
# SAS Project
# Statistical and Machine Learning

**Rodrigue Canivez**
**Laura Ung**
**Julie Daniaud**

University Paris 1 Pantheon-Sorbonne
January 2022

**Abstract :**

This paper reports on the findings of a comparison study between Statistical and Machine learning. After having tested their different applications on Monte Carlo simulations, the main result of our study is that the processes of Statistical learning are not really convincing when it comes to find the right model, however the data are structured. The Machine learning processes are more interesting. First, the quality of prediction is better. It is more contrasting when we add some correlation between the variables : some processes (LAR, LASSO) see their quality of prediction decreasing a lot when some other (like Elastic Net) still keep a quite good percentage of fitting. However, when it comes to add outliers, neither of these processes can have a decent probability to find the right model. Obviously, these results have to be taken with in hindsight since we evaluated these processes on a little scale.

# Contents

# 1 Introduction

The Big Data emergence of these past decades requires more and more advances technologies to take advantage of all of this available data. In this situation, the development of statistical learning and machine learning is becoming important. Indeed, these technologies were created with the goal of automate a lot of processes, especially concerning the selection of significant variables. It's especially efficient when the database is important : these new technologies can, more than past technologies because they are now quite independent, scan really big databases since everything is now automated according to what people want. To begin this study, it feels necessary to define what statistical learning and machine learning are and what makes them different. Both are indeed really close : they are processes designed to make the most accurate prediction by automatically selecting some variables in a database. The main difference between those two is that the statistical learning will use what we call inference when the machine learning won't. To explain what inference is, we can say that it's quite close to the notion of generalization : "it's about assigning to the population some known features that we find in a sample of this population". We must obviously translate that in our situation so that it makes sense. The statistical learning models will be faced by some distinctive situations (the sample) in such a way that it will train them to react in the future when they will be facing some database later. It's the main difference between the statistical learning and machine learning. In this study, we will try to answer the following question

Key question : How do automatic variable selection algorithms behave when fundamental ordinary least squares hypotheses are violated?

We will follow a three part plan. Firstly, we will introduce the main technologies of the statistical learning and machine learning like the stepwise regression, LASSO, LAR and Elasticnet... Then, in a second part, we will try these technologies to show their efficiency in a situation where all the hypothesis hold. Finally, we will try these technologies on some alternative situations (when there is correlation between the variables...).

# 2 Automatic selection tools

## 2.1 Selection criteria

We enumerate many techniques that allow automatic selection of variables; they are used to lead Statistical or Machine learning through taking decision to optimize our model. Each of these are statistical indicators, from statistical criteria to tests, and are more or less suitable for each situation we will describe afterwards.

In our study, we chose to only take in account selection criteria, which are AIC, AICc, BIC and PRESS. Let's describe these methods.

**$R^2$** or coefficient of determination represents the proportion of the explained variable's variance correctly predicted by the model.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \tag{1}$$

where $SSE$ is the sum of squares of errors (residuals), $SSR$ is the sum of squares of regression and $SST$ is the sum of squares total (total variance). The $R^2$ tends to increase at each added variable, but more terms in the model isn't often synonym of better fit; thus $R^2$ isn't really interpretable when we face a multiple regression made up with several independent variables. That's why, in this case, we usually take in consideration the adjusted $R^2$, which increases when the new variable has a great probability of prediction, and decreases when the new predictor is not significant for our model.

$$R^2_{adj} = 1 - \frac{(T-1)(1-R^2)}{T-k} \tag{2}$$

with $T$ the number of observations and $k$ the variables' one.
In this study we are not going to observe the movements of this criterion as a way to evaluate our model. Indeed, it doesn't tell anything on the bias of the predictors or the reliability of the global model. Moreover, an interpretation of the adjusted $R^2$ is complex, especially through the simulated data we worked on and that we will introduce in a second section.

**AIC** stands for Aikake Information Criterion. It fits for models estimated with a maximum likelihood method (multiple linear regression, logistic regression...). It is a compromise between bias (which decreases with the number of parameters) and model parsimony (willingness to describe data with the minimum number of parameters) :

$$AIC = T\log(\frac{SSE}{T}) + 2k \tag{3}$$

The best model has the lowest AIC value.

**AICc** is a corrected version of the AIC, which is preferable to the latter when $k$ is much higher than $T$. conventionally, $k$ is high when $\frac{T}{k} < 40$

$$AICc = 1 + \log(\frac{SSE}{T}) + \frac{2(k+1)}{T-k-2} \tag{4}$$

**BIC** stands for Bayesian Information Criterion. It is more parsimonious than AIC as it penalizes more the number of parameters of the model. It particularly allows to select the statistically significant ones.

$$BIC = T\log(\frac{SSE}{T}) + 2(k+2)\frac{T\hat{\sigma}^2}{SSE} - 2(\frac{T\hat{\sigma}^2}{SSE})^2 \tag{5}$$

**PRESS** stands for Predicted Residual Sum of Squares. It calculates the ability of prediction for a model. More precisely, it compares the regression model to the same regression model in which we would take off the observations one by one.

$$PRESS = \sum_{i=1}^{T}(y_i - \hat{y}_{(i)})^2 \simeq SSE \tag{6}$$

Where the index $(i)$ means that we removed the observation $i$.
The best model structure has the lowest PRESS value.

We can then choose to use these criteria as selection criteria, to accept or not new predictors, or even remove some of them, and as stop criteria, to end the process to an optimal model.

## 2.2 Statistical Learning

Statistical learning is mostly used to describe relationships between variables in a model. It also use inference to generalize results from a sample to a whole population.

There are three main selection methods that belong to the category of Statistical learning ; let's define them.

### 2.2.1 Forward selection

The idea of the Forward selection is to add predictors one by one, from a model which only contains a unique intercept equivalent to the mean of y. Thus, it selects one variable at a time, among the bunch of candidates, which provides the most statistically significance to the model according to the selection criterion we imposed.

In the same way, the selection is ended when no more predictor needs to be included, depending on the stop criterion we also specified.

### 2.2.2 Backward selection

The Backward selection works the exact opposite way as the Forward selection. Instead of starting with a model that only contain an intercept, the model contains all of the considered variables. Therefore, we do not add but remove gradually the ones that are less significant according to the selection criteria we have chosen. If that applies, the stop criterion will end the process before all of the candidates are included into the model.

### 2.2.3 Stepwise selection

Finally, the Stepwise selection perfectly combines the both previously described, as it is able to add and omit variables from the model. We start again from a very constrained model with a unique intercept. We add a variable at a time, while recalculating the model significance obtained. Hence, a primordial difference which defines the Stepwise selection is that it takes in consideration the variation of signification that appears when a variable is added or removed. Indeed, a significant one in a model of $k$ parameters can become non-significant as soon as a $k^{th}$ parameter is added, or even when one of the $k$ parameters is omitted. To sum up, it is a very dynamic method that allows, like the others, to obtain an optimal model, according to upstream fixed criteria.

Nevertheless, we will not use this method during this study; we will only focus on results of the Backward and Forward selection.

## 2.3 Machine Learning

The purpose of machine learning is to obtain a model that can make repeatable predictions ; but typically, interpretability is not its strong point. Indeed, machine learning is all about performance and results. There are four principal machine learning processes that we will introduce thereafter. All the predictors must be standardized and the explained variable is centered. Let's consider the following model:

$$y = X\beta + \varepsilon \tag{7}$$

### 2.3.1 LAR

The LAR regression, which stands for Least Angle Regression, is a procedure that do prediction by selecting the variables which are the most correlated with errors. It produces a shrinkage for each coefficient. The algorithm is given by the following sequence:

1. if predictors are already standardized, we start by making $\beta$ equal to zero and computing $r^i = y - \bar{y}$ for $i = 1$
2. we find the variable $x_j$ the most correlated with $r^i$
3. gradually vary $\beta_j$ towards its correlation coefficient with the obtained $r$ using ordinary least squares (OLS), until $\beta_l$ presents a higher correlation with the observed errors, for all $l$ different of $j$.
4. gradually vary $\beta_j$ and $\beta_l$ towards their OLS estimators with errors, until another $\beta_m$ presents a higher correlation with the observed errors.
5. iterate the algorithm until the $k$ predictors are included in the model.

The LAR regression presents however two disadvantages. First, this method is very sensitive to noise. Second, given that it is founded on correlation between predictors and residuals, it does not perform well in the presence of multi-collinearity.

This procedure is in a way the ancestor of the LASSO regression, which is introduced next. Let's see on what point it is best.

### 2.3.2 LASSO

In our study, we use the Least Absolute Shrinkage and Selection operator - abbreviated as LASSO - as a technique to value the regression's relevant variables by shrinking their associated coefficients. The coefficient shrinkage acts through a constraint, $\lambda$. It aims at selecting the relevant variables and therefore, drops the irrelevant ones by setting their coefficient value to zero, so that they do not appear in the estimated regression. With these features, we can consider this regression as an automatic procedure of variable selection.
Suppose a linear regression with standardized predictors $x_ij$ and centered response values $y_ij$. The LASSO regression minimizes the sum of squares, with a constraint in the form of an L1-norm penalty. The LASSO equation is as it follows:

$$\arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_1 \tag{8}$$

Where the L1-norm penalty is:

$$\|\beta\|_1 = \sum_{j=1}^{k} |\beta_j| \tag{9}$$

If $\lambda$ becomes smaller, the results given by the LASSO regression will converge to the results given by the OLS regression. On the contrary, when the $\lambda$ increases, the result given by the LASSO regression becomes parsimonious, which means that we minimize the number of variables in the model by penalizing more on the coefficients values. Regularization prevents from overfitting, i.e. accepting too much predictors in the regression. It works by adding a term of regularization (L1 or L2-norm according to the hypothesis) to the SSE which measures the model complexity.

This procedure presents two main defects. The first one is that if the number of variables is much higher that the number of observations, that is to say if $k >> T$, LASSO only selects $T$ predictors maximum. Moreover, it does not deal with correlation between explanatory variables: it will detect the one with the highest correlation with the target explained variable and hide the others. In conclusion, the possible influence of the other predictors will not be shown.

### 2.3.3 Ridge Regression

The LAR and LASSO regressions have a disadvantage : they do not support multicollinearity between variables nor extreme value. To overcome this problem, we may use the Ridge regression (RR). Indeed, the RR was developed as a way to solve the inaccuracy of OLS caused by highly correlated variables and so non-full rankness of $X$ ($RgX < k+1$) through regularization.

We therefore have to solve the following optimisation problem :

$$\arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \tag{10}$$

with $\|\beta\|_2^2$ a term of regularization for the linear regression, called L2-norm (or euclidian norm), and $\lambda$ a positive penalty coefficient; it is all about shrinkage again.

To obtain the best fit, the choice of $\lambda$ is crucial but complex. An option is to vary it and observe the evolution of the coefficient stability (Ridge coefficients paths). To evaluate the coefficients performance, we typically use cross-validation criterion.

At the end, the RR gives a unique solution for the estimator of $\beta$:

$$\hat{\beta}_{Ridge} = (X'X + \lambda I)^{-1} X'y \tag{11}$$

where the matrix $(X'X + \lambda I)$ is indeed invertible for $\lambda > 0$. The variance and the mean square error of (11) are then often smaller than the variance of the OLS estimator, which makes it a more efficient estimator.

Attention, the RR does not select predictors unlike LASSO regression, it tests their performance on a model predefined.

### 2.3.4 Elastic Net

Elastic Net is a recent and very used procedure combining LASSO, unreliable in the presence of collinearity, and Ridge regression, which fixes collinearity internally. All in all, taking in consideration the L1 and L2 penalty criteria, Elastic Net allows us to do Machine Learning with highly-correlated variables.

Estimations with Elastic Net are then given by the following formula :

$$\arg\min_{\beta} \|y - X\beta\|^2 + \lambda \left( \alpha \|\beta\|_1 + (1-\alpha)\|\beta\|_2^2 \right) \tag{12}$$

The second term of (12) is a penalty function; we notice that for $\alpha = 0$, we obtain the Ridge estimator and for $\alpha = 1$ it is the LASSO estimator that appears. Elastic Net estimator is actually used at the moment where $0 < \alpha < 1$.

# 3 Empirical application

Here finally comes the part where we test Statistical and Machine learning. To implement these, we simulated data on SAS Software and applied the processes described earlier on these data, with the ultimate aim to evaluate and compare their performance.

To consider all cases, we created four data generating processes (DGP) from our simulated data. The first one could be interpreted as a control group, as it follows ideal (and not very realistic) hypothesis. To the three other ones, we omitted some of these basic hypothesis to evaluate the comportment of the different technics when faced to data anomalies.

Each DGP contains fifty series of a hundred observations. We created a precise model:

$$y = 0.7x_1 + 0.6x_2 + 0.5x_3 + 0.4x_4 + 0.3x_5 + \varepsilon \tag{13}$$

The main purpose of this study is then to observe how reacts each regression technic, with four selection criteria (AIC, AICc, BIC and PRESS) faced to different type of data. To do that, we generated a hundred Monte Carlo simulations for each procedures and reported the rate of good prediction.Let's dive into it.

## 3.1 First DGP : the control group

The control group reunites all "perfect" hypothesis that we can find in a model : the data follows a gaussian law, with a null mean vector and an identity variance-covariance matrix, which means that variables are perfectly independant ; the error vector follows a gaussian law ; there is no outlier, etc. In short, there is no perturbation in this model that can damage the prediction.

Once this first DGP defined, we began with the Forward selection. For each select criteria, the output were approximately the same. Here are the results with the AICc criterion.

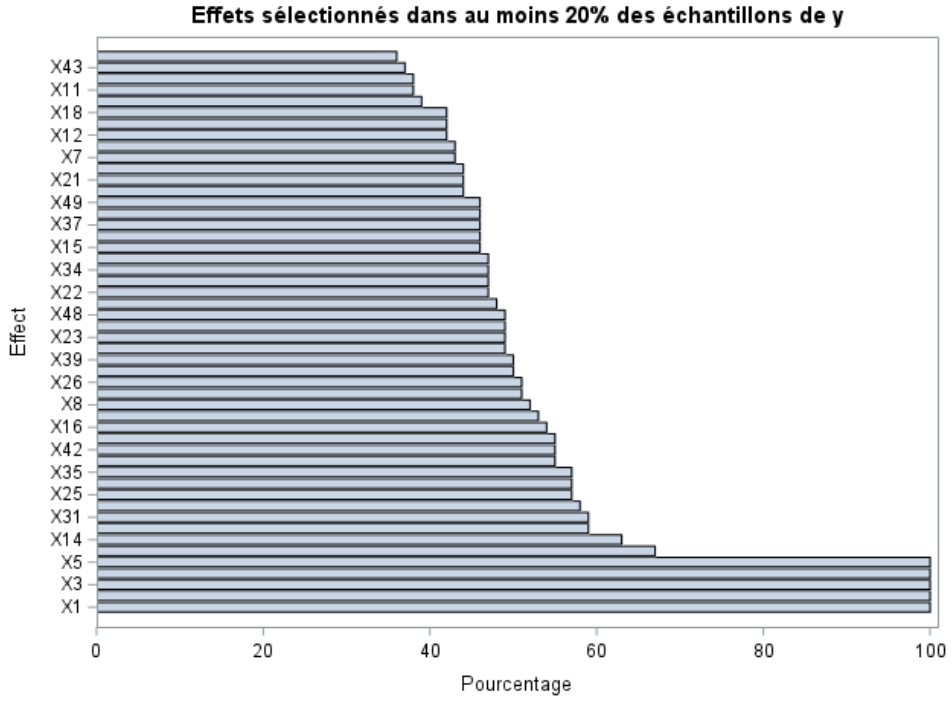


Figure 1: Probability to find variables individually with the forward selection

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 1 | 1.00 | 13 | 1.74 | Intercept X1 X2 X3 X4 X5 X7 X10 X19 X22 X23 X24 X35 |
| 1 | 1.00 | 14 | 1.73 | Intercept X1 X2 X3 X4 X5 X9 X10 X16 X22 X30 X31 X35 X36 |
| 1 | 1.00 | 14 | 1.72 | Intercept X1 X2 X3 X4 X5 X9 X20 X21 X22 X25 X28 X42 X46 |
| 1 | 1.00 | 16 | 1.69 | Intercept X1 X2 X3 X4 X5 X11 X18 X23 X27 X29 X32 X33 X38 X39 X50 |
| 1 | 1.00 | 19 | 1.68 | Intercept X1 X2 X3 X4 X5 X10 X12 X20 X22 X25 X31 X32 X33 X34 X35 X39 X42 X45 |
| 1 | 1.00 | 17 | 1.68 | Intercept X1 X2 X3 X4 X5 X6 X9 X19 X23 X27 X31 X33 X36 X37 X38 X39 |
| 1 | 1.00 | 19 | 1.68 | Intercept X1 X2 X3 X4 X5 X10 X14 X16 X19 X25 X29 X31 X32 X33 X39 X40 X43 X48 |
| 1 | 1.00 | 16 | 1.68 | Intercept X1 X2 X3 X4 X5 X10 X12 X18 X19 X20 X29 X31 X37 X42 X44 |

Figure 2: Probability to find variables jointly with the forward selection

We can first suggest that varying the select criteria doesn't make a significant difference on the Forward selection : the results have more or less the same shape. What is satisfying in these results is that, the Forward selection finds each of our fifth explanatory

variables at each time, as we can see on figure 1. However, it also finds other variables that we didn't defined by hand: it is a situation of overfitting. For instance, the variable $x_{16}$, that is not one of our expected variables, is selected by the forward selection with a probability of about 60% which is non-negligible. We can also see that on the figure 2 : the probability to find all five variables of the model at the same time is very high but a perfect fit of the model is impossible here.

Hence, we can deduce from this that the quality of prediction of the model that we obtain with the forward selection is not really good because of the really important risk of overfitting.

We can now try the backward selection :

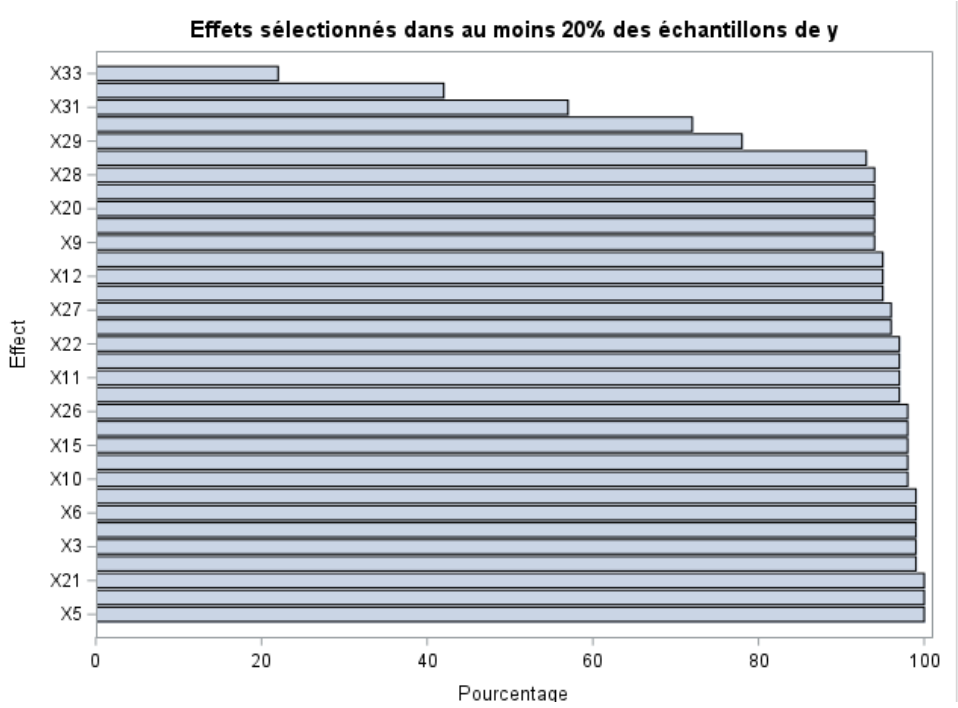


Figure 3: Probability to find variables individually with the backward selection

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 3 | 3.00 | 32 | 3.93 | Intercept X1 X2 X3 X4 X5 X6 X7 X8 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30 X31 X32 |
| 3 | 3.00 | 32 | 3.93 | Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X26 X27 X28 X29 X30 X31 X32 |
| 2 | 2.00 | 27 | 2.97 | Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X25 X26 X27 |
| 2 | 2.00 | 28 | 2.97 | Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 |
| 2 | 2.00 | 29 | 2.97 | Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 |
| 2 | 2.00 | 29 | 2.96 | Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 |

Figure 4: Probability to find variables jointly with the backward selection

We apply it the same way as the previous process to judge the quality of the prediction.

In a similar way, the results are not very convincing. Even though this process succeeded to predict the variables of the model, it also predicts some parasitic variables with an high probability. Worst than that, unlike the forward selection, this process will even select some variables that are not in the model with the same probability as the variables that are in the model : the overfitting is even worst here. So we can conclude that with the backward selection, the automatic selection of variables is not really effective and even less effective than the forward selection. These are not surprising results because, on average, the backward selection is more efficient in the case where the variables are correlated which is not the case here. Overall, we can finally say that the quality of prediction of the statistical learning process is quite limited, whether it be with the forward selection or with the backward selection. It will then be interesting to see if the machine learning process are more efficient in terms of predictions We begin by trying the LASSO regression.

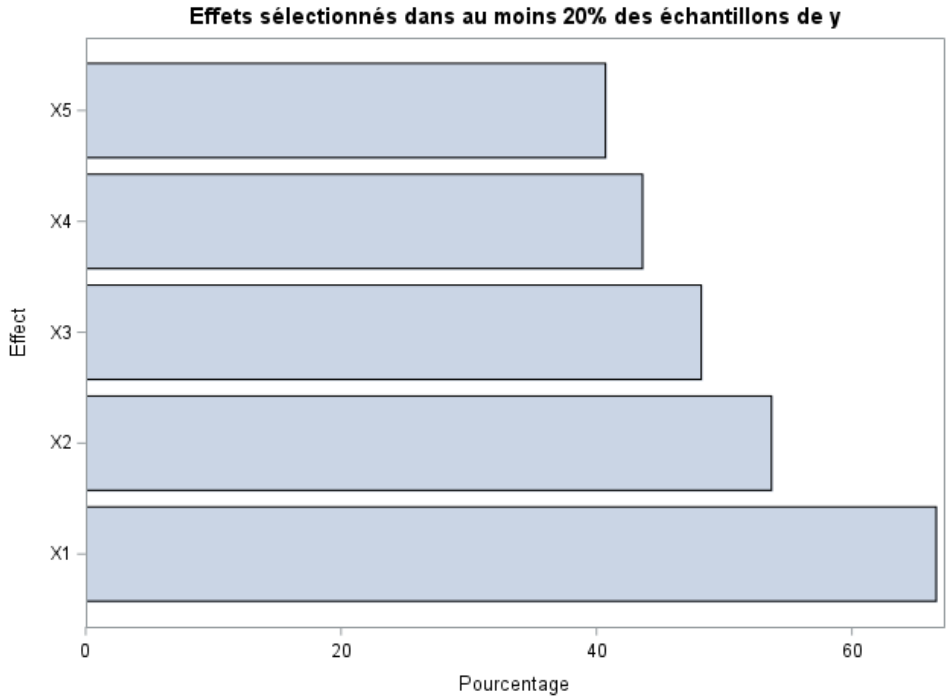


Figure 5: Probability to find variables individually with LASSO

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 318 | 31.80 | 1 | 319.0 | Intercept |
| 113 | 11.30 | 2 | 113.8 | Intercept X1 |
| 61 | 6.10 | 3 | 61.73 | Intercept X1 X2 |
| 40 | 4.00 | 6 | 40.59 | Intercept X1 X2 X3 X4 X5 |
| 29 | 2.90 | 4 | 29.67 | Intercept X1 X2 X3 |

Figure 6: Probability to find variables jointly with LASSO

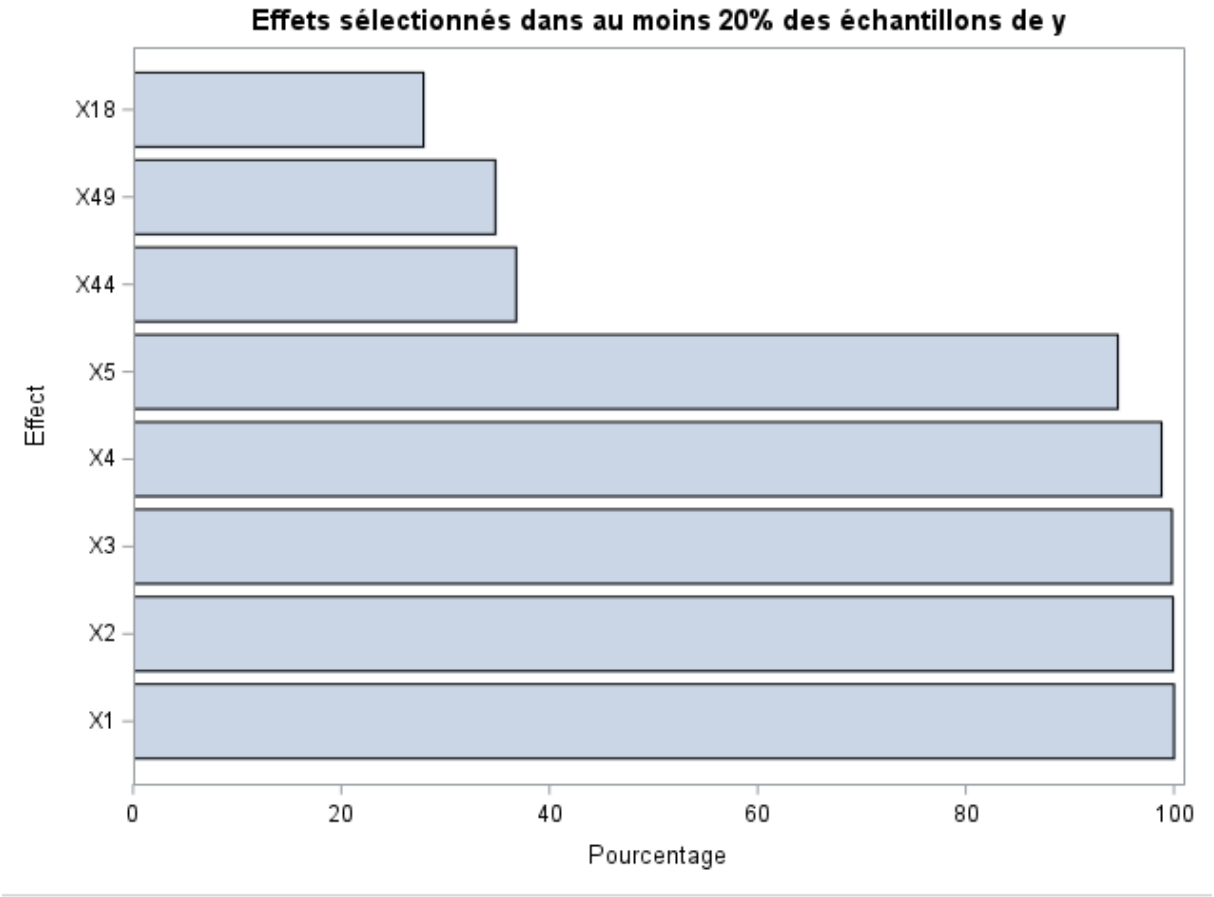


Figure 7: Probability to find variables individually with LASSO with a PRESS criteria

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 146 | 14.60 | 6 | 147.0 | Intercept X1 X2 X3 X4 X5 |
| 28 | 2.80 | 5 | 29.00 | Intercept X1 X2 X3 X4 |
| 24 | 2.40 | 7 | 24.90 | Intercept X1 X2 X3 X4 X5 X49 |
| 24 | 2.40 | 7 | 24.87 | Intercept X1 X2 X3 X4 X5 X36 |
| 23 | 2.30 | 7 | 23.90 | Intercept X1 X2 X3 X4 X5 X44 |
| 13 | 1.30 | 7 | 13.87 | Intercept X1 X2 X3 X4 X5 X21 |
| 13 | 1.30 | 7 | 13.87 | Intercept X1 X2 X3 X4 X5 X50 |

Figure 8: Probability to find variables jointly with LASSO with a PRESS criteria

We didn't do the graphics for the AICC criteria and the SBC criteria because the results were more or less the same in comparison with the BIC one so we will consider that the interpretation is the same too. If we want to read into these results, we can say that, overall, the variables seem to be quite correctly predicted, whether it be individually or jointly.

The process LASSO predicts only the variables of our original model that we've choose previously with very high probabilities and absolutely no waste (= when the process predicts variables that are not in our original model). The use of the PRESS criteria seems

even better because this one predicts all the variables of our original model individually with a probability of an hundred percent, which is perfect, and also with almost no waste. In comparison with the statistical learning, the improvement is huge. Indeed, the LASSO seems to have resolved the main problem of both processes of the statistical learning : there is now no (or almost) parasitic variables which were really damaging to the quality of the prediction of these previous selections because of the important risk of overfitting

We can logically see that in the probability to find the variables of our original model jointly. The main case that show this is when we use the PRESS criteria. Indeed, we get really interesting results. According to the PRESS criteria, the original model is indeed the model that is the most likely to be chosen by the LASSO with almost 14 percent of being chosen just as we choose it. It's also really significant, it really stand out because any others models that the LASSO predicts in this case has a probability of less than three to be percent to be chosen by the LASSO. So the LASSO with the help of the PRESS criteria as a stopping criteria allow a really good prediction of variables especially in comparison of what we have seen so far.

With all of the others criteria, it's more complicated. Indeed, the PRESS criteria seems to outperform them. In theses cases, the little overfitting seems to really hurt the quality of prediction of the model with only four percent of choosing the right model (the percentage are much higher only for some parts of the model like intercept alone has an around thirty percent chance of being chosen as a model).

To conclude this part about LASSO, we can conclude that this first process of machine learning make overall better predictions of variables in comparison with the statistical learning especially when the PRESS criteria is chosen with very little over fitting.

We will now try to predict our model with the help of LAR. In the same way as we did for LASSO, we will combine the results for the criteria because of their similarity.

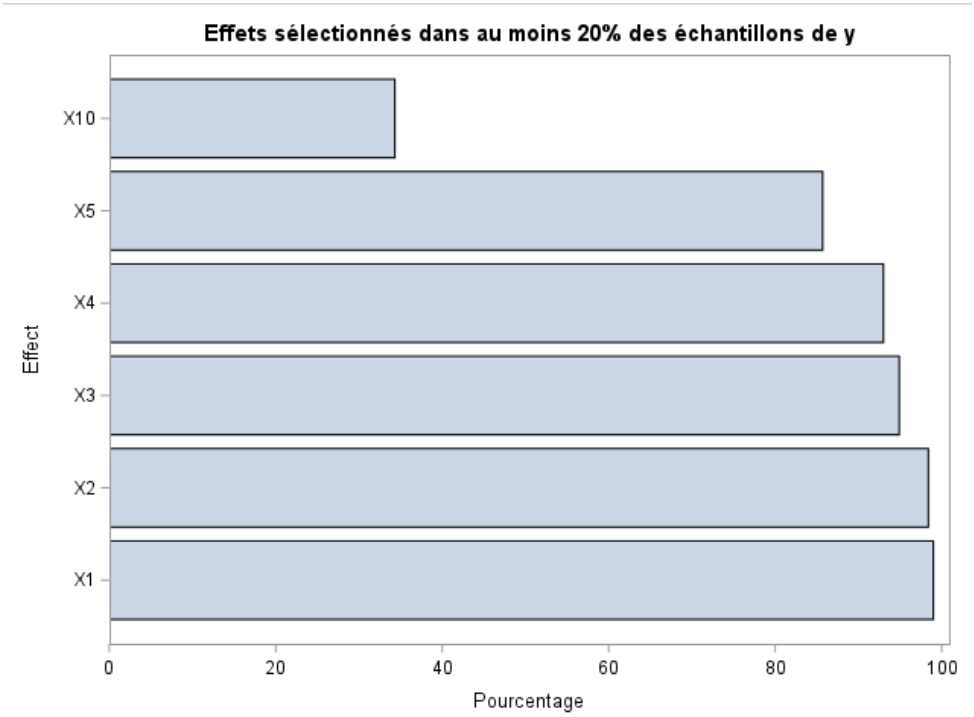


Figure 9: Probability to find variables individually with LAR

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 170 | 17.00 | 6 | 171.0 | Intercept X1 X2 X3 X4 X5 |
| 40 | 4.00 | 7 | 40.86 | Intercept X1 X2 X3 X4 X5 X10 |
| 28 | 2.80 | 5 | 28.97 | Intercept X1 X2 X3 X4 |
| 20 | 2.00 | 4 | 20.98 | Intercept X1 X2 X3 |
| 17 | 1.70 | 7 | 17.84 | Intercept X1 X2 X3 X4 X5 X46 |
| 14 | 1.40 | 7 | 14.82 | Intercept X1 X2 X3 X4 X5 X33 |
| 13 | 1.30 | 7 | 13.83 | Intercept X1 X2 X3 X4 X5 X42 |

Figure 10: Probability to find variables jointly with LAR

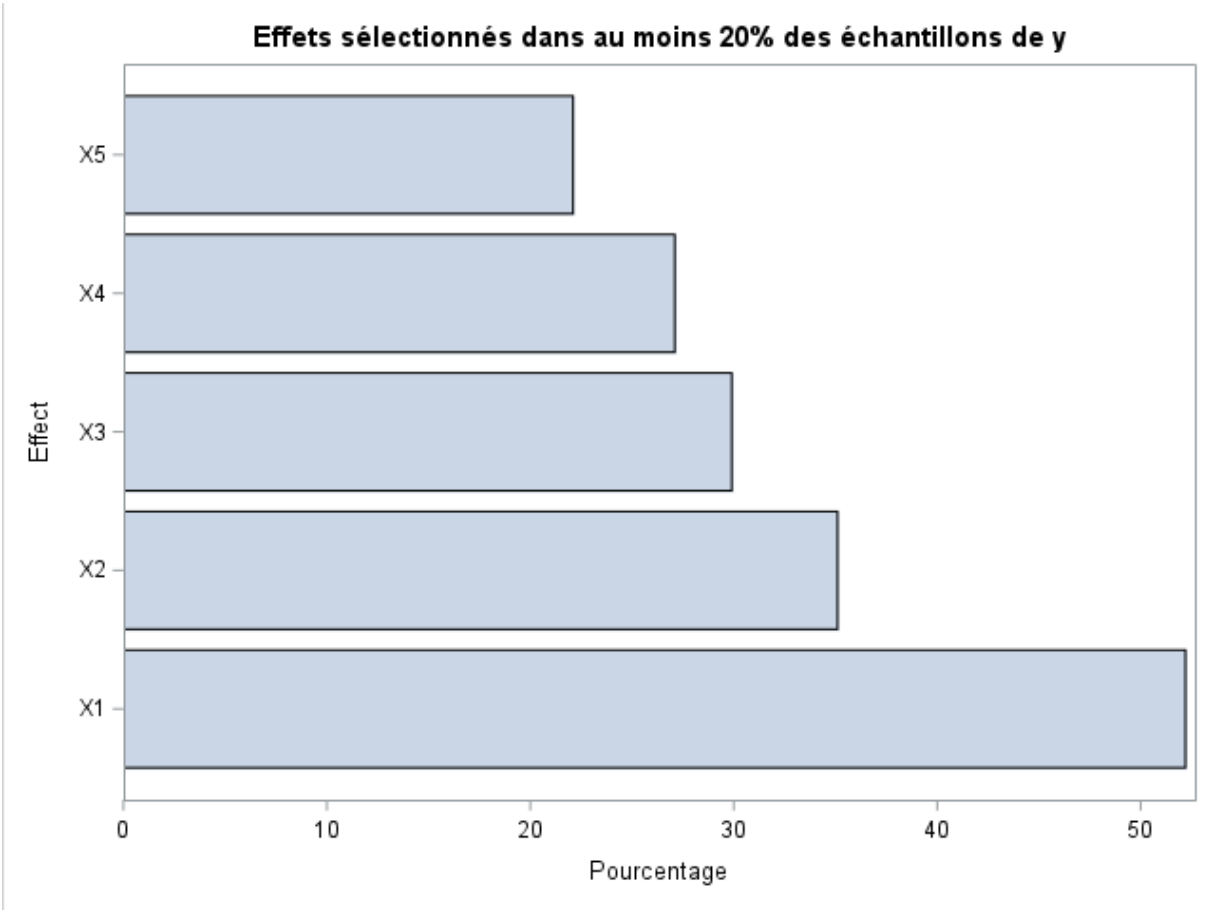


Figure 11: Probability to find variables individually with LAR with a PRESS criteria

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 429 | 42.90 | 1 | 430.0 | Intercept |
| 138 | 13.80 | 2 | 138.8 | Intercept X1 |
| 49 | 4.90 | 3 | 49.62 | Intercept X1 X2 |
| 24 | 2.40 | 4 | 24.54 | Intercept X1 X2 X3 |
| 20 | 2.00 | 3 | 20.61 | Intercept X1 X3 |
| 16 | 1.60 | 6 | 16.44 | Intercept X1 X2 X3 X4 X5 |

Figure 12: Probability to find variables jointly with LAR with a PRESS criteria

We find similar results as the LASSO case which is not really surprising because these process are really close In these case too, the PRESS criteria seems to stand out from the other in terms of joint probability. The quality of predictions seems also quite better than the statistical learning, with less predictions with parasitic variables in the estimated model.

We can then try Elastic Net for this group control.

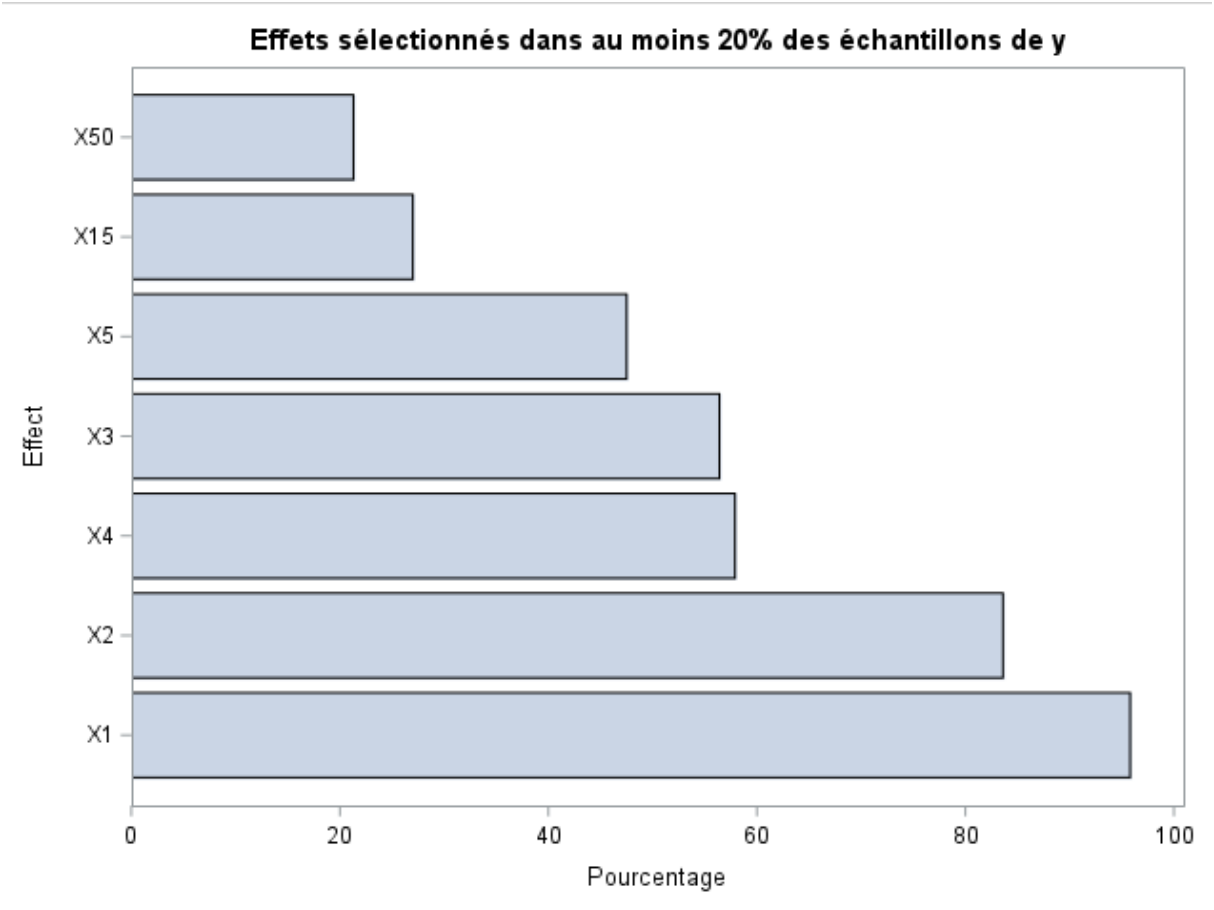


Figure 13: Probability to find variables individually with Elastic Net

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 209 | 20.90 | 3 | 209.9 | Intercept X1 X2 |
| 101 | 10.10 | 2 | 102.0 | Intercept X1 |
| 42 | 4.20 | 1 | 43.00 | Intercept |
| 38 | 3.80 | 4 | 38.84 | Intercept X1 X2 X4 |
| 37 | 3.70 | 6 | 37.74 | Intercept X1 X2 X3 X4 X5 |
| 33 | 3.30 | 5 | 33.79 | Intercept X1 X2 X3 X4 |
| 22 | 2.20 | 4 | 22.84 | Intercept X1 X2 X3 |

Figure 14: Probability to find variables jointly with Elastic Net with a PRESS criteria

With Elastic Net, we get results in line with LAR and LASSO with a AICC/BIC/SBC criteria : we get decent predictions, better than these from the statistical learning, but less good than the one we got with the PRESS criteria from LAR and LASSO. Indeed, we only get on average four percent of a joint probability to find all the variables of the model. It's not surprising because Elastic Net was made from the LASSO in order to solve other problems from the LASSO like the correlation between variables that we will see later in this study. In this case, we have also combine all the criteria together because the results were very close.

## 3.2 Second DGP: the variables are correlated with each other

Once we get these results, the second part of our empirical study will be about releasing some hypothesis of our original model. Indeed, in this model, a lot of hypothesis holds like the lack of correlation between the variables of our model which finally make the model quite restricted. So, it can be interesting to see the quality, or more precisely the evolution of the quality of the prediction of these different process that we saw before when some of these conditions don't hold. The first one hypothesis that we will try to release will be the total absence of correlation between the variables. Indeed, in our original, there was absolutely no correlation between the variables which we can verify by calculating the correlation matrix. To release this hypothesis, we will try to modify this correlation matrix and create some correlation between the variables. To do so, we will use an Iman Conover transformation. This will make all the variables highly correlated. In this case, we have chosen to use high coefficients to make sure that there was enough correlation between the variables. The new correlation matrix is now the following

We can then once again see the results of all our process in this new context Once again, we begin by the statistical learning. The first process that we will try is going to be the forward selection. Just like we could think, the results are not really convincing. We could really predict that because the forward selection was already struggling to predict accurately the model when the model had no issues. The results are overall quite similar whether there is correlation between the variables or no : whatever is the criteria, the variables are predicted decently individually but these results are tarnished by some high probabilities of overfitting, predicting variables that are not in the original model. The joint probability to find our model confirm that : all the variables of our original model are never predicted with a decent probability (the process predicts instead a lot of models with really low probabilities which seem almost random which is not really convincing in terms of predictions). As excepted too, the results are similar for the stepwise backward

selection in comparison with the results of this same process without correlation. Indeed, we find the same results that weren't convincing before and are not more convincing now. It can really make us doubt the overall efficiency of the statistical learning at our scale given that the results never seem to be good.

We can then try the machine learning.When we use LASSO, we get the following result :

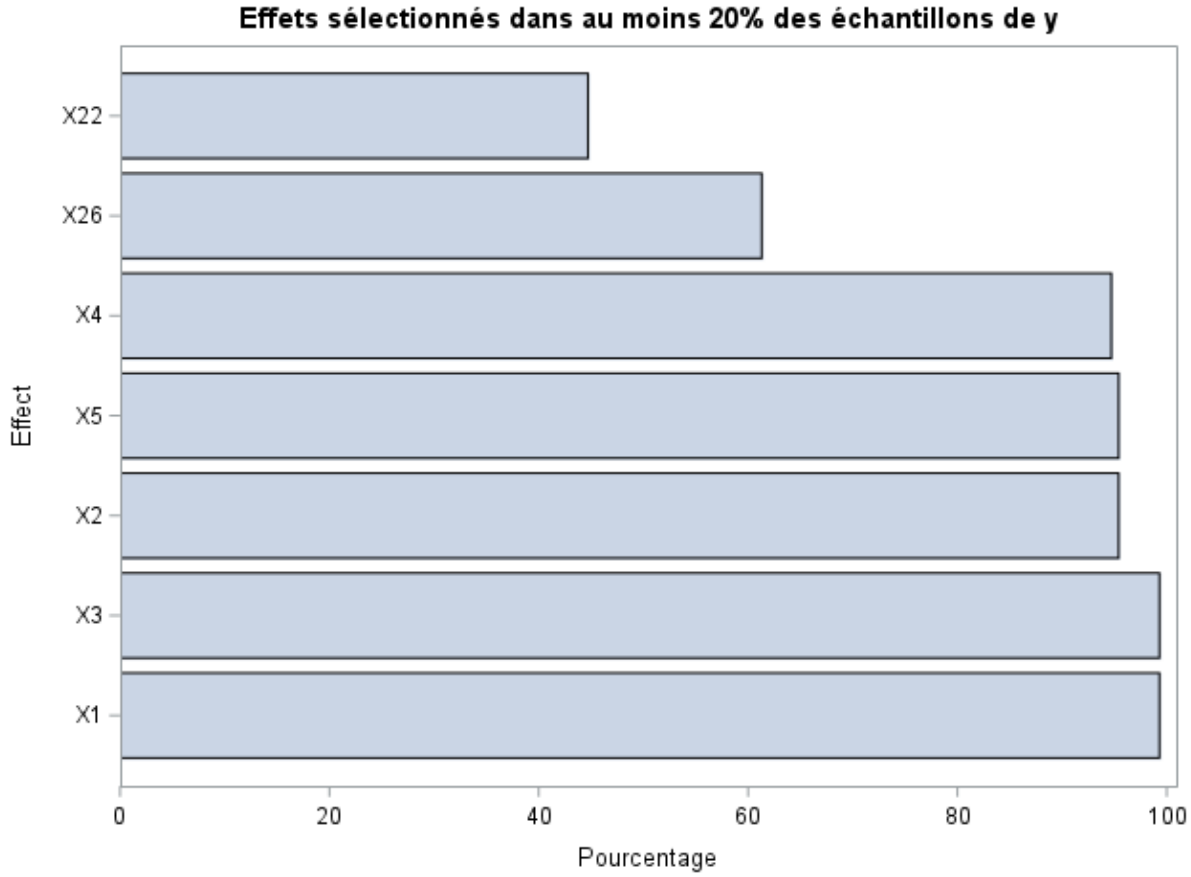


Figure 15: Probability to find variables individually with LASSO with a PRESS criteria

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 46 | 4.60 | 3 | 47.00 | Intercept X1 X3 |
| 42 | 4.20 | 7 | 42.92 | Intercept X1 X2 X3 X4 X5 X26 |
| 37 | 3.70 | 8 | 37.86 | Intercept X1 X2 X3 X4 X5 X22 X26 |
| 15 | 1.50 | 6 | 15.97 | Intercept X1 X2 X3 X4 X5 |
| 15 | 1.50 | 7 | 15.90 | Intercept X1 X2 X3 X4 X5 X22 |
| 11 | 1.10 | 7 | 11.86 | Intercept X1 X2 X3 X4 X5 X13 |

Figure 16: Probability to find variables jointly with LASSO with a PRESS criteria

In this case, we see that the results are really different from what we got before. Indeed, in this case, the joint probability to find all the variables of our model as well as

the intercept really dropped in comparison with the case where they were no correlation between variables, going from 14 percent to 1,50 percent only. So it implies that there is a really big decrease in terms of quality of prediction when we include some correlation between the variables of the model. When we add correlation and try the LASSO, we find overall results that are similar to what we find when we use the statistical learning : the machine learning has lost all of his predictive power. It's not a surprising result : the LASSO process is known for having trouble concerning the situations where there is correlation between the variables. It will be interesting to compare this situation with the Elastic Net process, because this process is in fact an extension of the LASSO developed especially with the goal of solving the LASSO limits concerning the correlation between variables To clarify, in this case, we only used the PRESS criteria, because it gave us the most significant results.

Now, it makes sense to continue this study by trying to use the Elastic Net process in this context. We then get the following results.

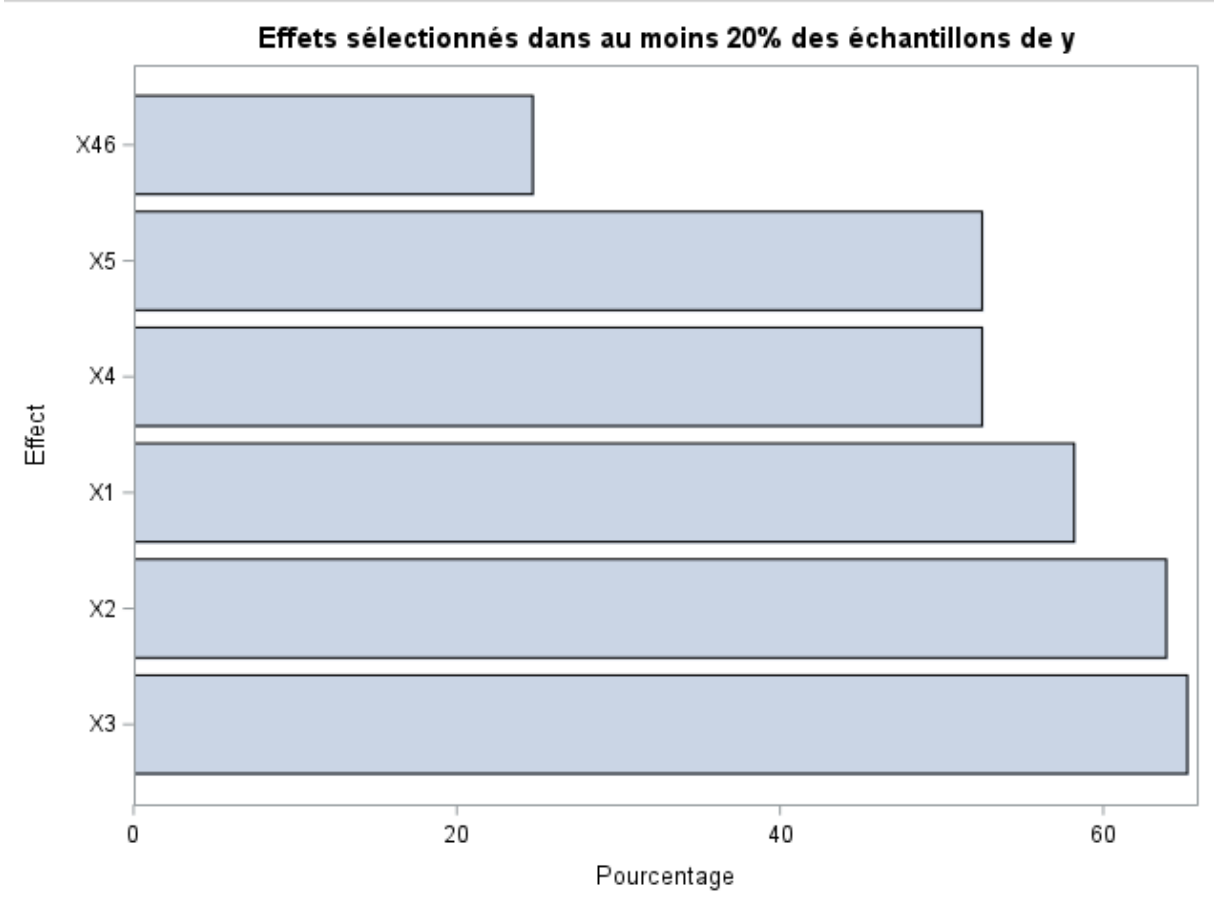


Figure 17: Probability to find variables individually with Elasticnet with a PRESS criteria

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 269 | 26.90 | 1 | 270.0 | Intercept |
| 132 | 13.20 | 6 | 132.7 | Intercept X1 X2 X3 X4 X5 |
| 72 | 7.20 | 7 | 72.60 | Intercept X1 X2 X3 X4 X5 X46 |
| 63 | 6.30 | 2 | 63.82 | Intercept X2 |
| 52 | 5.20 | 2 | 52.83 | Intercept X3 |
| 28 | 2.80 | 3 | 28.76 | Intercept X2 X3 |

Figure 18: Probability to find variables jointly with Elasticnet with a PRESS criteria

What we consider seems to happen here. Indeed, even though our original model is not the one that is predicted with the highest probability, the joint probability of finding all the right variables increases a lot (by 12 percent) in comparison with the LASSO process that we saw just previously. It's a significant increase to reach a good percentage of thirteen percent to find the right model which seems to be quite good in terms of quality of prediction. So we can say that the Elastic Net process has succeeded to handle the main limits of the LASSO that is the treatment of the correlation between variables : it doesn't affect anymore the quality of the prediction.

## 3.3 Third DGP : insertion of extreme values

We are going now to release another one of our initial hypothesis. In this case, we will introduce extreme values in our model to see if this has any impact on the quality of prediction of our processes. To do so, we will modify what we called the skewness. The skewness is a measure of the asymmetry of a distribution. Usually, the skewness is null but now we will modify it to make the skewness positive or negative but not null anymore.

Just like we did before, we will begin by studying the impact of adding extreme values on the statistical learning processes that are the forward selection and the backward selection

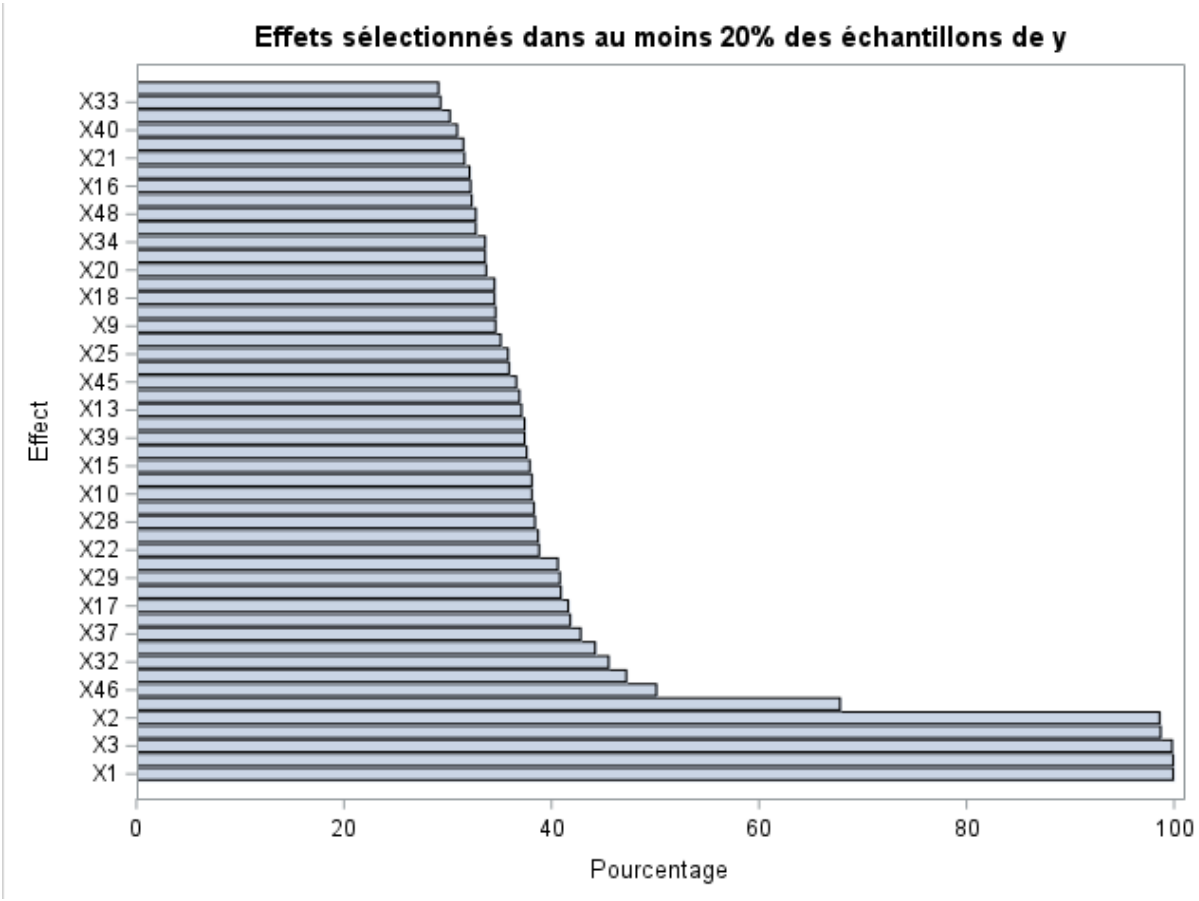


Figure 19: Probability to find variables individually with forward selection with a PRESS criteria

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 1 | 0.10 | 8 | 1.89 | Intercept X1 X2 X3 X4 X5 X14 X19 |
| 1 | 0.10 | 9 | 1.82 | Intercept X1 X2 X3 X4 X5 X12 X14 X18 |
| 1 | 0.10 | 10 | 1.79 | Intercept X1 X2 X3 X4 X5 X14 X19 X22 X29 |
| 1 | 0.10 | 10 | 1.77 | Intercept X1 X2 X3 X4 X5 X14 X26 X42 X44 |
| 1 | 0.10 | 11 | 1.73 | Intercept X1 X2 X3 X4 X5 X7 X17 X31 X43 X47 |
| 1 | 0.10 | 12 | 1.72 | Intercept X1 X2 X3 X4 X5 X7 X14 X17 X36 X39 X43 |
| 1 | 0.10 | 13 | 1.72 | Intercept X1 X2 X3 X4 X5 X14 X19 X37 X41 X43 X46 X49 |

Figure 20: Probability to find variables jointly with forward selection with a PRESS criteria

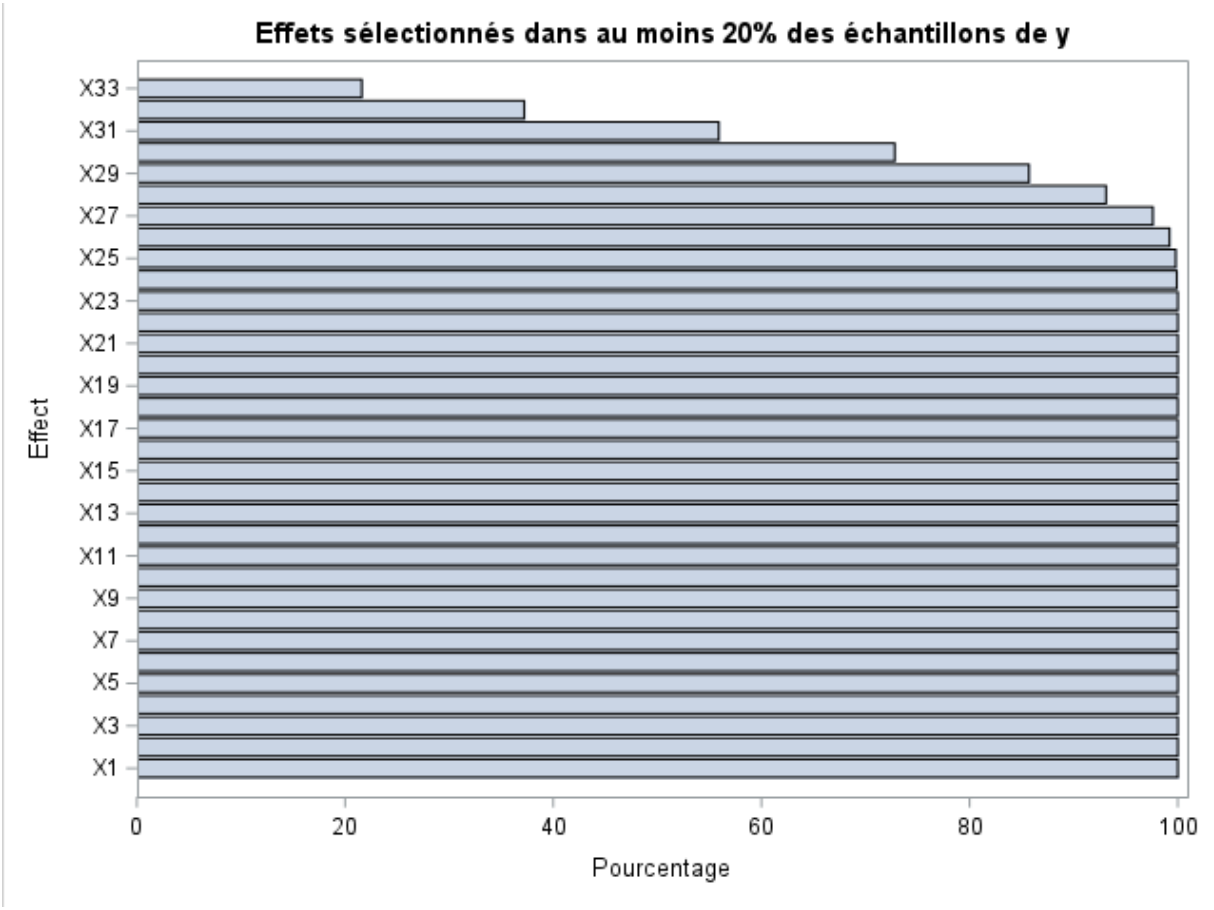


Figure 21: Probability to find variables individually with backward selection with a PRESS criteria

Fréquence de sélection du modèle

| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
|---|---|---|---|---|
| 187 | 18.70 | 32 | 188.0 | Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30 X31 |
| 169 | 16.90 | 31 | 170.0 | Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30 |
| 156 | 15.60 | 33 | 157.0 | Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30 X31 X32 |
| 129 | 12.90 | 30 | 130.0 | Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 |
| 99 | 9.90 | 34 | 99.93 | Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30 X31 X32 X33 |

Figure 22: Probability to find variables jointly with backward selection with a PRESS criteria

Just like what we saw about the statistical learning so far, the results of the prediction are not really good. The variables of the model are once again predicted one hundred percent of the time but the risk of overfitting is far too high to judge the quality of the prediction as satisfying. For example, in the case of the forward prediction represented by the figure 19b, we can see that almost thirty variables have a probability of forty percent to be add to the model by the process which is non negligible even though they shouldn't be there. All of this make the joint probability to find the right variables of the model really low to a point where it seems that it's almost random just like in the figure 22a. We could except that following that we saw before : if a process is not working correctly

when all the hypothesis hold, there is a little chance that it will get better when we decide to release some of the hypothesis making the prediction harder.

Following the results that we see in the last part, it seems more interesting to linger over the machine learning. We will begin with the LAR

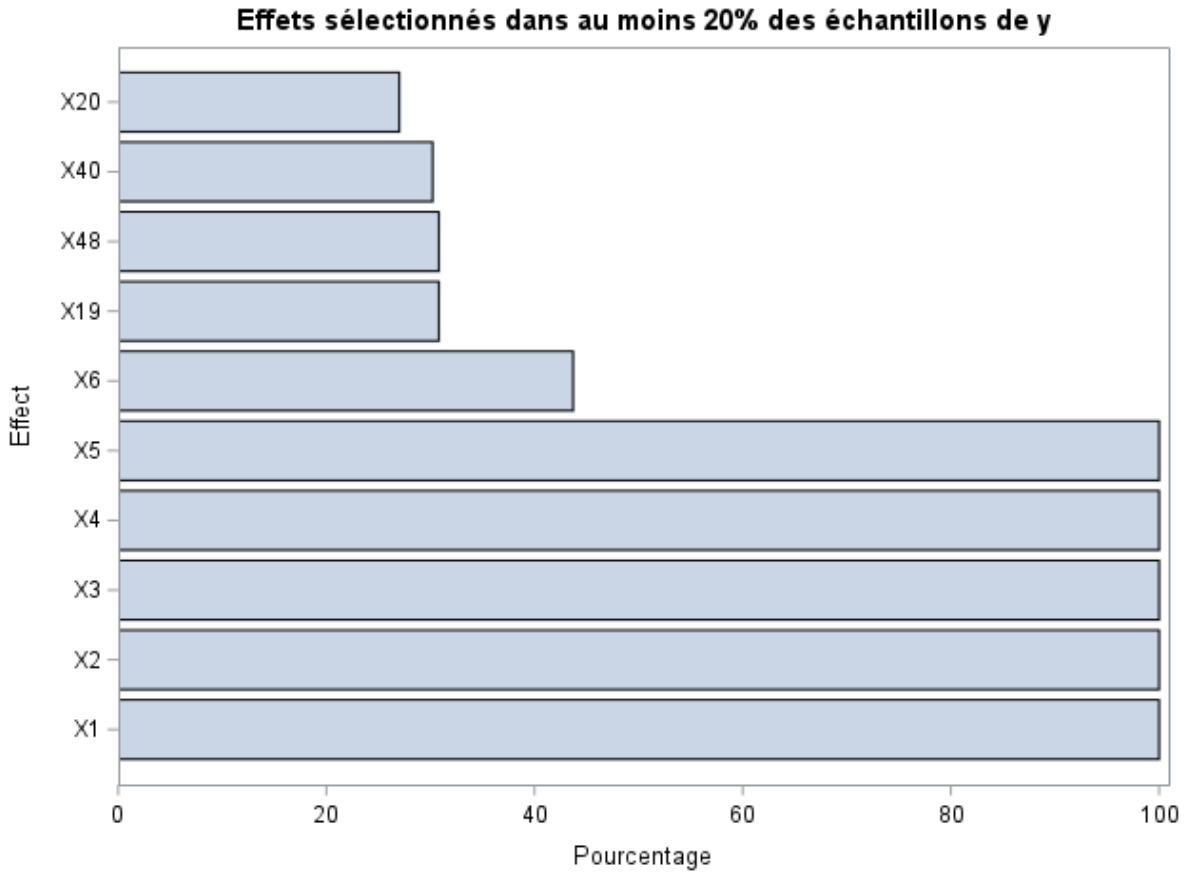


Figure 23: Probability to find variables individually with LAR with a PRESS criteria

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 37 | 3.70 | 6 | 38.00 | Intercept X1 X2 X3 X4 X5 |
| 30 | 3.00 | 7 | 30.92 | Intercept X1 X2 X3 X4 X5 X6 |
| 12 | 1.20 | 7 | 12.90 | Intercept X1 X2 X3 X4 X5 X19 |
| 12 | 1.20 | 7 | 12.90 | Intercept X1 X2 X3 X4 X5 X48 |
| 11 | 1.10 | 7 | 11.90 | Intercept X1 X2 X3 X4 X5 X40 |
| 9 | 0.90 | 7 | 9.90 | Intercept X1 X2 X3 X4 X5 X20 |

Figure 24: Probability to find variables jointly with LAR with a PRESS criteria

The results are indeed interesting. We see that the variables are accurately predicted by the model one hundred percent of the time individually. Despite that, we also see that there is a little bit of overfitting with other variables being predicted which tarnish the quality of prediction. Indeed, some variables, like X6 or X9 as we see in the figure

23b, have a non negligible probability to be predicted (around thirty percent). It has obviously an impact on the joint probability to predict all the variables together. The joint probability to find the model is now around three percent which is still better than in the case where there is correlation between the variables but it's still too low to be significant. So we cant consider the LAR process to be conclusive when we introduce extreme values in our model.

We will now analyze the results in the LASSO case. The results are the following :

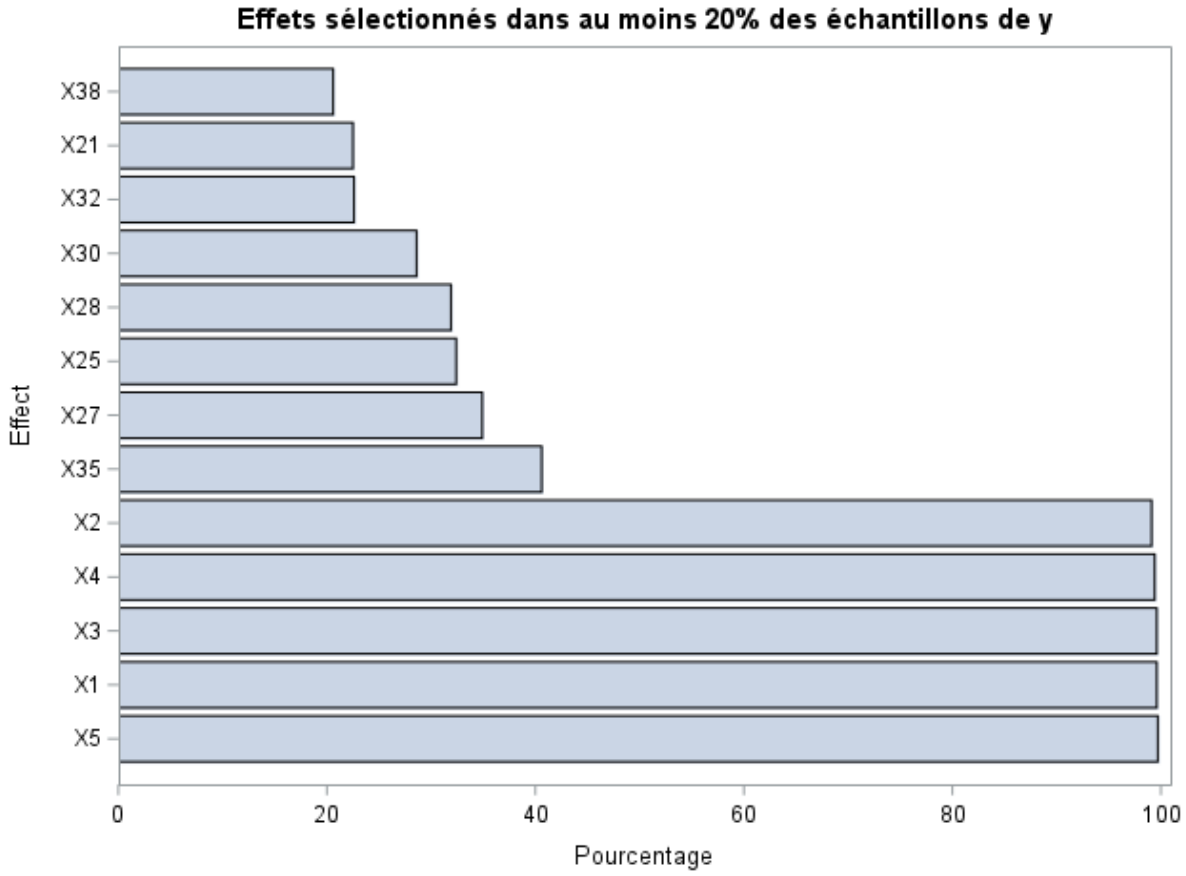


Figure 25: Probability to find variables individually with LASSO with a PRESS criteria

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 18 | 1.80 | 7 | 18.90 | Intercept X1 X2 X3 X4 X5 X27 |
| 13 | 1.30 | 7 | 13.91 | Intercept X1 X2 X3 X4 X5 X35 |
| 10 | 1.00 | 8 | 10.84 | Intercept X1 X2 X3 X4 X5 X25 X35 |
| 9 | 0.90 | 8 | 9.84 | Intercept X1 X2 X3 X4 X5 X27 X35 |
| 9 | 0.90 | 8 | 9.84 | Intercept X1 X2 X3 X4 X5 X28 X35 |
| 8 | 0.80 | 7 | 8.90 | Intercept X1 X2 X3 X4 X5 X28 |

Figure 26: Probability to find variables jointly with LASSO with a PRESS criteria

In this case, the LASSO doesn't really need more interpretation that what we said about the LAR. Indeed, the results are more or less the same : the variables of the model

are accurately predicted individually but there is some overfitting which will have an impact on the probability to predict exactly the right model which will, overall, make the quality of prediction low and not really conclusive. The LASSO doesn't seem more conclusive than the LAR when we add some extreme values to our model.

We must now try to use the Elasticnet process. It gives us the following results.

| Fréquence de sélection du modèle | | | | |
|---|---|---|---|---|
| Durées sélectionnées | Pourcentage de sélection | Nombre d'effets | Score de fréquence | Effets du modèle |
| 292 | 29.20 | 1 | 293.0 | Intercept |
| 173 | 17.30 | 6 | 173.6 | Intercept X1 X2 X3 X4 X5 |
| 54 | 5.40 | 3 | 54.77 | Intercept X1 X4 |
| 49 | 4.90 | 2 | 49.83 | Intercept X1 |
| 37 | 3.70 | 4 | 37.71 | Intercept X1 X4 X5 |

Figure 27: Probability to find variables jointly with Elasticnet with a AICC/BIC/SBC criteria

However, these results are one of the limits of our study. Indeed, in theory, we are not supposed to get good predictions with Elasticnet when we introduce extreme values. Yet, here, after introducing extreme values, the Elasticnet process keep predicting the right model with a good probability (around seventeen percent) in comparison with all the others probabilities. This will require a later development in another study to be solve.

## 3.4 Fourth DGP : the variables are correlated to each other and extreme values are inserted

In this subsection, we decide to introduce at the same time some correlation between the variables and extreme values in our model. To do so, we will simply combine both methods that we use in the two previous subsections. In the end, we don't really need to get in depth about this case. Indeed, the results are really similar to the case where we only introduce extreme values : the statistical learning processes are still not really convincing, and so does the machines learning processes. These results do not bring something new to our study.

# 4 Conclusion

After explaining the importance of tackling the question of automatic variable selection due to the emergence of Big Data, our paper starts by introducing the different techniques used in this study, from the selection criteria (e.g., AIC, AICC, BIC, PRESS, $R^2$) to the statistical and machine learning methods (e.g., LAR, LASSO, ElasticNet). It follows an empirical application through SAS, especially with the interactive matrix programming language IML. In this application, we were able to manipulate, simulate and distort data to operate four model selection processes, while introducing - starting from the second data generating process (DGP) - alterations to the fundamental OLS hypotheses. Following the extraction and the analysis of the results coming from each data generating process, we can highlight three main aftermaths:

1. **Machine learning processes are more efficient than statistical learning processes.**

   In the first data generating process, we wanted to replicate data in a "perfect setting", which is, in the absence of extreme values and correlation between the variables that were chosen in our model. To do so, we generated data from a multivariate normal distribution and created a model out of the generated data. Out of the fifty variables that were generated, only the first five variables were used in our model, and our goal was to see whether each statistical and machine learning methods are able to select the correct model with all five variables with a high probability, using "modelaverage". It appears that the statistical learning procedures, especially the Backward selection performs poorly by selecting almost every variables in our model, with a probability of 100%. The Forward selection performs better than the Backward selection, but it still selects more variables than it should, especially while looking at the joint probabilities, i.e., the probability that the software selects simultaneously different variables. Nonetheless, the LAR, LASSO and ElasticNet regressions seem to display more realistic results, with a joint probability of selecting the correct model with all the five variables that attains a high percentage (around 20-25% to choose the right model). However, this depends on the selection criterion used.

2. **The PRESS criterion works better while applied to machine learning processes.**

   For every data generating processes, the PRESS criterion as a selection criterion displays better results than other selection criterion such as the AIC, AICC, SBC or BIC for the LAR, LASSO and ElasticNet.

3. **Both types of algorithms perform poorly in the presence of hypotheses violation (e.g., introduction of correlation and extreme values.)**

   In the second data generating process, we used the Iman Conover transformation, as well as a Toeplitz matrix to inject correlations between the five first variables that are used in our model. In the third DGP, we injected extreme variables in our data so as to potentially challenge the predictive power of both statistical and machine learning methods. In the last DGP, we combined both effects (i.e., correlation and extreme values) as to distort the data and challenge even more the algorithms. We expect, for each three DGPs to lose in predictive power, i.e., either to detect poorly

the variables since they are correlated between each other, or to not detect at all the variables and to have poor joint probabilities to be fully detected. For each method, both machine and statistical learning methods poorly select all five variables (in a joint probability of approximately 1%). Only ElasticNet for the last two DGPs display all the five variables with a high joint probability, which is unexpected.

To conclude, the generation of a control group application allows us to compare a situation where all the variables are perfect, i.e., without any correlation or extreme variables to alter the model selection, to situations where we introduce distortion on the five main variables. The distortions created within the data generated almost destroys the ability of the algorithms to select the variables in our model. In practice, it is rare to obtain perfectly generated data, without any correlations or outliers, so it is important to train our data as a way to erase the biases brought by raw data. To do so, we can use cross-validation as a way to train our data, through the partition of the data sample into K parts of the same size and for each repetition, we remove one part of the sample in order to estimate the rest and compute the performance of the model without the removed part. Then, we add up all the performance estimation to get the final performance measure.

# 5 Bibliography


- Tibshirani, Robert. “Regression Shrinkage and Selection via the Lasso.” Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, [Royal Statistical Society, Wiley], 1996, pp. 267–88, http://www.jstor.org/stable/2346178
- Tibshirani, Robert. “Regression Shrinkage and Selection via the Lasso: A Retrospective.” Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 73, no. 3, [Royal Statistical Society, Wiley], 2011, pp. 273–82, http://www.jstor.org/stable/41262671.
- Derksen, S Keselman H.J. "Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables.", Volume 45, 1992, pp 265-282.
- Choueiry, G. "Understand Forward and Backward Stepwise Regression.", https://quantifyinghealth.com/stepwise-selection/
- Gaillac, C. L’Hour J. "Machine Learning for Econometrics", 2021
- Gruber, M. "Improving Efficiency by Shrinkage: The James - Stein and Ridge Regression", volume 156
- Larocque, Denis, “Analyse multidimensionnelle appliquée”, Version Automne 2019, Département de sciences de la décision, HEC Montréal.

# 6 Appendix

In this appendix, you can find the code used during our study.

```
proc iml;

sigma=i(50); *identité;
null=j(50,1,0);
mvn=randnormal(50,null,sigma); **pour faire tordre le modèle on change ça + changer la

/* DGP 1: hyptohèses de base vérifiées */
eps=normal(j(nrow(mvn),1,0))*0.1; **0 = seed;
Y=0.7*mvn[,1]+0.6*mvn[,2]+0.5*mvn[,3]+0.4*mvn[,4]+0.3*mvn[,5]+eps;
total=y||mvn;
nom='y'||('X1':'X50');

create total from total[colname=nom];
append from total;
close total;

submit;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=forward(select=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=forward(select=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=forward(select=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=forward(select=PRESS) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total plots=all;
model y=x1-x50 / selection=backward(select=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=backward(select=AIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=backward(select=BIC) ;
```

```
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=backward(select=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=backward(select=PRESS) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total plots=all;
model y=x1-x50 / selection=lasso(select=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=lasso(choose=AIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=lasso(choose=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=lasso(choose=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=lasso(LSCOEFFS choose=PRESS stop=SBC) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total plots=all;
model y=x1-x50 / selection=lar(choose=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=lar(choose=AIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=lar(choose=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=lar(choose=SBC) ;
modelaverage nsamples=1000;
run;
```

```
proc glmselect data=total plots=all;
model y=x1-x50 / selection=lar(LSCOEFFS choose=PRESS) ;
modelaverage nsamples=1000;
run;


proc glmselect data=total plots=all;
model y=x1-x50 / selection=elasticnet(choose=AIC l2=0.001) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=elasticnet(choose=BIC l2=0.001) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=elasticnet(choose=SBC l2=0.001) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=elasticnet(choose=AICC l2=0.001) ;
modelaverage nsamples=1000;
run;

endsubmit;

*******************************************************************************

proc iml;

sigma=i(50); *identité;
null=j(50,1,0);
mvn=randnormal(50,null,sigma); **pour faire tordre le modèle on change ça + changer la

/* DGP 2 : on ajoute de l'autocorrélation entre les variables explicatives */
T = toeplitz({1 0.9 0.8 0.75 0.85});
*print T;

start ImanConoverTransform(X, C);
   N = nrow(X);
   S = J(N, ncol(X));
   /* T1: Create normal scores of each column */
   do i = 1 to ncol(X);
      ranks = ranktie(X[,i], "mean");          /* tied ranks */
      S[,i] = quantile("Normal", ranks/(N+1)); /* van der Waerden scores */
   end;
   /* T2: apply two linear transformations to the scores */
   CS = corr(S);        /* correlation of scores */
   Q = root(CS);        /* Cholesky root of correlation of scores */
```

```
   P = root(C);          /* Cholesky root of target correlation */
   T = solve(Q,P);       /* same as  T = inv(Q) * P; */
   Y = S*T;              /* transform scores: Y has rank corr close to target C */

   /* T3: Permute or reorder data in the columns of X to have the same ranks as Y */
   W = X;
   do i = 1 to ncol(Y);
      rank = rank(Y[,i]);          /* use ranks as subscripts, so no tied ranks */
      tmp = W[,i]; call sort(tmp); /* sort column by ranks */
      W[,i] = tmp[rank];           /* reorder the column of X by the ranks of M */
   end;
   return( W );
finish;
store module=(ImanConoverTransform);

W = ImanConoverTransform(mvn[,1:5], T);

mvn2 = W||mvn[,6:50];
eps2=normal(j(nrow(mvn2),1,0))*0.1;
Y=0.7*mvn2[,1]+0.6*mvn2[,2]+0.5*mvn2[,3]+0.4*mvn2[,4]+0.3*mvn2[,5]+eps2;

total2=y||mvn2;
nom='y'||('X1':'X50');

create total2 from total2[colname=nom];
append from total2;
close total2;

submit;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=forward(select=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=forward(select=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=forward(select=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=forward(select=PRESS) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=backward(select=AICC) ;
```

```
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=backward(select=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=backward(select=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=backward(select=PRESS) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=lasso(choose=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=lasso(choose=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=lasso(choose=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=lasso(LSCOEFFS choose=PRESS) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=lar(choose=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=lar(choose=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=lar(choose=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=lar(LSCOEFFS choose=PRESS) ;
modelaverage nsamples=1000;
run;
```

```
proc glmselect data=total2 plots=all;
model y=x1-x50 / selection=elasticnet(choose=AIC l2=0.001) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=elasticnet(choose=BIC l2=0.001) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=elasticnet(choose=AICC l2=0.001) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total plots=all;
model y=x1-x50 / selection=elasticnet(choose=SBC l2=0.001) ;
modelaverage nsamples=1000;
run;
endsubmit;
*******************************************************************************
proc iml;

sigma=i(50);
null=j(50,1,0);
mvn=randnormal(50,null,sigma);

/* DGP 3: valeurs extrêmes */
mu= {-6, 3, 8, -4, 5};
sigma={0.5,  0.6, 2.5, 1.2, 1.9};
do t=1 to 50;
call randgen(u, "Uniform", 1, 6);
i=floor(u);
x1=x1//normal(0)*sigma[i]+mu[i];
x2=x2//normal(0)*sigma[i]+mu[i];
x3=x3//normal(0)*sigma[i]+mu[i];
x4=x4//normal(0)*sigma[i]+mu[i];
x5=x5//normal(0)*sigma[i]+mu[i];
x = x1||x2||x3||x4||x5;
end;

mvn3 = x||mvn[,6:50];
eps3=normal(j(nrow(mvn3),1,0))*0.1;
Y=0.7*mvn3[,1]+0.6*mvn3[,2]+0.5*mvn3[,3]+0.4*mvn3[,4]+0.3*mvn3[,5]+eps3;

total3=y||mvn3;
nom='y'||('X1':'X50');

create total3 from total3[colname=nom];
append from total3;
close total3;
```

```
submit;

proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=forward(select=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=forward(select=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=forward(select=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=forward(select=PRESS) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=backward(select=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=backward(select=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=backward(select=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=backward(select=PRESS) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=lasso(choose=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=lasso(choose=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=lasso(choose=SBC) ;
modelaverage nsamples=1000;
```

```
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=lasso(LSCOEFFS choose=PRESS) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=lar(choose=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=lar(choose=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=lar(choose=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total3 plots=all;
model y=x1-x50 / selection=LAR(LSCOEFFS choose=PRESS) ;
modelaverage nsamples=1000;
run;
endsubmit;

*************************************************************************************
proc iml;

sigma=i(50);
null=j(50,1,0);
mvn=randnormal(50,null,sigma);

/* DGP 4: corrélation + valeurs extrêmes */
mu= {-6, 3, 8, -4, 5};
sigma={0.5,  0.6, 2.5, 1.2, 1.9};
do t=1 to 50;
call randgen(u, "Uniform", 1, 6);
i=floor(u);
x_1=x_1//normal(0)*sigma[i]+mu[i];
x_2=x_2//normal(0)*sigma[i]+mu[i];
x_3=x_3//normal(0)*sigma[i]+mu[i];
x_4=x_4//normal(0)*sigma[i]+mu[i];
x_5=x_5//normal(0)*sigma[i]+mu[i];
x = x_1||x_2||x_3||x_4||x_5;
end;

mvn3 = x||mvn[,6:50];

T = toeplitz({1 0.75 0.5 0.45 0.2});
```

```
start ImanConoverTransform(X, C);
   N = nrow(X);
   S = J(N, ncol(X));
   /* T1: Create normal scores of each column */
   do i = 1 to ncol(X);
      ranks = ranktie(X[,i], "mean");          /* tied ranks */
      S[,i] = quantile("Normal", ranks/(N+1)); /* van der Waerden scores */
   end;
   /* T2: apply two linear transformations to the scores */
   CS = corr(S);        /* correlation of scores */
   Q = root(CS);        /* Cholesky root of correlation of scores */
   P = root(C);         /* Cholesky root of target correlation */
   T = solve(Q,P);      /* same as  T = inv(Q) * P; */
   Y = S*T;             /* transform scores: Y has rank corr close to target C */

   /* T3: Permute or reorder data in the columns of X to have the same ranks as Y */
   W = X;
   do i = 1 to ncol(Y);
      rank = rank(Y[,i]);          /* use ranks as subscripts, so no tied ranks */
      tmp = W[,i]; call sort(tmp); /* sort column by ranks */
      W[,i] = tmp[rank];           /* reorder the column of X by the ranks of M */
   end;
   return( W );
finish;
store module=(ImanConoverTransform);

W = ImanConoverTransform(mvn3[,1:5], T);

mvn4 = W||mvn3[,6:50];
eps4=normal(j(nrow(mvn4),1,0))*0.1;
Y=0.7*mvn4[,1]+0.6*mvn4[,2]+0.5*mvn4[,3]+0.4*mvn4[,4]+0.3*mvn4[,5]+eps4;

total4=y||mvn4;
nom='y'||('X1':'X50');

create total4 from total4[colname=nom];
append from total4;
close total4;

submit;

proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=forward(select=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=forward(select=BIC) ;
```

```
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=forward(select=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=forward(select=PRESS) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=backward(select=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=backward(select=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=backward(select=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=backward(select=PRESS) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=lasso(choose=AICC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=lasso(choose=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=lasso(choose=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=lasso(LSCOEFFS choose=PRESS) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=lar(choose=AICC) ;
modelaverage nsamples=1000;
```

```
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=lar(choose=BIC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=lar(choose=SBC) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=lar(choose=AIC) ;
modelaverage nsamples=1000;
run;

proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=elasticnet(choose=AICC l2=0.001) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=elasticnet(choose=BIC l2=0.001) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=elasticnet(choose=SBC l2=0.001) ;
modelaverage nsamples=1000;
run;
proc glmselect data=total4 plots=all;
model y=x1-x50 / selection=elasticnet(choose=AIC l2=0.001) ;
modelaverage nsamples=1000;
run;
endsubmit;
```