UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Paris 1 Panthéon-Sorbonne University

Applied Econometrics

Master 1 Econometrics and Statistics

Statistical Learning VS Machine Learning

Students:
Aya MOKHTAR
Carmen CRISTEA

Directed by:
Mr P. DE PERETTI

11 February 2024

Acknowledgments

We would like to express our sincere gratitude to Mr. De Peretti, our paper supervisor, for his rigorous supervision, valuable advice, and attentive guidance throughout the completion of this paper. His support has been instrumental in successfully completing this work.

Abstract

Feature selection is a fundamental step in statistical and machine learning modeling. This paper aims at explaining, evaluating and comparing the most widely used methods in both statistical and machine learning, while also highlighting their empirical limitations. Moreover, great emphasis is placed on the stopping criteria, as making a judicious choice can significantly improve the accuracy of the selected models. Finally, an empirical analysis based on the theoretical findings concludes this research paper.

The central insights of our paper indicate that machine learning techniques generally outperform statistical learning methods. However, regardless of the method under consideration, the more the data is altered with correlation patterns and/or outliers, the more pronounced the deterioration of the outcomes. Both statistical and machine learning appear to be sensitive to sample size changes, leading to a notable decline in the good fitting rate. Furthermore, the statistical learning techniques are more prone to severe overfitting patterns across all data scenarios.

Contents

Introduction

Statistical learning and machine learning are two closely related fields. Although most use these two terms interchangeably, it is crucial to acknowledge that certain differences exist. A nuanced insight into these techniques and their specific strengths is essential for synergistically combining them.

To begin with, machine learning is a field of Computer Science and AI, while statistics is encompassed within the realm of mathematics. Additionally, one of the main differences between statistical learning and machine learning is their purpose.

On the one hand, statistical learning has a well-established focus on inference and is known to be grounded on numerous assumptions. For instance, statistics usually presumes that predictors or features are known and additive, models are parametric, and testing of hypotheses and uncertainty are forefront (Breiman, 2001). Statistical learning is recognized for its proficiency in elucidating scientific questions related to the causal impact of a certain variable on a particular outcome of interest. Moreover, it is aimed at uncovering patterns and relationships in data using a wide range of methods: linear regression, logistic regression, Bayes' Rule. [3]

On the other hand, machine learning techniques are known to outperform alternative methods when it comes to prediction purposes. It brings optimal solutions to a wide range of intricate tasks, such as image recognition, medical diagnostics and fraud detection. In contrast with statistical learning, ML makes minimal assumptions: many ML models are based on non-parametric approaches, their structure is not specified, additivity is not expected, and assumptions about normal distributions, linearity or residuals, for example, are not needed for modeling (Carmichael and Marron, 2018). Hence, ML methods provide higher malleability and their requirements tend to be less rigid. [1]

Grasping both the strengths and weaknesses of each approach can help practitioners adapt the techniques to the specific needs of their research and assignments. It's important to note that machine learning and statistical learning techniques should not be seen as rival but rather complementary approaches as they serve different purposes. Statistical learning theory formalizes a model that makes a prediction based on observations and ML automates the modeling (von Luxburg and Scholkopf, 2011). Choosing between the two should hinge on the aim of the analysis, expected outcomes, available data, and additional considerations.

In order to address the specific question of model selection, we decided to assess the proficiency of both approaches: statistical learning and machine learning.

Chapter 1

Theoretical Approach

1.1 Statistical Learning Approaches

Statistical learning methods emerged in the late 1960s as a theoretical analysis addressing the challenge of estimating functions from a set of data. It constitutes a sophisticated approach to constructing predictive models using a set of features. In order to build the best model, the features are either added or removed based on their relevance. At each step, only statistically significant variables are retained. The process continues until there is no statistically valid reason to add or remove any other predictive variable. [2]

In this context, our study will focus on three specific feature selection methods: the forward method, the backward method and the stepwise method. These techniques are great tools for learning how basic models are built and evaluated.

1.1.1 Forward Selection

Forward selection is a method aimed at choosing variables that starts with a basic model, often referred to as the null model, containing only the intercept. It examines various ways to add one variable at a time to the model and selects the best variable based on specified criteria, such as lowest p-value, highest adjusted R squared, lowest Mallow's Cp, lowest AIC. More detailed explanations on these criteria will be provided in the next section. This process keeps going, adding one variable at a time, until the chosen measure stops getting better.

For instance, if we choose the AIC criterion, the forward process will stop when there is no feature left that can further decrease the AIC. The process then concludes at a certain iteration once the model no longer improves with the addition of new features. [11]

1.1.2 Backward Selection

Backward selection, in contrast to forward selection, is a variable selection method that begins with a full model, containing all available features. It removes systematically the least impactful variable in each iteration. The decision to remove a feature is based on the criteria mentioned in the preceding section. The regression model is re-estimated after each elimination and the process continues until either all remaining variables are significant or the chosen criterion can no longer be further optimized.

1.1.3 Stepwise Selection

Stepwise selection is a hybrid approach that integrates elements from both Forward and Backward selection methods (mentioned above). During each iteration, a new feature is added to the model (forward method). Next, the model is tested in order to verify whether the significance of the feature has decreased or increased. This is determined by a chosen selection criterion as it indicates whether the feature is detected as significant or not. In the case of non-significance of a feature, it will be removed from the model (backward method). The stepwise process continues until adding or removing features no longer has an effect on the chosen stopping criterion for the selection process. [16]

1.2 Machine Learning Approaches

Loss function and lambda-penalty

The machine learning processes considered in this paper are based on the minimization of a loss function, which measures the goodness of fit of the model. Generally, it serves to optimize and regularize the models by introducing the so-called lambda-penalty: the higher the lambda, the stronger the regularization. Its value can be determined using explanatory or predictive criteria, which will be further elaborated upon in the subsequent sections.

1.2.1 Least Absolute Shrinkage and Selection Operator (LASSO)

Introduced by Robert Tibshirani in 1996, the Least Absolute Shrinkage and Selection Operator (LASSO), also called L1 penalty function, is given by the following function:

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \sum_{i=1}^{T} \left(y_i - \left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \right) \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

The LASSO regression is a variable selection procedure. It is particularly useful when dealing with high-dimensional datasets. The aim of this method is to shrink the coefficients of the least important features and determine the most impactful ones. The closer lambda is to zero, the more it will converge towards the OLS estimator. The bigger the value of lambda, the more parsimonious the model will be. We can choose a specific criterion or opt for cross-validation to determine which model will be the best. [17]

1.2.2 Least Angle Regression LAR

The least-angle regression algorithm has positive results in the case of a high-dimensional data, as well as in cases where variables exhibit multicollinearity and strong correlation. Moreover, it is particularly used when the number of predictors surpasses the quantity of observations. The procedural steps of the algorithm are outlined as follows: [7]

- Normalize the variables (to be sure not to have too many extreme values), then calculate the residuals (r) and initialize $\beta(i)$ to zero.
- Identify X_j, which demonstrates the highest correlation with r_i.
- Gradually modify the coefficient $\beta(j)$ until its correlation with r exceeds that of $\beta(k)$ with the observed residuals, thus expanding the set of information.
- Repeat the process with these two features until another variable $\beta(m)$ exhibits a more robust correlation with the observed residuals. Repeat this step, further expanding the set of information.
- The process continues until all predictors are in the model (when the full least squares solution is reached).

LAR is computationally as fast as forward selection and potentially more accurate, especially with a high feature-to-instance ratio. However, it does come with a significant drawback. LAR is highly sensitive to noise, which can lead to unpredictable results in certain situations.

1.2.3 Elastic Net

The Elastic Net Regression combines both LASSO (L1 regularization) and RIDGE (L2 regularization) penalties. It is important to note that the Ridge regression doesn't use a variable selection method. It is aimed at handling multicollinearity.

The target of the Elastic Net regression is therefore twofold: to make the coefficients of the least relevant features tend towards 0, as well as to mitigate multicollinearity. [10]

The associated objective function is defined as follows:

$$\beta_{\text{Ela}} = \arg\min_{\beta} \sum_{i=1}^{T} \left(y_i - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \lambda \left(\alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2 \right)$$

The Elastic Net, compared to the LASSO method, handles correlated predictor features more efficiently. However, its performance heavily depends on the choice of the hyperparameters (α) and (λ), as their improper setting can lead to poor performance.

Chapter 2

Performance Evaluation - Criteria

2.1 Akaike Information Criterion: AIC

Akaike (1974) introduced the first information criterion, the Akaike Information Criterion (AIC), which weights model performance and complexity in a single metric. More specifically, AIC is based on Kullback-Leibler (K-L) distance (Kullback & Leibler, 1951), which is a way of conceptualizing the distance, or discrepancy, between two models. One of these models is the «true» or «generating» model from which actual observations emerge. The other model is a model specified in accordance with the researcher's knowledge about the «true» model. If the two models are exactly the same, then the K-L distance between the two models is zero; otherwise, the K-L distance is greater than zero. While the absolute distance to the «true» model cannot be computed, relative distances can be used to determine which model provides the most «reasonable» approximation of the «true» model.

Akaike's (1974) key insight was that maximum likelihood estimation, with some correction based on the number of parameters in the model, is the expected K-L distance, as follows:[13]

$$\mathrm{AIC} = -2\log(L_\theta) + 2k$$

where θ is the parameter vector and k is the dimension of θ. Per information theory, the Kullback & Leibler distance is understood as the information lost when a model is used to approximate the «true» one and must therefore be minimized. Thus, the basic principle underlying the AIC criterion is the assumption that the less information a model loses, the higher is its quality. [14]

- Akaike Information Criterion corrected AICc

One possible shortcoming of the AIC criterion is that it may perform poorly if the number of estimated parameters is large compared to the sample size. A second-order version of the AIC, which is valid for small sample sizes, was also derived and is called AICc.

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

where: k the number of parameters in the model and n: the sample size.

AIC and AICC (Hurvich & Tsai, 1989) are efficient information criteria as they will select the best model when the generating model is of infinite dimension.

2.2 Schwarz's Bayesian Criterion: SBC

SBC is a consistent information criterion developed by Gideon E. Schwarz and published in 1978. Its essence lies in Bayesian inference. Compared to the AIC criterion, as the size of the sample increases, SBC will select a true model of finite dimension as long as it is included in the set of candidate models. SBC is designed to maximize the posterior probability of a model given the data. In line with the AIC criterion, the SBC metrics are basically cost functions, which have to be minimized; they both favor models with a high likelihood but implement a penalty for complexity. In contrast, SBC penalizes model complexity more heavily than AIC. When optimizing the Bayesian information criterion, the probability of selecting a fixed most parsimonious true model tends to 1 as the sample size tends to ∞ (e.g., Nishii (1984) and Haughton (1988,1989)). [4]

The formula for SBC is given by:

$$\mathrm{SBC} = -2\ln(L) + k\ln(n)$$

where: L the maximized value of the likelihood function of the model, k the number of parameters in the model and n the sample size. [6]

- Sawa Bayesian information criterion BIC

The BIC is quite similar to the Schwarz's Bayesian Information Criterion (SBC). Both of them are Bayesian criteria, but the Sawa Bayesian Criterion incorporates adjustments to the penalty term for model complexity, as follows:

$$\mathrm{BIC} = n\ln\left(\frac{SSE}{n}\right) + 2(p+2)q - 2(q^2)$$

where: $q = \frac{n\sigma^2}{\mathrm{SSE}}$, n is the number of observations, SSE is the error sum of squares, and p is the number of model parameters.

2.3 Mallow's CP

In statistics, Mallows' CP is applied within the framework of model selection, specifically when there are multiple predictor variables to choose from. It's crucial to highlight that, given k independent variables, one can potentially generate 2^k models using different subsets and combinations of these variables. A model with too many predictors can be relatively imprecise while a model with too few predictors can produce biased estimates. Therefore, the aim of Mallows' statistic is to select the best model involving a subset of these predictors by balancing underfitting and overfitting the data (bias and variance).

The formula for Mallows' CP is given by:

$$Cp = \frac{\mathrm{SSE}_p}{\mathrm{MSE}_f} - (n - 2p)$$

where SSE_p is the sum of squared errors of the regression with p coefficients and MSE_f is the mean squared error corresponding to the full model, the model that contains all of the explanatory variables. [12] If model(p) is correct, then Cp will tend to be close to or smaller than p. If Cp is substantially larger than p, then there is a large bias component in the model.

2.4 K-folds Cross-Validation CV

Cross validation is a widely used method to evaluate the performance of predictive machine learning models. This technique involves partitioning the entire dataset into two segments : one used to train the model and the other one used to test its performance. Usually, the training and test sets must cross-over in successive rounds so that each data point has a chance of being tested against. One of the most ubiquitous cross-validation forms is k-folds. [8]

First, the dataset must be shuffled in a random manner. Then, all data is divided into k subsets (folds), ideally equal in size. For each fold, a model is trained on the $(k-1)$ folds, and then tested using the selected (hold-out) fold. In other words, we iterate k times with a different subset reserved for testing purposes each time. Therefore, the model is consecutively tested on every fold available. The skill of the model will further be summarized by averaging the k results obtained on the test sets. The selected model is the one that minimizes the average errors on the tested data. [9]

2.5 Leave-One-Out Cross-Validation (PRESS)

Leave-one-out cross validation (PRESS) is k-fold cross validation taken to its logical extreme, with k equal to N, the number of data points in the set. That means that N separate times, the model is trained on all the data except for one point and a prediction is made for the left-out point. As before, the average error is computed and used to evaluate the model. [15]

Chapter 3

Data Generating Process - DGP

After defining all the selection algorithms, both Statistical and Machine Learning techniques, as well as the stopping criteria, we will further elaborate on the methodology of our research.

It is noteworthy to mention that we focused on six techniques, 3 of which correspond to statistical learning (FORWARD, BACKWARD, STEPWISE), and the remaining 3 to machine learning (LASSO, LAR, ELASTIC NET). To achieve the best outcomes, we combined them with both explanatory criteria, such as AIC, AICc, BIC, SBC and Mallow's CP, and predictive stopping criteria, in particular CV, PRESS.

To assess the performance of multiple algorithms, we examined various scenarios using four distinct types of data: linearly independent data, data with multicollinearity, data with outliers, and data combining both multicollinearity and outliers. This will provide us with insight into their performance in real-world scenarios. The first dataset is the one with linear independence (DGP1), commonly referred to as «perfect» data, as it has no correlation among features and no outliers. In order to generate the 50 variables, we used a multivariate normal distribution. The second dataset (DGP2) introduces multicollinearity among the first 5 features, the ones used to generate the outcome variable Y. For the third dataset (DGP3), we intentionally introduced 5% of outliers to the first 3 variables. Finally, to assess the performance of the algorithms in a real-life scenario, we combined both multicollinearity among the first 5 variables and outliers in the first 3 variables. And so we have obtained our fourth dataset (DGP4).

Metrics

To see how well the chosen algorithms combined with the stopping criteria perform empirically, we constructed metrics. Their main purpose is to define and identify overfitting, underfitting, and good fitting by comparing the selected variables to the ones that were truly used to generate Y, the outcome variable. They play a crucial role in revealing the consistency of accurate outcomes across each algorithm and across the different datasets.

It's important to emphasize that the metrics are used to compare the set of variables selected by the algorithm to the true set of variables of our model, in this case {'intercept', 'X1', 'X2', 'X3', 'X4', 'X5'}.

We have chosen to define six metrics: good fit, almost good fit, type A overfit, type B overfit, type A underfit, and type B underfit. In fact, we used the XSECT (Intersection) and NROW (seemingly same to Card) functions to define our metrics. If the number of selected elements is 6 (intercept included) and the cardinality of the intersection (common elements between selected and true features) is also 6, then a good fitting case is recorded; otherwise, it is considered a failure.

The failure cases are further decomposed into 5 scenarios. The metric «almost gf» represents situations where the model returns 5 variables but not all of them are correct, for instance {'intercept', 'X1', 'X2', 'X3', 'X10', 'X11'}.

Next, «overfit$_a$», the so-called «good overfitting», corresponds to scenarios where the algorithm returns more than 5 variables but does include all the correct ones. This set {'intercept', 'X1', 'X2', 'X3', 'X4', 'X5', 'X8', 'X9'} could serve as an example. On the contrary, «overfit$_b$» is associated with poor algorithm performance: not only does it include more than 5 variables, but it also fails to include one or more true variable(s). For instance, we could imagine a selected set such as {'intercept', 'X1', 'X2', 'X8', 'X9', 'X10', 'X11'}.

Now, let's explore the underfitting metrics. The so-called «good underfitting», referred to as «underfit$_a$», occurs when the model selects fewer than 5 variables, all of which are correct, for example {'intercept', 'X1', 'X2', 'X4'}. In contrast, «underfit$_b$» refers to the specific cases where the algorithm selects fewer than 5 variables, and not all of them are accurate, for instance {'intercept', 'X1', 'X2', 'X10'}.

To conclude this part, these metrics provide a thorough analysis of the model's feature selection performance, capturing various aspects of both correct and incorrect selections. The evaluation includes assessing the frequency of feature selection for each predictor and examining the probabilities associated with each metric through the feature selection process.

3.1 DGP 1 : Linearly independent data

For all the DGPs that we used in our research, we chose to include 50 variables. First, we started by defining the characteristics of the dataset : the mean, the covariance matrix, and the sample size (N). The covariance matrix is, in this specific case, the Identity matrix: it has 1s in the diagonal and 0s elsewhere. The mean is defined as a vector of 0s, with each variable having a mean of 0. However, we allowed the sample size to vary throughout the selection processes in order to identify its impact on the robustness of the outcomes. In order to meet all the above-mentioned criteria, we generated data from a multivariate distribution using the RANDNORMAL function. The model that generates our outcome variable Y only includes the first five features, as follows:

$$Y = 0.8 \times X_1 - 0.7 \times X_2 + 0.5 \times X_3 - 1.2 \times X_4 + 0.9 \times X_5 + \epsilon$$

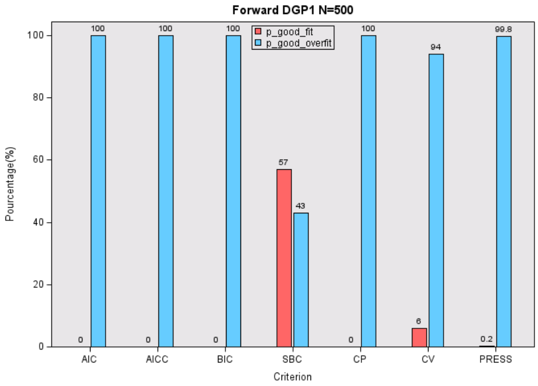
It's important to note that the coefficients associated with the five variables were chosen randomly. Moreover, we initially set the standard deviation of the error term to 0.1 but through iterative testing we concluded that diminishing it improved the performance of our selection process. Hence, we settled on 0.05. In order to mitigate the «randomness» of our results and to be able to conclude on the performance of each algorithm, we opted for an iterative approach. In other words, for each combination of algorithms and selection criteria, we generated datasets and performed the selection process 1000 times. This method allowed us to derive an empirical «probability» of the good fitting scenario for the analyzed algorithms.

We will now concentrate on the analysis of our results for the first type of data, DGP1, which corresponds to what we defined as «perfect data». It is crucial to give special consideration to this particular instance, as it will serve as a reference case for the remaining dataset scenarios.

Statistical Learning results

In this section, we evaluated forward, backward and stepwise selection methods, incorporating all the criteria outlined in the first section. The graphs below show that all statistical learning techniques tend to suffer from excessive overfitting.

Concerning the forward method, only the Schwarz Bayesian information Criterion (SBC) and K-fold Cross-Validation (CV) criteria exhibited respectively 57% and 6% of good fitting occurences, whereas the remaining criteria led to systematic good overfitting. In contrast, the backward method performed even worse, resulting in 100% overfitting for all considered stopping criteria except for SBC, which yielded 56.8% of good fitting.



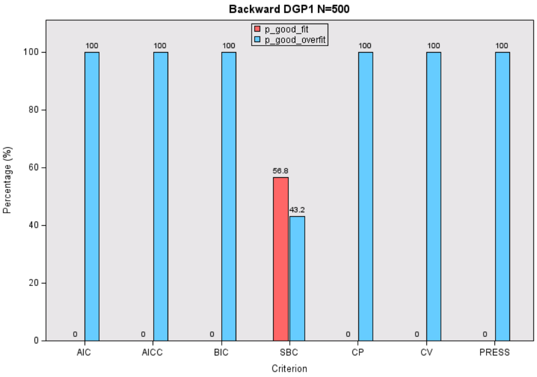


Fig. 1: Forward in Independence

Fig. 2: Backward in Independence

Nevertheless, the stepwise selection yielded much better results, achieving a good fitting rate between 50% and 60% across all criteria, with a relatively reduced inclination towards overfitting. These satisfactory outcomes can be attributed to the algorithm's ability to add and remove features. As a result, this process could enhance the efficiency of the variable selection procedure and contribute to an overall improvement in model fitting. A noteworthy observation in this specific case is the absence of occurrences identified as «bad_overfit». Thus, we can be confident that our five explanatory features are always included during the selection process, even if more variables than expected are selected.

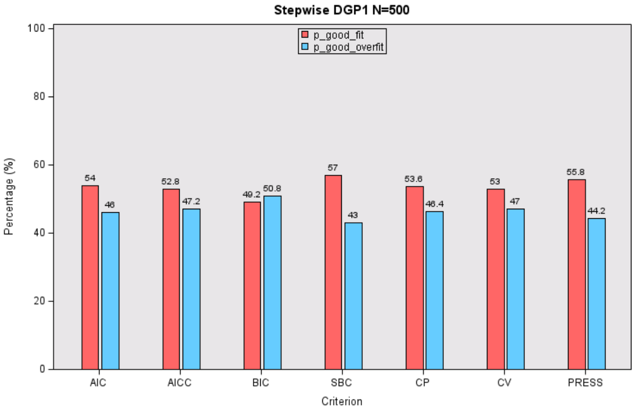


Fig. 1: Stepwise in Independence

The illustrations presented below indicate how often each variable appears across the selected models using the forward method. We used a comparison between SBC and AIC, with the latter being a classical example of a criterion leading to overfitting. As mentioned previously, the SBC criterion consistently includes the five relevant explanatory variables X1 to X5 along with other variables, but with low probability patterns (less than 1% for each variable). On the contrary, the AIC not only includes our five target features, but also incorporates the remaining variables, with much higher frequency patterns (between 15% and 25%).

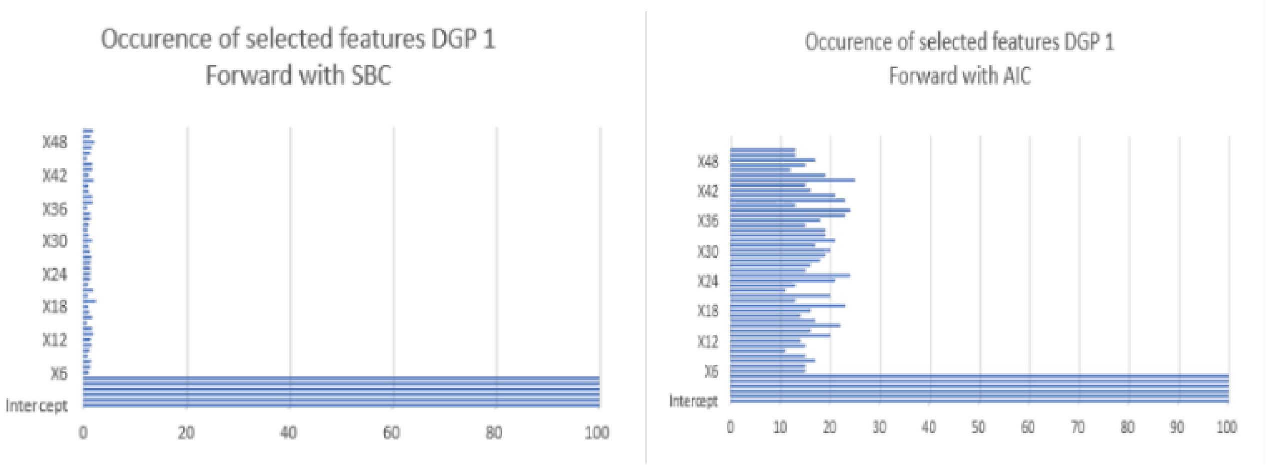


Fig. 1: Occurrences of selected features with Forward Selection

To conclude, in the case of linearly independent data, we observe the consistent presence of overfitting. All statistical learning methods manage to incorporate the five explanatory features, but they systematically introduce additional irrelevant variables. It is important to note that the Schwarz Bayesian criterion (SBC) stands out as the most effective criterion, achieving success in identifying the accurate model in more than 50% of instances.

Machine Learning Learning results

As for ML methods, we notice that both LAR and LASSO show similar results when it comes to model selection. For instance, both methods appear to select models with a large spectrum of fitting behaviors: good fitting, good overfitting, and good underfitting.

Our findings indicate that, when combining LAR and LASSO with all criteria but PRESS, the occurrence of good fitting ranges between 38% and 43%. Through a method of trial and error, we identified that PRESS, particularly when paired with the stopping criterion SBC in the glm select procedure (stop=SBC), yields the best performance in the selection process, achieving nearly 60% of good fitting. In contrast, the overfitting rate exceeds 50% for most criteria, except for PRESS, which again demonstrates improved performance with an overfitting score of 40.8%.

Regarding underfitting patterns, they range from 5% to 10% for both LAR and LASSO when combined with all criteria but PRESS, which registers no occurences of underfitting.

An important result to notice is the absence of occurrences categorized as «bad overfitting» and «bad underfitting» with the methods LASSO and LAR. In other words, in instances of overfitting, we are sure that, despite selecting more variables than necessary, the set of five true variables was consistently included. Likewise, in instances of underfitting, despite selecting less than 5 variables, we are certain that all the selected ones belong to the set of true variables.

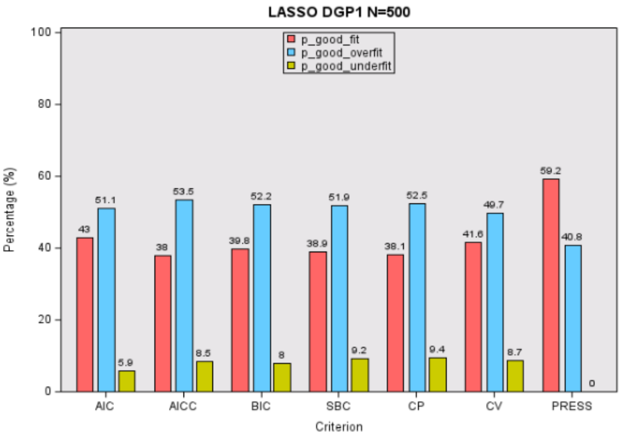


Fig. 2: LASSO in Independence

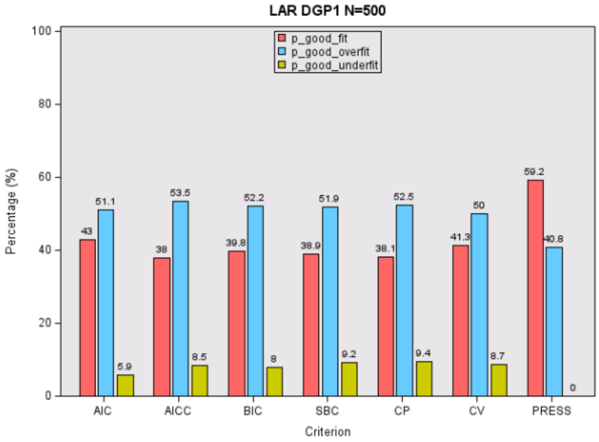


Fig. 3: LAR in Independence

Our results are corroborated by the graph illustrating the occurrence of selection of all the variables from our dataset, X1 to X50, with LASSO. We notice that, when using the PRESS criterion, the 5 relevant variables are consistently selected, which means that it exhibits no underfitting patterns. The remaining variables, X6 to X50, are rarely chosen individually but they contribute to the relatively high overfitting rate (40.8%).

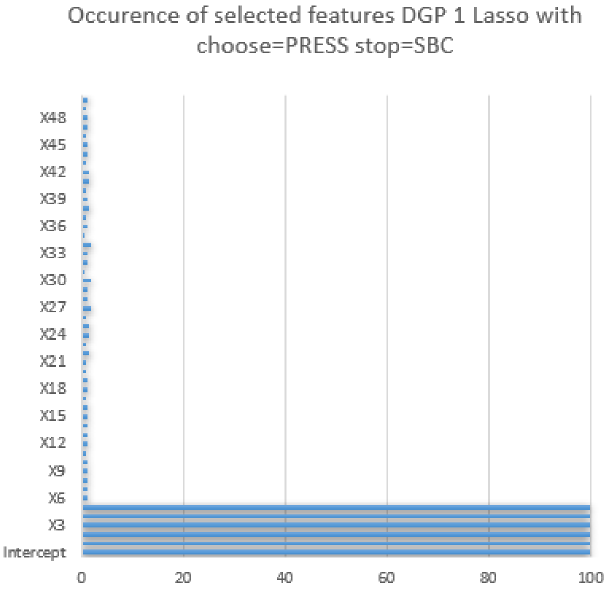


Fig. 1: Occurrences of selected features with LASSO

In contrast, the ELASTIC NET method exhibits different patterns, with a generally poorer performance compared to LASSO and LAR. The analysis highlights a good fitting rate ranging from 34% to 38% for all criteria, except for Mallow's Cp (CP) and Cross Validation (CV). The former shows satisfactory results with a good fitting rate of 51.2%, whereas the latter indicates much poorer performance with only 10.4% in terms of good fitting. Moreover, the CV criterion appears to be particularly prone to overfitting, with an overfitting rate exceeding 80%. Lastly, ELASTIC NET demonstrates a propensity to select more overfitted models, on average, than the previous two methods.

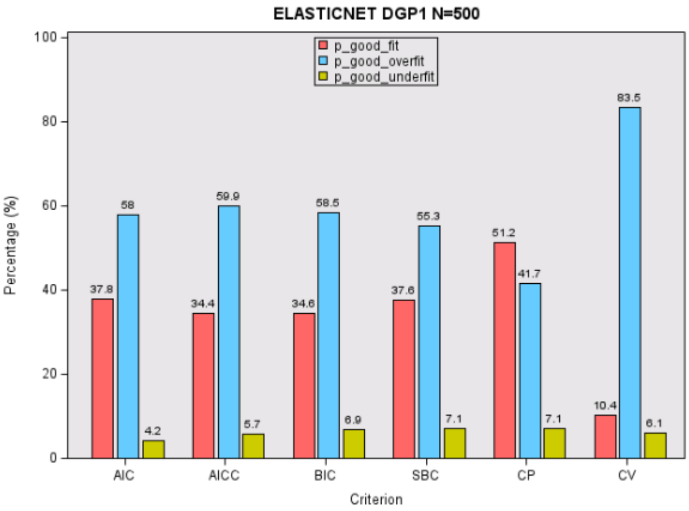


Fig. 1: Elastic Net in Independence

In the graph above, we see that the occurrence of selection of the irrelevant variables, X6 to X50, is relatively high for the CP criterion. These outcomes raise the question of the severity of the overfitting that stems from it.

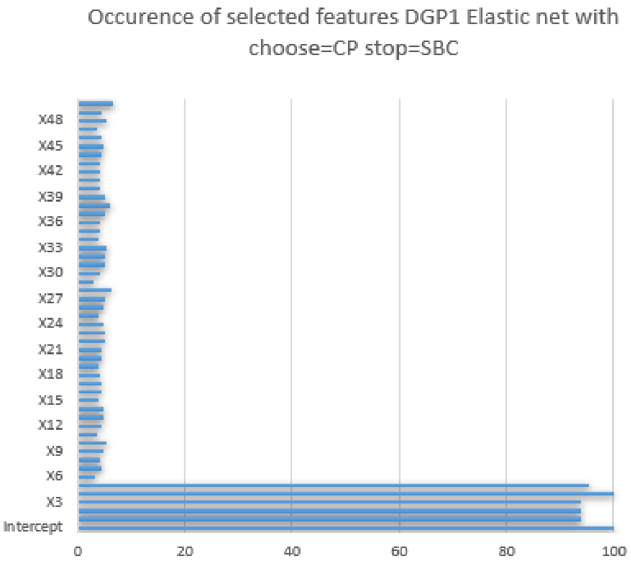


Fig. 1: Occurrences of selected features with Elastic Net

To better understand the magnitude of the overfitting pattern for the ELASTIC NET method, we brought some changes to the metrics. We broke the good overfitting down into two categories: «overfit_aa», which refers to cases where the selected models include the 5 true variables and 2 irrelevant variables at most, and «overfit_ab» which encompasses scenarios where selected models include more than 2 irrelevant variables alongside the 5 true ones. The good news is that, most of the time, the overfitting is moderate, as the ELASTIC NET method combined with the considered criteria tends to add 1 or 2 irrelevant variables at most.

These results further reinforce the reliability of the CP criterion: in nearly 96% of the total overfitting cases, at most 2 additional variables are selected. In contrast, we notice that the CV criterion yields mediocre results as it selects models with severe overfitting patterns.

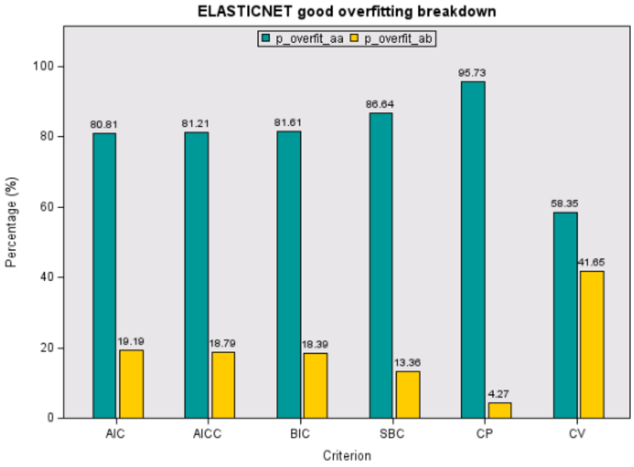


Fig. 1: Elastic Net with overfitting breakdown

3.2 DGP 2 : Data with correlation between explanatory variables

To gain a deeper understanding of how the algorithms perform with real-life data, we generated a dataset incorporating a blend of both correlated and non-correlated variables. We introduced multicollinearity in the first 5 features, whereas the remaining 45 were generated using a multivariate normal distribution with mean 0 and the covariance matrix - the Identity matrix. Aiming at introducing controlled correlation for the first five predictors, the ones used to generate the outcome variable Y, we employed the TOEPLITZ function. It allowed us to create a positive correlation matrix different from the Identity matrix for the first 5 variables, as follows:

Correlation Matrix

$$\begin{bmatrix} 1 & 0.8 & 0.6 & 0.4 & 0.5 \\ 0.8 & 1 & 0.8 & 0.6 & 0.4 \\ 0.6 & 0.8 & 1 & 0.8 & 0.6 \\ 0.4 & 0.6 & 0.8 & 1 & 0.8 \\ 0.5 & 0.4 & 0.6 & 0.8 & 1 \end{bmatrix}$$

Statistical Learning results

After incorporating correlation among our five explanatory features, we found that all statistical learning methods produced results quite similar to the initial Data Generating Process (DGP1), with a slight improvement ranging from 2% points to 3% points for both forward and backward methods. As illustrated in the graphs below, the forward selection method improved its good fitting rate by 1% point, going from 57% to 58% in terms of good fitting rate in the case of SBC. The CV criterion also exhibited a slight increase of 3% points, while all the remaining criteria led to 100% of good overfitting, as previously. Furthermore, backward selection also exhibited a marginal enhancement in the good fitting rate compared to the reference case (DGP1).

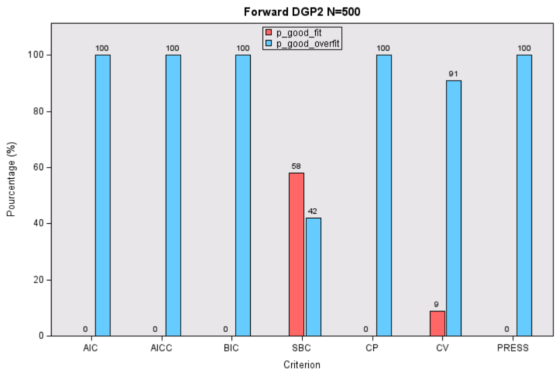


Fig. 2: Forward with Internal correlation

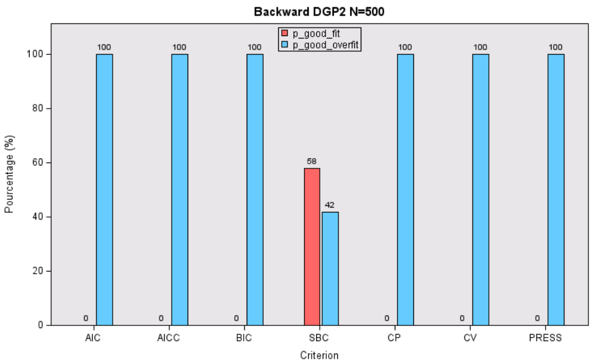


Fig. 3: Backward with Internal correlation

As for stepwise selection, it shows a subtle improvement in the accuracy of the selected models, ranging from 3% points to 13.2% points compared to the DGP1. These outcomes suggest that the correlation among the 5 explanatory variables had a positive impact on the statistical learning methods.

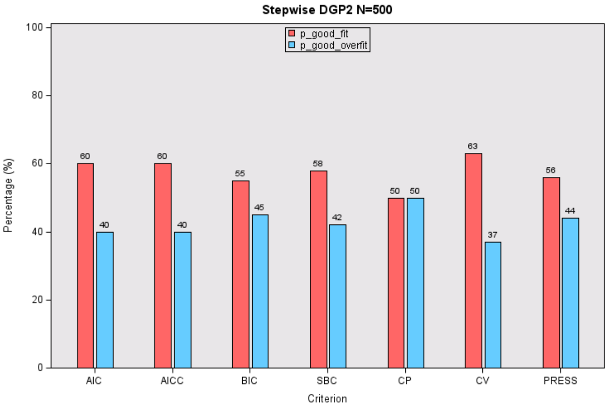


Fig. 4: Stepwise in Correlation

Once again, it is important to note the absence of bad overfitting across all the statistical learning methods, in spite of the correlation patterns. As a result, we can be confident that these algorithms always succeed in selecting the 5 relevant variables. However, in certain instances, they tend to choose additional irrelvant variables.

To conclude, incorporating correlation among the five explanatory variables slightly improves the results of statistical learning methods, in particular with the SBC stopping criterion for both forward and backward methods, and with the CV criterion for the stepwise method.

Machine Learning results

In line with the outcomes from the DGP1, both LASSO and LAR persistently deliver similar results. The introduction of correlation among the first five features did not alter this pattern. Firstly, we notice a sharp decline in the good fitting rate across most criteria, ranging from 6% to 13.4%. These results are in stark contrast to the approximately 40% good fitting rate observed for the same criteria in the case of independent data.

Next, we notice that the decrease in the good fitting rate is counterbalanced by a notable increase in the good overfitting rate, varying between 75% and nearly 84%. It is noteworthy that in the independence section, we identified a maximal overfitting rate of 53.5%. We observe a significant rise in underfitting levels as well.

As before, the PRESS criterion outperforms the others significantly. Surprisingly, we find an extremely high good fitting rate: 74.2% compared to only 59.2% in the case of «perfect» data. Neither bad overfitting nor bad underfitting is identified. Stated differently, both LAR and LASSO manage to select the relevant variables (X1 to X5) but add unnecessary variables in the case of good overfitting or select less than 5 true predictors in that of good underfitting.

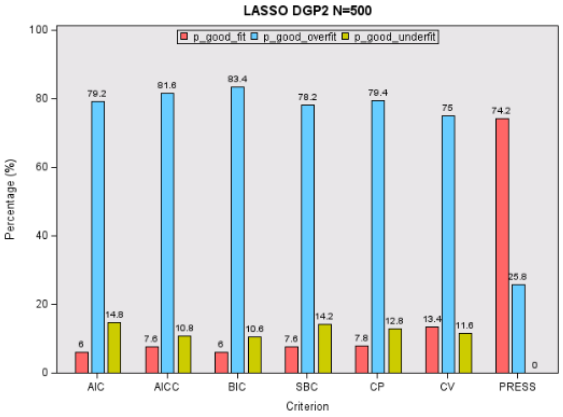


Fig. 5: LASSO with Internal correlation

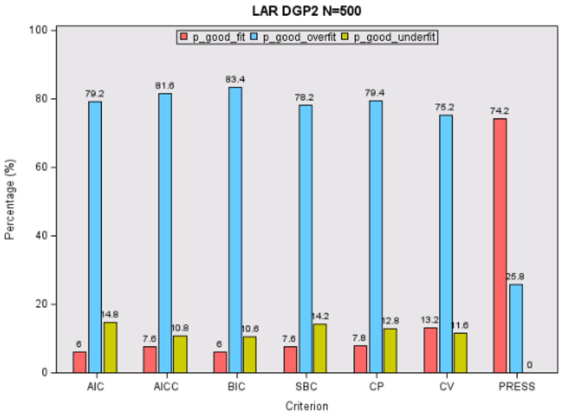


Fig. 6: LAR with Internal correlation

Significant changes can be noticed in the performance of ELASTIC NET in the presence of correlation between the first 5 features. Firstly, we see that the good fitting rate is extremely low, ranging from 0% to 1.4%, whereas it attained a maximum of 51.2% for the CP criterion in the independence case.

Most selected models using ELASTIC NET present either good overfitting or good underfitting. However, we observe the emergence of a new fitting behavior: bad overfitting. To put it differently, this metric refers to cases where the method selects more than 5 variables, and among them, there is at least one true variable that is missed. The bad overfitting rate is not negligible, going up to 22.4% while using the BIC criterion.

Lastly, our findings indicate 3% to 8.2% of almost good fitting. This metric captures the cases where the selected models contain exactly five variables but they are a combination of relevant and irrelevant predictors.

We conclude that the ELASTIC NET method yields a very poor performance in the presence of correlated variables.

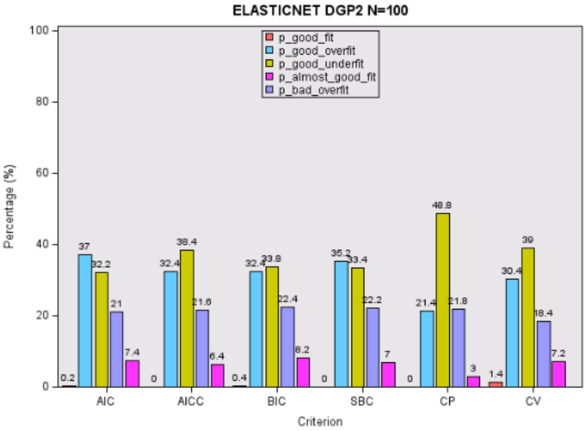


Fig. 7: Elastic Net in Independence

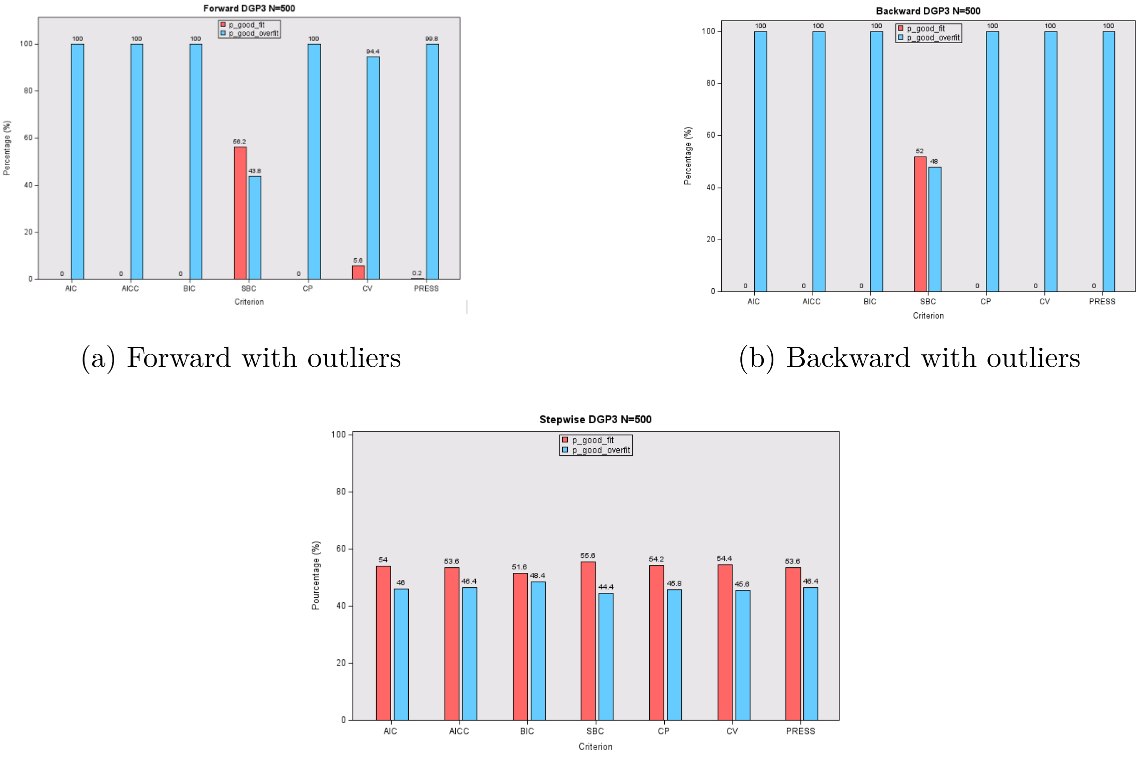
3.3 DGP 3 : Data with outliers

In this section, we intentionally introduced 5% of extreme values into the 3 first features. First, we used the «randnormal» function to generate the 50 non-correlated features, with no extreme values, following the exact same principle as in DGP1. Subsequently, we created a loop to iterate through all the observations of the first 3 predictors. Within this loop, we used the random uniform function as follows: if a value greater than 0.95 is generated, the corresponding observation was replaced with an extreme value drawn from a normal distribution with a mean of 4 and standard deviation of 1; otherwise, a normal distribution with a mean of 0 and a standard deviation of 1 was used.

Statistical Learning results

In this section, we introduced outliers to the first 3 explanatory features. However, their presence didn't cause notable differences when compared to DGP1, the reference case, indicating a consistent behavior of the statistical learning methods. Similarly to the preceding DGPs, the Schwarz Bayesian Criterion (SBC) systematically exhibited the highest percentage of good fitting for forward, backward, and stepwise methods, achieving rates of 56.2%, 52%, and 55.6% respectively.

To conclude, outliers don't seem to significantly affect the performance of Statistical Learning methods in our scenario, contrary to our initial expectations. We had expected a sharp drop in the good fitting rate due to outliers, but that wasn't the case with our dataset. Our focus will now shift to Machine Learning.



(a) Forward with outliers

(b) Backward with outliers

(c) Stepwise with outliers

Fig. 8: Model selection accuracy with Statistical Learning methods

Machine Learning results

Here, we extend the analysis by examining the graphical representation of the outcomes yielded by the considered ML methods : LAR, LASSO, ELASTIC NET, in the specific scenario of data containing outliers among the first 3 features. We will compare the performance of these algorithms in this specific instance to the one revealed in the reference case (DGP1).

We notice with little surprise that both methods LAR and LASSO exhibit comparable patterns in terms of good fitting, overfitting, and underfitting rates. In line with the reference case, no instances of bad overfitting or bad underfitting were identified. Moreover, the analysis reveals lower rates of good fitting compared to the reference case, ranging from 29% to nearly 36%, a result that totally aligns with our expectations.

Additionally, in contrast to the 1st type of data, our findings highlight lower rates of good overfitting counterbalanced by higher rates of good underfitting, up to nearly 26% in the case of AICC. The PRESS criterion remains the most optimal among the considered criteria, yielding a good fitting rate of 57.4%, just slightly less than that in the reference case.

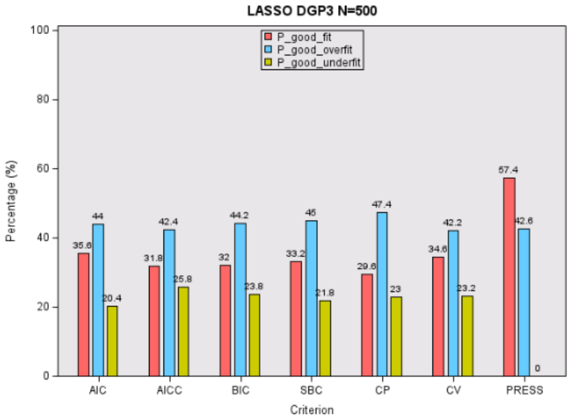


Fig. 9: LASSO with outliers

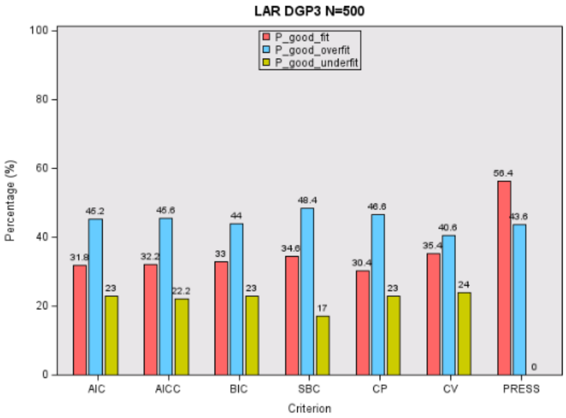


Fig. 10: LAR with outliers

The general patterns and their order remain similar to the reference case for all the considered criteria in the case of ELASTIC NET: overfitting is the most frequent, followed by good fitting, and lastly, underfitting. On average, we observe a decrease in the good fitting rate, accompanied by increases in both underfitting and overfitting. The only exception is the CP criterion combined with SBC as a stopping criterion: it exhibits a consistent good fitting rate and important adjustments between good overfitting and good underfitting rates. Notably, both reveal similar rates, 27.4% and 23% respectively, whereas the remaining criteria are much more prone to overfitting than underfitting.

Consequently, we can conclude that the considered algorithms combined with most criteria exhibit sensitivity to extreme values, leading to fluctuations in their performance through the model selection process.

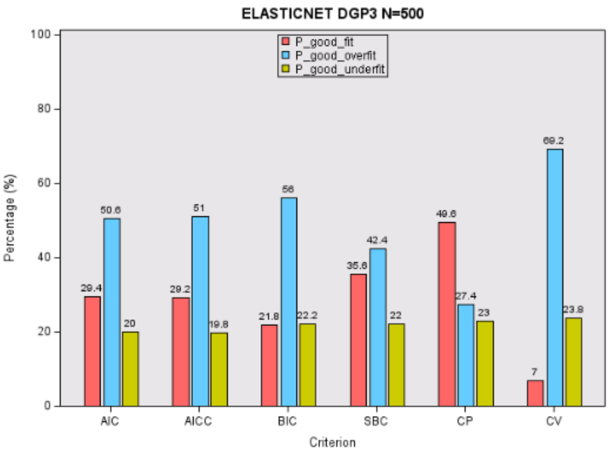


Fig. 11: Elastic Net with outliers

3.4 DGP 4 : Data with a combination of correlation among the explanatory variables and extreme values

The last DGP is aimed at introducing a mix between multicollinearity and outliers. It includes both correlation between the first 5 features, as presented in DGP2, and extreme values in the first 3 features, as seen in DGP3. For that matter, we used the Iman Conover transformation, known to approximately induce a specified correlation among component variables without changing the univariate distribution of the components.

Statistical Learning results

Focusing on both forward and backward methods, significant similarities are observed in their performance. The SBC stands out as the most effective criterion in terms of good fitting rate, with 55% for the forward method and 53.8% for the backward method. However, both demonstrate poorer performance when compared to the DGP1 scenario. We can notice that the good fitting rate for the forward method diminished from 57% to 55%, a decrease of 2% points. Similarly, there was a slight reduction of 0.2% points in terms of good fitting for K-fold Cross Validation (CV). The same trend was observed with the backward method, where a decrease of 3% points was noted as well. Otherwise, all the remaining criteria showed 100% of good overfitting. As for statistical learning methods, it's important to emphasize that incorporating both correlation and extreme values had a significant negative impact on model selection performance.

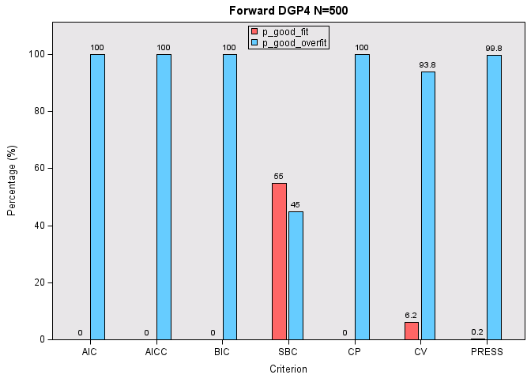


Fig. 12: Forward with a combination of correlation and outliers

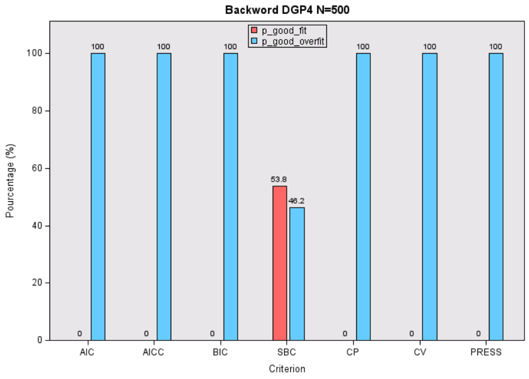


Fig. 13: Backward with a combination of correlation and outliers

Let's now turn our focus to stepwise selection where notable changes were observed, including the emergence of «good underfitting» for the first time with all criteria but SBC. We detected a slight decrease in the good fitting rate, conterbalanced by a marginal increase in the good underfitting rate (between 0.2% and 4.2%). It is important to highlight that the occurrence of good underfitting implies that the stepwise selection method misses certain true variables but does not include irrelevant ones.

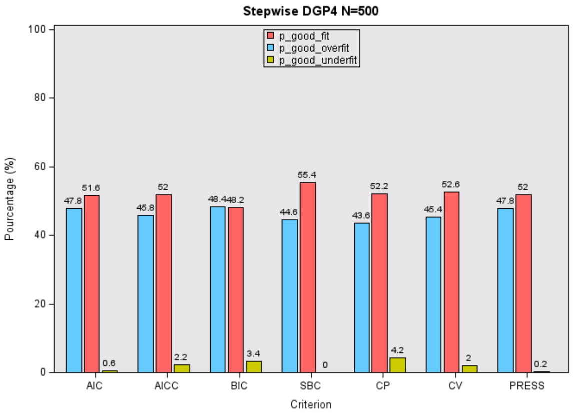


Fig. 14: Stepwise with a combination of correlation and outliers

Machine Learning results

Here, with nothing but a brief visual analysis of our graphs, we can see huge similarities in the performance of LASSO and LAR between the correlation case (DGP2) and this specific section. We could assume that correlation is at the core of their poor performance, as these methods had much higher success rates in the case of outliers.

When comparing the performance of LASSO and LAR in this section to that in the case of «perfect» data, we conclude that the presence of both correlation and outliers radically alters the quality of the provided outcomes. Let's further explore the LASSO method. Not only can we notice much higher good overfitting rates, going up to 75.8% with AIC, but also significant increase in the good underfitting rates. They range from 16.8% to 20.8% for most criteria. Once again, the PRESS criterion outperforms every other considered criterion, exhibiting 71.4% of good fitting, 27.8% of good overfitting and less than 1% of good underfitting. These results are all the more surprising taking into account that the success rate in the independent data scenario only attained 59.2%.

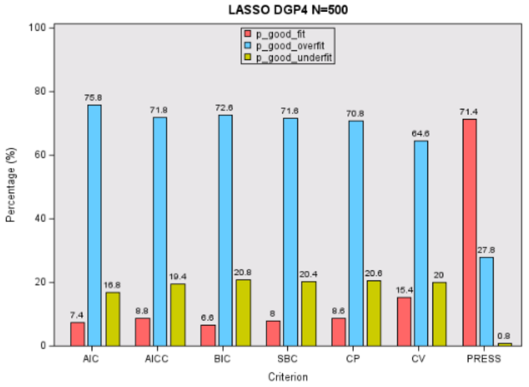


Fig. 15: LASSO with a combination of correlation and outliers

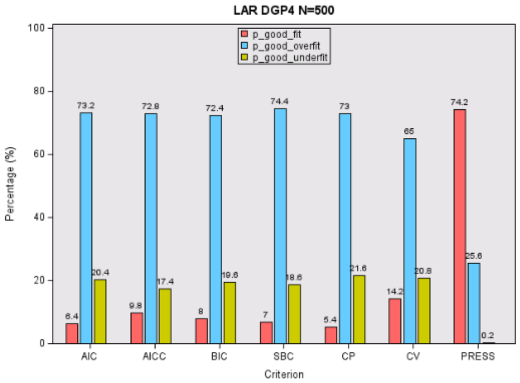


Fig. 16: LAR with a combination of correlation and outliers

We observe that the 5 relevant variables are selected consistently. However, in very few instances, the LASSO method fails to select X3, X4 and X5, thereby contributing to the observed underfitting rate (0.8%). The occurrence of selection of the remaining variables is relatively low and accounts for the overfitting rate (27.8%)

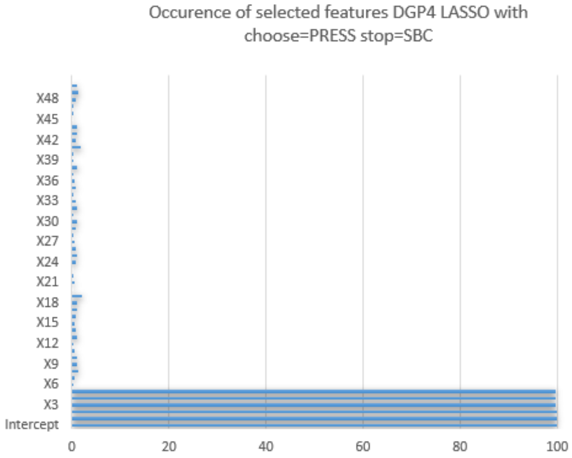


Fig. 17: Occurrences of selected features with LASSO

The ELASTIC NET method, on the other hand, yields consistently poor performance for all the considered criteria, with soaring rates of good overfitting. The CP criterion, which provided satisfactory outcomes in the first section, seems to be particularly sensitive to the incorporated correlation patterns. The CV criterion combined with the SBC stopping criterion appears to remain unaltered in the presence of the outliers-correlation mixture, with a good fitting rate of nearly 14%. Additionally, we notice the emergence of bad overfitting, most likely triggered by the correlation patterns, as they seem to more drastically affect the performance of ELASTIC NET than outliers.

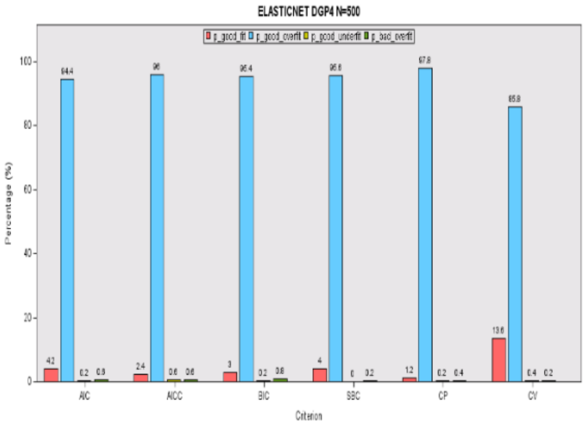


Fig. 18: Elastic Net with a combination of correlation and outliers

Let's now assess the severity of good overfitting in the case of ELASTIC NET. Unlike the patterns observed with the DGP1, the overfitting in this case is mostly severe, as the selected models that exhibit good overfitting contain 3 or more irrelevant variables at least 70% of the time.

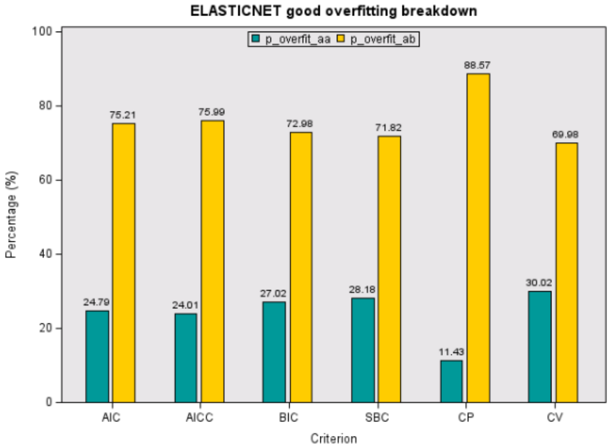


Fig. 19: Elastic Net with goodoverfitting breakdown

3.5 Impacts of the size of the sample

Last but not least, we will evaluate the fluctuations in the performance of the considered statistical learning and machine learning methods associated with changes in the sample size. In the previous sections, we conducted a comprehensive analysis of the success rates in model selection considering a large sample size, namely N=500. Let's now shift our focus to a real-life scenario of data scarcity by considering a smaller sample size: N=100.

DGP1 with Independence

Statistical Learning results

With Statistical learning methods, as the sample size diminishes, we observe a decrease in the good fitting rate across various criteria. As seen in the left graph, the forward method yielded much poorer results in terms of good fitting with the SBC criterion, dropping from 57% to 17%, which represents a 40% points decrease. Backward performed even worse than forward, undergoing almost 48% points of decrease with the SBC criterion. This trend was also noted with stepwise, as shown on the right graph, where a huge reduction in the good fitting rate was observed across all criteria compared to the DGP1 results. To conclude, the model fit was negatively impacted by the sample size, with a much stronger tendency towards good overfitting for all the considered statistical learning methods.

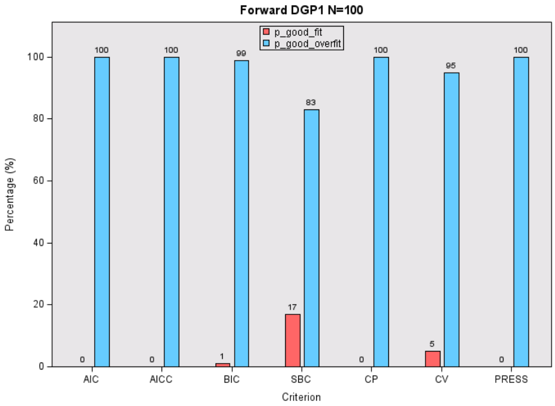


Fig. 20: Forward with independence using N=100

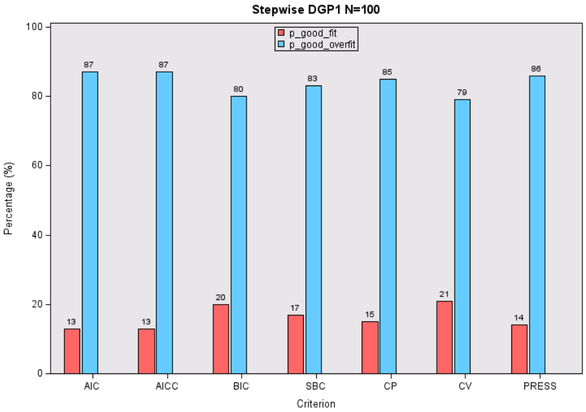


Fig. 21: Stepwise with independence using N=100

Machine Learning results

Both the LASSO and the ELASTIC NET methods are particularly sensitive to the sample size variation in the reference case (DGP1). Consequently, we notice a significant drop in the good fitting rates counterbalanced by a prominent rise in good underfitting rates. We conclude that limited data leads to significantly diminished performance of the considered ML methods in the case of data with perfect characteristics.

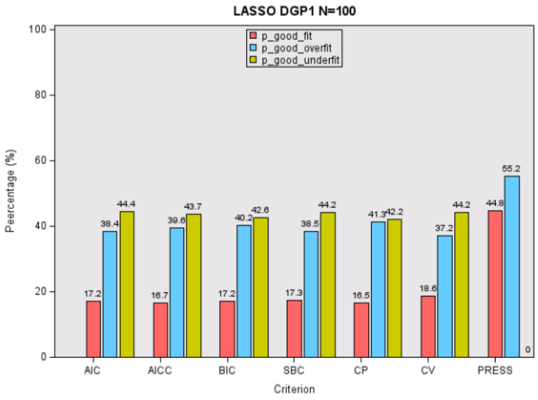


Fig. 22: LASSO with independence using N=100

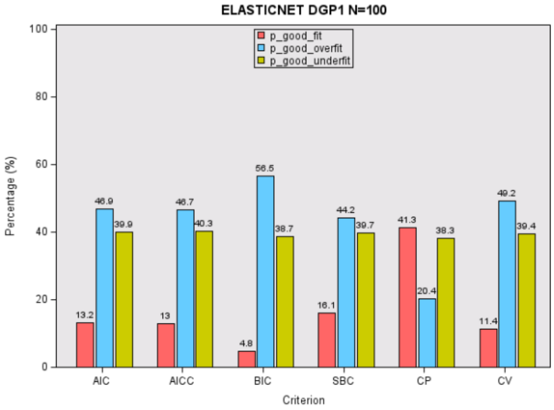


Fig. 23: Elastic Net with independence using N=100

DGP4 with correlation and outliers

To further explore this intricate question, we address the most extreme scenario, the one with both outliers and correlation combined with a small sample size.

Statistical Learning results

As previously noted with DGP1, both the forward and stepwise methods demonstrate sensitivity to variations in sample size. Specifically, with the forward method, we observe a decline in the good fitting rate from 54.8% when N=500 to 14% when N=100. However, we also note 4.8% points increase for cross-validation (CV) and the emergence of good fitting patterns for the BIC criterion (2%).

Concerning the stepwise method, a similar decrease in the good fitting rate is observed, along with a greater tendency towards good overfitting. However, by reducing the sample size, we observe the absence of instances of good underfitting for all the criteria. It suggests that stepwise is able to identify the true 5 explanatory variables along with other variables in the majority of cases.

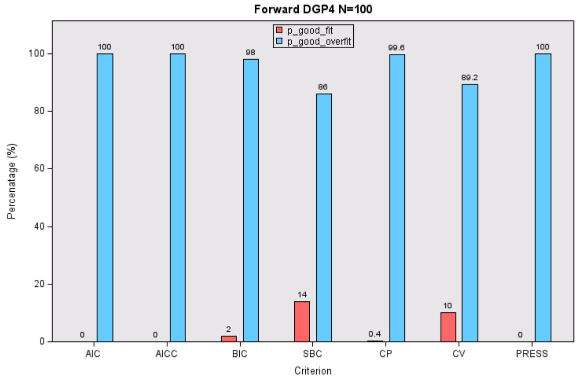


Fig. 24: Forward with correlation and outliers using N=100

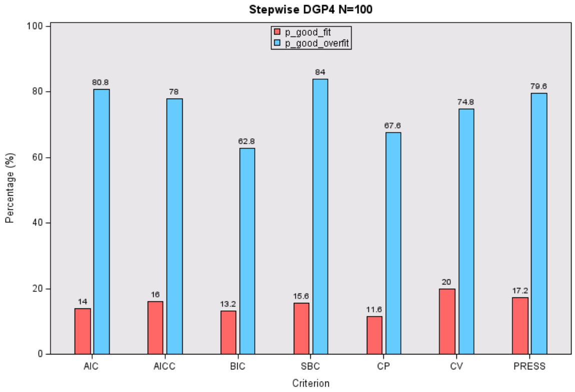


Fig. 25: Stepwise with correlation and outliers using N=100

Machine Learning results

Our results confirm once again that the ML methods are significantly affected by the sample size. As for the LASSO method, we notice the emergence of two new fitting behaviors: bad overfitting and bad underfitting, which means that it fails to select the true variables and results in a mix of relevant and irrelevant predictors. Moreover, the combination that gave the highest success rate in the DGP4 (71.4%), PRESS with SBC as a stopping criterion, no longer significantly outperforms the other criteria, exhibiting notable good underfitting and overfitting rates. To conclude, this graph illustrates constant mediocre outcomes for the LASSO method in the small sample size case.

Regarding the ELASTIC NET method, it is important to highlight its poor performance in the model selection process when the number of observations is set at 100. Its good fitting rate is the smallest of all, and is equal to 0% in the case of AICc, BIC, SBC and CP. Counterintuitively, the AIC criterion marginally outperforms the AICc criterion, which is expected to exhibit higher success rates as it is adapted to combat overfitting in limited data scenarios. Lastly, the emergence of bad underfitting is also a sign of outcome deterioration.

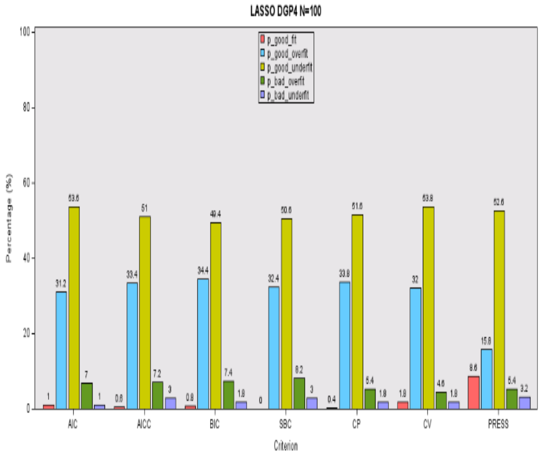


Fig. 26: LASSO with correlation and outliers using N=100

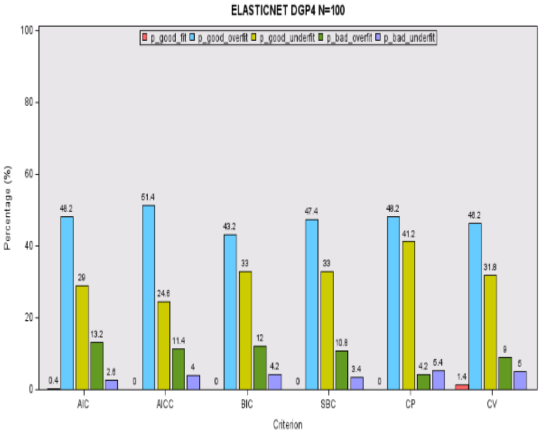
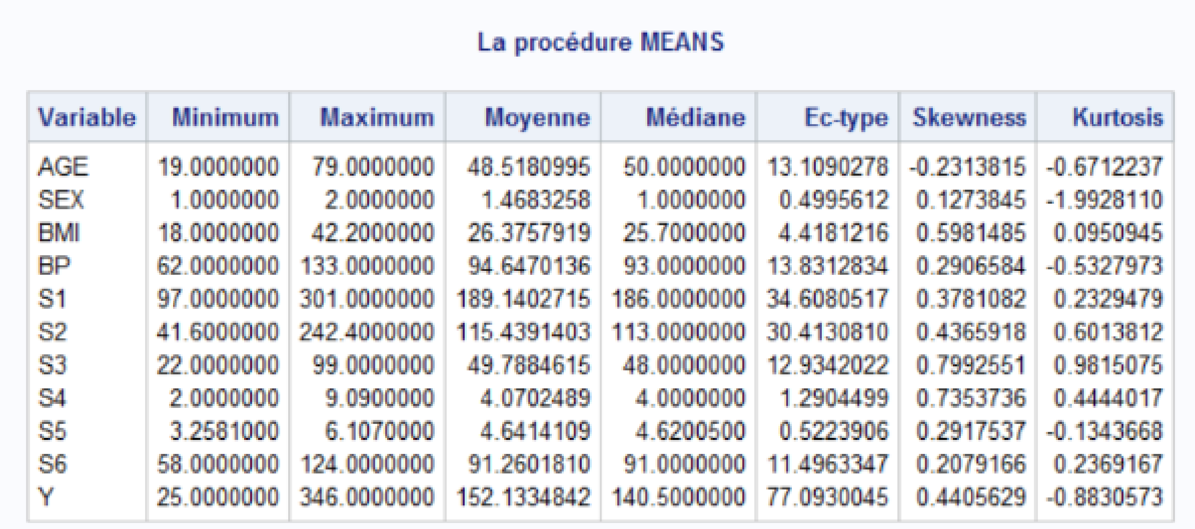


Fig. 27: Elastic Net with correlation and outliers using N=100

Chapter 4

Empirical Application

Finally, we will present an empirical analysis as the ultimate step of our research. In order to implement the findings derived from the research phase, we chose a dataset related to the study of diabetes containing 442 observations. The 10 variables of the database include both individual characteristics (such as age and sex) and health parameters (blood pressure, six blood serum measurements). The target variable, denoted as Y, is a quantitative measure that illustrates the progression of the disease one year after baseline.



La procédure MEANS

Variable	Minimum	Maximum	Moyenne	Médiane	Ec-type	Skewness	Kurtosis
AGE	19.0000000	79.0000000	48.5180995	50.0000000	13.1090278	-0.2313815	-0.6712237
SEX	1.0000000	2.0000000	1.4683258	1.0000000	0.4995612	0.1273845	-1.9928110
BMI	18.0000000	42.2000000	26.3757919	25.7000000	4.4181216	0.5981485	0.0950945
BP	62.0000000	133.0000000	94.6470136	93.0000000	13.8312834	0.2906584	-0.5327973
S1	97.0000000	301.0000000	189.1402715	186.0000000	34.6080517	0.3781082	0.2329479
S2	41.6000000	242.4000000	115.4391403	113.0000000	30.4130810	0.4365918	0.6013812
S3	22.0000000	99.0000000	49.7884615	48.0000000	12.9342022	0.7992551	0.9815075
S4	2.0000000	9.0900000	4.0702489	4.0000000	1.2904499	0.7353736	0.4444017
S5	3.2581000	6.1070000	4.6414109	4.6200500	0.5223906	0.2917537	-0.1343668
S6	58.0000000	124.0000000	91.2601810	91.0000000	11.4963347	0.2079166	0.2369167
Y	25.0000000	346.0000000	152.1334842	140.5000000	77.0930045	0.4405629	-0.8830573

Fig. 1: Descriptive statistics of our Database

Certain general trends can be noticed in the data. First, we see high variability in the outcome variable Y, with a standard deviation of 77.09.

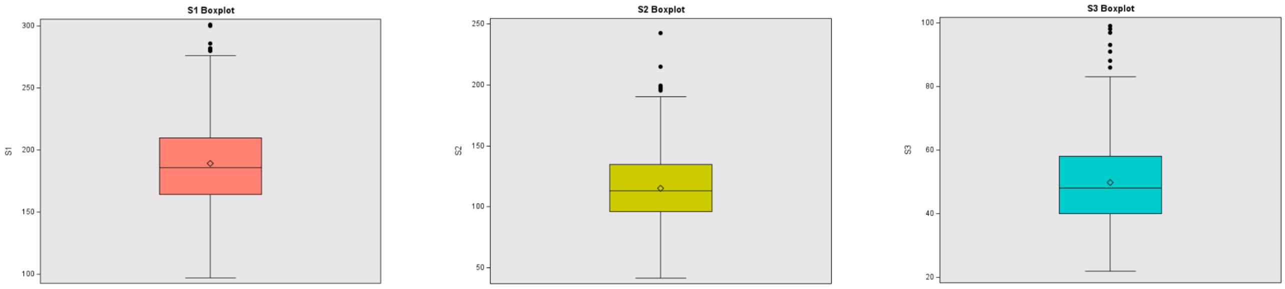
The Skewness provides additional information about the symmetry of the distribution and is, in this specific case, positive for most considered variables. Moreover, the mean tends to be slightly higher than the median. Hence, we could expect potential outliers in the higher end of the data range.

The S3 and S4 appear to exhibit the most pronounced rightward skewness. In contrast, the only left skewed variable is age but the associated asymmetry is moderate.

Kurtosis is a statistical measure that describes the shape of a distribution, specifically the degree to which a distribution's tails differ from those of a normal distribution.

In the last column, we notice the measure of excess kurtosis. Most health parameters, such as S2, S3, S4 yield a leptokurtic distribution, with more pronounced tails and a higher likelihood of outliers.

As we suspected the presence of outliers, let's introduce a graphical visualization to confirm it:



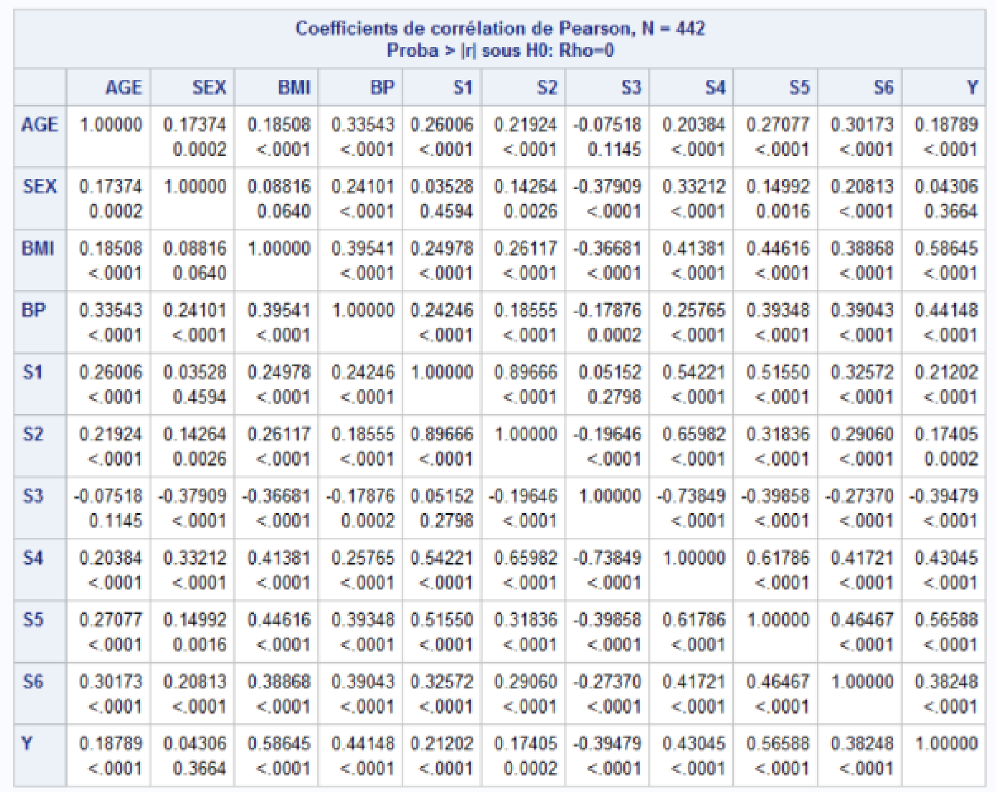
(a) S1's Boxplot

(b) S2's Boxplot

(c) S3's Boxplot

Fig. 2: Three Boxplots

Next, we will delve into the correlation analysis. Let's take a look at the correlation matrix.



Coefficients de corrélation de Pearson, N = 442
Proba > |r| sous H0: Rho=0

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
AGE	1.00000	0.17374 0.0002	0.18508 <.0001	0.33543 <.0001	0.26006 <.0001	0.21924 <.0001	-0.07518 0.1145	0.20384 <.0001	0.27077 <.0001	0.30173 <.0001	0.18789 <.0001
SEX	0.17374 0.0002	1.00000	0.08816 0.0640	0.24101 <.0001	0.03528 0.4594	0.14264 0.0026	-0.37909 <.0001	0.33212 <.0001	0.14992 0.0016	0.20813 <.0001	0.04306 0.3664
BMI	0.18508 <.0001	0.08816 0.0640	1.00000	0.39541 <.0001	0.24978 <.0001	0.26117 <.0001	-0.36681 <.0001	0.41381 <.0001	0.44616 <.0001	0.38868 <.0001	0.58645 <.0001
BP	0.33543 <.0001	0.24101 <.0001	0.39541 <.0001	1.00000	0.24246 <.0001	0.18555 <.0001	-0.17876 0.0002	0.25765 <.0001	0.39348 <.0001	0.39043 <.0001	0.44148 <.0001
S1	0.26006 <.0001	0.03528 0.4594	0.24978 <.0001	0.24246 <.0001	1.00000	0.89666 <.0001	0.05152 0.2798	0.54221 <.0001	0.51550 <.0001	0.32572 <.0001	0.21202 <.0001
S2	0.21924 <.0001	0.14264 0.0026	0.26117 <.0001	0.18555 <.0001	0.89666 <.0001	1.00000	-0.19646 <.0001	0.65982 <.0001	0.31836 <.0001	0.29060 <.0001	0.17405 0.0002
S3	-0.07518 0.1145	-0.37909 <.0001	-0.36681 <.0001	-0.17876 0.0002	0.05152 0.2798	-0.19646 <.0001	1.00000	-0.73849 <.0001	-0.39858 <.0001	-0.27370 <.0001	-0.39479 <.0001
S4	0.20384 <.0001	0.33212 <.0001	0.41381 <.0001	0.25765 <.0001	0.54221 <.0001	0.65982 <.0001	-0.73849 <.0001	1.00000	0.61786 <.0001	0.41721 <.0001	0.43045 <.0001
S5	0.27077 <.0001	0.14992 0.0016	0.44616 <.0001	0.39348 <.0001	0.51550 <.0001	0.31836 <.0001	-0.39858 <.0001	0.61786 <.0001	1.00000	0.46467 <.0001	0.56588 <.0001
S6	0.30173 <.0001	0.20813 <.0001	0.38868 <.0001	0.39043 <.0001	0.32572 <.0001	0.29060 <.0001	-0.27370 <.0001	0.41721 <.0001	0.46467 <.0001	1.00000	0.38248 <.0001
Y	0.18789 <.0001	0.04306 0.3664	0.58645 <.0001	0.44148 <.0001	0.21202 <.0001	0.17405 0.0002	-0.39479 <.0001	0.43045 <.0001	0.56588 <.0001	0.38248 <.0001	1.00000

Fig. 3: Correlation Matrix

To enhance clarity, we could take a look at the heatmap. The dark red shades that persist in the graph below suggest high positive correlation patterns between the variables, whereas the dark blue ones indicate strong negative correlation. For instance, S1 and S2 are highly positively correlated, while S3 and S4 seem to be strongly negatively correlated.

To summarize the descriptive statistics section, we conclude that this database is associated with both the presence of outliers and correlation, resembling the real-life data scenario DGP4.

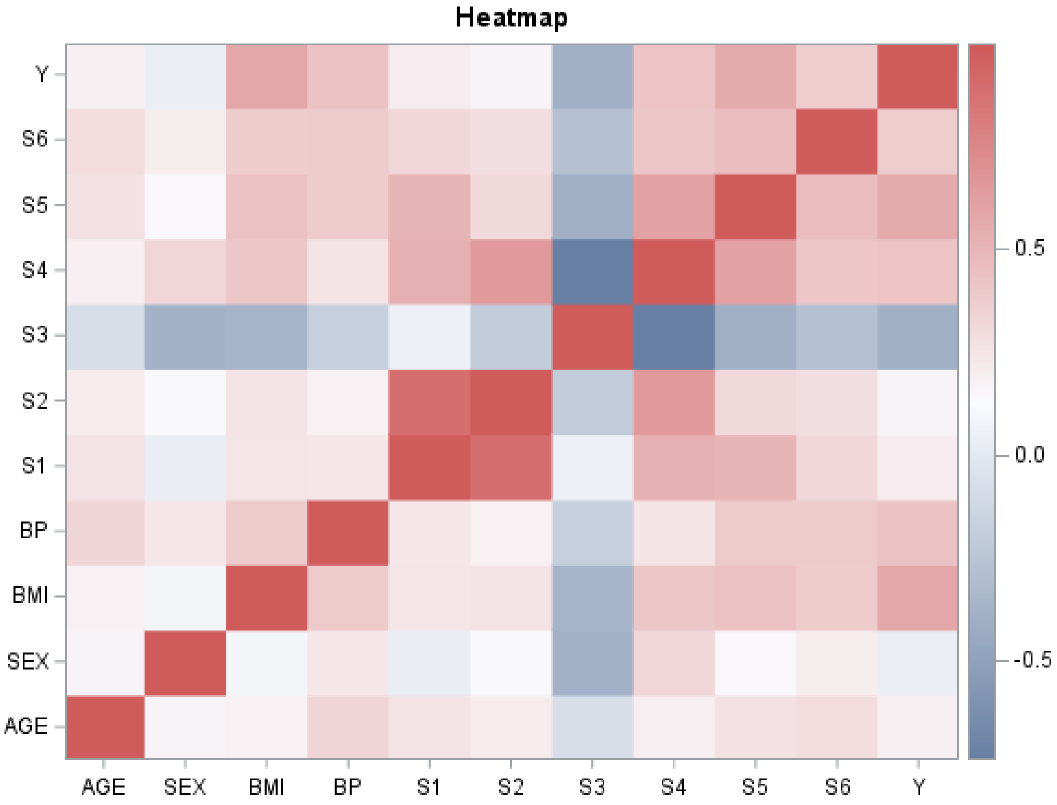
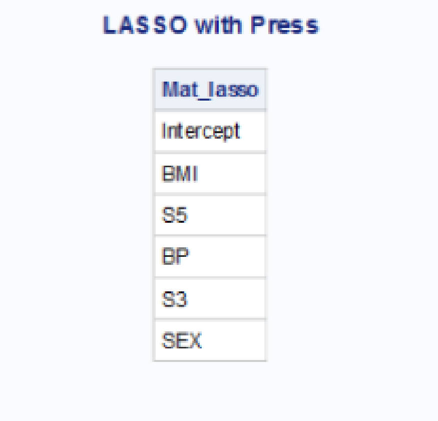


Fig. 4: Heatmap

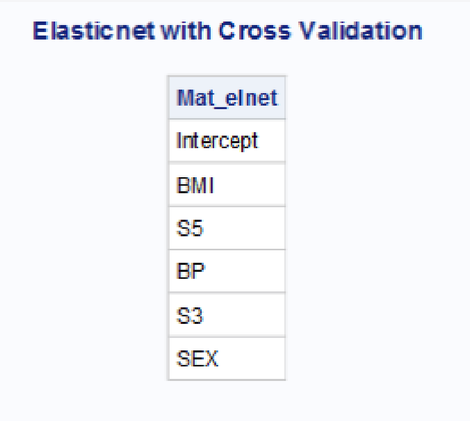
As we are in a case with both outliers and correlation between our explanatory variables (as seen in DGP4), we expected that machine learning methods would perform better in terms of selected variables than the statistical learning techniques. We notice that both LASSO (combined with PRESS) and ELASTIC NET (combined with CV) methods yield the same outcomes in terms of model selection. They choose the same variables: SEX, BMI, BP, S3, and S5.



LASSO with Press

Mat_lasso
Intercept
BMI
S5
BP
S3
SEX

Fig. 5: results of Lasso's selection

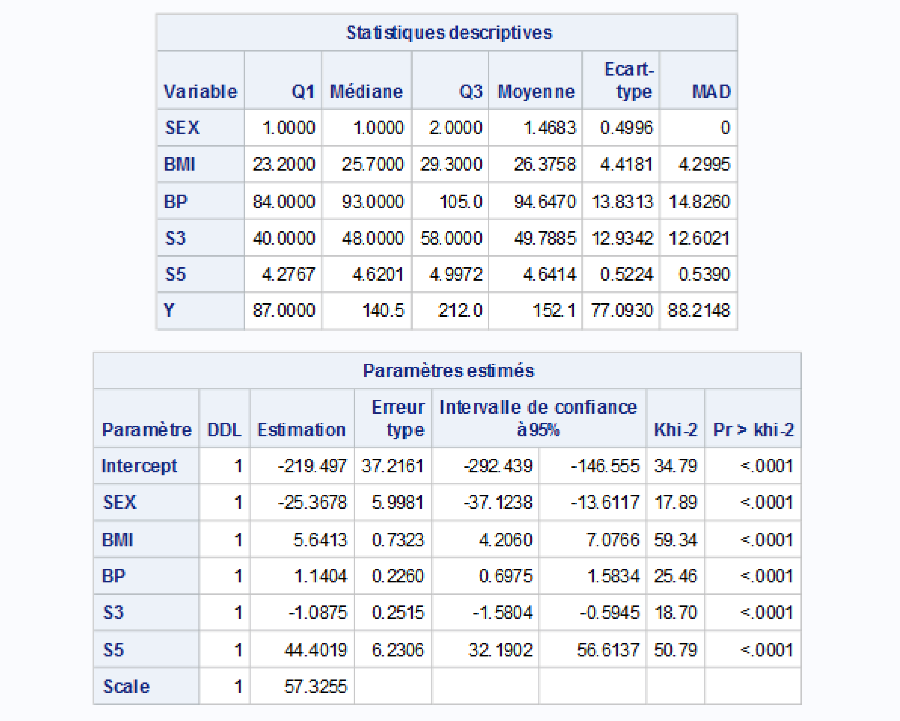


Elasticnet with Cross Validation

Mat_elnet
Intercept
BMI
S5
BP
S3
SEX

Fig. 6: results of Elastic Net's selection

Moreover, while examining the data, numerous irregularities that could alter the outcomes were identified: outliers, correlation, heteroskedasticity. Therefore, we chose not to preprocess the data but instead proceeded with the ROBUSTREG procedure applied to the model selected by LASSO combined with the PRESS stopping criterion. Indeed, the robust regression technique is esentially aimed at providing robust results in the presence of outliers. [5] As you can notice in the result table below, all the variables are highly significant at the 1% threshold, and even at more stringent levels with a systematic p-value < 0.0001.



Statistiques descriptives

Variable	Q1	Médiane	Q3	Moyenne	Ecart-type	MAD
SEX	1.0000	1.0000	2.0000	1.4683	0.4996	0
BMI	23.2000	25.7000	29.3000	26.3758	4.4181	4.2995
BP	84.0000	93.0000	105.0	94.6470	13.8313	14.8260
S3	40.0000	48.0000	58.0000	49.7885	12.9342	12.6021
S5	4.2767	4.6201	4.9972	4.6414	0.5224	0.5390
Y	87.0000	140.5	212.0	152.1	77.0930	88.2148

Paramètres estimés

Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance à 95%		Khi-2	Pr > khi-2
Intercept	1	-219.497	37.2161	-292.439	-146.555	34.79	<.0001
SEX	1	-25.3678	5.9981	-37.1238	-13.6117	17.89	<.0001
BMI	1	5.6413	0.7323	4.2060	7.0766	59.34	<.0001
BP	1	1.1404	0.2260	0.6975	1.5834	25.46	<.0001
S3	1	-1.0875	0.2515	-1.5804	-0.5945	18.70	<.0001
S5	1	44.4019	6.2306	32.1902	56.6137	50.79	<.0001
Scale	1	57.3255					

Fig. 7: PROC ROBUSTREG results

Conclusion

As we come to the final stages of our examination, it is important to highlight the broad conclusions drawn from our research findings.

Reiterating the fundamentals, it is worth mentioning that our research encompasses both Statistical and Machine Learning techniques in the context of model selection. As we perfectly controlled the data generation processes, we were able to precisely evaluate the performance of the considered methods across four data set scenarios: «perfect» data, data with multicollinearity, data with outliers, and data combining outliers and multicollinearity. With little surprise, our findings revealed that real-world data commonly presents with outliers and multicollinearity, as observed in DGP4.

To begin with, a significant tendency towards overfitting was found across statistical learning techniques. For instance, the BACKWARD method demonstrated poor results, with most criteria resulting in complete overfitting patterns, except for SBC. As for the FORWARD technique, it yielded mixed results, with some criteria exhibiting good fitting patterns (SBC and k-fold Cross validation), while all the remaining ones exclusively resulted in overfitting. In contrast, the STEPWISE selection delivered improved performance, achieving good fitting across all criteria, and exhibiting a relatively reduced inclination towards overfitting. Moreover, statistical learning methods exhibited decreased performance when multicollinearity and outliers were simultaneously incorporated (case of DGP4) for most stopping criteria, and showed particular sensitivity to changes in the sample size.

To continue with Machine Learning approaches, they yield a much better performance in the specific context of model selection, compared to Statistical Learning. The only challenge is to properly adapt the stopping criteria in order to achieve the highest possible good fitting rate. Our findings indicate similar results for LASSO and LAR methods, whereas the ELASTIC NET technique has its own particularities in terms of outcomes.

As for LASSO and LAR, they exhibit high good fitting rates in the case of «perfect» data with both SBC and PRESS criteria. However, as we alter the data, by incorporating outliers and multicollinearity, most criteria fail to consistently select the accurate model and result in severe overfitting patterns. In fact, the PRESS criterion is the only one to remain robust across all data scenarios, consistently yielding good results. It is noteworthy that LASSO and LAR seem to present particular sensitivity to changes in the sample size, as the results significantly worsen.

The ELASTIC NET method seems to be naturally inclined towards overfitting patterns across all types of data. Nevertheless, the degree of severity of good overfitting differs, ranging from moderate in case of «perfect» data (at most 2 irrelevant variables are selected) to extremely severe across altered data (3 or more irrelevant variables are selected). When combined with the stopping criterion CP, the ELASTIC NET method yields satisfactory outcomes in the case of «perfect» data and data incorporating outliers. We also observe that multicollinearity has a more devastating impact on the outcomes than outliers for all the considered Machine Learning methods and for most criteria.

Discussion

Statistical and Machine Learning techniques both present specific strengths and weaknesses. Therefore, a broad understanding of their mechanisms is fundamental for comprehensive analysis and modeling using real-life data. The empirical application emphasized the importance of conducting exploratory data analyses prior to using model selection techniques. This imperative step provides insight into the particularities of the data patterns, thereby enabling us to discern which procedures are more susceptible to yield satisfactory results.

In light of our research paper, there are several areas where improvements could be made. As for the data patterns, we could have explored an alternative database case featuring a structural break in order to assess the performance of Statistical and Machine Learning methods in the context of model selection. Moreover, it would have been interesting to analyze the case of correlation between all the features or to incorporate outliers in more than 3 variables. Finally, we could have combined these data irregularities in one single DGP in order to evaluate the accuracy of the outcomes. Another limitation could be addressed by introducing randomly chosen values of β. This would bring insight into the potential impact of the β-vector values on the selection of the associated features. The values could be chosen in an arbitrary manner using normal distributions with different standard deviation patterns.

Lastly, we could assess the performance of the considered methods by systematically adjusting the standard deviation of the error term. In our paper, we limited our analysis to only two scenarios (0.1 and 0.05), but exploring a wider range could provide valuable insights.

As for the considered methods, it would be intriguing to introduce the Incremental Forward Stagewise method, as well as the XGBoost approach in order to evaluate and compare their performances in the context of model selection.

Bibliography

[1] Michele Bennett, Karin Hayes, Ewa J Kleczyk, and Rajesh Mehta. Similarities and differences between machine learning and traditional advanced statistical modeling in healthcare analytics. *arXiv preprint arXiv:2201.02469*, 2022.

[2] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer, 2003.

[3] Iain Carmichael and JS Marron. Data science vs. statistics: two cultures? *Japanese Journal of Statistics and Data Science*, 1:117–138, 2018.

[4] Joseph E Cavanaugh and Andrew A Neath. Generalizing the derivation of the schwarz information criterion. *Communications in Statistics-Theory and Methods*, 28(1):49–66, 1999.

[5] Colin Chen. Paper 265-27 robust regression and outlier detection with the robustreg procedure. In *Proceedings of the Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*. Citeseer, 2002.

[6] Douglas Bates and Martin Mächler and Ben Bolker and Steven Walker. Linear Mixed-Effects Models using Eigen and S4. https://www.rdocumentation.org/packages/lme4/versions/0.6-5/topics/BIC, 2017.

[7] GeeksforGeeks. Least Angle Regression (LARS). https://www.geeksforgeeks.org/least-angle-regression-lars/, 25 Jan, 2023.

[8] Omid Ghorbanzadeh, Hejar Shahabi, Fahimeh Mirchooli, Khalil Valizadeh Kamran, Samsung Lim, Jagannath Aryal, Ben Jarihani, and Thomas Blaschke. Gully erosion susceptibility mapping (gesm) using machine learning methods optimized by the multi-collinearity analysis and k-fold cross-validation. *Geomatics, Natural Hazards and Risk*, 11(1):1653–1678, 2020.

[9] Harvard IACS. Model selection with cross-validation. Online presentation, 2021.

[10] Trevor Hastie. Elastic net talk. Online presentation, Date de la présentation.

[11] R. R. Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[12] Masahito Kobayashi and Shinichi Sakata. Mallows' cp criterion and unbiasedness of model selection. *Journal of Econometrics*, 45(3):385–395, 1990.

[13] SAS. SAS Enterprise Miner Documentation. https://documentation.sas.com/doc/en/emref/14.3/n0ovmb476la6kon1iyyfkwb2qx8r.htm, 2017.

[14] SAS Institute Inc. GLMSELECT SAS/STAT User's Guide. https://support.sas.com/documentation/onlinedoc/stat/131/glmselect.pdf, 2022.

[15] Jun Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

[16] Mary L. Thompson. Selection of variables in multiple regression: Part i. a review and evaluation. *International Statistical Review / Revue Internationale de Statistique*, 46(1):1–19, 1978.

[17] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.