UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Monday 31st March, 2025

Empirical Evaluation of Machine Learning Procedures

DIABIRA Ibrahim
HAMZAOUI Marwan
GYURJYAN Tigran

Directed by: Phillipe DE PERETTI

Acknowledgements

We would like to express our sincere gratitude to Professor Philippe DE PERETTI, Director of the Master's program in Econometrics and Statistics and lecturer in Applied Econometrics of Linear Models.

His invaluable guidance, availability, and insightful advice have greatly contributed to the completion of this thesis. Always attentive, he patiently answered our questions and provided essential support, both in theoretical aspects and practical implementation. His assistance in writing the code and interpreting the results has been crucial to the success of our work.

We sincerely thank him for his dedication, high standards, and constant support throughout this project.

Abstract

Technological advancements have made data collection easier than ever before, raising the question of variable selection in econometrics and machine learning more prominently today than in the past. Particularly when the dataset contains a large number of predictors, it is essential to identify the most relevant ones to explain the target variable, in order to achieve a parsimonious model and good predictive performance. This paper revisits various methods, distinguishing between two categories: Statistical Learning methods (Forward, Backward, Stepwise) and methods used in Machine Learning (LASSO, LARS, Elastic Net). Several stopping criteria are considered, ranging from statistical criteria such as the F-test, to information criteria like AIC, and finally predictive criteria such as K-fold cross-validation and Press. The different methods are first studied under the assumption of independence among explanatory variables, then in the presence of correlation, outliers, and finally structural breaks. The results show that Machine Learning methods generally have a better ability to identify the correct model, while Statistical Learning methods tend to overfit. While Lasso and LARS often yield very similar results, Elastic Net stands out and appears to be more robust in the presence of correlation. Finally, while outliers seem to have a limited impact, structural breaks in the data tend to lead to incorrect models.

Keywords : Forward, backward selection, LARS, LASSO, Statistical learning, Machine learning, Variable selection, AIC, BIC, F-STAT, k-fold, cross-validation, SAS, proc glmselect, proc iml.

Contents

List of Figures

1 Introduction

As we increasingly deal with high-dimensional databases containing many potential explanatory variables, the construction of a good linear model to explain or predict a phenomenon relies on effective variable selection. Indeed, it is essential to identify the most relevant variables to have an interpretable model with good predictive capability. As early as the 20^{th} century, several authors began exploring this field, and criteria such as Mallows' C_p, AIC, and BIC emerged, aiming to compare and select models by considering both their goodness of fit and their parsimony. Tibshirani's 1994 article marked the birth of the Lasso, which introduced a change compared to Ridge regression by shrinking the coefficients of the least relevant variables to zero.

Thus, several variable selection methods exist and can be classified into two categories: Statistical Learning methods and Machine Learning methods. Both approaches aim to build models with predictive and explanatory capabilities based on data, with the difference being that Statistical Learning relies on statistical inference.

This thesis aims to provide an empirical evaluation of the different methods using generated data. More specifically, we will consider Forward, Backward, and Stepwise methods for Statistical Learning, and Lasso, LARS, and Elastic Net for Machine Learning.

To achieve this, the second section will revisit precise definitions of the various concepts. We will particularly focus on the theoretical functioning of the algorithms, the different stopping criteria present in the literature, and will distinguish between statistical, information-based, and predictive criteria. In the third section, we will define our metrics to evaluate the goodness of fit, as well as our data generation processes, considering the null hypothesis of independence in the data, followed by correlation, the injection of outliers, and finally structural breaks. We will see that Statistical Learning methods seem less effective at predicting the correct model, often leading to similar overfitting results. On the other hand, Machine Learning methods often yield very similar results between Lasso and LARS. Elastic Net, which combines Lasso and Ridge, often produces quite different results, which we will analyze in detail. Finally, the thesis will conclude with an empirical study on the Diabetes dataset, which we will analyze initially and then attempt to propose a variable selection based on our results obtained at the end of the data generation process.

All outputs presented are derived from SAS, with data generated using the IML procedure, and variable selection performed using the glmselect procedure, considering different selection and stopping criteria. The definitions of the algorithms and criteria are primarily drawn from the original articles (see Bibliographiy for full references).

2 Theoretical Approach

2.1 Statistical Learning

2.1.1 Forward Selection

One of the earliest formal descriptions of the forward selection method can be traced to M.A. Efroymson in his 1960 paper titled "Multiple Regression Analysis", published in *Mathematical Methods for Digital Computers*. [3]
It is a sequential process that is relatively simple to understand. Specifically, the forward selection technique begins with only the intercept and no explanatory variables, known as a null model. Progressively, the features that are most correlated with the dependent variable and that most improve the model's fit are added. So at each successive step, the variable among those not yet included in the model that contributes the most to the reduction of the residual sum of squares (RSS) will be added to the model. Typically the F statistic is used to gauge the improvement in fit, and the process stops when the significance level for adding any effect is greater than some specified entry significance level, as we will see in the section on selection criteria. Whitout any stopping criteria, Forward selection should continue until all variables have been included in the model.

2.1.2 Backward Selection

The functioning of backward elimination [15] is quite similar to the forward selection method, with the difference that the process begins with the full model that includes all the explanatory variables. Then, at each step, the feature that contributes the least to the model is eliminated. The process continues until a specific stopping criterion is met or no variables remain in the model. As with forward selection, the decision to remove a variable can be based on its significance level. For example, at each step, the feature with the highest *pvalue* of the F-test or T-test is removed. Other criteria like the R^2_{adj} or AIC will be discussed later.

2.1.3 Stepwise Forward

The problem with the two selection methods presented above is that the addition (Forward) and the elimination (Backward) are made without taking into consideration their effect on the other features already included in the model. In other words, these methods do not take into account the correlation between features, which can lead to erroneous decisions in the case of multicollinearity. Indeed, a variable previously added by the forward method may become statistically insignificant after adding a new variable. Similarly, variables removed by the backward method may become significant after removing other variables. The fact that forward and backward operate in only one direction is not optimal.
Then, Stepwise Forward Selection combines the previous methods and partially addresses their limitations. The process begins similar to forward selection, with an empty model, and the feature that is most strongly correlated with the dependent variable is added to the model. A key innovation is that after adding a variable, the stepwise method reassesses the statistical significance of all variables already included in the model. Once non-significant variables are eliminated, the forward process restarts, and this cycle continues until a stopping criterion is met.
This method is more computationally demanding, but it has the advantage of testing interactions

between features during the selection. Thus, stepwise selection is more likely to produce a better subset compared to backward or forward methods, though it does not guarantee the optimal subset of each size (Rawlings and al. 1988).[7]

2.2 Machine Learning

2.2.1 LASSO Regression

The Least Absolute Shrinkage and Selection Operator (LASSO) was first introduced by Robert Tibshirani in 1996 in his paper titled "Regression Shrinkage and Selection via the Lasso", published in the *Journal of the Royal Statistical Society*[13]. To achieve a parsimonious, interpretable model and reduce overfitting when the number of predictors is large, Lasso shrinks some regression coefficients toward zero and exactly sets the coefficients of irrelevant predictors to zero. This dual mechanism of shrinkage and automatic feature selection makes it particularly effective for sparse modeling. More precisely, the Lasso minimizes the SSE as in classical linear regression, but with the addition of an "L1 penalty", which explains its ability to produce interpretable models, as we have already said. The lasso can be defined in two equivalent ways, as follows :

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

or,

$$\arg\min_{\beta} \sum_{i=1}^{T} \left(y_i - \left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \right) \right)^2 \quad \text{s.c.} \quad \sum_{j=1}^{p} |\beta_j| \leq \tau$$

We can see that a hyperparameter λ, which controls the strength of the penalization is introduced. When it is large (τ small), the penalty is strong resulting in a parsimonious model. On the other hand, if λ is small (τ large), the Lasso regression will approach the Ordinary Least Squares (OLS) solution. We can therefore understand the importance of choosing this parameter. In practice, LASSO regression must be estimated for a set of values taken by λ and the final model will be selected based on criteria or cross-validation. Even though the Lasso is widely used, it has some limitations. Especially when the number of predictors is larger than the number of observations (p>n), the Lasso selects at most n variables, because of the nature of the convex optimization problem. Morevoer, if the features are highly correlated, Lasso will tend to arbitrarily select one of them, reducing the others to zero.

2.2.2 The Elastic-Net regression

The Elastic net regression was introduced by Hui Zou and Trevor Hastie in 2005 in their paper called "Regularization and variable selection via the elastic net" published in the *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*[16]. To address the limitations of the lasso that we discussed earlier, the authors proposed this new method, which combines Ridge regression and the Lasso. In addition to the two cases we mentioned earlier, R. Tibshirani also found that in the case where n >p (where the number of observations is greater than the number of variables) and there is high correlation among the variables, empirically, the Lasso tended to be

outperformed by Ridge regression. Before going into more detail about the Elastic Net, let's first recall Ridge regression :

$$SSE_{\text{ridge}}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left(\mu + \beta^T X_i\right)\right)^2 + \lambda\|\beta\|^2$$

Despite its similarity to the Lasso, Ridge regression has a different function, as it does not enable variable selection but only shrinks the coefficients.This comes from the fact that Ridge uses an L_2 penalty (Euclidean norm) rather than an absolute norm like the Lasso. Among its advantages, we can notably mention that Ridge regression is effective in handling multicollinearity, reducing prediction variance, and stabilizing coefficient estimates.
We can now define the Elastic Net as follows :

$$\beta_{Ela} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{T}\left(y_i - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|l_1 + (1-\alpha)\|\boldsymbol{\beta}\|{l_2}^2\right)$$

With, $\lambda\left(\alpha\|\beta\|l_1 + (1-\alpha)\|\beta\|{l_2}^2\right)$, the penalty function. We can see that it combines the L1 norm et the L2 norm. λ controls the strength of regularization, and α determines the mix between the two norms. If $\alpha = 1$, we obtain the Lasso, and if $\alpha = 0$, we obtain the Ridge. Another way to write it is :

$$\arg\min_{\beta} \sum_{i=1}^{T}\left(y_i - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 \text{ u.c. } \quad \alpha\sum_{j=1}^{k}|\beta_j| + (1-\alpha)\sum_{j=1}^{k}|\beta_j|^2 \leq \tau$$

So the Elastic net simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables. Several studies have shown that the elastic net often outperforms the lasso in terms of prediction accuracy (Zou and Hastie 2005)[16].

2.2.3 Incremental Forward Stagewise

Before presenting LARS, let's briefly define Forward Stagewise. The Forward Stagewise algorithm (FS) is a type of boosting algorithm for the linear regression problem. It produces a coefficient profile by repeatedly updating the coefficient of the variable, which is the most correlated with the current residuals, towards the OLS estimator until the moment when another feature is more correlated with the recomputed residuals. The operating steps are as follows :

Step 1 : Start with the residuals $r^i = y - \overline{y}$, $\beta = 0$, $(\beta_1^i, \beta_2^i, \ldots, \beta_k^i)$, $i = 1$,
Step 2 : Find the variable x_j most correlated with r^i,
Step 3 : Update the coefficient β_j as follows: $\beta_j^{i+1} = \beta_j^i + \delta_j^i$
with $\delta_j^i = \epsilon \cdot \text{sign}(\text{corr}(r^i, x_j))$, ϵ being a small positive scalar,
Step 4 : Update the residuals as follows: $r^{i+1} = r^i - \delta_j^i x_j$ and repeat steps 2 and 4 until no correlation can be found between the residuals and any other explanatory variable in the set.

2.2.4 LARS

The Least Angle Regression (LARS) was introduced in 2004 by B. Efron, T. Hastie, I. Johnstone and R. Tibshirani [2]. The Lars algorithm is a stylized version of the Stagewise Procedure, which is a "useful and less greedy version of the traditional forward selection methods" (Effron and al. 2004). It adds one predictor at a time, selecting the one which reduces the prediction error the most. So the Lars algorithm takes up the main points of the Incremental Forward Stagewise Regression and can be described as follows :

- Step 1 : Center and scale the predictors, apply additional normalization so that the sum of squares of each regressor equals 1. Start with the residuals $r^i = y - \overline{y}$, and set $\beta = 0$, $(\beta_1^i, \beta_2^i, \ldots, \beta_k^i)$, $i = 1$.
- Step 2 : Find the variable x_j most correlated with r^i.
- Step 3 : Gradually move β_j toward its correlation coefficient with r obtained using OLS, until a variable β_l exhibits a stronger correlation with the observed residuals. Increase the information set.
- Step 4 : Gradually move (β_j, β_l) toward their OLS estimators with the residuals, until a variable β_m exhibits a stronger correlation with the observed residuals. Increase the information set.
- Continue until all k predictors are included in the model.

The Lars-Lasso relationship

In contrast to Lasso, the LARS algorithm does not enforce restrictions on the coefficients. Without any constraints, the two methods generally lead to different results. However, a modification of LARS, presented in the original paper, can make LARS produce the same results as Lasso. Simply put, this modification involves ensuring that the selected features respect a sign restriction, such that the sign of the coefficient is the same as the sign of the correlation as follows :

$$\text{sign}(\hat{\beta}_j) = \text{sign}(\hat{c}_j) = s_j$$

If this condition is violated, the Lars is modified by moving out the problematic variable before continuing. In the same way, with another conditions the Lars algorithm could reproduce the IFS results.

2.3 Feature Selection Criteria

In the previous section, we began discussing the fact that different selection criteria are used in the context of variable selection. This section will focus on enumerating several criteria: statistical tests, explanatory criteria, and predictive criteria.

2.3.1 The T-test

The T-test was introduced in 1908 by William Gausset, known as Student[12]. It is used to determine if there is any significant difference between the means of two groups. More precisely it

is commonly used to evaluate if an explanatory variable has a significant impact on the dependent variable. Then we have :

- $H_0 : b_i = 0$ (the explanatory variable has no significant effect)
- $H_1 : b_i \neq 0$ (the explanatory variable has a significant effect)

And we have :

$$\frac{\hat{b}}{\hat{\sigma}} \sim T(N - P - 1)$$

With N the size of the samble, P the number of features in the linear regression and σ the estimated standard error of the model. When N is large enough, the Student Law can be approximated with a Gaussian Law. Once the T-stat has been calculated, it is compared to a quantile of the Student distribution for a chosen error threshold α. We accept H_0 if $t < \mathrm{q}_{1-\alpha/2}$ and we reject it if $t > \mathrm{q}_{1-\alpha/2}$. We can also use the $p - values$. The T-test evaluates the variables separately, without considering the joint effect of other variables. In the case of multicollinearity, the T-test may incorrectly estimate the individual importance of the variables.

2.3.2 The Partial F-test

The partial F-test, is a specific version of the F-test introduced by Fisher in 1935[4]. It is used to compare two models and to determine whether there is a statistically significant difference between a regression model and a constrained version of the same model.

F-to-enter

For example, in the Forward Selection method discussed earlier, the typical criterion used is the ratio of the reduction in the residual sum of squares (RSS) achieved by adding the next candidate variable to the residual mean square (RMS) of the model that includes that variable. This criterion is expressed by Rawlings and al. (1988) [7] as a critical 'F-to-enter' or in terms of a critical 'significance level to enter' (SLE), where F is the F-test of the partial sum of squares of the variable being considered. At the first step the F-stat will be :

$$F = \frac{\frac{\mathrm{SSE}_{\mathrm{null}} - \mathrm{SSE}_{\mathrm{model}}}{1}}{\frac{\mathrm{SSE}_{\mathrm{model}}}{N-2}}$$

with $\mathrm{SCR}_{\mathrm{null}}$: the model which contains only the intercept, and $\mathrm{SCR}_{\mathrm{model}}$: the residual sum of squares which contains the intercept and the candidate variable. At the second step we will use the residual sum of squares of the model with the intercept and the variable that have been added.

F-to-remove

In the same way, in the backward method, we start from the full model which contains the intercept and all the features, and we compute for each variable the partial F-stat, which can be expressed as an F-to-remove or "significance level to remove" :

$$F = \frac{\frac{\mathrm{SSE}_{\mathrm{restricted}} - \mathrm{SSE}_{\mathrm{full}}}{1}}{\frac{\mathrm{SSE}_{\mathrm{full}}}{N-(p+1)}}$$

So an "acceptance threshold" (significance level) could be set for adding or removing a variable. Typically, we refer to the p-values of this test to make a decision. However, according to Berk (1978), this 'F-test' uses the data to select the most favorable variables, which introduces biases that invalidate these ratios as tests of significance. Therefore, we should consider them as stopping rules rather than as classical tests of significance.

2.3.3 R^2 and R^2_{adj}

Before defining R^2[14], we need to briefly recall the analysis of variance equation. Simply put, the variance in the dependent variable y is the sum of the variance explained by the model and the residual variance. The coefficient of determination (R^2) is an indicator that measures the proportion of the variance in the dependent variable (y) explained by the explanatory variables (X) in a regression model. Mathematically, it is calculated as follows:

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}}$$

with SSE : Sum of Squared Errors and TSS : The Total Sum of Squares.
Ranging between 0 and 1, the R^2 is used to evaluate the quality of a model, with 0 when the variables included in the model don't explain the dependent variable and 1 when they explain perfectly the dependent variable. But in many cases, the R^2 is not suffisant to judge the quality of a model, because of its defaults, as the fact that it grows mechanically when we add an additional variable. That is why we introduce the R^2_{adj}, which can be computed as follows :

$$R^2_{\text{adj}} = 1 - \frac{\text{MS(Res)}}{\text{MS(Total)}} = 1 - \frac{(1-R^2)(n-1)}{n-p}.$$

With MS : Mean squares, n : the number of observations, p : the number of predictors.
Then, the primary objective is to correct the mechanical effect described above. Indeed, the adjusted R^2 removes the impact of the degrees of freedom and gives a quantity that is more comparable than R^2 over models involving different numbers of parameters (Rawlings al. 1988). So the simplest model with the adjusted R^2 near its upper limit is chosen as the "best" model. As the R^2, the adjusted R^2 is not perfect, as it can be negative and difficult to interpret.

2.3.4 Mallow's Cp

The C_p statistic was introduced by Mallows in 1973[6], and is a statistical measure used to evaluate the fit of a regression model estimated via ordinary least squares (OLS). The C_p statistic is computed as :

$$C_p = \frac{\text{SSE}_p}{s^2} + 2p - n$$

where SSE_p is the sum of squares errors from the p-variable subset model being considered and s^2 is an estimate of σ^2, from the model containing all independent variables. When all important independent variables are included, the sum of squares errors is an unbiased estimate of (n-(p+1))σ^2, and the C_p is close to p, the number of predictors. When important variables are omitted from the model, C_p is greater than p. So a value of C_p close to p indicates a little bias in

Ms(Res) as an estimate of σ^2. To summarize :

- The ideal value of the C_p is approximately p, which means that the model includes all the relevant features.
- If the value of C_p is higher than p, then some important variables have been omitted, and there is underfitting.
- The goal is to minimize C_p, while keeping it close to p.

2.3.5 The Maximum of Likelihood

Before defining the explanatory criteria, such as AIC first, followed by BIC, we will first recall the definition of the Maximum Likelihood Estimation, which consists of maximizing the Likelihood function written as follows :

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

with f, the theoretical density function, θ the vector of parameters. The θ which maximizes this function is the one for which the observed data is the most likely, given the theoretical distribution. To simplify calculations, the logarithm of the likelihood (log-likelihood) is often used, as it transforms the product into a sum, making the optimization process more computationally efficient and numerically stable.

2.3.6 AIC and AIC$_c$

The Akaike Information criterion (AIC) was introduced by Hirotugu Akaike in 1973 [1] and is used to compare several models by evaluating the relative quality of a statistical model. The AIC selects the model with the best compromise between quality of fit and model complexity. In other words, it trades off precision of fit against number of parameters in the model and the one with the lowest AIC will be the most appropriate. The AIC is defined as follows :

$$\text{AIC} = -2\ln(L) + 2p$$

- L: The model's maximum likelihood.
- p: The number of parameters estimated in the model.

So the more variables there are, the greater the penalty. However, the AIC can be biased in small samples. To address this, the AIC_c [5] is used, which helps to avoid the overfitting caused by the AIC when the number of observations (N) is smaller than the number of variables (p).

$$\text{AICc} = -2\ln(L) + 2p + \frac{2p(p+1)}{N-p-1}$$

2.3.7 BIC and SBC

The AIC, that we presented above is widely used, however it tends to select models with more variables than the true model. That's why other criteria have been developed, which impose a heavier penalty on complexity. Among them, there are notably :

- The Sawa Bayesian Information Criterion (BIC) [10] :

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

- The Schwarz Bayesian Criterion (SBC) [11]:

$$\text{SBC} = -2\ln(L) + p\ln(N)$$

or :

$$\text{SBC} = n \ln\left(\frac{\text{SSE}}{n}\right) + p\ln(n)$$

We can see that the SBC assigns the coefficient ln(n) to the number of variables included in the model, as opposed to the coefficient 2 used in the AIC. This demonstrates the heavier penalization mentioned earlier, since the term ln(n) increases with the sample size. Note that the Sawa's BIC and SBC are really closed and often confused.

2.3.8 Predictive Criteria : K-Fold and PRESS

We have seen that in Machine Learning methods, especially for the Lasso, a criterion is needed to choose the value of λ. The previously discussed criteria can be used to determine that value by choosing the one which minimizes the AIC, for example. We can also use another type of criterion : predictive ones. In particular, we have cross validation (k-fold) and PRESS. In k-fold cross validation, the data is split into k roughly equal-sized parts. One of these parts is set aside for validation, and the model is fitted on the remaining $k-1$ parts. The fitted model is then used to compute the predicted residual sum of squares on the validation set. The process is repeated for each of the k-folds, and the sum of the predicted residual sum of squares is used to estimate the prediction error.The value of λ, which minimizes this prediction error is then selected.
So, we need to choose the value of K. The extreme case occurs when K=N, meaning that the model is estimated N times and validated on each of the N observations in the sample, and is known as leave-one-out cross validation. In that case the estimator of the prediction error is approximately unbiased, but it may have high variance because the training sets, which consist of nearly all the observations, are very similar to each other (Hastie and al., 2001). On the other hand, when K=5 (a typical value), the situation is reversed : cross validation has lower variance, but bias may become an issue. Thus, there is a trade-off between variance and bias.

3 Data Generating Process

To compare and analyze the different methods and criteria for variable selection, we generated a dataset and used our own metrics to evaluate the performance of variable selection. Specifically, we created a dataset with 100 observations and 50 variables. The true model, however, relies on 7 relevant variables and can be written as follows :

$$y = \alpha - 0.7X_1 + 0.8X_2 + 0.4X_3 + 0.67X_4 + 0.77X_5 - 0.25X_6 + 0.44X_7 + \epsilon$$

In this section, we analyse several specific characteristics of our data, including independence, multicollinearity, breaks, and the presence of outliers.

3.1 Metrics

Before going into more detail about the various data generation processes, we will briefly define our four metrics. Each procedure was evaluated over 1000 repetitions. We initialized counters for each metric, which were updated after each iteration. Following this, we were able to calculate the probabilities of the following four cases:

- **Perfect fitting** : Refers to the case where the procedure correctly selects all the variables of the true model that we defined above. In other words, it will be considered a perfect fit if the method used selects exactly the variables X_1 to X_7, neither more nor less.
- **Overfitting** : corresponds to the case where the procedure correctly identifies all the relevant variables but also includes additional irrelevant features in the model. The interpretation of the model can be compromised, and its performance on unseen data may suffer due to increased variance. For example, if the method correctly selects the variables X_1 to X_7 but also includes X_8 and X_9, it will be considered an overfit.
- **Underfitting** : corresponds to the case where the variable selection correctly identifies some, but not all, of the relevant variables. In other words, the selection is smaller than the actual model, identifying only a subset of the correct variables. Then, the resulting model is too simplistic to capture the true underlying relationships. For example if the method used selects only X_1, X_2, X_3, X_4, X_5, ignoring X_6 and X_7, it will be considered as an underfit.
- **Wrong Model** : If the selection procedure results in a model that does not correspond to any of the three cases described above, it will be considered a wrong model. Then, the resulting model is either a mix of relevant and irrelevant variables (but not all relevant variables, as that would indicate overfitting), or it contains no relevant variables at all, or it is an empty model with no variables selected except for the intercept.

3.2 DGP with Independance

To generate data under the assumption of independence, we used a multivariate normal distribution along with an identity matrix of order 50 as the variance-covariance matrix for our explanatory variables. The choice of the identity matrix ensures that the variables are uncorrelated (and standardized, by the way). Next, the value of y was computed, and a dataset was created containing y and the matrix of predictors. The next step involved regressing y on X_1 to X_{50} using the various algorithms and criteria, and then calculating the metrics. As we have already mentioned, these steps were repeated 1000 times, which allowed us to estimate the probabilities for each scenario and gain a clear overview of the efficiency of the different methods.

3.3 DGP with correlation between the variables

After studying the case of independence, we will now introduce collinearity into our data. More specifically we will consider :

- **Internal correlation** : When the colinearity concerns only the explanotory variables in the true model (X_1 to X_7)
- **External correlation** : Which includes Internal correlation but also introduces collinearity involving external variables to the true model.

Internal Correlation

To generate data with Internal correlation and compute the metrics, the steps remained the same as in the independence case, but we modified the variance-covariance matrix using a Toeplitz matrix :

Intern_Correlation							
	X1	X2	X3	X4	X5	X6	X7
X1	1	0.8	0.75	0.7	0.65	0.6	0.55
X2	0.8	1	0.8	0.75	0.7	0.65	0.6
X3	0.75	0.8	1	0.8	0.75	0.7	0.65
X4	0.7	0.75	0.8	1	0.8	0.75	0.7
X5	0.65	0.7	0.75	0.8	1	0.8	0.75
X6	0.6	0.65	0.7	0.75	0.8	1	0.8
X7	0.55	0.6	0.65	0.7	0.75	0.8	1

Figure 1: Internal Correlation among 7 explanatory variables

This allowed us to introduce correlation between the first 7 variables (so only the variables Internal to the model), which decreases linearly.

For the remaining variables that are not part of the true model, we constructed an identity matrix of order 43 to avoid any correlation. We therefore have:

- A matrix called $Xcorr$, containing the 7 correlated variables in the model.
- A matrix called $Xind$, containing the 43 other external variables, which are not correlated.

We then created a dataset by concatenating y, X_{corr}, and X_{ind}, and followed the same steps as before.

External Correlation

To introduce external correlation, we followed the same procedure as for the Internal correlation, but with an extended Toeplitz correlation matrix of 15 variables.

This matrix includes the 7 model variables and 8 other external variables. As before, the variance-covariance matrix of the remaining variables was defined as an identity matrix of order 35.

Extern_Correlation															
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
X1	1	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5	0.45	0.4	0.35	0.3
X2	0.95	1	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5	0.45	0.4	0.35
X3	0.9	0.95	1	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5	0.45	0.4
X4	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	0.5	0.45
X5	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	0.5
X6	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
X7	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
X8	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
X9	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
X10	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
X11	0.5	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
X12	0.45	0.5	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
X13	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1	0.95	0.9
X14	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1	0.95
X15	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1

Figure 2: Correlation involving extern variables

We therefore have :

- A matrix $Xcorr$ containing the 15 correlated variables.
- A matrix $Xind$ containing the other 35 uncorrelated variables.

We then created a dataset by concatenating y, X_{corr}, and X_{ind}, and followed the same steps as before.

3.4 DGP with Outliers

The next step of our study was to test the performance of the different algorithms in the presence of outliers. To generate data with outliers, we drew X from two normal distributions. We obtained the mixture of the two distributions by first drawing a random number u from a uniform distribution $U(0, 1)$. If $u \leq 0.9$, the observation is drawn from $N(0, 1)$; otherwise, it is drawn from $N(4, 1)$. This process ensures that approximately 90% of the data comes from $N(0, 1)$ and 10% from $N(4, 1)$, introducing outliers into the dataset. Mathematically, this can be expressed as:

$$X_{i,j} = \begin{cases} N(0,1) & \text{if } u \leq 0.9, \\ N(4,1) & \text{if } u > 0.9, \end{cases}$$

where $u \sim U(0, 1)$.

We then created a dataset by concatenating y and X, and followed the same steps as usual. This configuration makes it possible to test the robustness of our analysis methods in the face of data contaminated by outliers.

3.5 Data Generating Process with Structural Break

Finally, the last data generation process was conducted under the assumption of a structural break in the data. A structural break, often referred to as a structural change, describes a significant alteration, disruption, or discontinuity in the relationships between variables within a statistical model. To model this, we split our observations into two parts and used a second set of coefficients (β). The structural break occurs at the 50th observation, which means that all observations before this point are modelled using the first set of coefficients, whereas all

observations after this point are modelled using the second set of coefficients. Using a loop, the first 50 values of y were computed with β as follows:

$$\text{For } N = 1 \text{ to } 50 :$$

$$\beta = (-0.7, 0.8, 0.4, 0.67, 0.77, -0.25, 0.44)$$

and the remaining fifty were computed as follows:

$$\text{For } N = 51 \text{ to } 100 :$$

$$\beta = (0.67, -0.25, 0.8, -0.25, 0.77, -0.7, 0.4)$$

The resulting dataset allows us to analyze the impact of the structural break and evaluate the performance of methods designed to detect such changes.

3.6 DGP mixing the different scenarios

After testing the various variable selection procedures in the 6 previous cases separately, we combined the different scenarios.

External correlation & structural break

To generate data under these two assumptions, we combined the steps described earlier. We specified two variance-covariance matrices : a Toeplitz matrix for the true variables of the model and an identity matrix for the remaining variables. Next, we split the data into two parts, with the coefficient vector β changing at the 50th observation. As always, we created the dataset by concatening y and X, and then proceeded with variable selection.

Structural break & outliers

Once again, the data generation process simply combines the steps described in each scenario. Here, 90% of the observations are drawn from a $\mathcal{N}(0, 1)$ distribution, while the remaining 10% are drawn from a $\mathcal{N}(4, 1)$ distribution to introduce outliers. The dataset is then split into two parts, with the coefficient vector β changing at the 50th observation to simulate a structural break. The subsequent steps remain unchanged.

Internal correlation & outliers

In this scenario, the data generation process differs slightly from previous cases. To maintain the normality of the data while introducing both correlation and outliers, we used the Iman-Conover transformation [8]. First, we generated independent variables from a multivariate normal distribution with an identity matrix as the variance-covariance matrix. We introduced then outliers in the same way as described previously. Next, we defined a Toeplitz correlation matrix to induce a decreasing correlation structure among the first 7 variables.

The Iman-Conover transformation is applied to these variables, rearranging the data to match the specified correlation while preserving the original distributions, including outliers. As a result, the final dataset exhibits both the desired correlations and the characteristics of the initial distributions, including the presence of outliers.

4 Results

The section below is dedicated to the results obtained from the various data generation processes. To do this, we used the GLMSELECT procedure in SAS [9], which allows variable selection by specifying both a choice criterion and a stop criterion. The criterion specified in CHOOSE is used to select the final model. It is evaluated at each stage, and the procedure selects the model that gives the optimal value for the chosen criterion. The STOP option is used to end the selection process when it is no longer possible to improve the specified criterion. For the statistical learning procedures, we combined the criteria presented previously in Choose with two Stop criteria corresponding to the significance to stay/enter, and for the machine learning methods we tested all the possible combinations.

4.1 DGP with independance

Statistical Learning

The results we obtained were very similar for the three methods. There is a very high probability of overfitting (between 80 and 99% depending on the case) and very little perfect fitting. Despite this, our models show neither underfitting nor wrong selection.

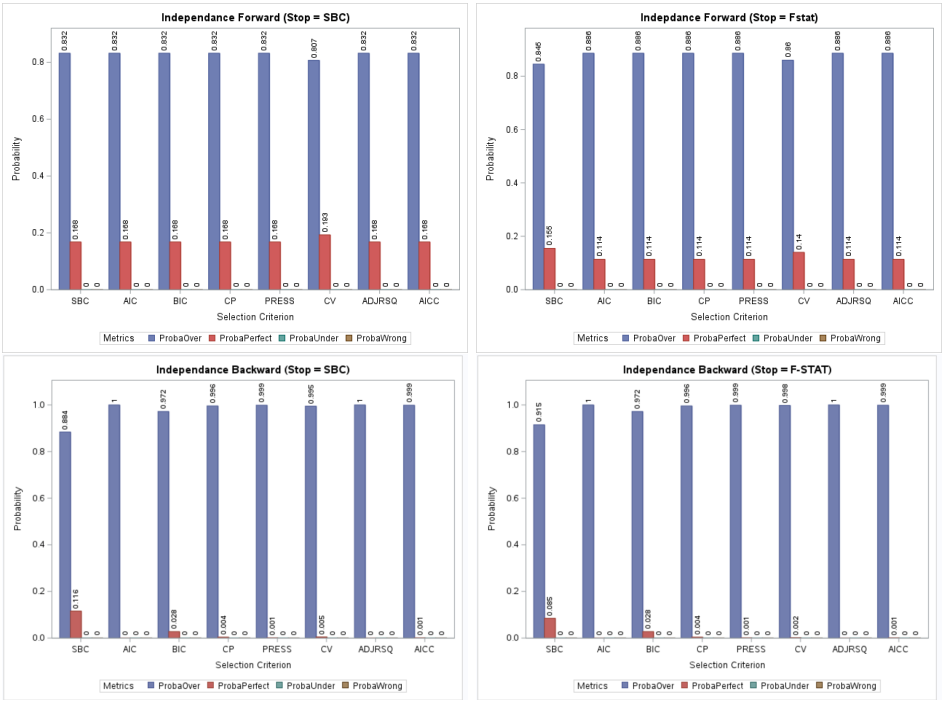


Figure 3: Forward & Backward with Independance

Indeed, the forward and stepwise methods show the highest probability of perfect fitting between 15% and 20% depending on the case, while the backward method has a very low probability, close to 0 and reaching a peak of 10% with SBC as the choice criterion. We can see that the SBC and the CV as selection criteria often give better results, as they have a marginally lower overfitting rate and a higher perfect fitting rate than the others.

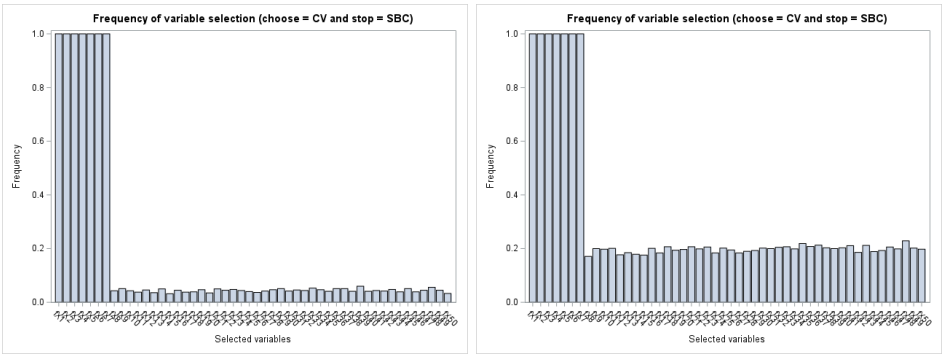


Figure 4: Variable Selection frequency with Forward & Backward (Stop = SBC)

The frequency graphs show that our 7 variables are always selected, along with additional variables, which explains the high probability of overfitting. However, we can see that the additional variables do not have a high frequency for forward, which leads us to believe that models with overfitting may have few additional variables, and therefore that some cases may be very close to the true model.

Machine Learning

For all criteria, the results appear very similar between LASSO and LAR. However, there is still an issue of overfitting, as almost all criteria specified in the STOP option exhibit overfitting at around 75%, with perfect fitting for the remainder, except for R^2_{adj}, which reaches 90% overfitting. Below are some results using three stopping criteria: k-fold Cross-Validation SBC and the R^2_{adj} .

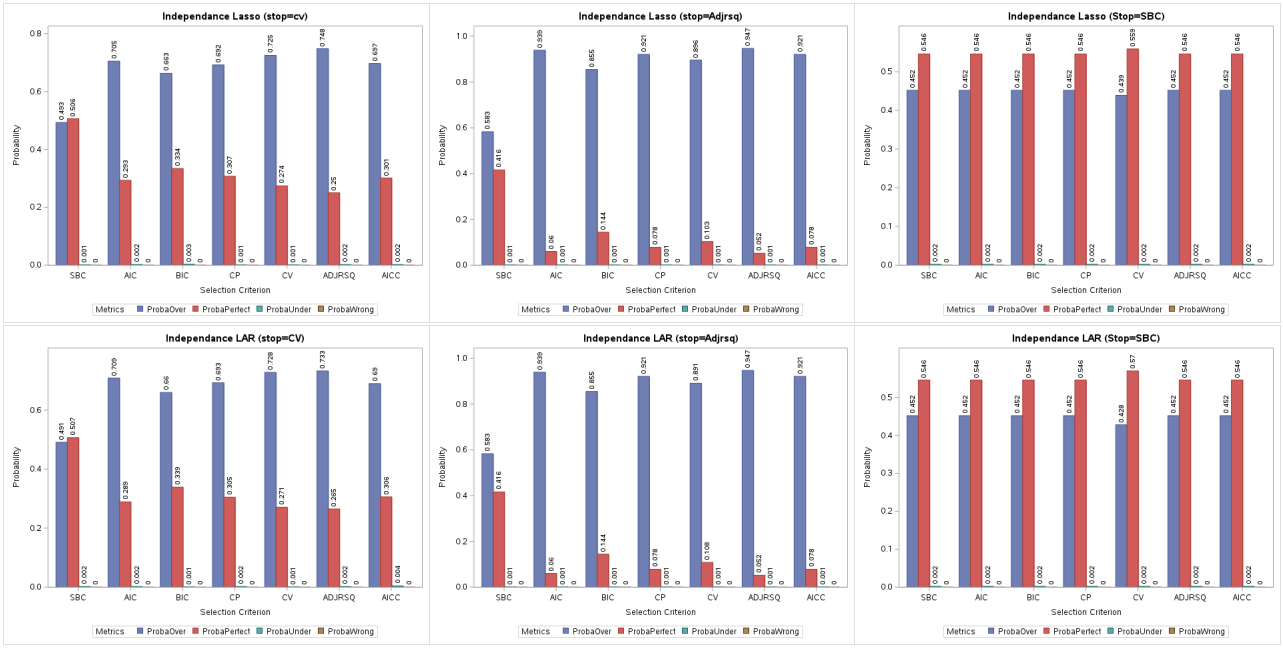


Figure 5: LASSO & LAR with Independance

We can see that the SBC as a stop criterion is better, because it allows to reach higher perfect fitting than overfitting for all choose criteria.The combination with choose = CV and stop=SBC gives the best results, reaching almost 56% of perfect fit. For all the other criteria, the SBC as a Choose option is always the one which allows the highest probability of perfect fitting. This is likely due to its stricter penalty that limits overfitting.

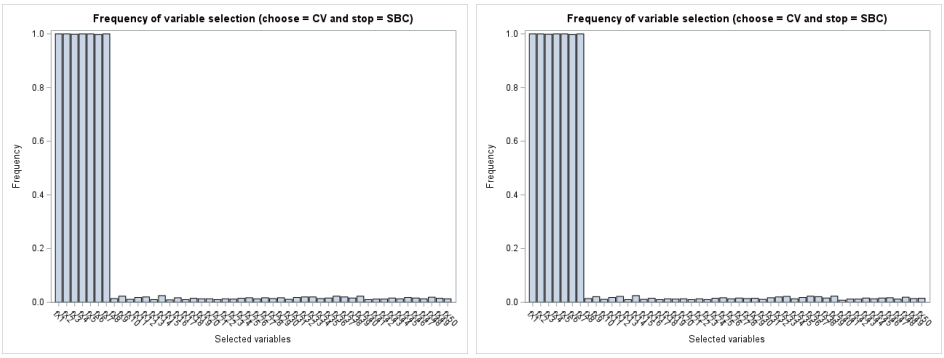


Figure 6: Variable Selection frequency with LASSO & LAR (Stop = SBC)

The frequency graphs above clearly show that both methods always correctly select all relevant variables but also include additional variables outside the model (left for Lasso and right for LAR).

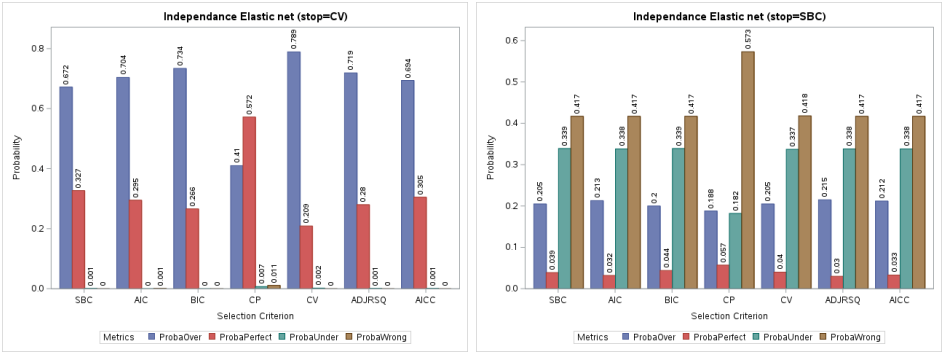


Figure 7: Elastic Net with Independance

For the Elastic Net, the results are quite different. For almost all stopping criteria, there is still an issue of overfitting, but underfitting and wrong models are also present, especially for the information criteria, with very few perfect models. The best results are obtained when using Cross-Validation as the stopping criterion, while SBC seems to produce the worst results, as it mainly leads to wrong models for all choose criteria. The combination with choose = C_p and stop = CV gives the best results, reaching 57% of perfect fit.

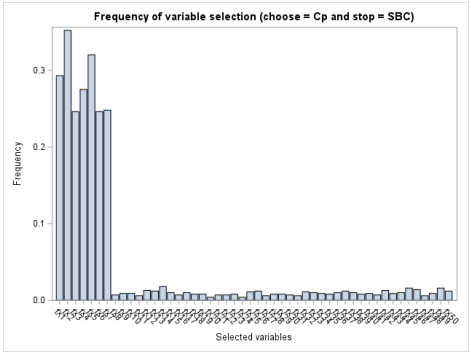


Figure 8: Variable selection frequency with Elastic Net (Stop = SBC)

We can see that with SBC as the stopping criterion, Elastic Net selects the true features with a probability lower than 0.5, and external features are rarely selected. This explains the fact that the wrong models produced are mainly models containing some of the true variables along with a few external features. We can also see that the true variables that are selected the most are the ones associated with the highest coefficient β.

4.2 DGP with Correlation

Internal & External Correlation

Statistical Learning

We observe that in the context of Internal correlation, the results we obtain are very similar for each of the selection methods and each of the combinations of criteria. We obtain the same trends as in the case of independence, with a very high probability of overfitting, a lower probability of perfect fitting and an absence of underfitting and wrong models. Here too, the CV and SBC choice criteria seem to be better.

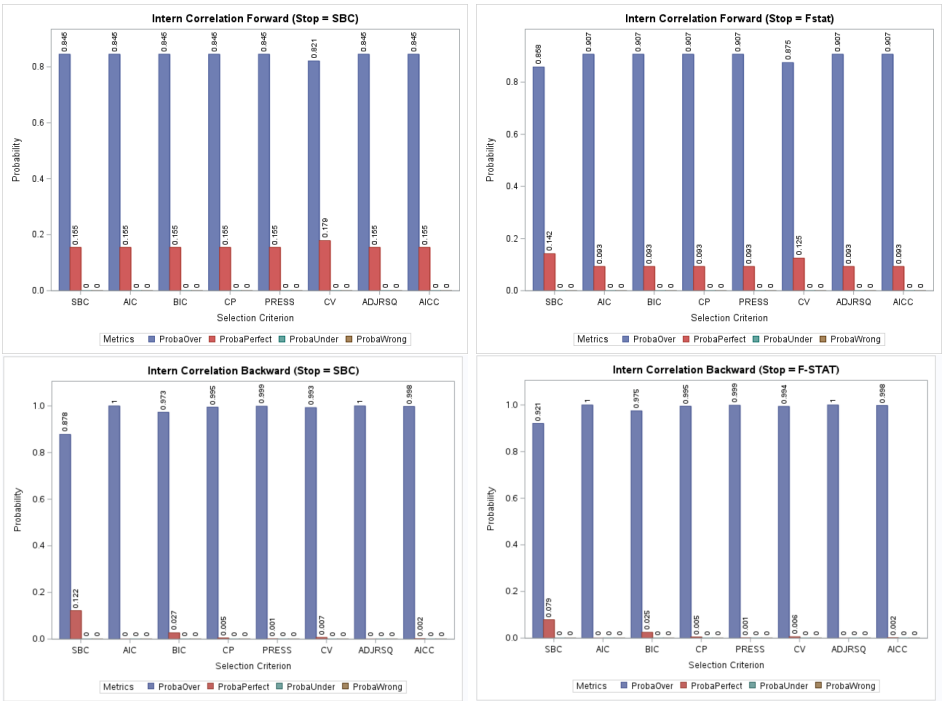


Figure 9: Forward & Backward with Intern Correlation

For external correlation, the results are once again repetitive; there is no significant difference in statistical learning between Internal correlation, external correlation, and independence. We are still faced with an over-representation of overfitting and a weaker representation of perfect fitting. The Forward and Stepwise methods remain the best in terms of the probability of achieving a perfect fit, and the CV and SBC selection criteria once again appear to be the best.

Internal Correlation

Here again, the results are very similar between LASSO and LAR :

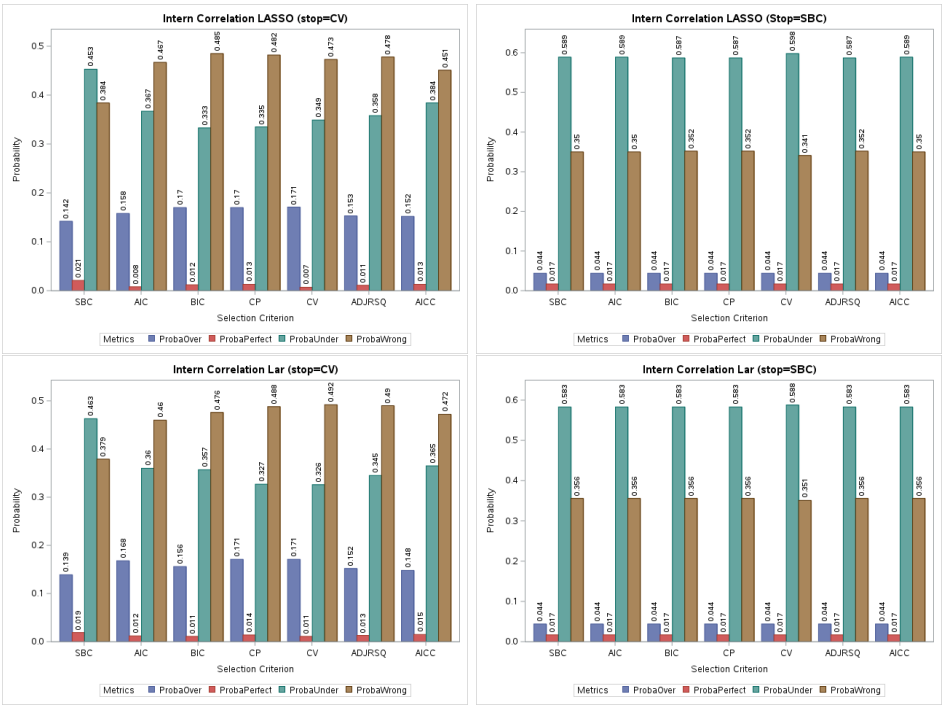


Figure 10: LASSO & LAR with Internal Correlation

We can immediately see a striking difference from the independence case, as there is much more underfitting and a higher probability of selecting a wrong model. For Cross-Validation, wrong models are almost always dominant, except when the selection criterion is SBC. With SBC, underfitting is dominant, as there is significantly less overfitting compared to CV. We can also observe that the probability of obtaining a perfect model is marginal.A possible explanation could be that, in the presence of correlation, espacially the Lasso arbitrarily selects one of the correlated variables, leading to underfitting.

It is interesting to note that C_p leads to quite different results :

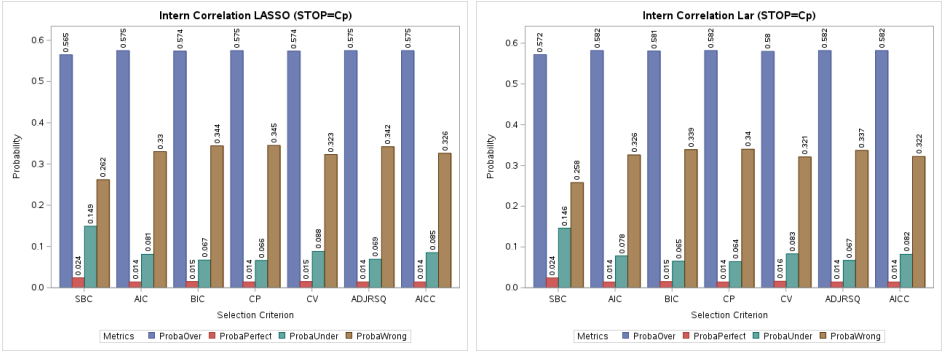


Figure 11: Lasso & Lar with Internal Correlation (Stop=C_p)

Contrary to the other criteria, there is significantly more overfitting than underfitting or wrong models. The probability of selecting perfect models remains very low. There is then a bias-variance tradeoff to consider, as C_p results in higher variance, while the other criteria lead to greater bias.

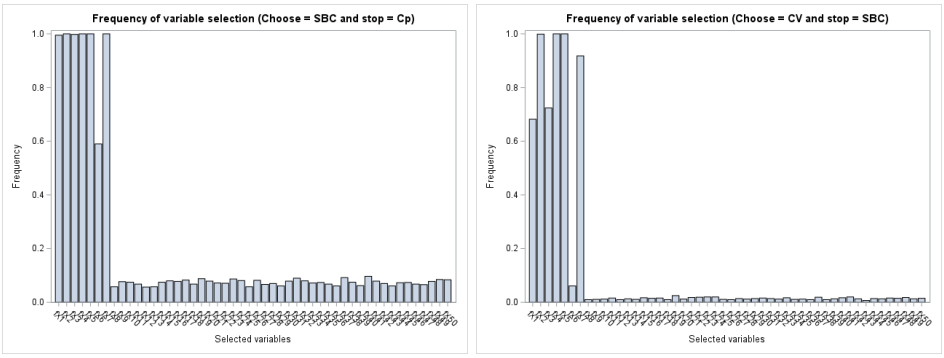


Figure 12: Variable Selection frequency with LASSO (Stop = C_p & SBC)

The results discussed above are shown in the frequency figure, as with SBC, the features X_1, X_3, and X_6 are selected with a probability lower than 0.8, while in the case of C_p, only X_6 is selected with a low probability.

Here again, the results with Elastic net are different :

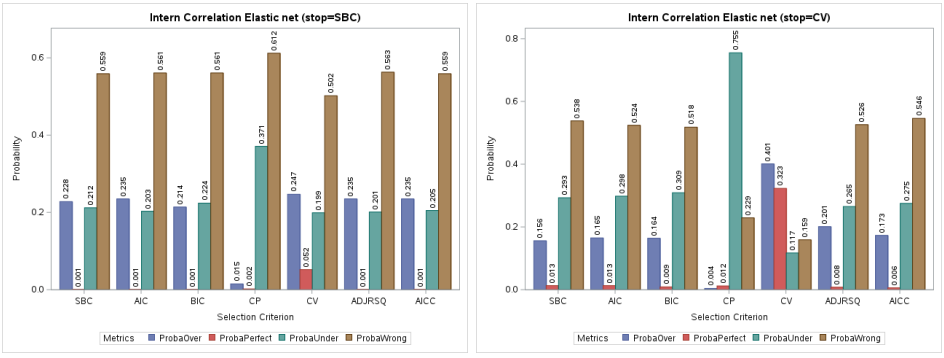


Figure 13: Elastic Net with Internal Correlation

Firstly, we can see that, as in the case of independence, when SBC is the stopping criterion, Elastic Net mainly leads to wrong models. With CV as the stopping criterion, the introduction of Internal correlation results in underfitting and wrong models, with significantly less overfitting and perfect models compared to the case of independence.

An interesting result is that when CV is both the selection and stopping criterion (so the final model is the one that minimizes the prediction error), there are far fewer incorrect models, and nearly 33% of the models are perfect. Another observation is that the combination of Choose = C_p and Stop = CV results in almost 76% underfitting and a lower probability of selecting a wrong model.

Thus, Elastic Net seems to be slightly better at handling Internal correlation, probably due to the L_2 penalty.

We can clearly see that the probability of selecting the true features is higher when both the choosing and stopping criteria are CV. However, X_6, which has the smallest coefficient, has a lower probability of being selected. In the figure on the right, we observe that when the stopping criterion is SBC, the true variables have a lower probability of being selected, while external variables have a slightly higher probability.

External Correlation

The results comparing Internal and external correlation are very interesting.Here are some results with CV, SBC and C_p as stopping rules :

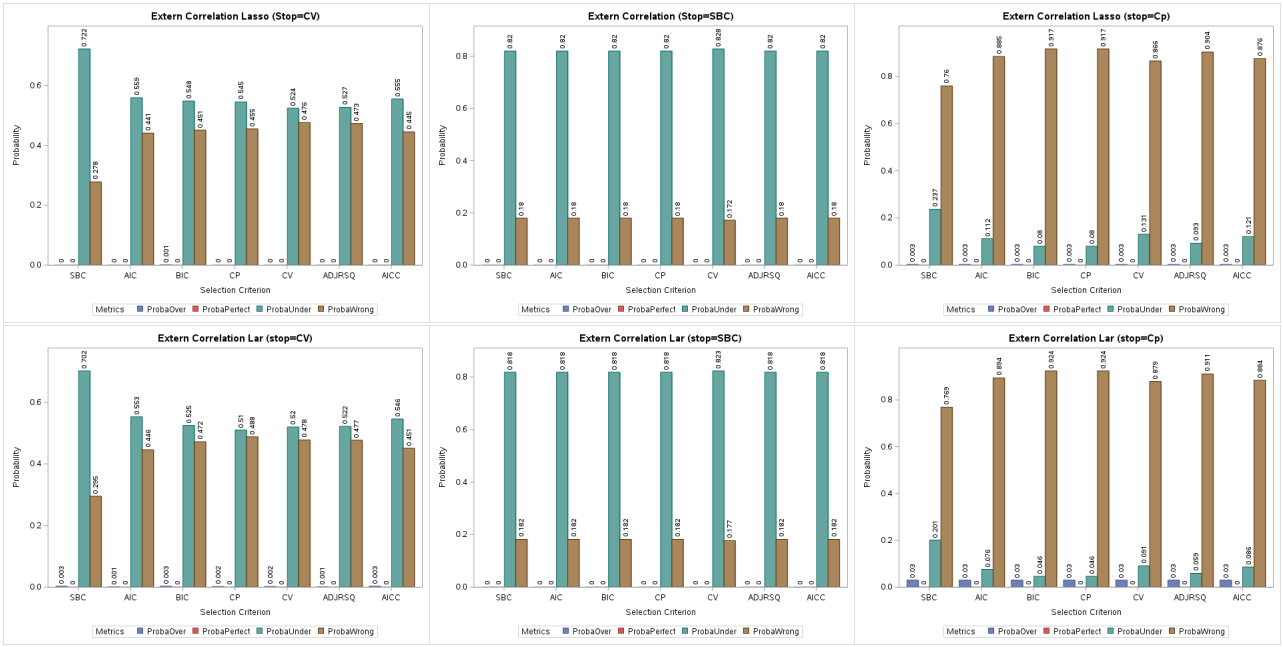


Figure 14: LASSO & LAR with External Correlation

When transitioning from internal to external correlation, we observe that the perfect fit rate drops to 0%. Likewise, there is no overfitting, and the only two remaining situations are underfitting and wrong models. The results for the other criteria remain quite similar to those of SBC and CV. Mallows' C_p, on the other hand, exhibits a significantly higher rate of wrong models and a slight increase in overfitting.

The figures above show that X_1 and X_6 are almost never selected by both LASSO and LAR, with X_2 having only a 0.35 probability of being selected. External correlated variables, especially X_8 has a higher probability to be selected than X_1.

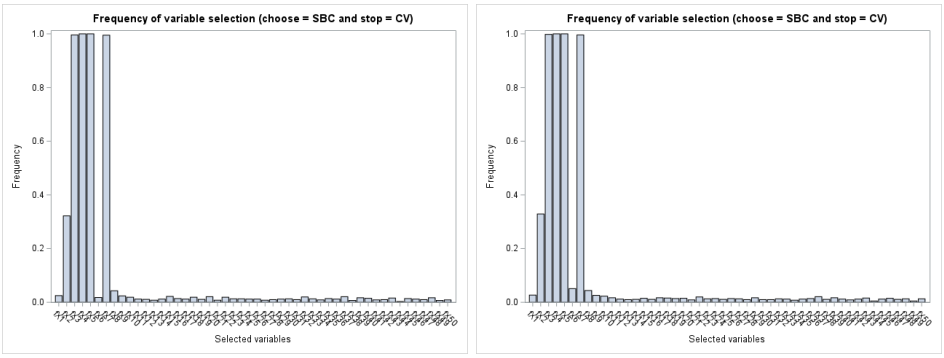


Figure 15: Variable Selection frequency with Lasso & Lar (stop=CV)

Here are the results for Elastic Net :

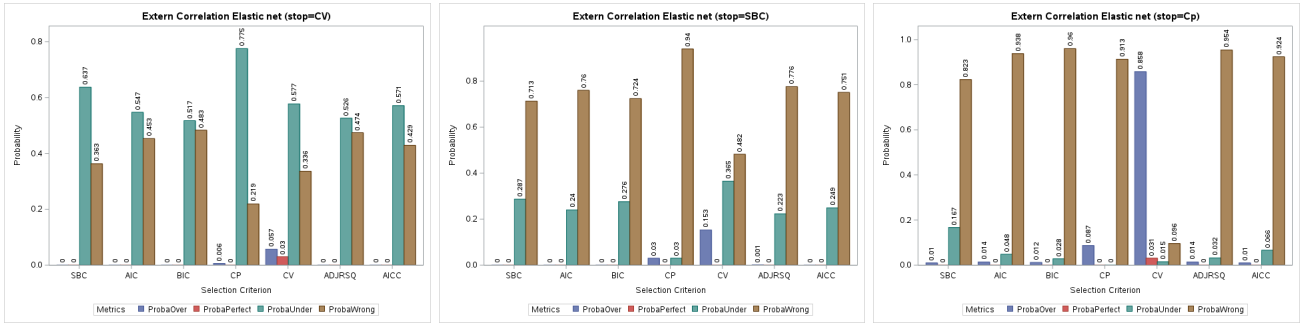


Figure 16: Elastic Net with external correlation

Here again, the results mainly show underfitting and wrong models in the presence of external correlation. The difference is that with a strict stopping criterion like SBC, there are more wrong models than with LASSO and LAR. We can also notice that when the final model is selected by cross-validation, overfitting occurs, with a small proportion of perfect models, especially when C_p is used as the stopping criterion. Cross-validation seems to be the best stopping criterion here, as it leads to a lower probability of selecting incorrect models.

Contrary to the case of Internal correlation, Elastic Net does not seem to perform better than LASSO and LAR here.

4.3 DGP with Outliers

Statistical Learning

For Outliers, there are no major differences compared to independence, Internal and external correlation cases. Overfitting remains prevalent across all three methods, while perfect fitting is rare or absent with certain choice criteria for Backward, particularly with the SBC or F-stat stop criterion. Forward and Stepwise, on the other hand, achieve slightly more perfect fitting, at around 15%. When the stop criterion is SBC, there is marginally more perfect fitting for Forward and Backward.

Machine Learning

Theoretically, we expect that introducing outliers into the data will increase the Sum of Squared Residuals (SSR) and thus lead to a greater penalty in LASSO. However, across all criteria, we found that the introduction of outliers does not seem to affect LASSO and LAR and the results with Elastic Net are also similar to the case whitout outliers.

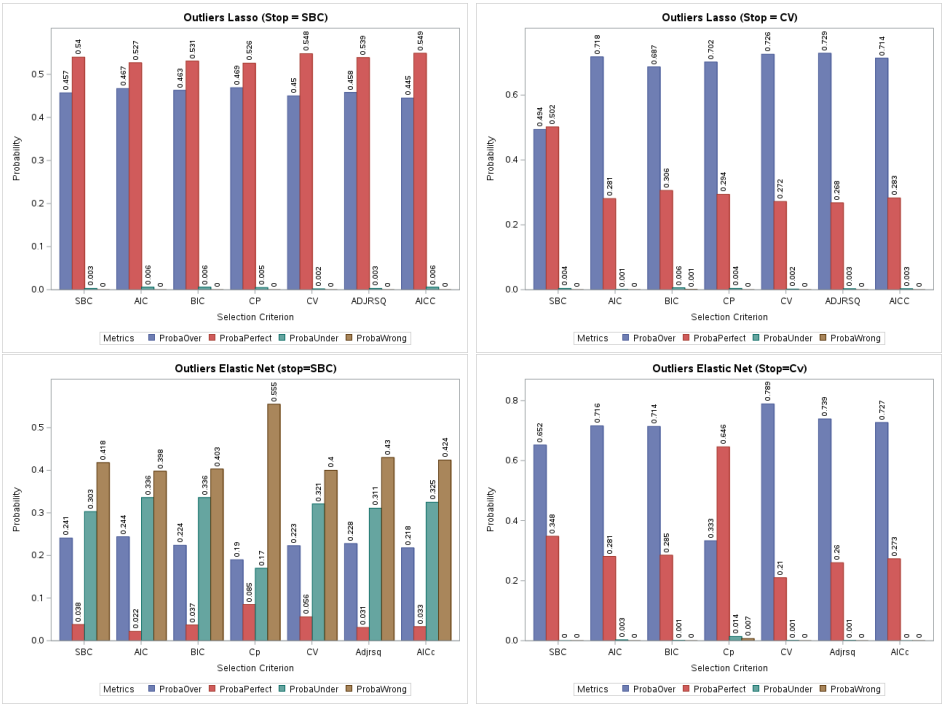


Figure 17: LASSO & Elastic Net with Outliers

Indeed, the results are quite similar to the independence case, as there are only two possible outcomes: overfitting or perfect fitting for Lasso and Lar with SBC and CV as stopping criterion. The best criterion appears to be SBC with Lasso and Lar, as it yields the highest probability of obtaining a perfect model. The combination of Choose = C_p and Stop = CV gives the best results for Elastic net with almost 65% of perfect fitting.

4.4 DGP with Structural Break

Statistical Learning

When considering the structural break in our data, a clear trend appears across all three methods: a high number of wrong models. Compared to independence, there is almost no perfect fitting, while the proportion of errors exceeds 80%.

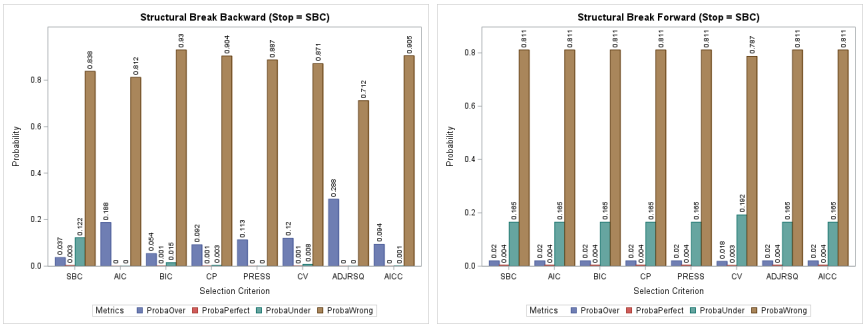


Figure 18: Statistical Learning with Structural Break

Stepwise and Forward have very similar results, especially in terms of underfitting. Backward, on the other hand, shows little or no underfitting, but some overfitting. These results can be explained by the fact that we don't consider the same vector of coefficients, which shows that statistcial learning methods are irrelevant in the presence of structural break.

Machine Learning

In general, when there is underfitting, all the combinations we tested with the three Machine Learning methods presented wrong models and underfitting in fairly similar proportions, with a slight superiority for wrong models. Perfect fitting and over-fitting are poorly represented, with probabilities close to 0.

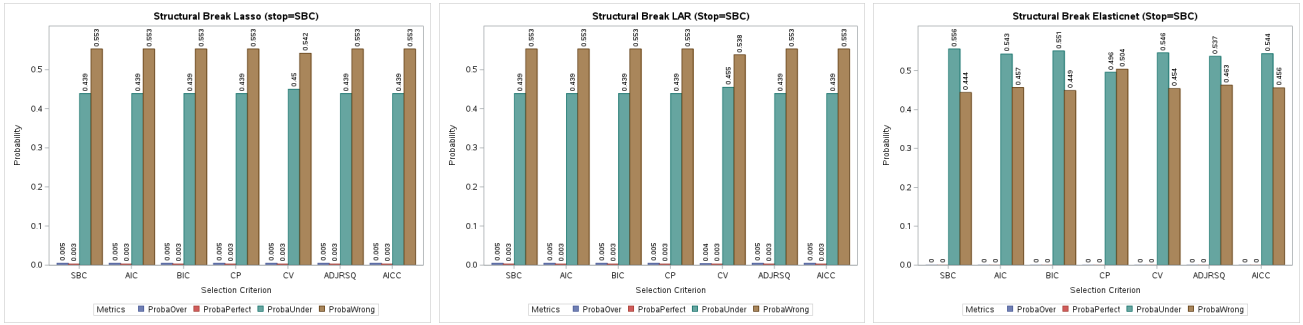


Figure 19: Machine Learning with Structural Break

The only difference between our different methods is that when the stopping rules is SBC, the elastic net shows a reversal of probabilities, with underfitting becoming predominant compared with wrong models.

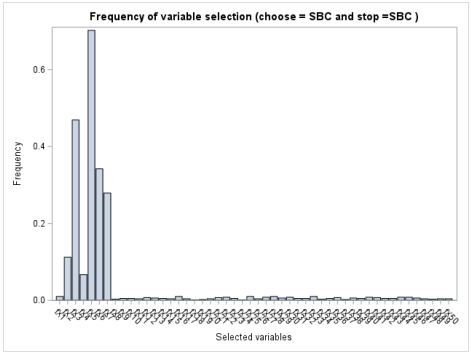


Figure 20: Variable Selection Frequency with Elastic Net (stop = SBC)

We can clearly see that X_1 is very poorly selected, which explains the high number of false models and the absence of perfect fitting and overfitting. For the other variables such as X_2 X_3 X_4, they are not always chosen, which may partly explain the underfitting. The only variable that seems to be always selected is X_5.

4.5 DGP With Mixing

External Correlation & Structural Break

• For the statistical learning, adding the external correlation to the structural break, we observe the same results as with structural break alone : many wrong models and some underfitting for Forward and Stepwise. This is because the structural break remains the dominant factor, making it difficult for selection methods to identify the right variables. Unlike the situation where only external correlation was present, there is no longer any perfect fitting and very little overfitting. Thus, the structural break limits the overfitting but introduces underfitting.

• About machine learning methods, the presence of external correlation leads to similar trends, notably a strong dominance of underfitting for Lasso and LARS when the stop criteria are SBC and CV. With R^2_{adj}, we observe more wrong models than underfitting, suggesting that this criterion leads to less optimal variable selection. Unlike to Lasso and LARS, ElasticNet shows a strong presence of wrong models with R^2_{adj} and SBC, indicating instability in variable selection. Finally, when CV and Cp are used as stopping criteria, Elastic Net mainly produces wrong models, but with underfitting. These results appear to be sub-optimal, since they show almost no perfect fitting compared with independence, where there is a high proportion of perfect fitting.

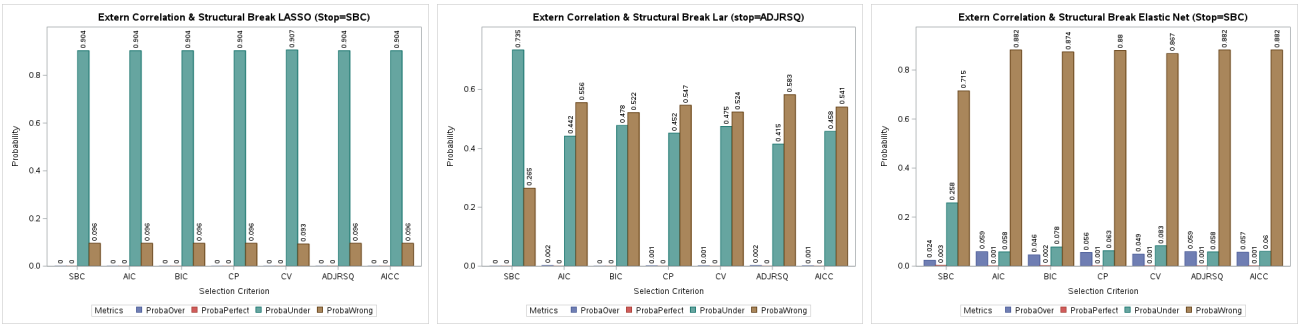


Figure 21: Machine Learning with External Correlation Structural Break

Structural Break & Outliers

- For the statistical learning methods, the results show a very high probability of wrong models, around 80% on average for each case, and a low proportion of underfitting. In addition, perfect fitting and overfitting are almost non-existent. We are therefore very close to the results obtained in the case of structural break (with less overfitting for the SBC stopping criterion), and quite far from the results obtained in the case of outliers, which showed a lot of overfitting.
- For Machine Learning, we also obtained results very close to what we had for structural break, with a high probability of wrong models and a lower proportion of underfitting for LASSO and LARS. The elastic net, for its part, has a very low probability of underfitting compared to when there was only the break. Once again, the probability of perfect fitting is almost zero.

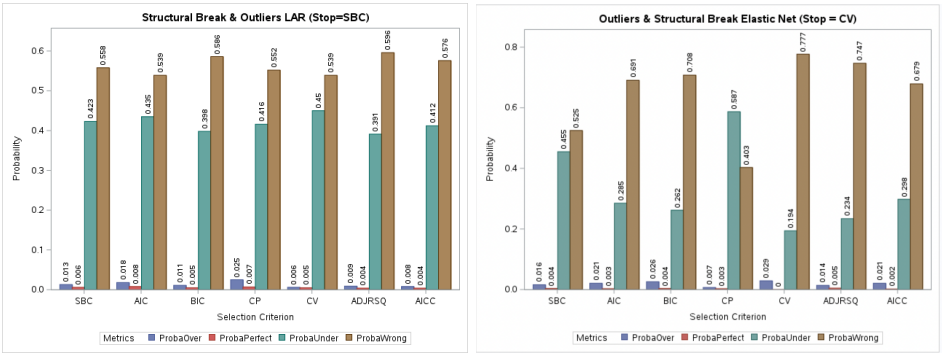


Figure 22: Machine Learning with Structural Break & Outliers

Internal Correlation & Outliers

For the three Statistical Learning methods, when we mix outliers and internal correlation, we obtain the same results as when they were generated separately, i.e. a very high probability of overfitting and a lower probability of perfect fitting. Underfitting and wrong models also have null probabilities. Machine Learning methods lead to more interesting results.

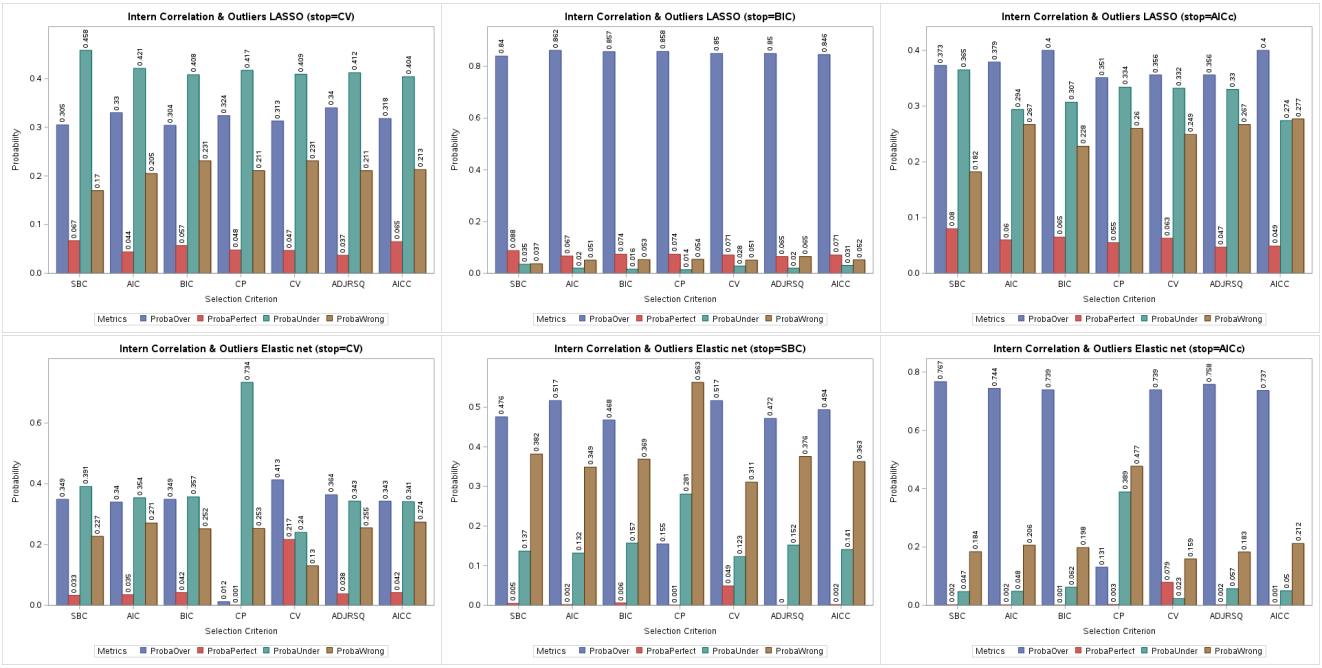


Figure 23: LASSO & Elastic Net with Outliers and Internal Correlation

Two key observations can be drawn from these graphs. The results are quite surprising: there are slightly more perfect models and fewer wrong models than the DGP with Internal correlation. The combination of SBC as a selection criterion and cross-validation as a stopping criterion produces the fewest wrong models compared with AIC_c. However, although the BIC produces fewer incorrect models, it leads to an over-fitting of almost 85%. For Elastic net, the combination of SBC as a selection criterion and CV as a stopping criterion seems to be the best.

5 Empirical Analysis

For our empirical study we will use the diabete dataset, which was also used by Effron & al in 2004.

Patient	AGE x1	SEX x2	BMI x3	BP x4	S1 x5	S2 x6	S3 x7	S4 x8	S5 x9	S6 x10	Response y
1	59	2	32.1	101	157	93.2	38	4	4.8598	87	151
2	48	1	21.6	87	183	103.2	70	3	3.8918	69	75
3	72	2	30.5	93	156	93.6	41	4	4.6728	85	141
4	24	1	25.3	84	198	131.4	40	5	4.8903	89	206
5	50	1	23	101	192	125.4	52	4	4.2905	80	135
6	23	1	22.6	89	139	64.8	61	2	4.1897	68	97
7	36	2	22	90	160	99.6	50	3	3.9512	82	138
8	66	2	26.2	114	255	185	56	4.55	4.2485	92	63
9	60	2	32.1	83	179	119.4	42	4	4.4773	94	110
10	29	1	30	85	180	93.4	43	4	5.3845	88	310

Figure 24: Diabete study

The dataset contains N=442 patients who were measured on 10 baseline variables. As we can see, the predictors include age, sex, Body Mass Index (BMI), and blood pressure. The variables S1 to S6 represent various blood serum measurements. The dependent variable Y is a quantitative measure of disease progression.

Before proceeding with variable selection, we will conduct a data analysis.

La procédure MEANS

Variable	N	Moyenne	Ec-type	Minimum	Maximum	Médiane	Skewness	Kurtosis
AGE	442	48.5180995	13.1090278	19.0000000	79.0000000	50.0000000	-0.2313815	-0.6712237
SEX	442	1.4683258	0.4995612	1.0000000	2.0000000	1.0000000	0.1273845	-1.9928110
BMI	442	26.3757919	4.4181216	18.0000000	42.2000000	25.7000000	0.5981485	0.0950945
BP	442	94.6470136	13.8312834	62.0000000	133.0000000	93.0000000	0.2906584	-0.5327973
S1	442	189.1402715	34.6080517	97.0000000	301.0000000	186.0000000	0.3781082	0.2329479
S2	442	115.4391403	30.4130810	41.6000000	242.4000000	113.0000000	0.4365918	0.6013812
S3	442	49.7884615	12.9342022	22.0000000	99.0000000	48.0000000	0.7992551	0.9815075
S4	442	4.0702489	1.2904499	2.0000000	9.0900000	4.0000000	0.7353736	0.4444017
S5	442	4.6414109	0.5223906	3.2581000	6.1070000	4.6200500	0.2917537	-0.1343668
S6	442	91.2601810	11.4963347	58.0000000	124.0000000	91.0000000	0.2079166	0.2369167
Y	442	152.1334842	77.0930045	25.0000000	346.0000000	140.5000000	0.4405629	-0.8830573

Figure 25: Descriptive Statistics

We can see that almost all features have a positive skewness ($S > 0$) which means that their distributions have outliers on the right because of a longer right tail. Only the variable Age is skewed on the left. However the skewnesses are close to zero in most of the cases which means that they have nearly symmetrical distribution.

Secondly we can see that some variables are leptokurtic($K > 3$) whith heavier tails and more extreme values than a normal distribution while the others are platykurtic ($K < 3$)[1].The graph showing the histogram of each variable's distribution with the kernel density estimate clearly indicates that most variables are relatively close to a normal distribution, except for a few (see Appendix : figure 38).

[1]The results displayed by SAS subtract 3 from the Kurtosis.

Let's now analyze the correlation between the predictors.

La procédure CORR

9 Variables :	AGE BMI BP S1 S2 S3 S4 S5 S6

Coefficients de corrélation de Pearson, N = 442

	AGE	BMI	BP	S1	S2	S3	S4	S5	S6
AGE	1.00000	0.18508	0.33543	0.26006	0.21924	-0.07518	0.20384	0.27077	0.30173
BMI	0.18508	1.00000	0.39541	0.24978	0.26117	-0.36681	0.41381	0.44616	0.38868
BP	0.33543	0.39541	1.00000	0.24246	0.18555	-0.17876	0.25765	0.39348	0.39043
S1	0.26006	0.24978	0.24246	1.00000	0.89666	0.05152	0.54221	0.51550	0.32572
S2	0.21924	0.26117	0.18555	0.89666	1.00000	-0.19646	0.65982	0.31836	0.29060
S3	-0.07518	-0.36681	-0.17876	0.05152	-0.19646	1.00000	-0.73849	-0.39858	-0.27370
S4	0.20384	0.41381	0.25765	0.54221	0.65982	-0.73849	1.00000	0.61786	0.41721
S5	0.27077	0.44616	0.39348	0.51550	0.31836	-0.39858	0.61786	1.00000	0.46467
S6	0.30173	0.38868	0.39043	0.32572	0.29060	-0.27370	0.41721	0.46467	1.00000

Figure 26: Correlation matrix between the predictors

We can see that several variables are for the most part correlated rather weakly with a correlation below 0.5 in absolute value with a few exceptions such as between S2 and S1 where the correlation reaches 0.89 or S3 and S4 where the correlation reaches 0.73 in absolute value.

In conclusion, the data seems to be very similar to our data generation process, with correlation and outliers. We will therefore continue with variable selection using the Lasso with cross validation as the stop criterion and SBC as the choice criterion, as well as the elastic net with cross validation for both choose and stop.

Effets :	Intercept X2 X3 X4 X7 X9

Analyse de variance				
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F
Modèle	5	1296887	259377	85.41
Erreur	436	1324122	3036.97748	
Total sommes corrigées	441	2621009		

Racine MSE	55.10878
Moyenne dépendante	152.13348
R carré	0.4948
R car. ajust.	0.4890
AIC	3994.18808
AICC	3994.44615
SBC	3574.73594
CV PRESS	1316817

Paramètres estimés		
Paramètre	DDL	Estimation
Intercept	1	-218.613988
X2	1	-7.140599
X3	1	5.511416
X4	1	0.806139
X7	1	-0.624800
X9	1	41.080918

(a) Elastic net

Effets :	Intercept X2 X3 X4 X7 X9

Analyse de variance				
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F
Modèle	5	1333128	266626	90.26
Erreur	436	1287881	2953.85586	
Total sommes corrigées	441	2621009		

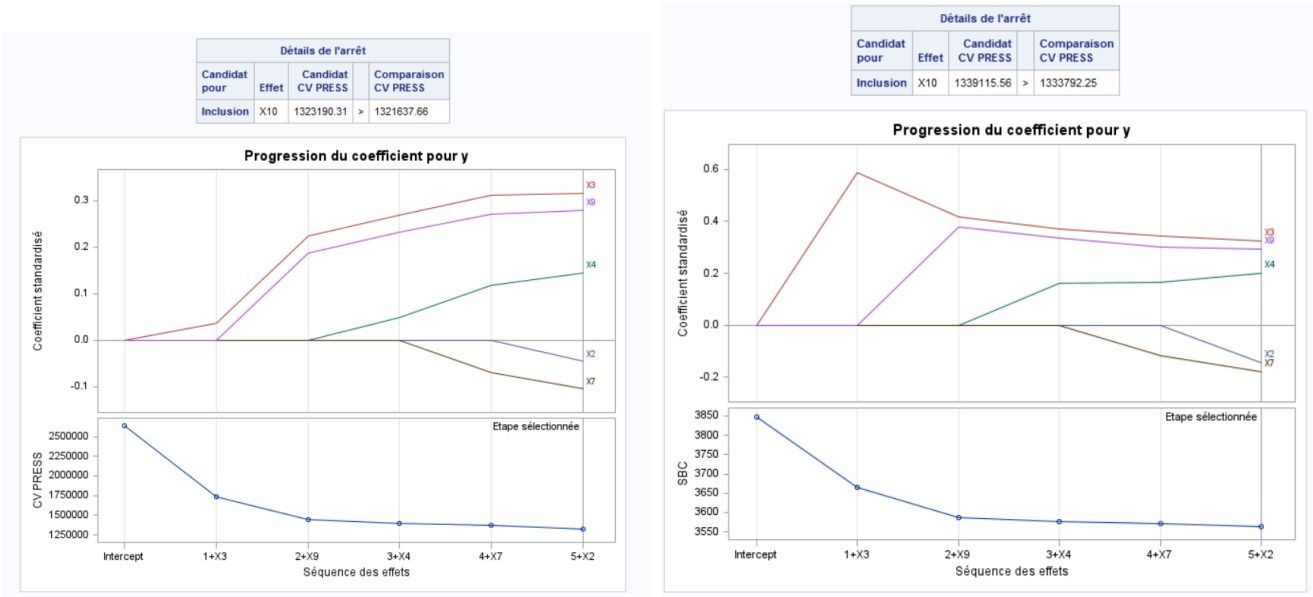
Racine MSE	54.34939
Moyenne dépendante	152.13348
R carré	0.5086
R car. ajust.	0.5030
AIC	3981.92197
AICC	3982.18004
SBC	3562.46983
CV PRESS	1333831

Paramètres estimés		
Paramètre	DDL	Estimation
Intercept	1	-217.684869
X2	1	-22.474240
X3	1	5.643077
X4	1	1.123165
X7	1	-1.064416
X9	1	43.234413

(b) Lasso

Figure 27: Variable Selection in the Diabete Dataset

Both Lasso and Elastic Net select the same variables, but we can see that the model selected by Lasso has a slightly better R^2 and R^2_{adj}. Similarly, AIC, AICc, and SBC are better for the model selected by Lasso. Thus, Lasso associated with cross-validation as a stopping criterion seems slightly better here.



(a) Elastic net Coefficient Path

(b) LASSO Coefficnet Path

Figure 28: Variable Selection in the Diabete Dataset

The Figure 28 shows the coefficient Path of the two processes. We can see that X_3 is the most important and first selected feature for both. As already we have seen before, the two methods are quite similar in this case except the fact that the model obtained by the Lasso have greater coefficients for almost all features.

6 Conclusion

To conclude, we have observed that there is no universal method leading to a perfect fit. Depending on the case, results are more favorable for one model than for the other. The same applies to stopping criteria: while the SBC often seems to be the best criterion for Lasso, Lars, and Statistical Learning methods, Elastic Net tends to yield better results with predictive criteria such as K-fold cross-validation.

The second important point to note is that, even in the purely theoretical and perfect case where variables are drawn from a multivariate normal distribution and are independent of each other, the fit is never truly perfect. It is therefore essential to consider the results relative to this ideal case to assess the quality of the fit. For Statistical Learning methods based on statistical inference, the results seem to hold even when the simulated data do not correspond to the "perfect" scenario.

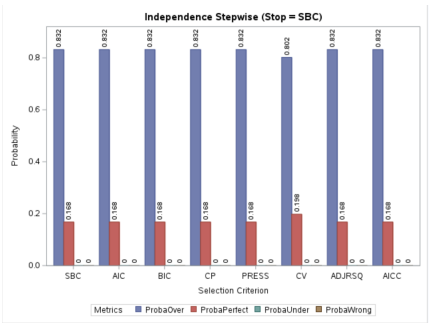
In the presence of correlation, Elastic Net based on cross-validation performs better, thanks to the L2 penalty, which, as previously mentioned, handles correlation more effectively. External correlation appears to disrupt the algorithms more significantly, leading to a predominance of underfitting and wrong models, where external variables tend to overshadow the internal variables of the models. However, the algorithms seem relatively robust to outliers. The noise introduced by structural break is the scenario that most disrupts all methods globally, resulting in a majority of wrong models.

By comparing our results to an empirical analysis with correlated data containing outliers, we attempted a variable selection based on our findings. Elastic Net and Lasso led to similar selections, with slightly better interpretability for the model selected by Lasso.

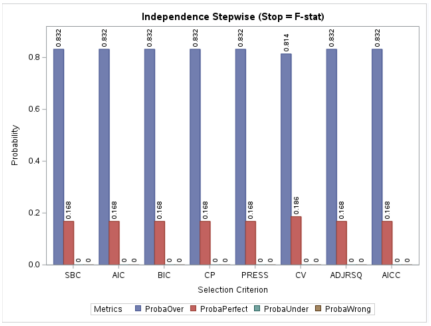
However, our results may be somewhat limited, as they only consider individual cases or pairs of cases. An improvement to better generalize our findings could involve a simulation closer to reality, incorporating the many other cases present in real-world situations.

7 Appendix

This appendix presents the results of the Stop and Choose combinations that were not included in the main body of the thesis. These additional analyses provide a more detailed examination of the different configurations and their impact on the studied models.

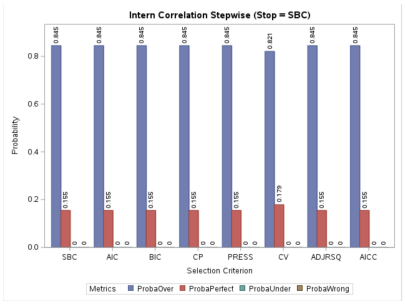


(a) Stepwise Stop = SBC

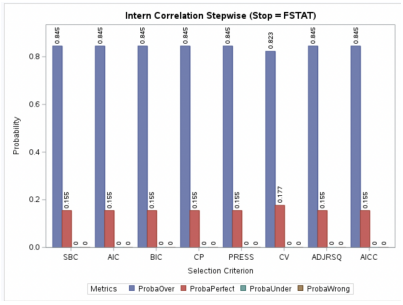


(b) Stepwise Stop = F-Stat

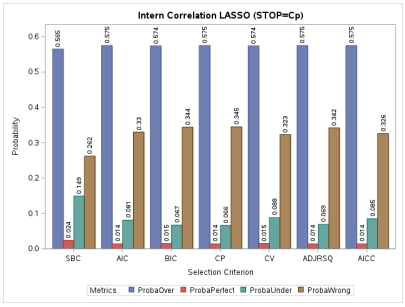
Figure 29: Independence Results



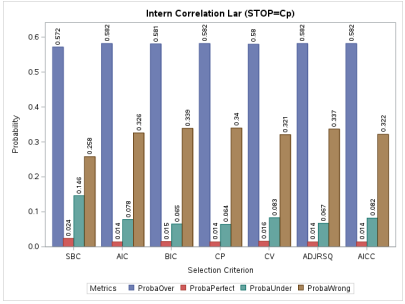
(a) Stepwise Stop = SBC



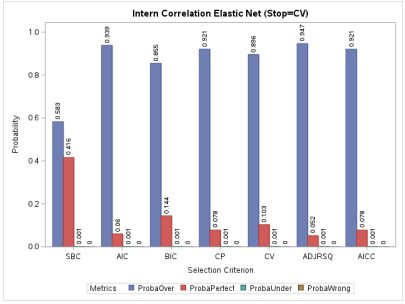
(b) Forward Stepwise = F-Stat



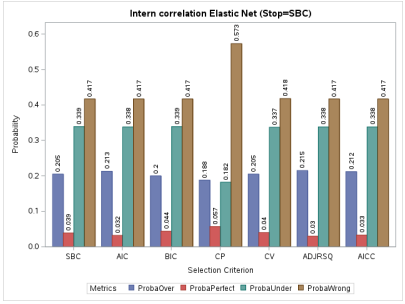
(c) Lasso Stop = SBC



(d) Lars Stop = C_p

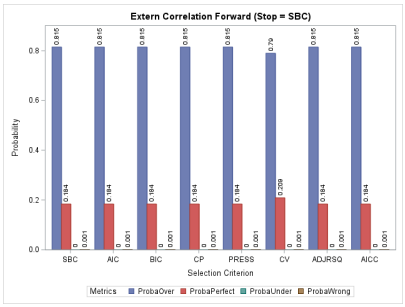


(e) Elastic Net Stop = CV

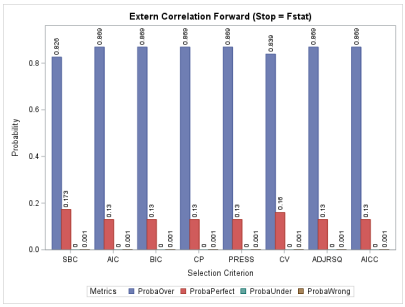


(f) Elastic Net Stop = SBC

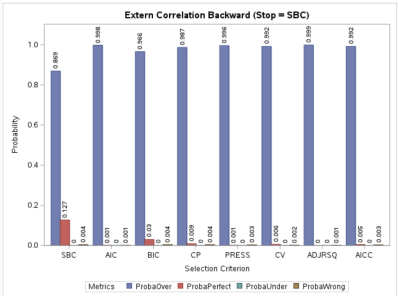
Figure 30: Intern Correlation Results



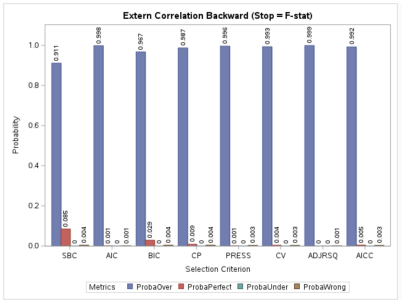
(a) Forward Stop = SBC



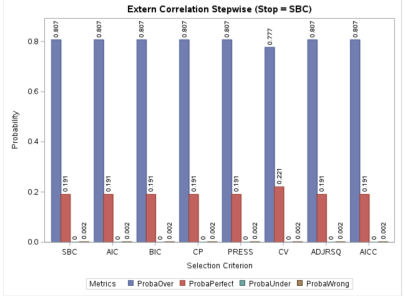
(b) Forward Stop = F-Stat



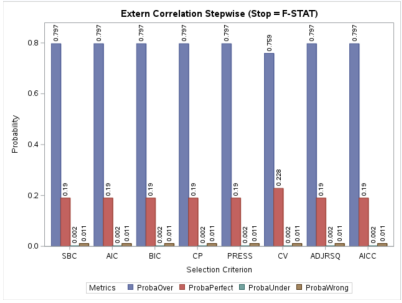
(c) Backward Stop = SBC



(d) Backward Stop = F-Stat

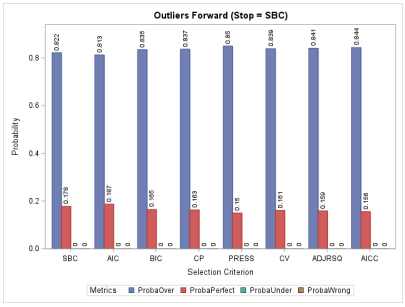


(e) Stepwise Stop = SBC

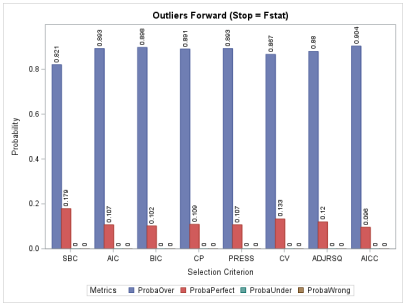


(f) Stepwise Stop = F-Stat

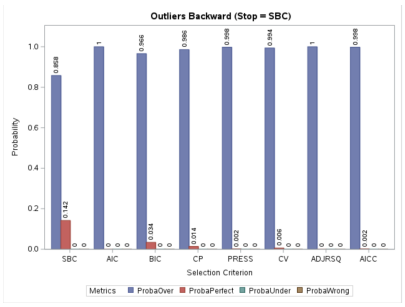
Figure 31: External Correlation Results



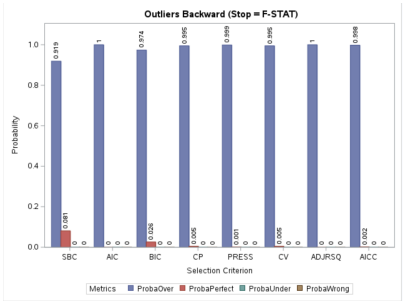
(a) Forward Stop = SBC



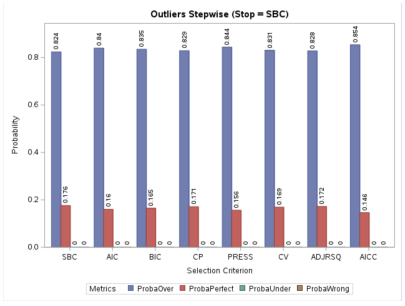
(b) Forward Stop = F-Stat



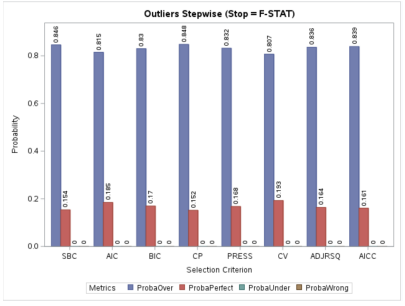
(c) Backward Stop = SBC



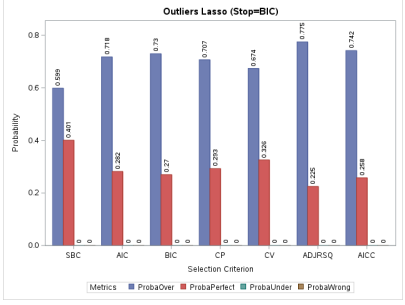
(d) Backward Stop = F-Stat



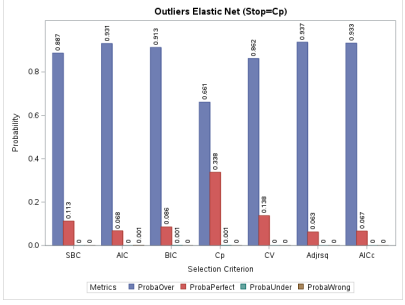
(e) Stepwise Stop = SBC



(f) Stepwise Stop = F-Stat

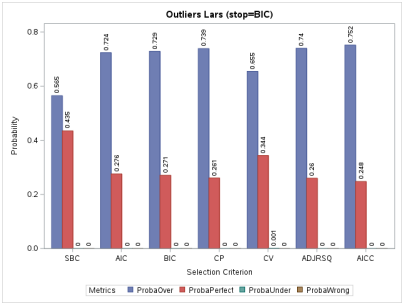


(g) Lasso Stop = BIC

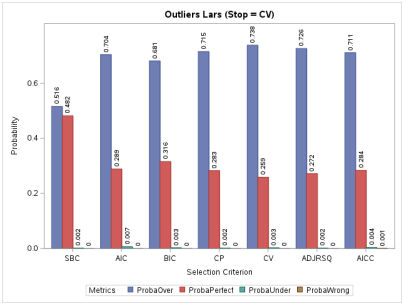


(h) Elastic Net Stop = C_p

(i) Elastic Net Stop = SBC

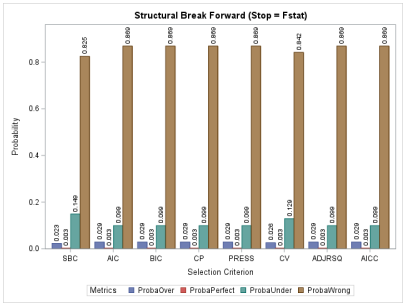


(j) lars Stop = BIC

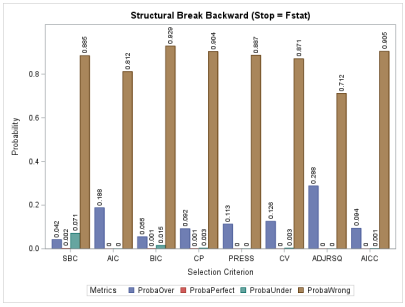


(k) Lars Stop = CV

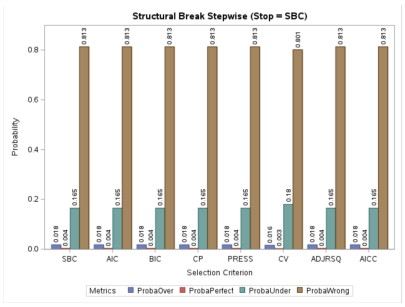
Figure 32: Outliers Results



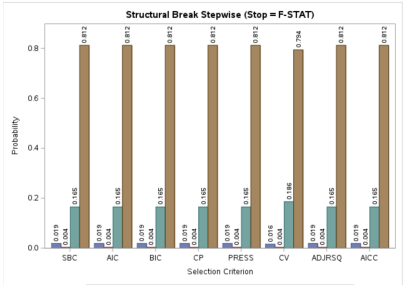
(a) Forward Stop = F-stat



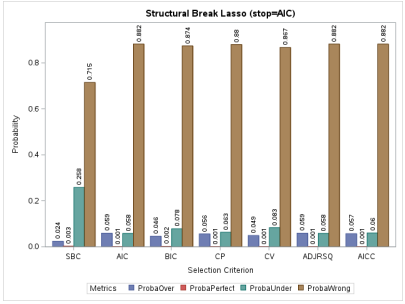
(b) Backward Stop = F-stat



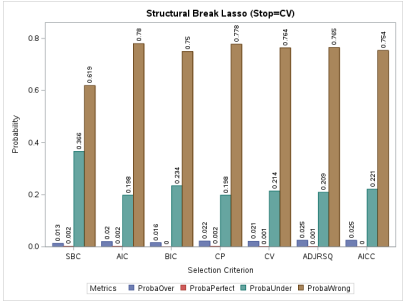
(c) Stepwise Stop = SBC



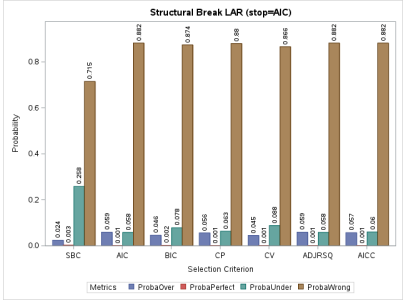
(d) Stepwise Stop = F-stat



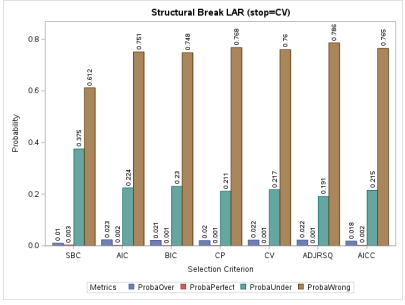
(e) Lasso Stop = AIC



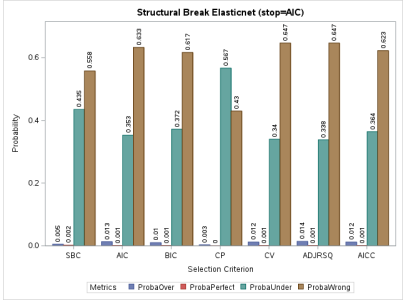
(f) lasso Stop = CV



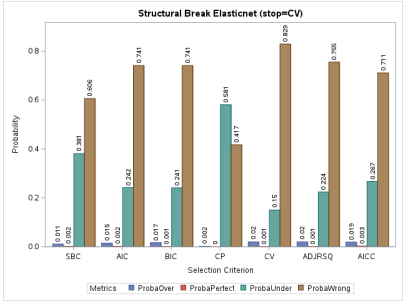
(g) Lars Stop = AIC



(h) Lars Stop = CV

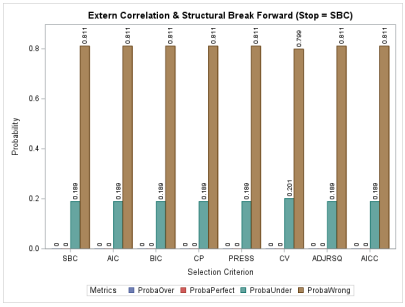


(i) Elastic Net Stop = AIC

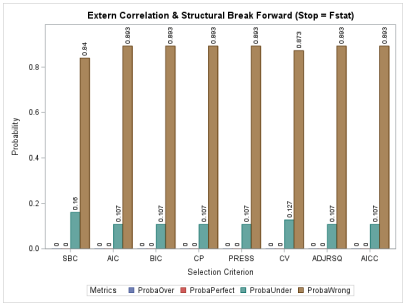


(j) Elastic Net stop = CV

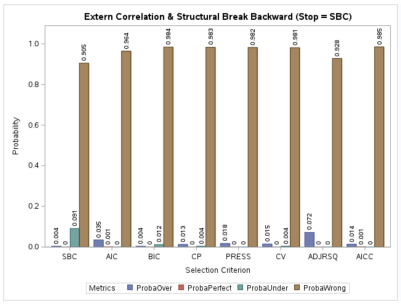
Figure 33: Structural Break Results



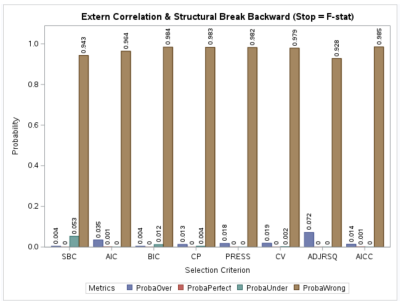
(a) Forward Stop = SBC



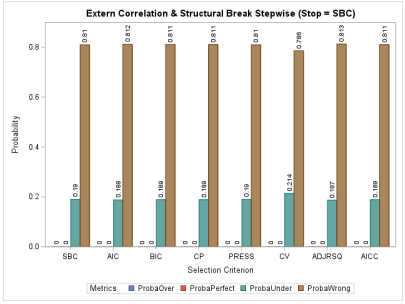
(b) Forward Stop = F-stat



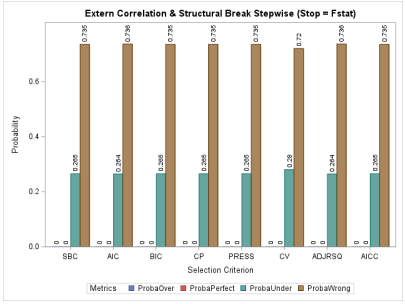
(c) Backward Stop = SBC



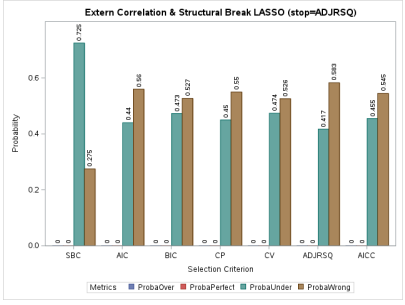
(d) Backward Stop = F-Stat



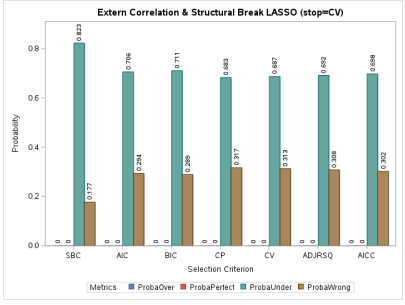
(e) Stepwise Stop = SBC



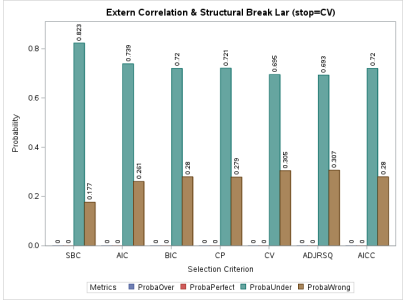
(f) Stepwise Stop = F-stat



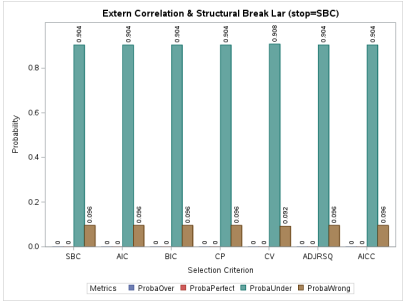
(g) Lasso stop = R^2_{adj}



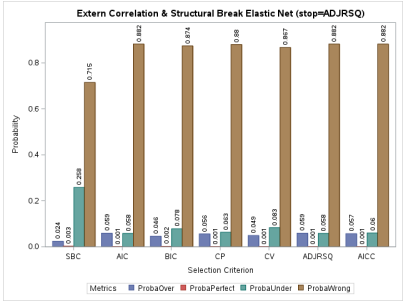
(h) Lasso Stop = CV



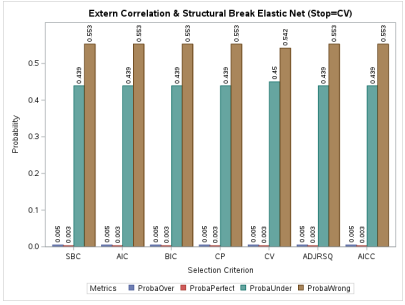
(i) Lars Stop = CV



(j) Lars Stop = SBC

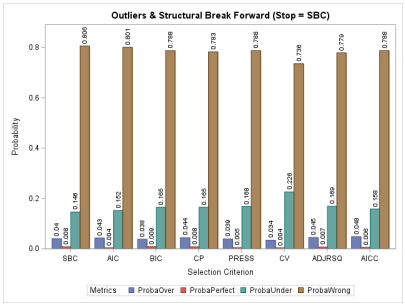


(k) Elastic Net Stop = R^2_{adj}

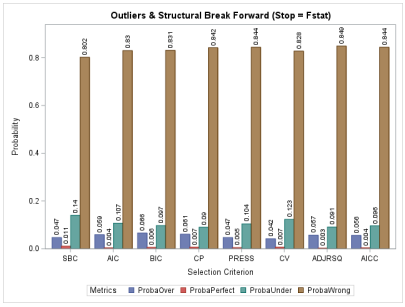


(l) Elastic Net Stop = CV

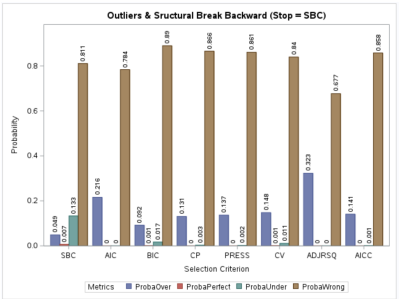
Figure 34: External Correlation & Structural Break Results



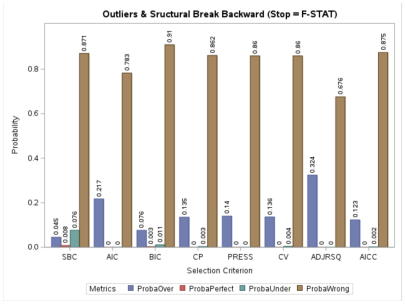
(a) Forward Stop = SBC



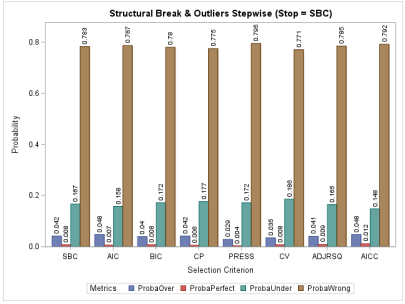
(b) Forward Stop = F-stat



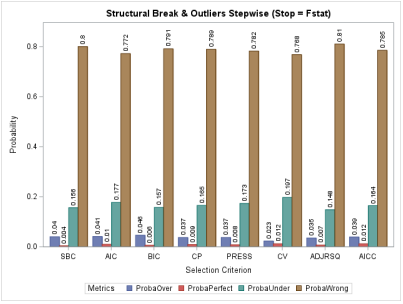
(c) Baxkward Stop = SBC



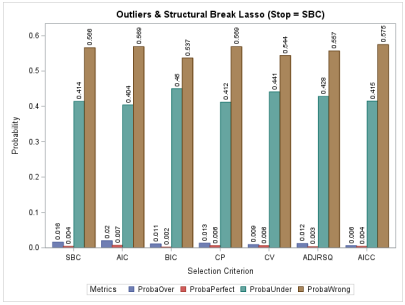
(d) Backward Stop = F-Stat



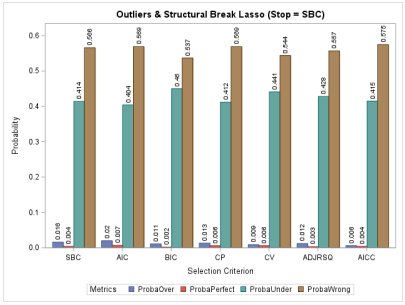
(e) Stepwise Stop = SBC



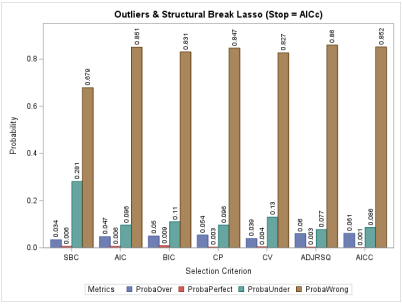
(f) Stepwise stop = F-stat



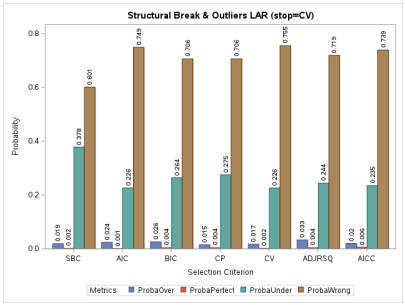
(g) Lasso Stop = SBC



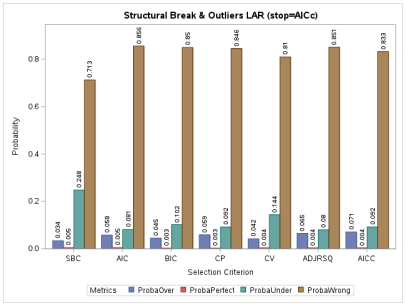
(h) Lasso Stop = SBC



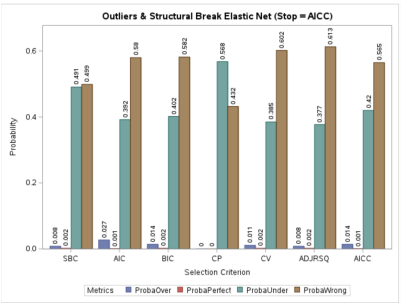
(i) Lasso Stop = AICC



(j) Lars = CV

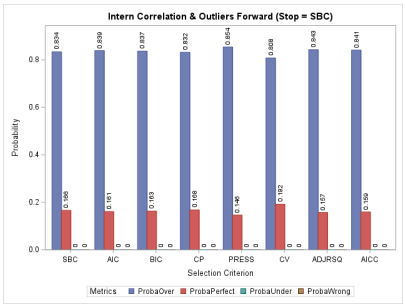


(k) Lars = AIC_c

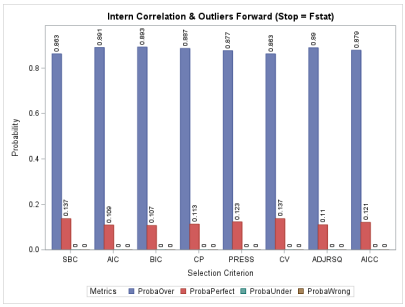


(l) Elastic Net Stop = AIC_c

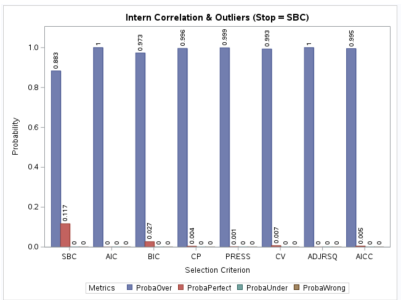
Figure 35: Structural Break & Outliers Results



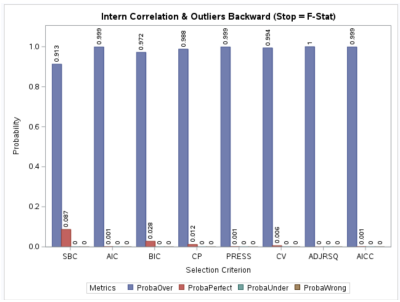
(a) Forward Stop = SBC



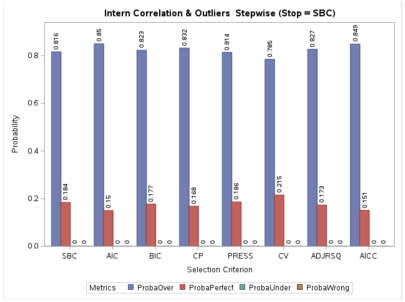
(b) Forward Stop = F-stat



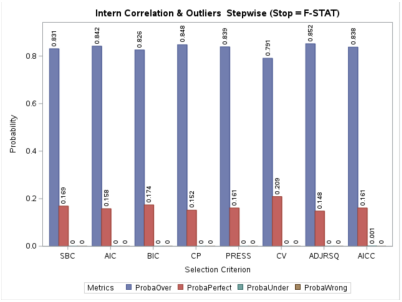
(c) Backward Stop = SBC



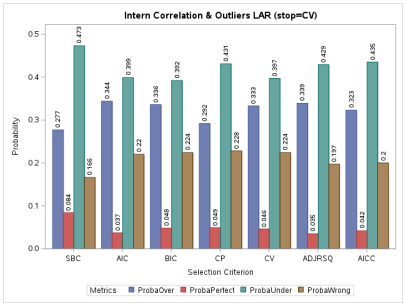
(d) Backward Stop = F-Stat



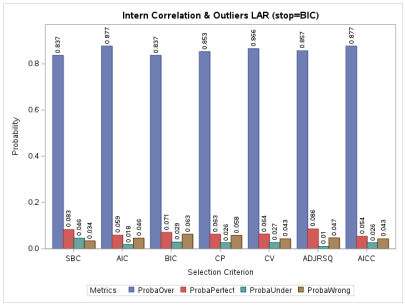
(e) Stepwise Stop = SBC



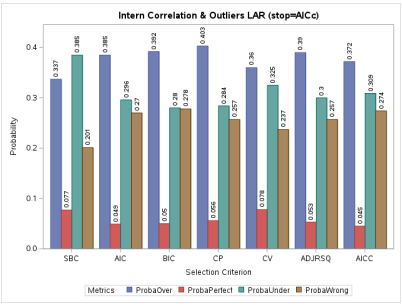
(f) Stepwise Stop = F-Stat



(g) Lars Stoop = CV

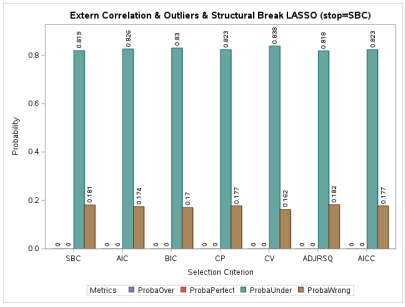


(h) Lars Stop = BIC

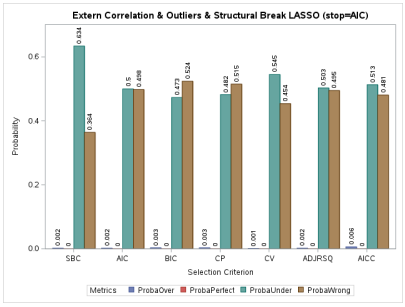


(i) Lars Stop = AIC_c

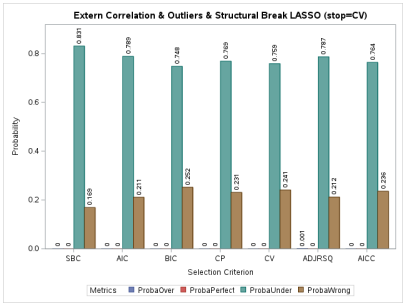
Figure 36: Internal Correlation & Outliers Results



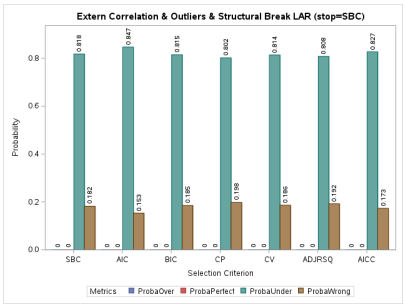
(a) Lasso Stop = SBC



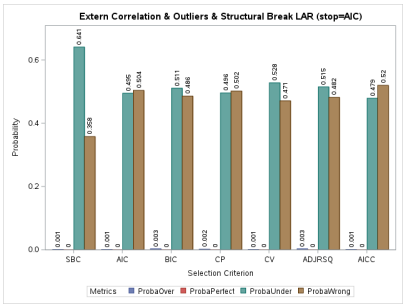
(b) Lasso Stop = AIC



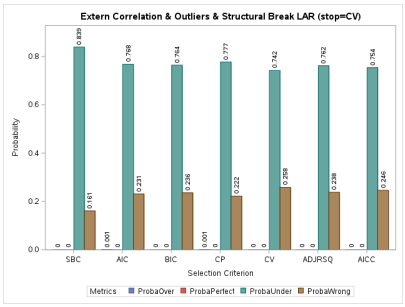
(c) Lasso Stop = CV



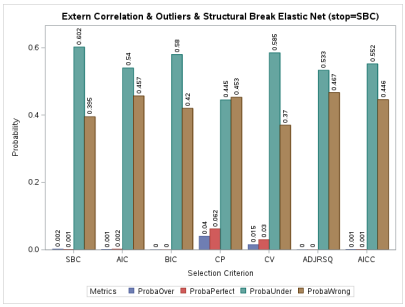
(d) Lars Stop = SBC



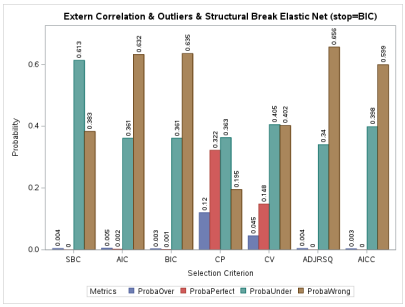
(e) Lars Stop = AIC_c



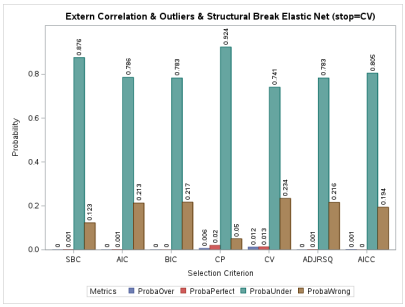
(f) Lars Stop = CV



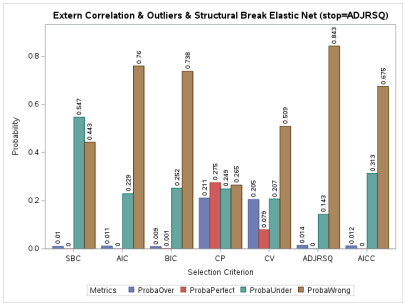
(g) Elastic Net Stop = SBC



(h) Elastic Net Stop = BIC

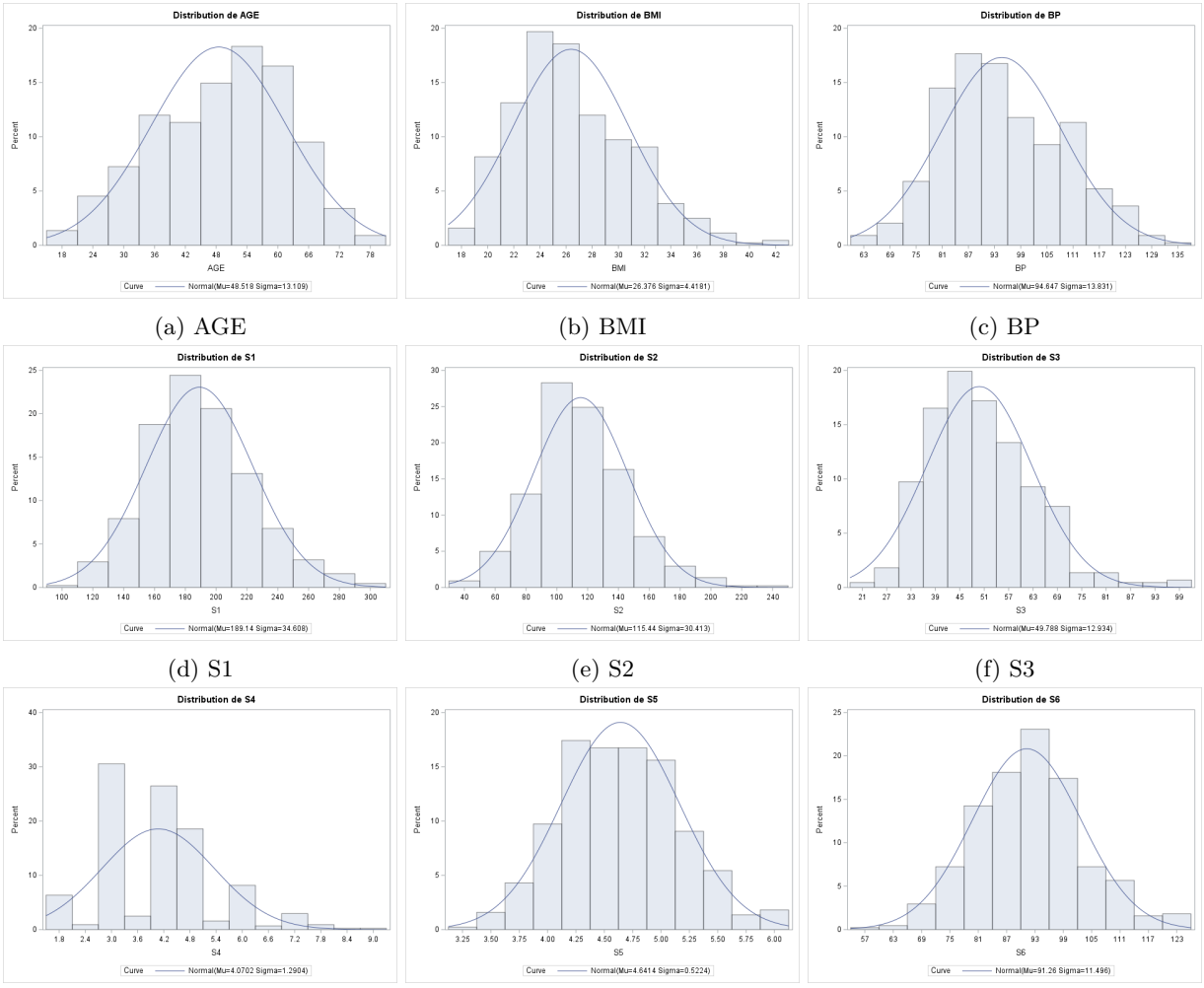


(i) Elastic Net Stop = CV



(j) Elastic Net Stop = R^2_{adj}

Figure 37: Intern alCorrelation & Structural Break & Outliers Results



(a) AGE (b) BMI (c) BP

(d) S1 (e) S2 (f) S3

(g) S4 (h) S5 (i) S6

Figure 38: Distribution of the variables

8 Bibliography

References

[1] Hirotugu Akaike. "Information theory and an extension of the maximum likelihood principle". In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[2] Bradley Efron, T. Hastie, I. Johnstone, and R. Tibshirani. "Least angle regression". pages 407–451, 2004.

[3] M. A. Efroymson. "Multiple Regression Analysis". In A. Ralston and H. S. Wilf, editors, *Mathematical Methods for Digital Computers*. John Wiley, New York, 1960.

[4] Ronald Aylmer Fisher et al. "On a Distribution Yielding the Error Functions of Several Well Known Statistics". 1924.

[5] C. M. Hurvich and C.-L. Tsai. "Bias of the Corrected AIC Criterion for Underfitted Regression and Time Series Models". *Biometrika 48*, pages 499–509, 1991.

[6] Colin L. Mallows. "Some comments on Cp". *Technometrics 42.1*, pages 87–94, 2000.

[7] J. O. Rawlings. *"Applied Regression Analysis: A Research Tool"*. Wadsworth & Brooks/Cole, Pacific Grove, 1988.

[8] SAS Blogs. "Simulate correlated variables by using the Iman-Conover transformation". URL : `https://blogs.sas.com/content/iml/2021/06/14/simulate-iman-conover-transformation.html`.

[9] SAS Help Center. "The GLMSELECT Procedure". URL : `https://documentation.sas.com/doc/en/statug/15.2/statug_glmselect_syntax01.htm`.

[10] Takamitsu Sawa. "Information criteria for discriminating among alternative regression models". *Econometrica: Journal of the Econometric Society*, pages 1273–1291, 1978.

[11] Gideon Schwarz. "Estimating the Dimension of a Model". *The Annals of Statistics 6.2*, pages 461–464, March 1978.

[12] Student. "The probable error of a mean". *Biometrika 6.1*, pages 1–25, 1908.

[13] Robert Tibshirani. "Regression shrinkage and selection via the Lasso". *Journal of the Royal Statistical Society Series B: Statistical Methodology 58.1*, pages 267–288, 1996.

[14] Sewall Wright. "Correlation and Causation". *Journal of Agricultural Research 20.7*, pages 557–585, 1921.

[15] Long Zhang and Kang Li. "Forward and backward least angle regression for nonlinear system identification". *Automatica 53*, pages 94–102, 2015.

[16] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society Series B: Statistical Methodology 67.2*, pages 301–320, 2005.