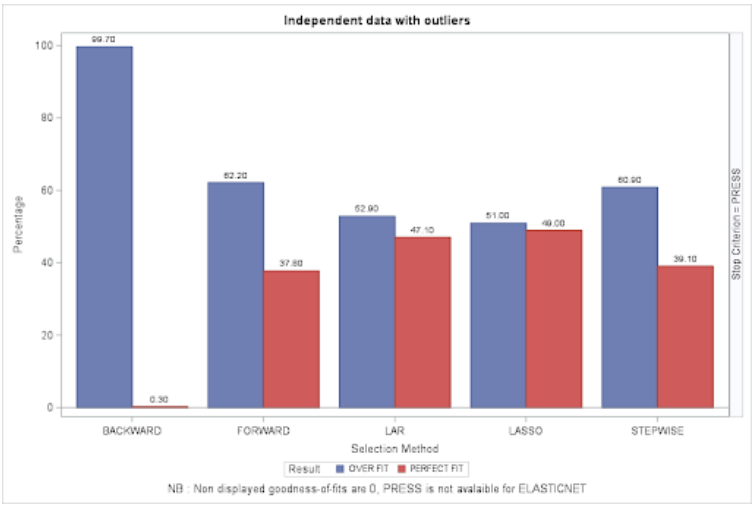

Figure 28: Details of selection performance associated with the best combination of criteria (Independent data with outliers).

3.3.3 Correlated data

Given stopping criterion (stop=)	Best selection criterion (choose=)	Best method	Perfect fit percentage (%)
SBC	PRESS	Forward	40.80
BIC	AICc	Stepwise	40.05
AIC	CV	Stepwise	39.60
AICc	SBC/BIC	Stepwise	40.10
CV	SBC	Stepwise	43.10
PRESS	CV	Stepwise	42.30

Table 4: Best combinations for a given criterion (Correlated data)

After analyzing all possible combinations for each method in the case of correlated data, we observed certain trends. For 80% of the combinations (24 out of 30), the method most frequently yielding the correct model (perfect fit) is Stepwise, followed by Forward 20% of the time. The table above illustrates the combinations that result in the method most likely to achieve the highest percentage of perfect fitting. The interpretation of the first row is as follows: When the stopping criterion is SBC, the best selection criterion is PRESS since this combination yields the method with the highest percentage of perfect fitting. The second observation is that combinations of statistical learning criteria are poorly represented among the best combinations. Ultimately, we can conclude that the optimal combination to use in the case of correlated data is STOP = CV and CHOOSE = SBC. In other words, with this combination, variable selection will stop once SBC is no longer minimized, and the model with the smallest PRESS will be chosen. This criterion combination also allows the forward selection method to find the correct model 40.20% of the time. Below is the graph providing a more detailed view of the results associated with using the STOP = CV and CHOOSE = SBC combination.

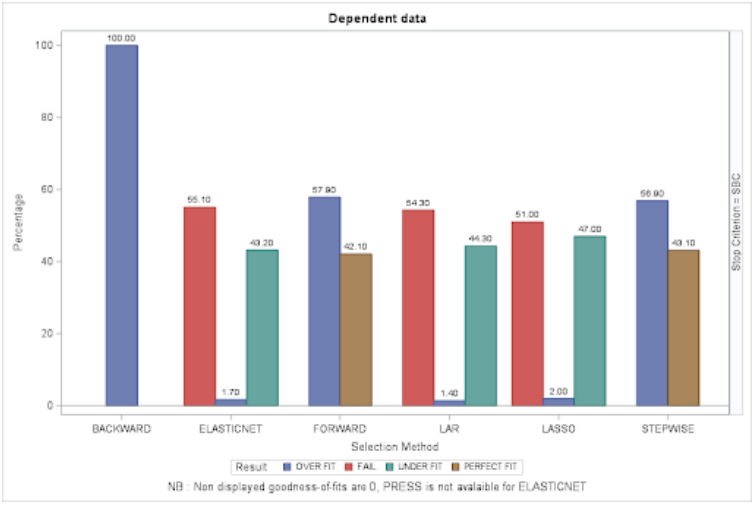

Figure 29: Details of selection performance associated with the best combination of criteria (Correlated data).

3.3.4 Correlated data with outliers

Outliers can distort model performance estimates and influence estimated coefficients, thereby affecting variable selection. This partly explains the low chances of obtaining the correct model regardless of the combination used. In the case of correlated data with outliers, regardless of the combinations, underfitting (with additional variables) is overrepresented. Additionally, the best combination in this context is AICc as the stopping criterion and SBC as the selection criterion, with a chance of obtaining the correct model of 5.40%, which remains very low. Below is the graph providing a more detailed view of the results associated with using the STOP = AICc and CHOOSE = SBC combination.

Given stopping criterion (stop=)	Best selection criterion (choose=)	Best method	Perfect fit percentage (%)
SBC	BIC	Elasticnet	3.70
BIC	AIC	Elasticnet	5.20
AIC	SBC	Elasticnet	5.10
AICc	SBC	LARS	5.40
CV	SBC	Elasticnet	3.80
PRESS	SBC	Stepwise	1.50

Table 5: Best combinations for a given criterion (Correlated data with outliers)

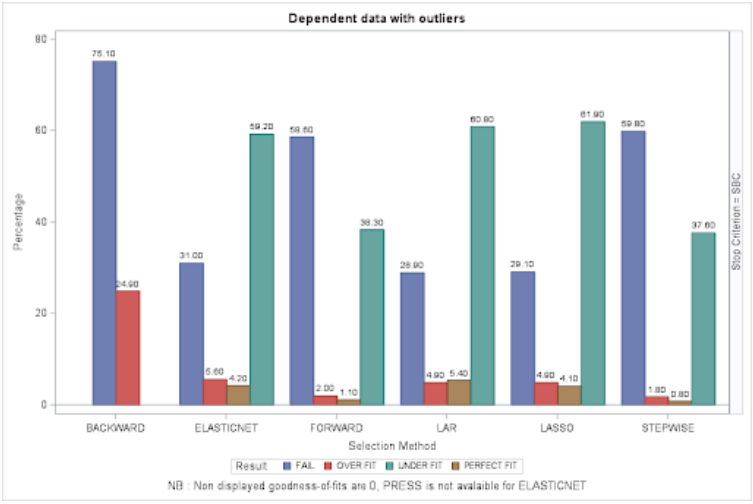

Figure 30: Details of selection performance associated with the best combination of criteria (Correlated data with outliers).

4 The sensitivity of the variable selection process.

In this section, we will focus on a sensitivity analysis of the data. This analysis is crucial as the robustness and reliability of variable selection models are of paramount importance in the fields of statistical learning and predictive modeling. However, the effectiveness of these models can be strongly influenced by the nature of the data on which they are applied. It is therefore important to understand the sensitivity of the models to variations in data characteristics, such as sample size, the presence of outliers, the distribution of variables, or the variation in parameter values. This sensitivity evaluation helps to understand the limits and specific performance of each selection method in diversified scenarios. By analyzing the models' reaction to these variations, we can not only identify situations where they excel but also determine conditions under which they might show signs of weakness.

To achieve this, we will consider two of the four types of data presented below, namely inde-

pendent data and dependent data with outliers (the ideal (type of data) and the "worst"), and we will conduct this analysis in two parts:

1. The first part will focus on variations in parameter values.
2. The second part will examine the impact of changes in sample size.

4.1 Variations in parameter values.

This part involves changing the values of the parameter Beta from these values (1.1,1,0.5,-0.7,0.3) to those (1,0.5,0.4,-0.8,0.1):
Indeed, we have chosen values close to the original Beta in order to understand how experimental changes truly impact model performance. Since the goal of the analysis is to observe how performance is affected by variations in parameter values, we will consider the best possible combination depending on the type of data.

4.1.1 Independant Data

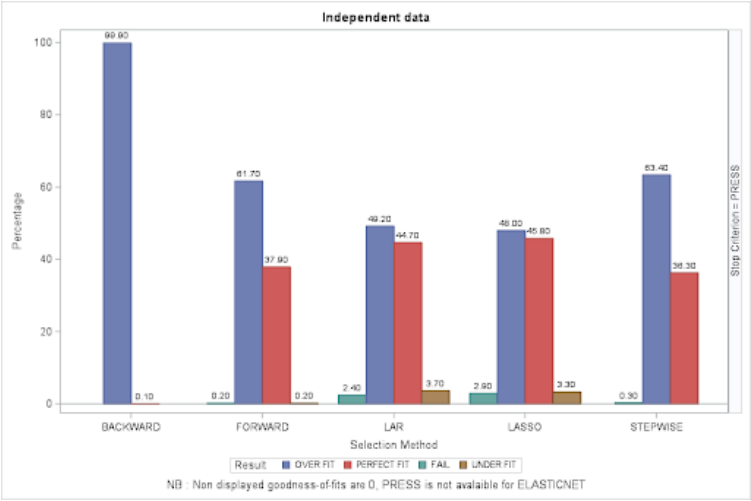

Figure 31: Details of selection performance associated with the best combination of criteria (Independent data).

It appears that despite variations in parameters, the performance of model selection methods remains relatively constant. Particularly, the Backward selection method shows a tendency towards overfitting, with performances similar to those observed for the original data. On the other hand, other selection methods demonstrate a fairly high probability of precisely identifying the true model, suggesting stability in their performances.
However, despite these observations, LASSO remains the preferred selection method for independent data with the Schwarz Bayesian Criterion (SBC) as the stopping criterion. This preference may stem from several factors, such as LASSO's ability to penalize coefficients and promote sparsity, which can be advantageous in variable selection.

4.1.2 Correlated data with outliers

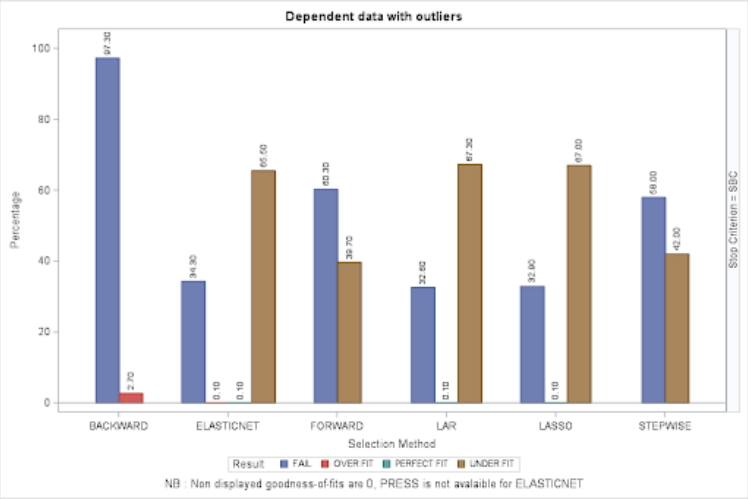

Figure 32: Details of selection performance associated with the best combination of criteria (Correlated data with outliers).

However, for dependent data with outliers, we observe a significant change in performance. If the probability of achieving perfect fitting was considerably low for the original data, here it is completely nonexistent. This can be explained for various reasons:
1. Sensitivity to Outliers in Dependent Data: Dependent data can be more sensitive to outliers, especially when dependencies between variables are considered. Outliers can have a more pronounced impact on the relationships between variables, thus resulting in more significant changes in model performance.
2. Impact on Performance Measures: Performance measures, such as the probability of achieving perfect fitting, can be heavily influenced by the presence of outliers. Outliers can lead to significant errors in model prediction, which is reflected in less favorable performances.

4.2 Change in sample size

In this sub-section, we have chosen to vary the size of the observations (from 250 to 1000).

4.2.1 Independent data

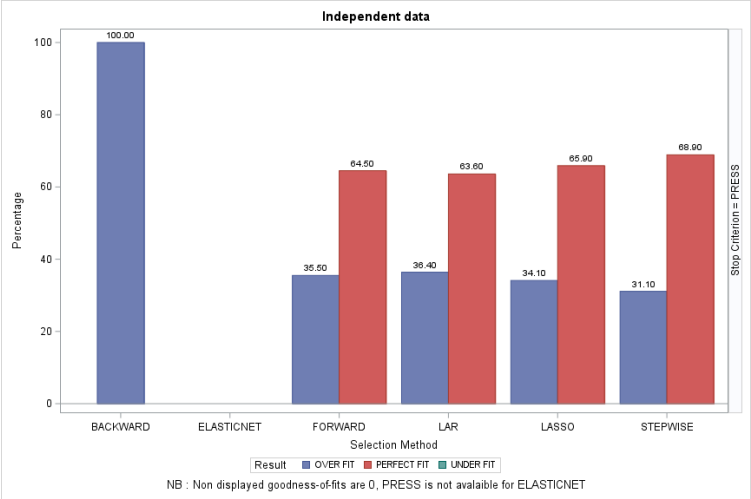

Figure 33: Performance following sample variations

Following the variation in the number of observations, we observe an increase in the rate of obtaining a perfect fit. This improvement results from the increase in the number of observations from 250 to 1000, which potentially favored a more precise convergence towards the true model. Furthermore, it is plausible that the variable selection algorithm achieved a more precise identification of relevant variables, leading to a better fit with the true model. Increasing the number of observations thus offers a wealth of additional information to guide the variable selection process more effectively.

4.2.2 Correlated data with outliers

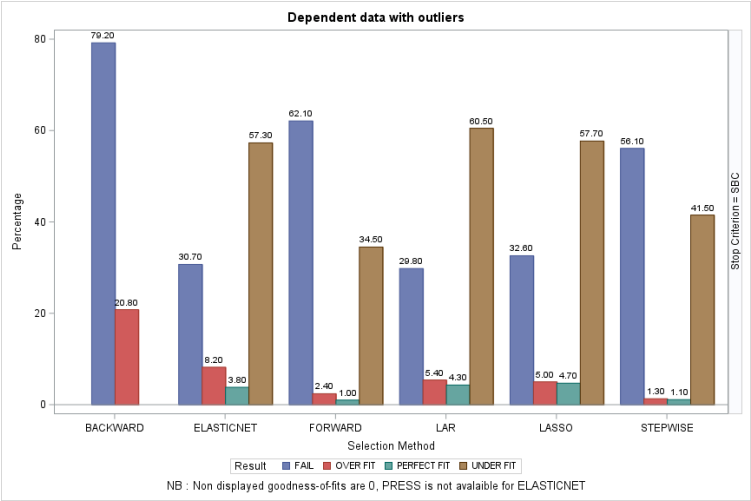

Figure 34: Performance following sample variations

Despite fluctuations in the number of observations, the propensity to select perfect fitting in a very limited way has not changed. It is plausible that the correlation between the variables gives the model greater complexity, which can make variable selection more difficult, even

with a higher number of observations. Moreover, the presence of outliers can also complicate the variable selection process. Outliers can significantly influence coefficient estimates and introduce additional challenges in variable selection.

In conclusion, the results indicate that the performance of model selection methods varies according to experimental conditions, such as data type (independent or dependent with outliers) and sample size. For independent data, despite variations in parameters, performance remains relatively constant, with a tendency towards overfitting observed for the Backward selection method. LASSO with the Schwarz Bayesian Criterion (SBC) stands out as the preferred method, underlining its effectiveness in variable selection.

On the other hand, for dependent data with outliers, performance undergoes significant changes, with a total absence of probability of obtaining perfect fitting. The increased sensitivity to outliers in dependent data may explain these results, suggesting that outliers have a more pronounced impact on the relationships between variables

5 Discussion/conclusion

This article is devoted to evaluating the performance of different variable selection methods in four distinct data contexts. To do so, we first established a theoretical framework outlining the various methods in statistics and machine learning, as well as the evaluation criteria we chose. We then initialized empty result tables, to be added to as the variables were selected. To carry out this study, we used algorithms that generated four different types of data, under various assumptions (H0: data independence and H1: data dependence). These data were created using Monte Carlo simulations. We specified that the true model corresponds to the intercept with all variables from X1 to X5. In parallel, we calculated performance metrics to assess the ability of each method to identify the correct model. These metrics include perfect fit (matching the true model), overfit (obtaining the true model with additional variables), underfit (not obtaining all the variables in the true model), and fail (not obtaining all the variables in the true model with incorrect selection of additional variables).

The results from section 3 highlight that adopting the same criterion for both stopping and selection is not optimal. In an ideal context with normal and independent data, the Stepwise method with PRESS as the criterion stands out, achieving a perfect fit score of 41.40%. However, the introduction of outliers alters these results, leading to the Stepwise method with BIC as the criterion displaying the best result with a perfect fit score of 40.60When the assumption of data dependency is considered, the best method is Stepwise with the AIC criterion, while under this same assumption with the presence of outliers, the optimal method becomes LASSO with BIC, achieving a perfect fit score of 4.20%. These observations highlight the significant influence of specific data conditions, such as data dependency and the presence of outliers, on the choice of variable selection method and associated criterion.

At this stage, our research work was not yet complete, and we undertook combinations of criteria to explore possible performance improvements in various contexts. This new approach revealed some notable differences from section 3 of our study. Firstly, in the context of "ideal" data, the best model was no longer Stepwise, but rather LASSO with SBC as the stopping criterion and PRESS as the choice criterion, showing a significant increase in the percentage of perfect fit (50.10%). For independent data with outliers, the best method remained LASSO with SBC as stop and PRESS as choice criterion (49% perfect fit). Although these results were lower than those obtained without outliers, they were still higher than those with the same criterion in STOP= and CHOOSE=. Under the correlation hypothesis, the best method was Stepwise, but this time with the combination STOP=CV and CHOOSE=SBC. The introduction of outliers led to the LARS method with the AICc and SBC combination, with a perfect fit rate of 5.40%. The major conclusion drawn from this part of our study was that using combinations of criteria improved variable selection performance, irrespective of the type of data considered.

To conclude, our study explored the sensitivity of variable selection processes in ideal and unfavorable data scenarios (correlated with outliers), using the best combinations of criteria identified. We examined the reaction of these processes to two major factors: change in Beta coefficients and change in sample size.

Concerning changing the Beta coefficients, we found that, for normal independent data, overall performance was similar to that without changing the coefficients. The LASSO method with STOP=SBC and CHOOSE=PRESS criteria remained the best. On the other hand, for correlated data with outliers, significant changes appeared, with a percentage chance of obtaining the correct model close to zero, irrespective of the method. This underlines the difficulty of variable selection in the context of correlated data with outliers.

About the change in sample size, we observed divergent results. For normal independent data, increasing the sample size (from 250 to 1000) led to a percentage chance of obtaining the correct model exceeding all other metrics. The Stepwise method performed particularly well, with 68.90%. perfect fit. On the other hand, for data correlated with outliers, performance did not evolve positively. This observation underlines the significant impact of the combination of correlation and the presence of outliers, complicating the variable selection process.

To improve this study, options such as integrating the SELECT option to manually choose the selection criterion, using additional criteria such as Mallow's Cp or adjusted R^2, and applying the algorithms to real data could have enriched our results and offered a more complete perspective on the performance and limitations of variable selection methods.

References

[1] Hastie, T, James, G., Tibshirani, R. and Witten, D. (2021), "An introduction to Statistical Learning ", Springer, 2nd Edition.

[2] Hastie, T. et al (2007), "Forward stagewise regression and the monotone lasso", *Electronic Journal of Statistics*,1,1-29.

[3] Draper, N. R., Smith, H. (1998), "Applied Regression analysis.", *John Wiley Sons Inc*, 3d Edition.

[4] Hastie, T., Tibshirani, R. and Friedman, J. (2009), "The Elements of Statistical Learning: Data mining Inference and Prediction", *Springer*, 2nd Edition.

[5] Vladimir N. Vapnik (1999), "An Overview of Statistical Learning Theory", *IEEE Transactions on neural networks, vol. 10, no 5.*

[6] Wicklin, R. (2013), "Simulating data with SAS", *SAS Institute.*

[7] Higham, N. J. (2002), "Computing the Nearest Correlation Matrix- a problem from finance", *IMA Journal of Numerical Analysis*, 22, 329-343.

[8] Bradley Efron , Trevor Hastie , Iain Johnstone and Robert Tibshirani (2004), "LEAST ANGLE REGRESSION", *Institute of Mathematical Statistics*, Vol. 32, No. 2, 407–451.

[9] GLMSELECT SAS Documentation

[10] Toepliz Matrix

[11] A Quick Intro to Leave-One-Out Cross-Validation