M1 Econometrics and Statistics 2023-2024

Machine Learning
VS
Statistical Learning

Issame ABDELJALIL, Brice SOPGUOMBUE,
Moustafa MARZOUK

Directed by : Philippe DE PERETTI

February 14, 2024

Machine Learning vs Statistical Learning

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Abstract

Variable selection is a crucial step when building predictive models, as it enables the best possible predictions to be made by retaining the best variables and removing any that appear unnecessary for the model. In this paper, we focused on comparing variable selection methods between Machine Learning and Statistical Learning techniques. To this end, we were able to test the performance of several of these techniques on simulated data, and with the help of several selected stopping criteria we were able to compare the latter in order to obtain the best possible results.
The aim of this thesis was also to gradually increase the difficulty of our analysis in order to show the limits of variable selection methods. We started with restrictive assumptions about the simulated data, which we gradually relaxed to distinguish the evolution of results according to the techniques, and to understand which algorithm performs best according to the assumptions made. Then, using the lessons learned from the simulated data, we decided to apply the selected techniques to data from a diabetes database, in order to draw conclusions.
In fine, we have seen that it was not possible to choose between Machine Learning and Statistical Learning variable selection methods. Indeed with Statistical Learning, the algorithms strike a balance between overfitting and perfect fitting. On the other hand, Machine Learning algorithms are more varied, encompassing both underfitting (good or bad underfitting) and overfitting, depending on the criteria considered.

Master 1 Econometrics and Statistics

February 14, 2024

Contents

Introduction

Econometrics, an essential tool for analyzing relationships between variables, is the result of a fusion between the disciplines of economics and statistics. The emergence of this discipline at the beginning of the 20th century was honoured by the Nobel Prize in Economics, in recognition of the pioneering work of Ragnar Frisch and Jan Tinbergen in 1969. Based on economic theory, econometrics aims to quantify causal relationships and understand underlying mechanisms, validating these models through the use of empirical data. This discipline, which is constantly evolving to incorporate theoretical advances, remains crucial to understanding global economic dynamics.

At the same time, the advent of machine learning and statistical learning has enriched the landscape of economic analysis. These fields, focused on prediction and optimization, draw on the foundations of econometrics while using innovative algorithmic approaches to extract complex patterns from large datasets.

Rooted in the development of applied statistics, statistical learning has made significant progress over the course of the 20th century. This field offers powerful tools for solving classification and regression problems by modeling relationships between variables using statistical methods. The models, often formulated under assumptions such as normality, aim to establish relationships between variables, thereby drawing conclusions and interpretations.

Arthur Samuel, the American computer scientist who pioneered artificial intelligence, introduced the term "machine learning" in 1959 after creating his checkers program for IBM in 1952, which successfully beat a self-styled American checkers champion. This research marked a milestone in the development of artificial intelligence. The main aim of machine learning is to create models capable of learning from data, and predicting, enabling computer systems to generalize and make decisions without explicit programming, paving the way for automation, i.e. "learning by doing".

Thus, the goal of machine learning is to predict; interpretability is often not a major concern. To evaluate the performance of a machine learning model, we mainly use metrics that validate the model's predictions, in order to minimize prediction errors. In contrast, statistical learning models are evaluated using tests, confidence intervals and so on.

Although statistical learning models can make predictions, this is not their ultimate goal. Conversely, machine learning models can enable interpretation, but this is at the expense of the emphasis on prediction. To increase model accuracy, both fields rely on iterative techniques, massive volumes of data and powerful computing resources.

In this paper, we focus on these two categories of variable selection methods. On the statistical learning side, we will examine forward, backward and stepwise methods, while on the machine learning side, we will look at LARS, LASSO, Elasticnet, using in particular the

GLMselect procedure integrated into SAS. Our aim is to understand which methods and stopping criteria, by exploiting their statistical strengths, will best succeed in finding the variables of our later model, by calculating the probabilities of perfect-fitting, over-fitting and under-fitting. We'll start with fairly restrictive assumptions, such as normality and the absence of collinearity between variables. Gradually, we will challenge them, introducing collinearity, outliers and a combination of the two, in order to handle data reflecting the complexity of our everyday environment.
In the first section, we will explain in detail the different methods, the stopping criteria, our metrics (which will be used to judge the quality of our results) and present the formulation of our five types of data-generating processes. Section 2, which will be the most challenging, will enable us to compare the methods with each other, and to identify those best suited to each type of data. Finally, our paper will conclude with an empirical study of the diabetes database, so that we can apply our results obtained in section 2.

Chapter 1

Presentation of the theoretical framework: the different methods used

1.1 Statistical Learning

The aim of these three techniques is to identify the simplest model that best fits the data. The coefficient of determination (R-squared) alone is not a sufficient indicator of model quality. Excessive inclusion of regressors carries the risk of overfitting.
As econometricians and statisticians, we select the input or output criterion for a variable using various metrics such as R-squared, p-value, F-test, AIC, BIC and so on.
Let's take the BIC as an example: it defines a variable's entry/exit criterion as a function of sample size. The larger the sample, the smaller the p-value defined by the BIC.

1.1.1 Forward

Once the variable is integrated into the model, it remains within it. The process starts with an initial model, $y = constant$.
Then, the first independent variable most correlated with y is added to the model. It is crucial to note that the increase in R2, while logical with the addition of variables, does not necessarily guarantee their relevance. If a variable is not significant, it is automatically excluded from the model.
At the end of the first step, the model is formulated as follows: $y = constant + b_i X_i$.
In the next step, we identify another variable that is highly correlated with y, i.e. one that increases the R2 or reduces the AIC criterion, etc., and include it in the model. This process is repeated iteratively.
It is important to note that the contribution of each variable to the variance may change during the process, as some variables may influence each other. The addition of variables ceases as soon as a variable is no longer significant, or no longer reduces the AIC, etc.

1.1.2 Backward

The Backward regression is the exact opposite of forward regression, as we start with the complete model. In the first step, we evaluate the model's behavior by removing an explanatory variable (testing each explanatory) and eliminating the variable with the least significant p-value. We examine whether the change in the coefficient of determination (R-squared) is significant. If no non-significant change is observed (i.e. if the variable explains virtually nothing), this variable is permanently excluded from the model, and the model is updated. We repeat this step as long as the removal of variables from the model does not significantly impact the R-squared, or as long as there are no more insignificant explanatory variables in the model, and so on.

1.1.3 Stepwise Selection

This method combines the two models previously presented, with one major difference: a newly added variable can also be removed, just as an excluded variable can be reintegrated. A concrete example of how this approach works is as follows: starting with an empty model, we introduce the regressor with the lowest p-value. In the second step, each model comprising two regressors is evaluated and the regressor with the lowest p-value is integrated. This process is repeated iteratively.

However, it is crucial to emphasize that at each stage, i.e. each time a new variable is added, it is necessary to check whether the previously added variables remain significant. Indeed, the addition of variables can influence the significance of others, and if they are no longer significant, they must be removed from the model.

The process ends when no more significant variables can be added.

1.2 Machine Learning

1.2.1 Ridge Regression

The idea behind this method is to offer an alternative to OLS when faced with high variance due to multi-collinearity between variables (=risk of overfitting).

By using Ridge regression, we reduce the variance and increase the bias to obtain a better MSE than that obtained with OLS in the presence of collinearity.

To the OLS expression, we add the SCR multiplied by a lambda term that controls the intensity of the penalty applied:

- If $\lambda = 0$, we end up with the OLS expression.
- If λ tends towards infinity, the coefficients will tend towards 0 without ever reaching it, leading to under-fitting.

Thanks to Ridge regression, we have regularization but no variable selection.

$$RSS = \textstyle\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

1.2.2 Lasso Regression

This technique is similar to Ridge Regression, except that the penalty term changes: instead of $\sum_{j=1}^{p} \beta_j^2$, we have $\sum_{j=1}^{p} |\beta_j|$.

$$L_{lasso} = \sum_{i=1}^{n} (y_i - \sum_j x_{i_j} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

In addition, in Ridge regression the coefficients may tend towards 0 but never reach it. Lasso regression, on the other hand, allows coefficients to be equal to 0, resulting in more parsimonious models that allow for variable selection (coefficients that we don't want are set to 0 so that we can get rid of them).

1.2.3 Elastic Net

This method combines the properties of Ridge (avoiding overfitting, combating multicollinearity) and Lasso (parsimonious, more interpretable model), while ensuring that the variable selection performed by Lasso is not too strong.

$$Elastic_{Net} = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{i_j} \beta_j)^2 + \lambda (\tfrac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j|)$$

Here, λ controls the overall intensity of the penalty applied. In addition, the α parameter measures how much of the penalty comes from ridge regression versus lasso regression.

- If λ=0, the penalty comes solely from ridge regression.
- If 0<λ<0.5, we have a stronger penalty for ridge regression.
- If λ=0.5, we have the same penalty for lasso and ridge.
- If 0.5<λ<1, the penalty is higher for lasso regression.
- If λ=1, the penalty comes solely from lasso regression.

1.2.4 Incremental Forward Stagewise Regression

IFS is an iterative algorithm for finding an estimator of the regressors that converges to the OLS. This is done by incrementing at each step the regressor most correlated to the model error, and repeating these steps until no correlation is found between the error and the regressors. The rate of convergence depends on the value of the increment (usually 0.01 or 0.001).

- Step 1: Initialize the coefficients at 0. The explained variable then depends solely on the constant (=its empirical mean), giving a residual such as: $r_1 = y - \bar{y}$.
- Step 2: We look for the explanatory variable most correlated with the residual from step 1. We increment the coefficient of this variable with our increment*correlation sign. This variable is included in the regression, giving us a new residual: $r_2 = r_1 - S_x$.
- Step 3: We repeat this step n times until we find no more x correlated with the residual. If the x we find is the same as in the previous step, we increment its coefficient, otherwise we increment the coefficient of the coefficient concerned. This is how we obtain our model.

1.2.5 Least Angle Regression

The LARS (Least Angle Regression) algorithm is similar to the previous one in its operation, but here there is no incrementing. At each step, we look for the variable most correlated with the residual, then progressively bring the coefficient value of this variable closer to its OLS value. As soon as we find a variable with a higher correlation level, we integrate it into the process, thus increasing the information set. In the case where several variables have the highest level of correlation, we increase the coefficients so as to have a convergence angle with the same angle between the variables.

- Step 1: As with IFS, we initialize our coefficients to 0 and obtain our residual: $r_1 = y - \bar{y}$.
- Step 2: We look for the x most correlated with the residual, take its coefficients and gradually increase them until we reach the correlation value for x obtained with OLS. As a result, our r_1 changes (becomes r_2), and we continue until we obtain another variable with a higher correlation level.
- Step 3: We then move our coefficients progressively towards the OLS by moving our line into an angle between our variables. Repeat these steps until you have a complete model.

1.2.6 Maximum Likelihood

Maximum likelihood is derived from the likelihood function. This function is constructed by taking the product of the individual probabilities of each observation of a random phenomenon, and associating with this phenomenon the probability that it will follow a certain probability distribution. Since the parameters of the distribution are unknown, we look for an asymptotically convergent estimate of these parameters that maximizes the likelihood function. By replacing this estimate within the likelihood function, we find an estimate of the maximum likelihood.

1.3 Stopping Criteria

1.3.1 AIC criterion

The AIC, or Akaike Information Criterion, is a statistical criterion used to compare models in order to select the most appropriate one. It is defined as :

$$AIC = 2k - 2ln\hat{L}$$

This criterion is a compromise between the notion of model parsimony and the bias present with a low number of variables, by penalizing the log-likelihood according to the number of variables (the more there are, the greater the penalty) in order to limit overlearning. The aim when using this criterion to select a model is to minimize it in order to obtain a parsimonious model that explains the data well.

1.3.2 AICC criterion

The Akaike Information Criterion corrected is a correction of the AIC, the latter being asymptotic it tends to misbehave and overfitter on small samples, hence the advent of a correction to be able to use it on this type of data:

$$AICC = AIC + \frac{2k^2+2k}{n-k-1}$$

1.3.3 BIC criterion

The BIC or Bayesian Information Criterion is a statistical criterion in the same vein as the AIC, with the difference that the penalty applied here is greater:

$$BIC = kln(n) - 2ln\hat{L}$$

However, the meaning and use of the BIC are not quite the same as those of the AIC. The BIC will allow you to find a model by minimizing the criterion (like the AIC), but it will tend to select statistically significant variables, which makes it less suitable for forecasting than the AIC. So, generally speaking, BIC selects fewer variables than AIC.

1.3.4 K-Fold Cross validation (CV)

A machine learning model needs to be tested on new data in order to reliably assess its performance. This evaluation step uses the model's performance against unknown data to determine whether the model is under-fitted, over-fitted or correctly specified. Cross-validation, a resampling technique, is a frequently used method for this purpose, and proves effective in the presence of a large dataset. Cross-validation (also known as K-fold validation) involves dividing the dataset into different subsets called "folds". In this process, the model is trained on the K-1 folds and evaluated on the remaining fold. Each "fold" functions in turn as a test set. Before starting cross-validation, part of the dataset is reserved for use in the end as test data, thus enhancing the reliability of the model evaluation.

1.3.5 Leave-One-Out (Press)

The Leave-One-Out (PRESS) method is a variant of cross-validation. Each observation in the dataset is a fold, and is used as a test set at each iteration, with the remaining folds used to train the model. This method is particularly effective when the dataset is very small, but can be very demanding when the dataset is large.

1.4 Metrics definition

In order to measure the performance of the procedures and criteria on our model, we needed suitable measurement tools, which is why we built 7 metrics to give the probability of occurrence of the different possible behaviors of the procedures. First of all, we have perfect fitting which represents the probability that the model as we have defined it will be selected. Then, we also have over-fitting, illustrating the probability that the model will select variables in addition to those defined, and large over-fitting, which measures the number of times over fitting is greater than 4 additional variables. Next, we also created an Under-fitting metric, which corresponds to the case where the chosen model contains fewer variables than the reference model. It is distinguished into 2 types:

- Good under-fitting, which selects only variables from the reference model, but not all of them.
- Bad under-fitting, which selects only variables outside the model or in addition to some of the model's variables.

We also distinguish the severity of under-fitting, measuring whether the under-fitting is too severe (model selected less than or equal to 3 variables).

Finally, our last metric measures the probability of failure, i.e. that the procedure over-fits while not selecting all the variables in the reference model.

Chapter 2

Comparison of results between stopping criteria for each DGP

2.1 Definition of DGPs

2.1.1 Independent DGP

Using simulations based on reduced-centered normal distributions (multivariate distribution), we have developed two data-generating processes (DGP) comprising 100 and 500 observations respectively. Our process includes 50 variables, and the associated variance-covariance matrix has the structure of an identity matrix, ensuring independence between the variables. Based on our data set, we decided to create a model encompassing the first six variables. This model is formulated as follows:

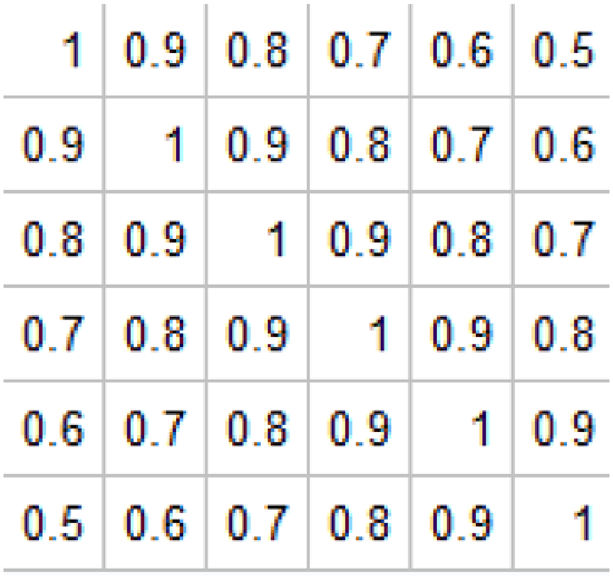
$$Y = 1.5 * X1 + 1.3 * X2 - 1.7 * X3 + 1.6 * X4 - 1.4 * X5 + 1.2 * X6 + \epsilon * 0.01$$

where ϵ represents white Gaussian noise, adjusted according to the value we wish to assign to our signal-to-noise ratio (SNR). Our final data set thus consists of 100 or 500 Y values calculated in accordance with our specifically defined model, combined with our initial matrix comprising the 50 variables. This data-generating process is repeated 1,000 times, and is incorporated at each iteration into each of the variable selection methods we have previously defined, with a specified stopping criterion, the metrics enabling us to judge the quality of our results. To extend our analysis, we decided to generate different types of processes generating atypical data, in order to evaluate the behavior of each method when faced with data that is no longer as "clean" as in the case of independence.

2.1.2 DGP with colinearity

When several independent variables in a regression model are highly correlated, multicollinearity occurs, making the estimation of regression coefficients often uninterpretable. We have considered two types of multicollinearity: internal and external, the former defined as a correlation between all the variables in the model we have defined, the latter between the variables in the model and variables external to the model (9 variables external to the model in our case). To specify these correlations, we decided to use a toeplitz variance-covariance matrix.

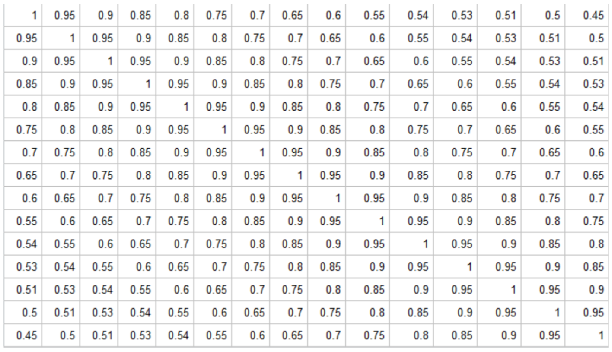
For the internal case, an identity variance-covariance matrix of size 44 has been introduced to ensure independence between the 44 variables not included in the model, and a toeplitz matrix of size 6x6 (2.1), which defines strong correlations between the 6 variables (ranging from 1 to 0.5) and which will be the variance-covariance matrix for the first 6 variables. It has been defined as follows:



1	0.9	0.8	0.7	0.6	0.5
0.9	1	0.9	0.8	0.7	0.6
0.8	0.9	1	0.9	0.8	0.7
0.7	0.8	0.9	1	0.9	0.8
0.6	0.7	0.8	0.9	1	0.9
0.5	0.6	0.7	0.8	0.9	1

Figure 2.1: Toeplitz matrix 6x6

For the external case, an identity variance-covariance matrix of size 35 has been introduced to ensure independence between the last 35 variables of our dataset, as well as a toeplitz matrix of size 15x15 (2.2), which also defines strong correlations between the first 15 variables, and which will also be the variance-covariance matrix for the first 15 variables. The latter has been defined as follows:



1	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.54	0.53	0.51	0.5	0.45
0.95	1	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.54	0.53	0.51	0.5
0.9	0.95	1	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.54	0.53	0.51
0.85	0.9	0.95	1	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.54	0.53
0.8	0.85	0.9	0.95	1	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.54
0.75	0.8	0.85	0.9	0.95	1	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55
0.7	0.75	0.8	0.85	0.9	0.95	1	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6
0.65	0.7	0.75	0.8	0.85	0.9	0.95	1	0.95	0.9	0.85	0.8	0.75	0.7	0.65
0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1	0.95	0.9	0.85	0.8	0.75	0.7
0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1	0.95	0.9	0.85	0.8	0.75
0.54	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1	0.95	0.9	0.85	0.8
0.53	0.54	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1	0.95	0.9	0.85
0.51	0.53	0.54	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1	0.95	0.9
0.5	0.51	0.53	0.54	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1	0.95
0.45	0.5	0.51	0.53	0.54	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1

Figure 2.2: Toeplitz matrix 15x15

For the 2 DGPs, the observations for each variable will then each be generated using a centered-reduced normal distribution. We then repeat the last step, similar to the DGP with independence, where we create the model and concatenate the dependent variable with the matrix of independent variables. This operation is repeated 1000 times for each variable selection method, each stopping criterion, and the results are evaluated using our metrics.

2.1.3 DGP with outliers

Outliers are defined as significantly atypical observations within a data set, which can disrupt our analyses. Their detection and management are essential to guarantee the reliability of our results.

We decided to incorporate outliers into the observations of the first 45 variables in the following way: we generate a random number between 0 and 1 from the uniform distribution; if the value generated is strictly less than 0.9, the observation will be drawn from a centered reduced normal distribution; conversely, it will be generated from a normal distribution, but with an expectation equal to 5.

The observations of the last 5 variables in the dataset are all generated from a centered reduced normal distribution.

Then, we repeat the last step as in the DGP with independence, i.e. the creation of the model and the concatenation of the dependent variable associated with the matrix of dependent variables.

Finally, we iterate this operation 1000 times within each variable selection method, each stopping criterion and we judge our results through our metrics.

2.1.4 DGP with colinearity and outliers

Our last data generation process involves the presence of outliers and internal multicollinearity. To implement it, we started by generating data with outliers. Then, our dataset was divided into two parts: the first comprising the data generated for the first six variables, and the second containing the rest. The first dataset was subjected to the Iman Conover transformation, a method forcing multicollinearity between variables. To specify the correlation structure, we used the Toeplitz matrix, previously employed in data generation with internal multicollinearity. Once the Iman Conover transformation had provided us with the embedded correlation matrix, we concatenated it with the matrix of the other 44 observations, then built our final model. This operation was repeated 1000 times for each variable selection method and each stopping criterion. The results were evaluated using our metrics.

2.2 Results for Statistical Learning methods

Now that we have presented the structures of all our data-generating processes, we are going to concentrate on the most important part of our paper: interpreting our results to determine the best methods and criteria for retrieving the model we've previously defined. Before beginning this section, we'll define some of the procedures we've used under the glmselect procedure, which have enabled us to control how our methods will select variables, and which we feel are relevant to our readers' understanding of our results. Under SAS, the various variable selection methods generate models with effects selected in distinct ways according to two criteria, namely the choose criterion and the stop criterion:

- The "choose" criterion selects the final model.
- The "stop" criterion is used to interrupt the search as soon as a minimum information criterion is reached.

When the "choose" criterion differs from the "stop" criterion, the software initiates the search, giving priority to the best model according to the "stop" criterion, without going beyond it. For all models identified using the "stop" criterion, the software re-evaluates the criterion chosen for "choose". If the "choose" criterion leads to the selection of a more restrictive model, this is the one that will be retained, otherwise it will be the one found by the "stop" criterion.

We'll start by looking at statistical learning methods. For the forward and backward methods, our data were initially generated with 100 observations. In the course of our analysis, we increased the number of observations to 500 in an attempt to improve the significance of our results. The results presented for these first two methods will be based on a set of n=500 observations, as no major change was observed by increasing the number of data. However, for the stepwise method we will detail the results for n=100 as well as n=500 according to each DGP, as some differences were observed.

2.2.1 Forward selection

Considering each type of data-generating process and applying the same criterion both as a stopping criterion and as a selection criterion, we observed similar results on average. This translates into an overfitting rate of around 100 percent (with an average of 50-60 percent severe overfitting) for the pairs AIC-AIC, AICC-AICC, BIC-BIC, CV-CV, Press-Press. In contrast, the SBC-SBC pair performs remarkably well, with an average of 50 percent perfect fitting and 50 percent overfitting, although the latter is not severe.

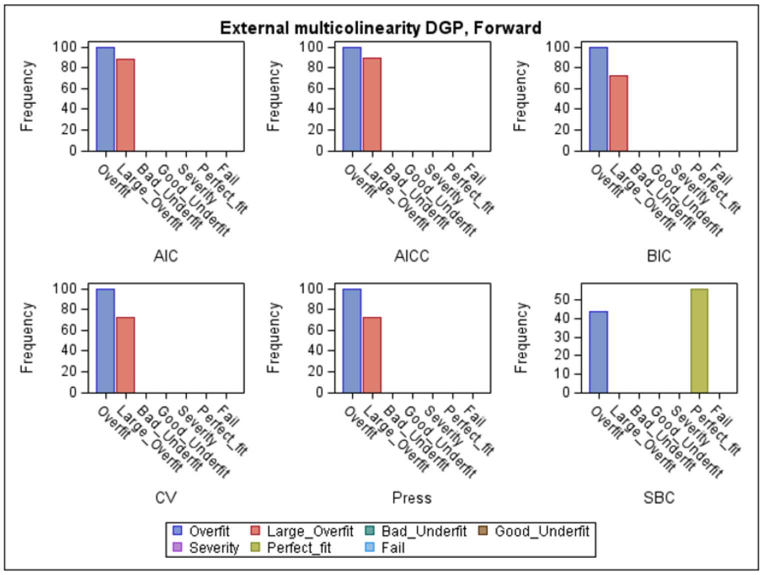


Figure 2.3: Forward, Same Criterion, N=500

Taking our analysis a step further, we found that whatever the DGP a trend can be observed, one that highlights the effectiveness of the SBC criterion for this method. In fact, when we mixed the stop and selection criteria, we obtained more or less the same results for all data-generating processes (the same results seen above), except for one case which surprised us. All else being equal, taking the stopping criterion as the SBC, our results improved significantly, with an average of 50-60 % perfect fitting and 40-50% overfitting without an ounce of severe overfitting.

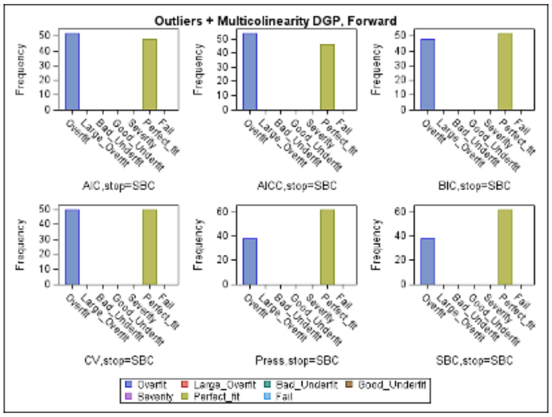


Figure 2.4: Forward, Outliers and Multicolinearity, N=500

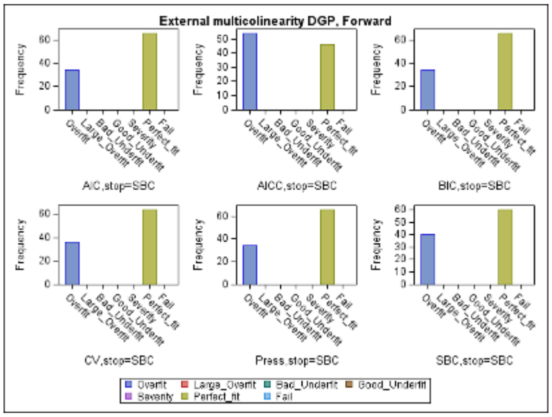


Figure 2.5: Forward, External Multicolinearity, N=500

2.2.2 Backward selection

In a similar way to the forward method, for all types of data-generating processes and applying the same criterion both as a stopping and selection criterion, we obtained roughly identical results. Once again, we can see an overfitting rate of 100 percent for the AIC-AIC, AICC-AICC, BIC-BIC, CV-CV and Press-Press pairs. Furthermore, the SBC criterion stands out, averaging 50 percent perfect fitting and 50 percent overfitting. Unfortunately, mixing criteria proves unnecessary for this method, as the results remain the same.

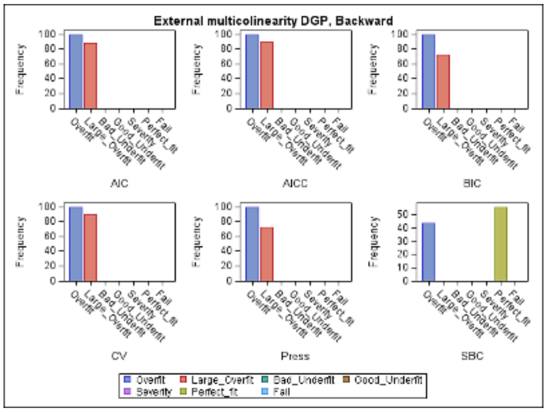


Figure 2.6: Backward, Same criterion, N=500

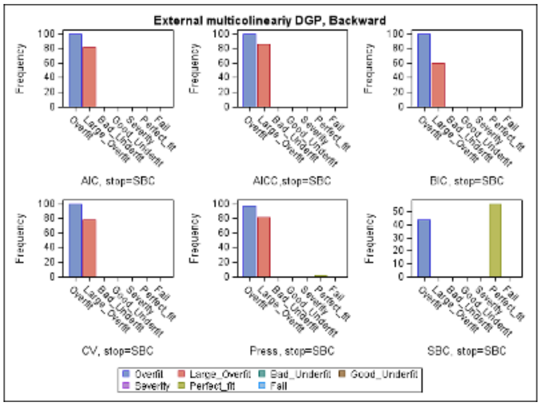


Figure 2.7: Backward, Mixing criteria, N=500

2.2.3 Stepwise selection

Independent DGP

When we compare the results of the Stepwise procedure with independently generated data, we obtain exactly the same results for all criteria (13% perfect fit and 87% overfit) except for cross validation, which yields 16% perfect fit and 84% overfit.When we perform combinations, we reach the same conclusions with the cv and press combination, i.e. 16% perfect fit.

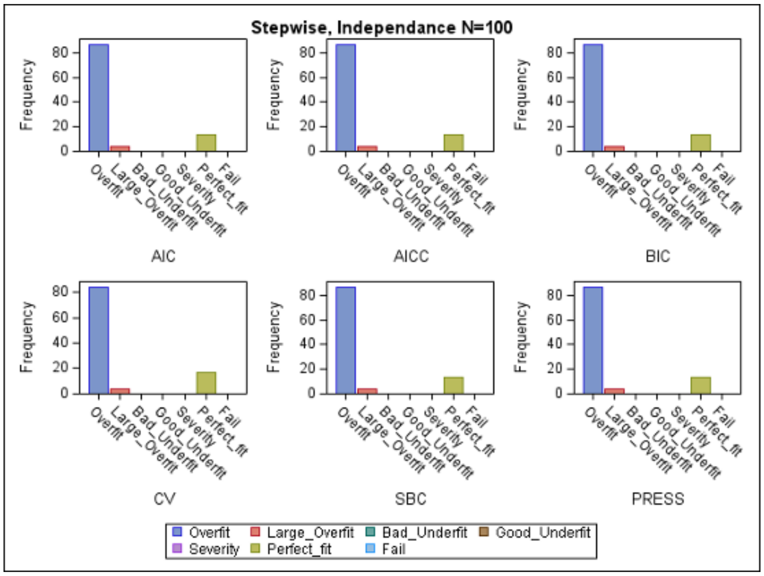


Figure 2.8: Stepwise, Independence, N=100

By increasing the number of observations to N=500, we also obtain the same results as before, whether using the same criteria in stop and choose, or using combinations of criteria.

Internal Multicolinearity DGP

No single criterion really stands out as in the previous DGP. On average, there are small increases, with the CV criterion gaining 9 percentage points over the independent DGP when looking at the perfect fit metric. In addition, all other criteria increase by 4 percentage points. Increasing the number of observations to N=500 produces very good results. These are the same for all criteria (42% overfit and 58% perfect fit). In addition, the combination of cv and press criteria produced a slightly higher increase, with 59% perfect fit.

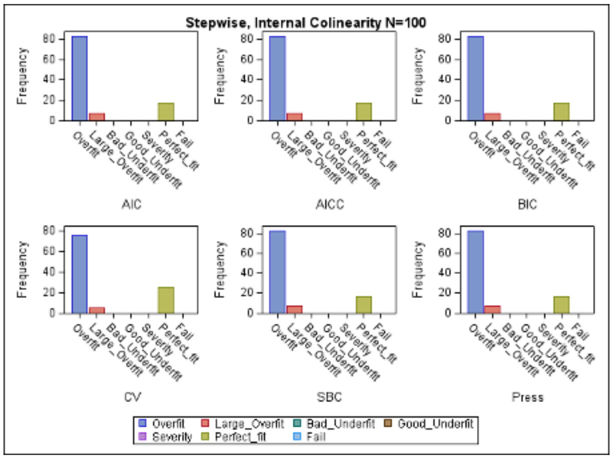


Figure 2.9: Stepwise, Same criterion, N=100

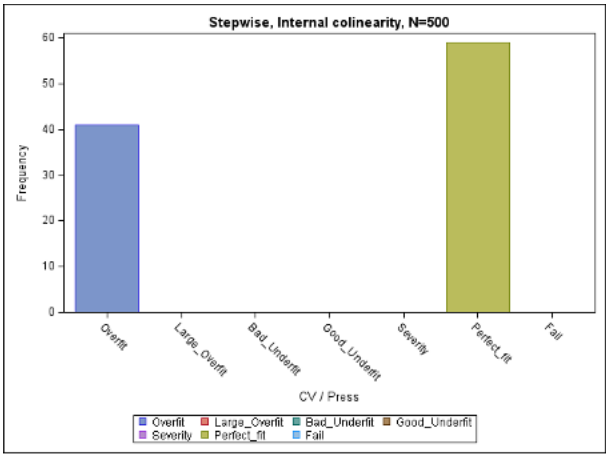


Figure 2.10: Stepwise, Mixing criteria, N=500

External Multicolinearity DGP

We still have similar results for all the criteria at N=100. However, combinations are better than single criterion. For example, we obtain 32% perfect fit for stop=cv choose=press and 30% for stop=sbc choose=cv, as opposed to 29% perfect fit for stop=cv and choose=cv.

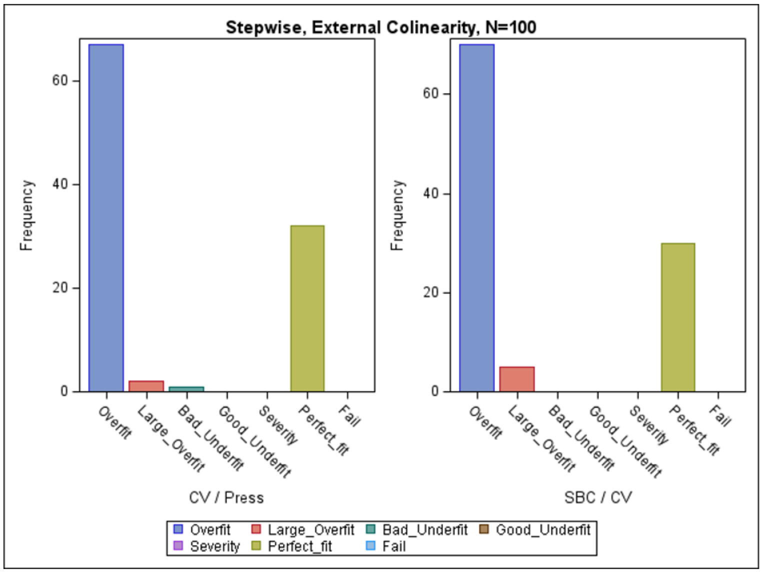


Figure 2.11: Stepwise, External Multicolinearity, N=100

We can see that by increasing the number of observations, as in the DGP of internal colinearity, we obtain very large changes in the criteria. In fact, they all give 59% perfect fit (e.g. the SBC criterion), except for 2 which stand out: The CV criterion alone and the SBC / CV combination return 60% perfect fit and 40% overfit.

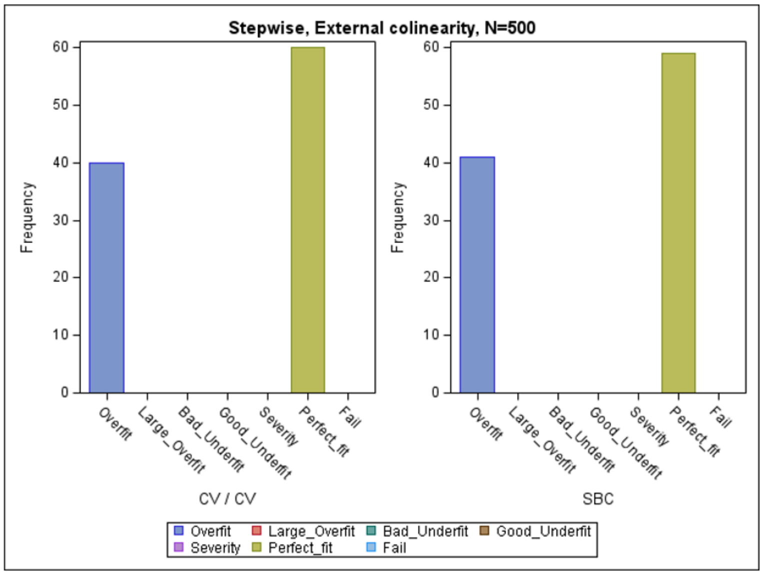


Figure 2.12: Stepwise, External Multicolinearity, N=500

Outliers DGP

The Stepwise procedure performs well overall on a dataset with extreme values, achieving perfect fitting levels ranging from around 15% to 30% on all criteria, without fail. No criteria stand out clearly from the others, although the highest level of perfect fitting is achieved with the CV/CV pair, which approaches 30%. Analyzing the same procedure with a larger number of observations, the proportion of successes for the various criteria climbs to between 50% and 60%, which represents a significant increase in the perfect fitting rate. This is illustrated in the graph (2.13), by combining CV with other criteria, since it is with this stopping criterion that we once again obtain the highest perfect fitting rates:

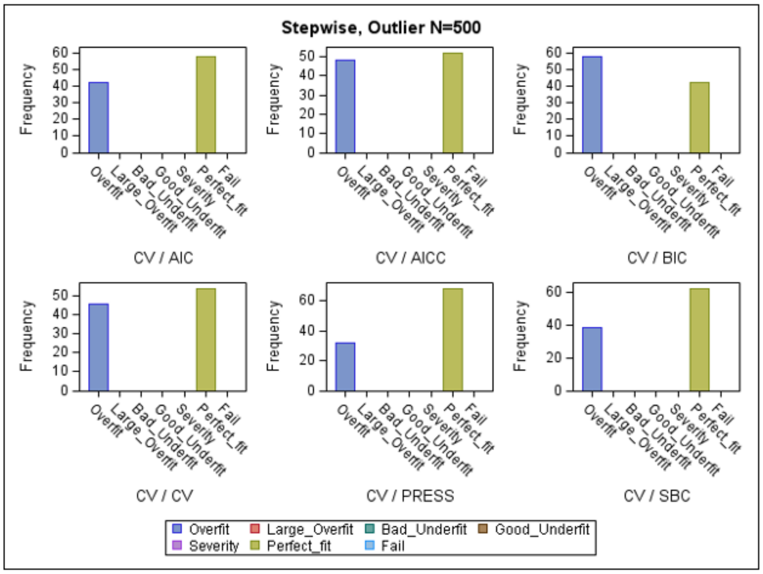


Figure 2.13: Stepwise, Outliers, N=500

The CV/SBC and CV/PRESS combinations stand out the most in terms of perfect fitting, with success rates of 62% and 68% respectively.

Outliers and Multicolinearity DGP

By adding multicolinearity to our data with outliers, we obtain more heterogeneous results, with the criteria not failing, with a tendency towards overfit and perfect fit. Under these conditions, we manage to get up to around 30% perfect fit by combining certain criteria such as BIC / CV or PRESS / CV, but we realize that these results are relatively volatile, since by repeating the tests on these combinations we can end up with perfect fit values close to 10%, which does not make them robust to this procedure. After increasing the number of observations, we observe the same conclusions as before, but the range of the perfect fit is greater, of the order of 50%-70%, so we find a certain volatility in the results. The graph below shows the combination that gave us the highest rate of perfect fitting.

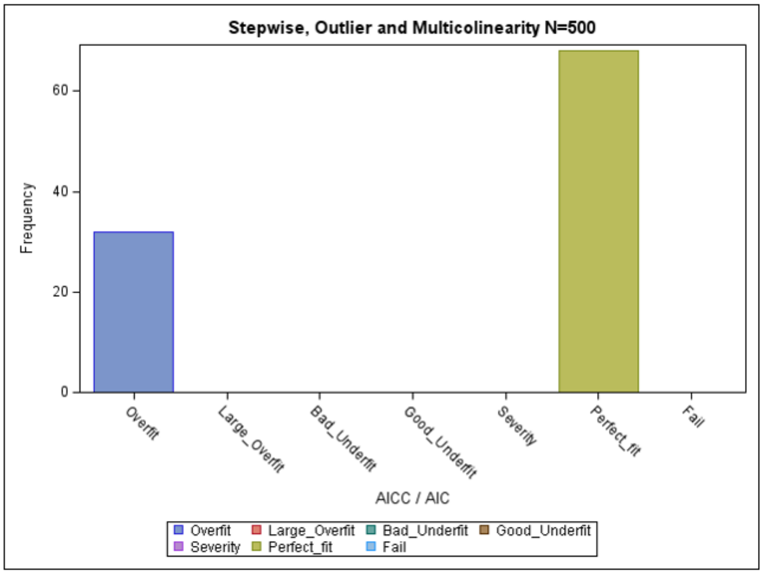


Figure 2.14: Stepwise, Multicolinearity and Outliers, N=500

2.3 Results for Machine Learning methods

2.3.1 Lasso selection

Independent DGP

Following on from our analysis of Statistical Learning algorithms, we now turn our attention to Machine Learning algorithms, starting with Lasso. Using an independent dataset and the same model as before, our various tests on the multiple stop and selection criteria show that the 2 combinations that are generally the least wrong are : SBC/PRESS (stop criterion / selection criterion) and the inverse combination. This gave us the following graph aggregating our metrics:

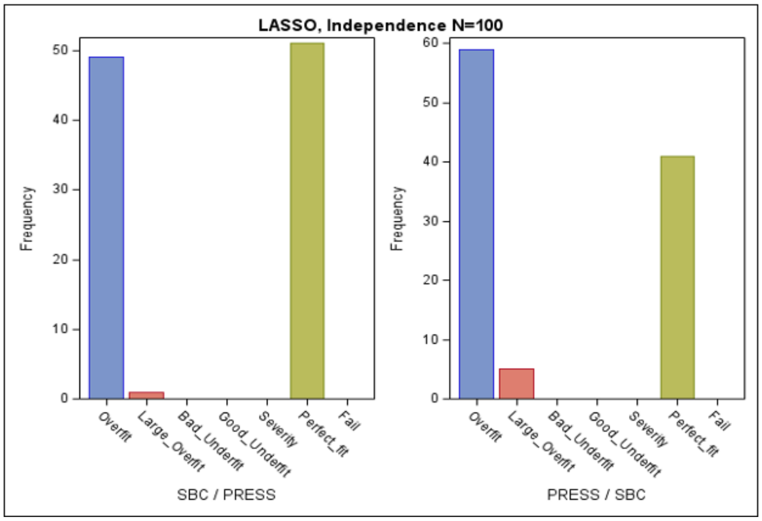


Figure 2.15: Lasso, Independence, N=100

As we can see, the first combination gives us the best result, with perfect fitting at around 50%, while the second is more around 40%. At the same time, we note a high rate of overfitting (50%), but this overfitting is not considered wide in the sense of our metrics. The other criteria have a lower probability of perfect fitting, with AIC and AICC tending to underfit severely, while BIC tends to overfit. On the other hand, as we increase the number of observations, we notice that the 2 best combinations remain the same, but that overall the probability of success has increased for all criteria.

Internal Multicolinearity DGP

With this type of DGP, the procedure behaves much worse than with the previous one: whatever the stopping criterion, you can't get a perfect fitting. However, there is one that is less wrong than the others, and it is the SBC, with a failure level of around 2% no matter what selection criterion was chosen.

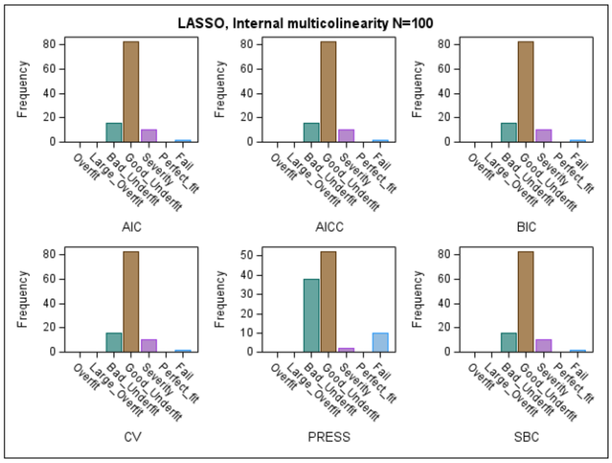


Figure 2.16: Lasso, Internal Multicolinearity, N=100

In addition, increasing the number of observations considerably improves the performance of the criteria. Under this condition, the Lasso is no longer wrong for all criteria, and it is by using SBC and PRESS that we achieve the best performance in terms of perfect fitting with the same method as for independence, i.e. by combining these 2 criteria:

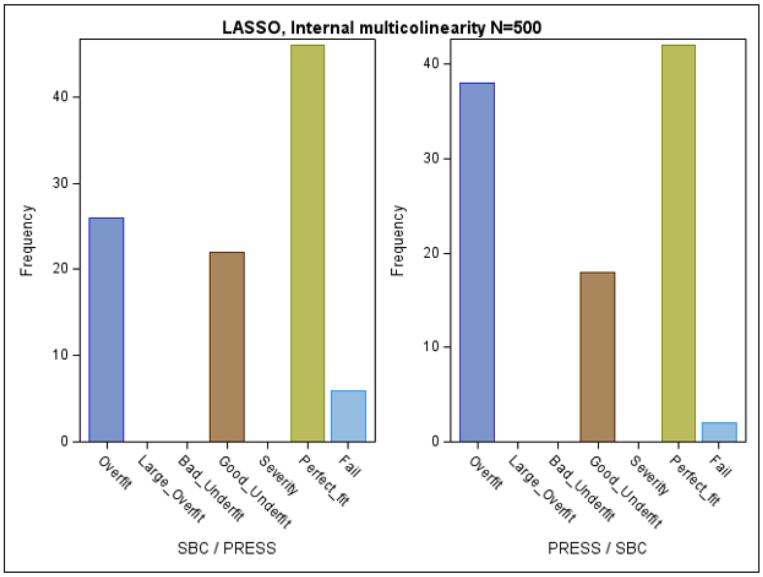


Figure 2.17: Lasso, Internal Multicolinearity, N=500

External Multicolinearity DGP

The conclusions are similar to those obtained for internal multicolinearity in the sense that the procedure behaves badly : whatever the criterion we don't get perfect fitting, except that rather than fail, the criteria tend to over-fit or under-fit. Furthermore, in this case, increasing the number of observations doesn't improve our results: the criteria continue to over-fit or under-fit.

Outliers DGP

As for the outlier dataset, we have several stop criteria that behave well, such as CV, PRESS and BIC. However, when we test our results with several combinations, we find the SBC / PRESS and PRESS / SBC pairs to be the most successful. Furthermore, after increasing the number of observations, the performance of the SBC / PRESS pair increases even further to 64% perfect fitting, as does the frequency of perfect fitting for PRESS / SBC, even if this remains minimal.

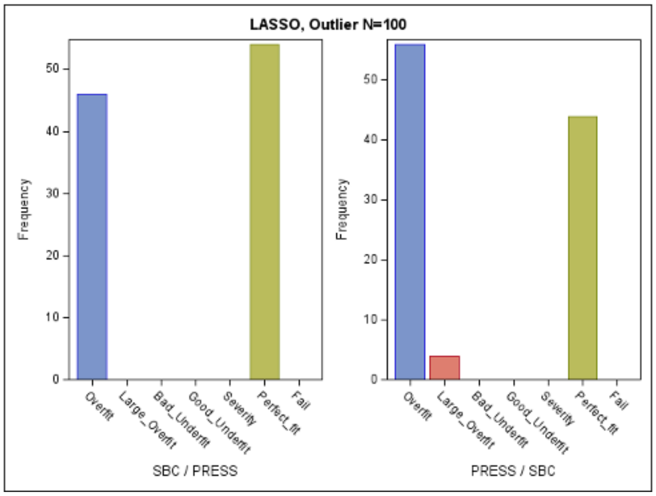


Figure 2.18: Lasso, Outliers, N=100

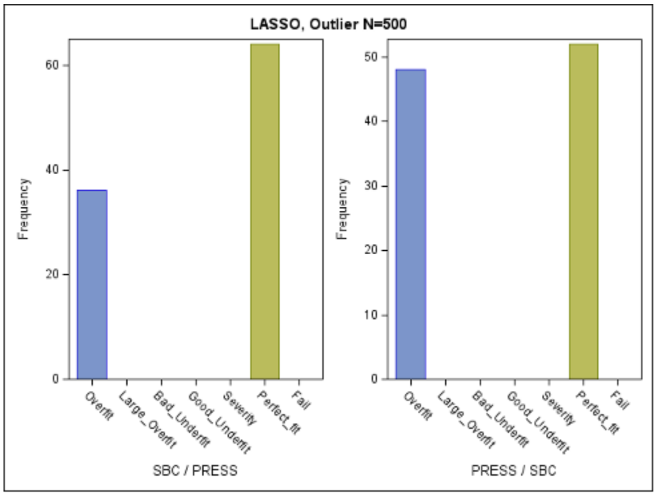


Figure 2.19: Lasso, Outliers,N=500

Outliers and Multicolinearity DGP

Combining outliers with colinearity in our data produces bad results in a similar way to multicolinearity. SBC is the least misleading criterion here, with fail levels below 5%. As expected, increasing the number of observations allows us to obtain better results by considerably reducing the probability of failure while increasing the probability of perfect fitting, whatever the criterion. As a result, SBC and PRESS are once again found to be the criteria giving the best results when combined, with the SBC / PRESS combination achieving a perfect fitting probability of 52%.

2.3.2 Elastic Net selection

Independence DGP

We start with a set of independently generated data. When we test the AIC or AICC for elastic net, we obtain very few perfect fits but many good underfits with very severe underfitting. The cross-validation criterion stands out with around 20% perfect fit (2.20), but we prefer to choose the mixing criteria CV/ SBC (32% of perfect fit).

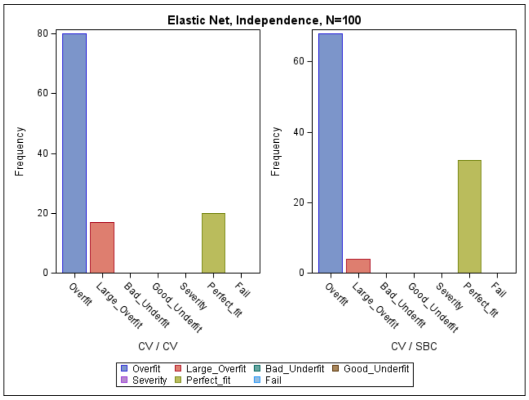


Figure 2.20: ElasticNet, Independence, N=100

Finally, when we increase the sample size to N=500, we notice significantly more overfitting for AIC. In addition, the perfect fit rate improves markedly for AICC (13%) and especially for SBC, which rises from 2% to 16% just by varying the sample size. We therefore retain SBC and AICC as variable selection criteria when the sample size is large.

Internal Multicolinearity DGP

When we look at the data generated with internal colinearity, we notice failures that we didn't have in the independent DGP for any of the stopping criteria. There are numerous fails for the AIC (23%) and AICC (15%) criteria, with a high rate of underfit (whether good or bad). As with the independent DGP, the BIC has 100% overfitting. However, cross validation does extremely well, with 67% perfect fit and only 26% overfitting.

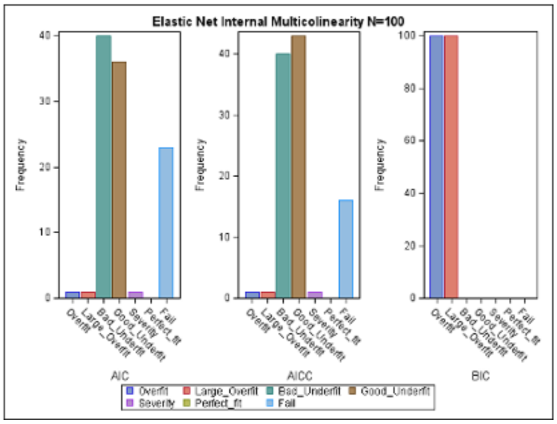


Figure 2.21: ElasticNet, Internal Colinearity, N=100

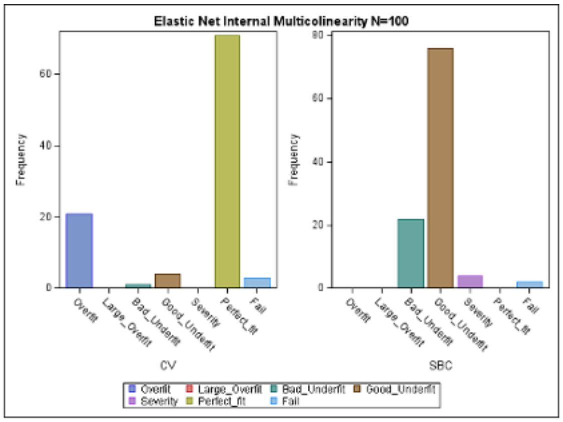


Figure 2.22: ElasticNet, Internal Colinearity, N=100

Finally, when we move on to N=500, all criteria without exception overfit and cross validation has 35% perfect fit. The best criterion is stop=cv choose=cv (44%) followed by the combination stop=sbc choose=cv (25%).

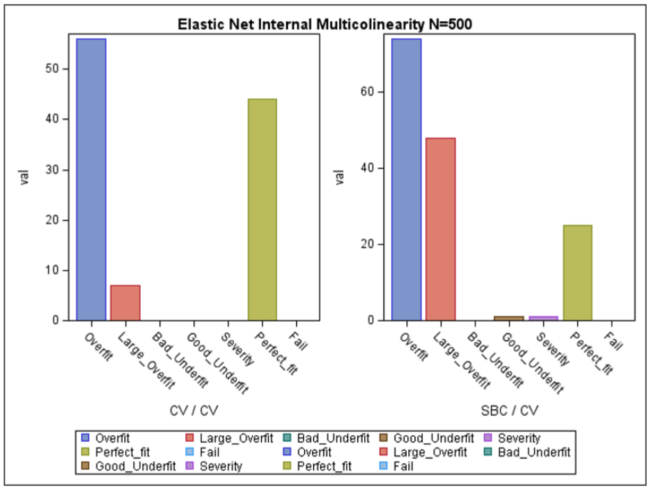


Figure 2.23: ElasticNet, Internal Colinearity, N=500

External Multicolinearity DGP

We then move on to data with external collinearity. Here, we have the same trends as the DGP for internal collinearity when N=100 and the error term is varied downwards (with the error term equal to 0.01 or 0.5).
However, when we go to N=500 we have overfitting, underfitting and fails for the AIC, AICC and SBC criteria. BIC is now only 100% overfitting, which makes it a poor criterion. We therefore select the cross validation with 69% perfect fitting.

Outliers DGP

In data with outliers, we no longer find fails in the data. Thus we have practically equivalent over-fitting and under-fitting rates, combined with a small but still present percentage of perfect fitting for AIC (2%) and AICC (7%). The SBC criterion is also beginning to show low percentages of perfect fitting (2%). The BIC criterion still shows 100% overfitting, so we opt for cross-validation, which nevertheless has a lower rate of perfect fit (15%) than in previous DGPs. However, with the combination of criteria stop=cv choose=sbc we find a good rate of perfect fitting (34%) on 100 observations.

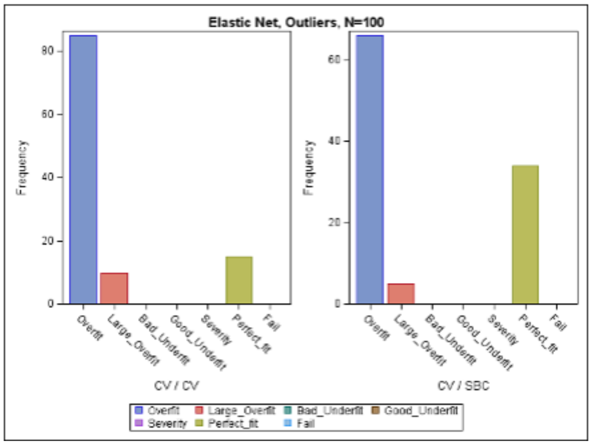


Figure 2.24: ElasticNet, N=100

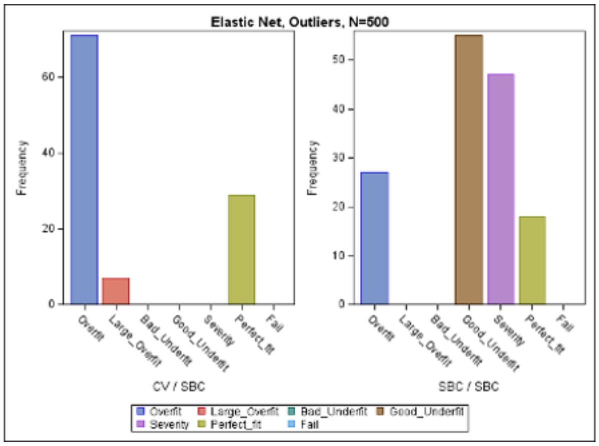


Figure 2.25: ElasticNet, N=500

Finally, by increasing the sample size to N=500, we achieve a much higher percentage of perfect fit for the AIC (10%), AICC (9%) and especially the SBC (18%), which perform better than the cross validation, this time at 3% perfect fit. On the other hand, the BIC is still 100% overfit. However, we'll keep the stop=cv and choose=sbc criteria combination, which has 29% perfect fit.

Outliers and Multicolinearity DGP

When we look at multicolinearity with the outliers, we find relatively similar failures for AICC and AICC, probably due to the colinearity of the data. The BIC is still 100% overfitting and the SBC no longer overfits at all compared with the DGP with outliers. We can also see that cross validation is clearly more underfitting than DGP outliers, which were not at all, but cross validation is still 10% perfect fit. However, we choose the stop=sbc choose=cv combination, which returned 16% perfect fit.

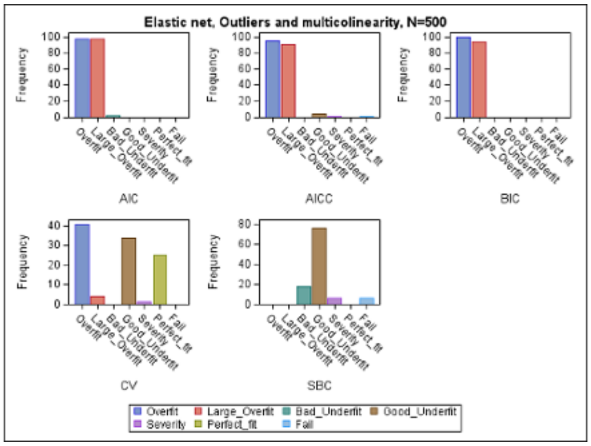


Figure 2.26: ElasticNet, Same criterion, N=500

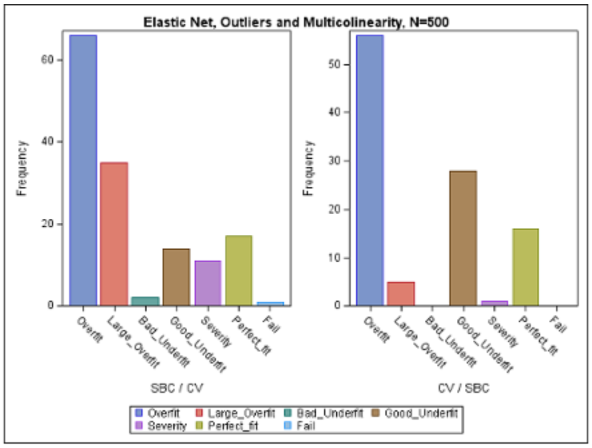


Figure 2.27: ElasticNet, Mixing Criteria, N=500

With an increase in the number of observations (N=500), we can see that the criteria do a lot of overfitting, except for the SBC. However, they are less wrong than in N=100. For example, the CV criterion increased its perfect fit from 10% to 25%. This criterion alone is even better than the stop=SBC choose=CV combination, which is at 17%, and the inverse combination, which is at 16%.

2.3.3 LARS selection

Independence DGP

For this last Machine Learning algorithm, we maintain the same analytical approach as before. We begin our analysis with our dataset with independence, under which the CV and PRESS criteria do rather well overall, but just as with Lasso, it's by setting SBC as the stopping criterion and then combining it with the other criteria that we achieve the best success:

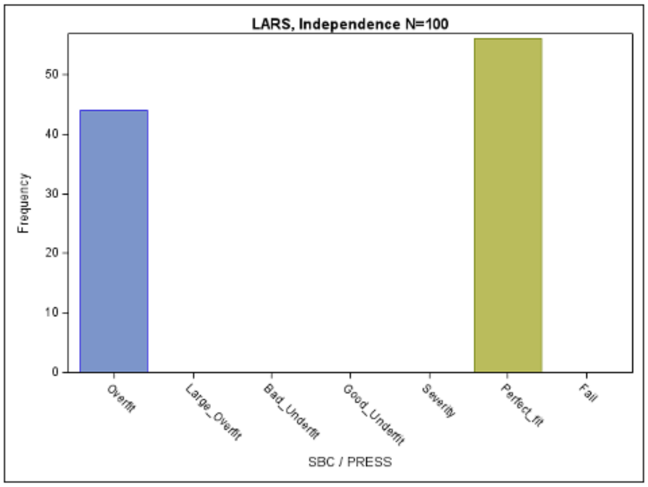


Figure 2.28: Lars, Independence, N=100

The other selection criteria do not perform as well with SBC. As for the impact of increasing the number of observations, it's not significant: although it increases the overall performance of the criteria, this combination remains the best, without gaining in precision here.

Internal Multicolinearity DGP

The data being "less clean" than the previous ones, our criteria have difficulty in finding the pattern we have defined, essentially failing for all criteria. We find SBC to be the stopping criterion with the lowest failure rates. As we move on to a higher number of observations, we obtain a better performance of the criteria, particularly for the SBC/-PRESS combination, which offers a perfect fitting rate of 46% for 6% failure, the rest being divided between good under-fitting and over-fitting.

External Multicolinearity DGP

Switching to external collinearity yields conclusions identical to those obtained for Lasso, the overfit or underfit procedure for each criterion, so it's wrong in a way. One might expect better results with a higher number of observations, as before. However, this is not the case here, as there is no change in the behavior of the criteria.

Outliers DGP

Once again, our results converge with those obtained through the Lasso for the data set with extreme values: SBC / PRESS is the combination giving the most perfect fitting, with a rate of 56%, followed by PRESS / SBC with a rate of 35%. As for the other stopping criteria, BIC always overfits, while AIC and AICC tend to underfit. Increasing the number of observations to 500 doesn't affect the first combination, but it does boost the second to a success rate of 62%. This is illustrated in the graph below:

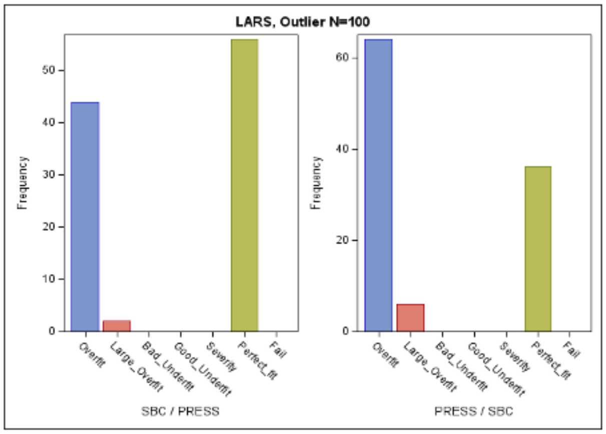


Figure 2.29: Lars, Mixing Criteria, N=100

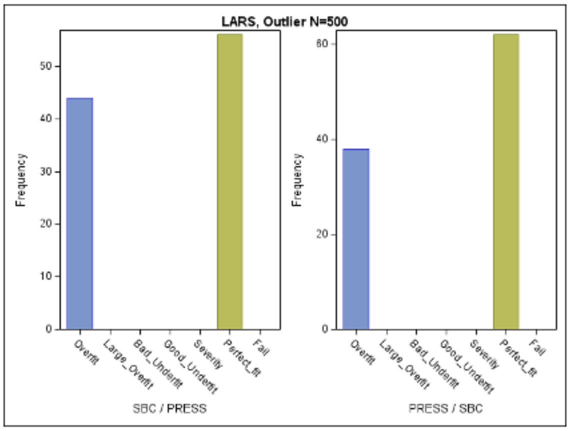


Figure 2.30: Lars, Mixing Criteria, N=500

Outliers and Multicolinearity DGP

We can see here that, whatever the criterion, we don't get good results in terms of perfect fitting, with a lot of underfitting and the presence of failures for all procedures (except BIC). Finally, increasing the number of observations leads to better results, since the criteria make fewer errors, particularly CV and PRESS, which give good levels of perfect fitting. Nevertheless, the SBC / PRESS pairing once again stands out for its higher level of success than the others. The evolution of this pairing following the increase in our number of observations is shown in the following graph:

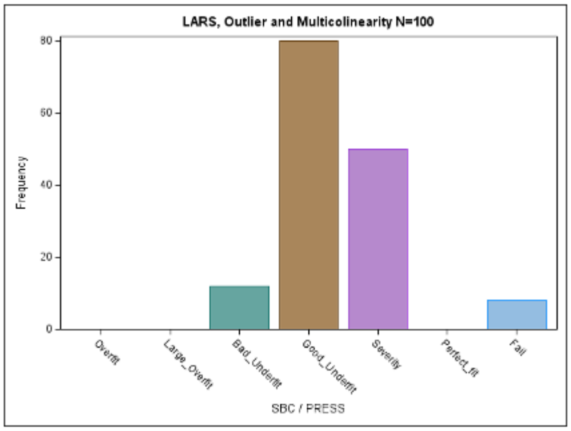


Figure 2.31: Lars, Mixing Criteria, N=100

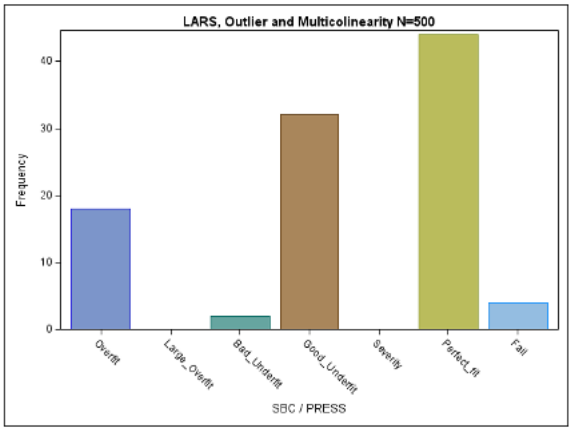


Figure 2.32: Lars, Mixing Criteria, N=500

2.4 Summary

We have decided to base our summary mainly on results obtained after increasing the quantity of our observations, since interpretations are clearer under these conditions, but also because in reality we are less confronted with sample sizes of 100 as considered here.

2.4.1 Independent DGP

Under the first data set we've established, the procedure that stands out from the others in terms of its success through our results turns out to be LASSO, and it's by calibrating it with SBC as the stopping criterion and PRESS as the selection criterion that we optimize its probability of success.

2.4.2 Multicolinearity DGP

Here we have 2 sub-cases to deal with: internal multicollinearity and external multicollinearity.

- In the first case, we find stepwise to be the best-performing procedure, particularly with the CV / PRESS pair, even though the CV criterion generally works well as a stopping criterion in this case.

- In the second one, in line with our results, we obtain the most conclusive findings with the cross-validation for Elastic Net.

2.4.3 Outliers DGP

Here, 3 methods stand out for their performance on data sets containing outliers.

- For the LASSO method, the combinations SBC / PRESS and PRESS / SBC demonstrated the best fit, with a perfect fitting rate reaching 64% after increasing observations.
- In the case of the Lars method, the combination SBC / PRESS also showed the best result, with a rate of 56%, closely followed by PRESS / SBC at 35%. The addition of observations increased the success rate of the second combination to 62%.
- With regard to the Stepwise method, the CV/SBC and CV/PRESS combinations emerged as the best performers, with success rates of 62% and 68% respectively after the addition of observations.

In conclusion, the stepwise method seems to offer the best results.

2.4.4 Multicolinearity and Outliers DGP

Here again, 3 methods were compared.

- For LASSO, the SBC / PRESS combination showed the best fit, with a perfect fitting rate reaching 52% after increasing the number of observations.
- For LARS, the SBC / PRESS combination stood out with a success rate of around 50% with 500 observations.
- For the stepwise method, increasing the number of observations amplifies the perfect fitting rate (around 70 percent) under the AICC/AIC criteria.

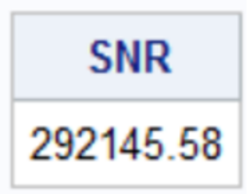
Once again, the stepwise method under the criteria we have defined seems to be the best method for this type of data.

2.5 Signal to Noise Ratio

In this section, we'll focus on the role played by white Gaussian noise, which acts as an error term within our model. To do this, we have manipulated the weight attached to it, so that the standard deviation of the noise adjusts and varies in the same direction as the weight, with the effect that the results also vary. In order to quantify this change in weight, we have constructed an indicator, which tends towards $+\infty$ when the weight decreases and towards 0 in the opposite case. It is called the Signal-to-noise ratio (SNR), the formula for which is given below:

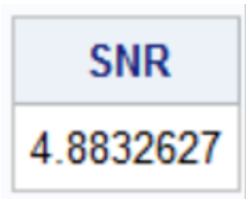
$$SNR = \frac{Var(X\hat{\beta})}{\hat{\sigma}^2}$$

Our analysis consists in repeating LASSO (SBC / PRESS) in the framework of data with independence, except that we set a weight equal to 2.5. By running the procedure, we therefore iterate over 2 SNR levels, one with a relatively low weight (initial case) and the other with a high weight (2.34).



SNR
292145.58

Figure 2.33: SNR of the model, last iteration, Weight = 0.01



SNR
4.8832627

Figure 2.34: SNR of the model, last iteration, Weight = 2.5

We can notice a clear difference between the 2 models. The results show a clear downward trend with regard to perfect fitting when we increase the weight of our white noise. The probability of perfect fitting drops by 15 percentage points compared with the initial case, and we also notice the appearance of fail (7%) and underfitting :

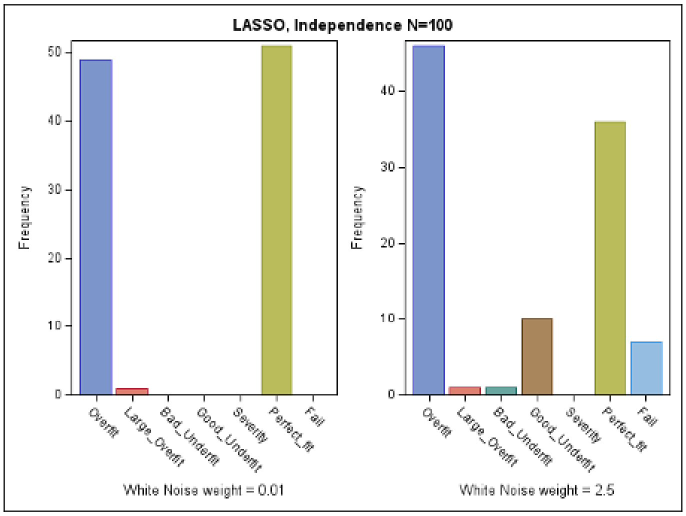


Figure 2.35: Lasso, Changes in Weight, N=100

Chapter 3

Empirical analysis

3.1 Description of the Dataset

The Diabetes dataset, extracted from the Wisconsin Diabetes Data Set, is a compilation of medical data designed to explore the links between various parameters and the development of diabetes. The variables included in this dataset provide insight into patient characteristics. A brief presentation of each of the dataset's variables is necessary to understand how it works:

- AGE: This variable represents patients' age in years.
- SEX: Coded as 1 for male and 2 for female.
- BMI: Body mass index, a measure of weight in relation to height.
- BP: Represents mean blood pressure.
- S1 to S6: These biochemical variables provide information on the metabolic aspects of diabetes, including lipoprotein, triglyceride and insulin levels, as well as serum peptide function.
- Y: used to represent the measure of diabetes progress caused by disease-related characteristics.

These data, made up of 442 observations, provide a vast field for testing our variable selection methods on something concrete.

3.2 Dataset Analysis

Following on from the description of our dataset, we used the univariate proc, the sgplot proc and the means proc to see if our data were polluted by outliers.

3.2.1 Definition of the kurtosis and the skewness

In statistics, skewness and kurtosis are two ways of measuring the shape of a distribution.

Skewness

Skewness measures the degree of asymmetry of a distribution, informing us of the direction of outliers, but not their frequency.

- Negative skewness indicates that the tail of the distribution is on the left.
- A positive skewness indicates that the tail of the distribution is on the right-hand side.
- A value of zero indicates that there is no asymmetry in the distribution, which means that the distribution is perfectly symmetrical.

Kurtosis

Kurtosis measures the degree of flatness of a distribution compared to a normal distribution.

- The kurtosis of a normal distribution is 0.
- If a distribution has a kurtosis of less than 0, it is said to be playkurtic, meaning that it tends to produce fewer and less extreme outliers than the normal distribution.
- If a distribution has a kurtosis greater than 0, it is said to be leptokurtic, meaning that it tends to produce more outliers than the normal distribution.

3.2.2 Shape of the Distribution

Using the univariate procedure first, we can observe that variables such as s3 or BMI are not normally distributed, but rather have a distribution tail that tends to the right and a rather low degree of kurtosis, so we could assume that these 2 variables are likely to produce outliers.

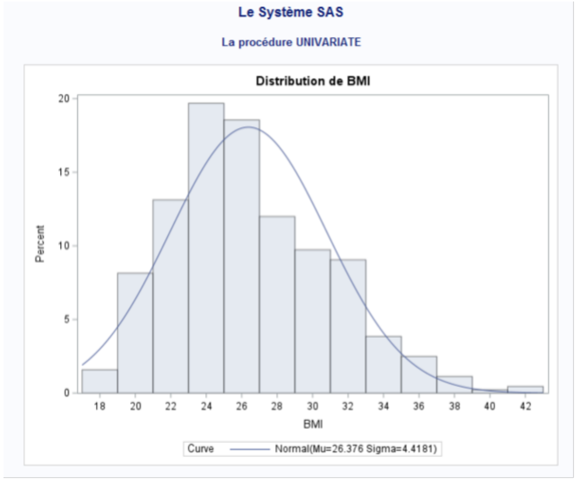


Figure 3.1: Distribution of BMI

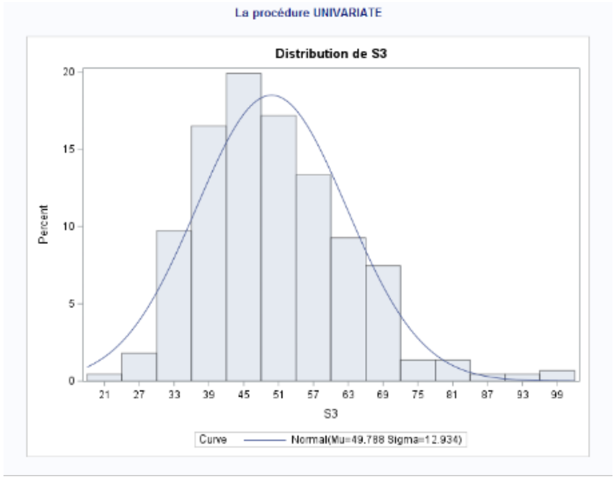


Figure 3.2: Distribution of S3

3.2.3 Boxplot and proc means

In 3.3 and 3.4, we can see the outliers in the boxplot for BMI and S3 respectively.

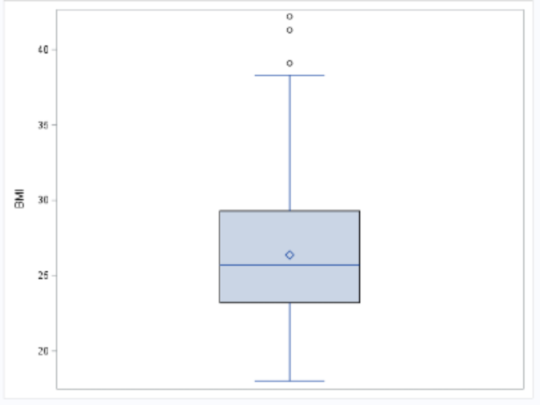


Figure 3.3: Boxplot of BMI

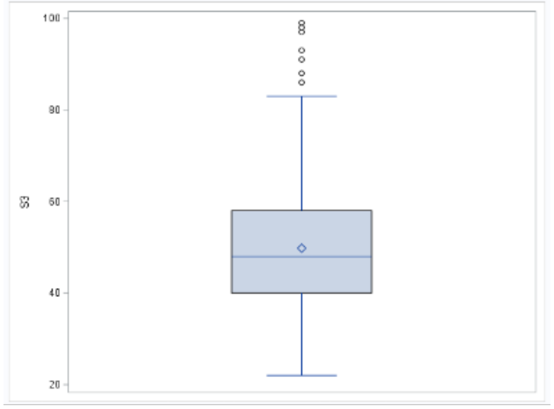
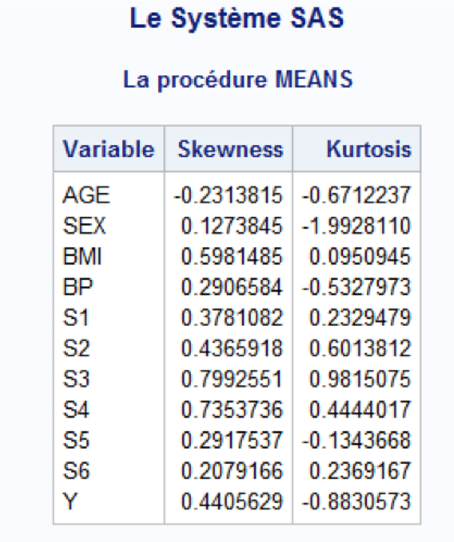


Figure 3.4: Boxplot of S3

Furthermore, according to the definition of the kurtosis, we notice outliers for others variables than S3 and BMI in 3.5 . Thus, the Proc sgplot and proc means confirmed our intuitions, and showed us that other variables such as S1, S2 , S3 or S6 are also polluted by outliers.



Le Système SAS

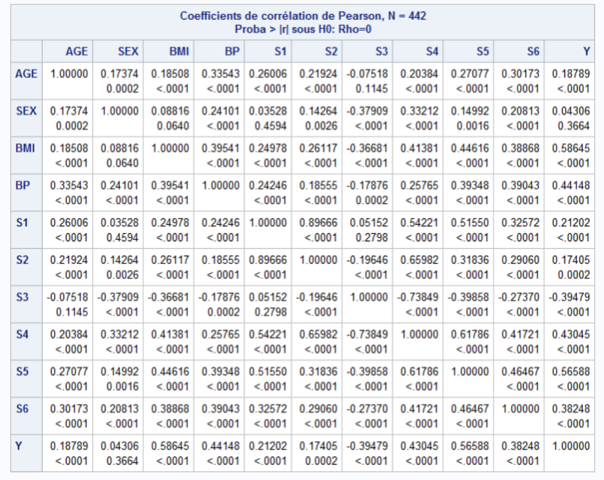
La procédure MEANS

Variable	Skewness	Kurtosis
AGE	-0.2313815	-0.6712237
SEX	0.1273845	-1.9928110
BMI	0.5981485	0.0950945
BP	0.2906584	-0.5327973
S1	0.3781082	0.2329479
S2	0.4365918	0.6013812
S3	0.7992551	0.9815075
S4	0.7353736	0.4444017
S5	0.2917537	-0.1343668
S6	0.2079166	0.2369167
Y	0.4405629	-0.8830573

Figure 3.5: Proc Means

3.2.4 Correlation

To see if there is a possible multicollinearity problem, we used the proc corr. Overall, we can observe strong correlations in both directions between all the biochemical variables from S1 to S6 (3.6), so there is potentially a multicollinearity problem in this dataset. The variables most correlated with the dependent variable are BMI, S5 and BP respectively, so we can assume from our results in (2.14)that the latter will be selected by the



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

Figure 3.6: Proc Corr

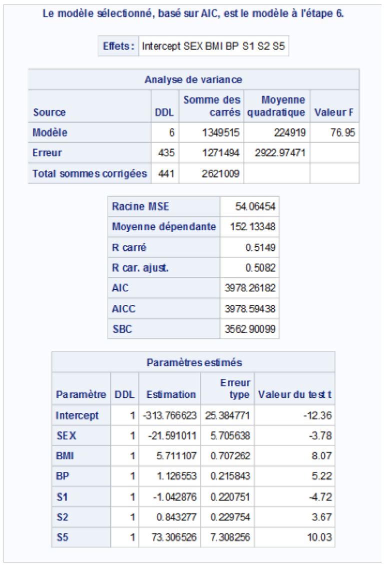
Stepwise method with the AICC/AIC pair, a method and criteria that are the most suitable for this type of data.

3.2.5 Regression

Before applying the stepwise method to our dataset, we ran a regression with the reg procedure. The results obtained reveal a coefficient of determination (R^2) of around 52%, indicating a fairly good specification of the model. However, among the six variables, AGE, S1, S2, S3, S4, and S6, none showed significance at the 1% threshold.

3.2.6 Proc GLM Select

If we now run the glm select procedure, we get the following results: 3.7



Le modèle sélectionné, basé sur AIC, est le modèle à l'étape 6.

Effets :	Intercept SEX BMI BP S1 S2 S5

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F
Modèle	6	1349515	224919	76.95
Erreur	435	1271494	2922.97471	
Total sommes corrigées	441	2621009		

Racine MSE	54.06454
Moyenne dépendante	152.13348
R carré	0.5149
R car. ajust.	0.5082
AIC	3978.26182
AICC	3978.59438
SBC	3562.90099

Paramètres estimés

Paramètre	DDL	Estimation	Erreur type	Valeur du test t
Intercept	1	-313.766623	25.384771	-12.36
SEX	1	-21.591011	5.705638	-3.78
BMI	1	5.711107	0.707262	8.07
BP	1	1.126553	0.215843	5.22
S1	1	-1.042876	0.220751	-4.72
S2	1	0.843277	0.229754	3.67
S5	1	73.306526	7.308256	10.03

Figure 3.7: GLM Select procedure

The stepwise method returns a model composed of the intercept, SEX, BMi, BP, S2 and S5. All of these are significant at the 1% level. The R-squared of the model is about 52%, which means that about 52% of the variation in the dependent variable is explained by 6 variables, which is pretty good for a complex problem like Diabetes.
Finally, other factors not included in the dataset, such as genetics, hypertension, diet and ethnic origin, could be taken into account to reinforce the credibility of the selected model.

Conclusion

In conclusion, the primary aim of this thesis was to highlight the different behaviors of algorithms depending on the data they face. In doing so, we gradually changed the shape of the data to highlight possible changes. This enabled us to gradually identify the various trends.

Ultimately, we can see that there is no absolute superiority between machine learning and statistical learning algorithms, despite the fact that the conclusions drawn by statistical learning algorithms are much more one-sided than those drawn by machine learning algorithms. Statistical Learning algorithms offer a compromise between overfitting and perfect fitting, whereas Machine Learning algorithms are more nuanced (with underfitting, overfitting, etc.) for all criteria. This opposition actually illustrates the very definition of these types of algorithms: Statistical Learning algorithms look for parameter interpretability, which enables them to be more radical in their choices, whereas Machine Learning aims for a model with good predictive capabilities, which in some ways means taking more precautions with the variables.

In addition to the procedures themselves, we set out to analyze the various statistical criteria using combinations. Of course, this is relative to the algorithm chosen, but we did see certain criteria come back more than others (SBC, CV, PRESS). We had the opportunity to play with our sample size, so it seems clear to us that it's important to have a large number of observations available for the algorithm to perform better.

Furthermore, when we looked at the noise in our model, we realized that it wasn't just our dataset that could cause our results to vary, but an "external" element to our variables, which could short-circuit the results depending on its size.

Finally, the conclusions of this study highlight the need to focus on the development of hybrid methodologies, integrating both the interpretable features of statistical learning algorithms and the predictive power of machine learning models.

Bibliography

[1] Azencott Chloé-Agathe. *Introduction au Machine Learning.* Feb. 2022. URL: https://cazencott.info/dotclear/public/lectures/IntroML_Azencott.pdf.

[2] Robert Cohen. *Introducing the GLMSELECT PROCEDURE for Model Selection.* Jan. 2006. URL: https://facweb.cdm.depaul.edu/sjost/csc423/documents/glmselect-summary.pdf.

[3] Josiane Confais and Monique Le Guen. *Premiers pas en régression linéaire avec SAS®.* Ed. by Centre d'Economie de la Sorbonne - CES UMR 8174. Documents de travail du Centre d'Economie de la Sorbonne 2007.47 - ISSN : 1955-611X. Oct. 2007. URL: https://shs.hal.science/halshs-00180861.

[4] M. Kumar and S.K. Rath. *Chapter 4 - Signal-to-Noise Ratio.* 2013. URL: https://www.sciencedirect.com/science/article/pii/B9781597497404000046.

[5] Tristan Mary-Huard. "Une introduction au critère BIC : fondements théoriques et interprétation". In: *Journal de la Societe Française de Statistique* 147.1 (2006), pp. 39–58. URL: https://hal.inrae.fr/hal-02654870.

[6] Christèle Robert-Granié and Bertrand Servin. "Modèle Linéaire mixte gaussien". Lecture - Formation Génopole de Toulouse et Interbio Niveau de la formation : Communauté Scientifiques. 2012. URL: https://hal.inrae.fr/hal-02803389.

[7] Deanna Schreiber-Gregory, M Jackson Foundation Henry, and Bader Karlen. "Regulation Techniques for Multicollinearity: Lasso, Ridge, and Elastic Nets". In: Jan. 2018. URL: https://www.pharmasug.org/proceedings/2019/ST/PharmaSUG-2019-ST-059.pdf.

[8] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 00359246. URL: http://www.jstor.org/stable/2346178.

[9] Hui Zou and Trevor Hastie. "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320. ISSN: 13697412, 14679868. URL: http://www.jstor.org/stable/3647580.