X Congreso Galego de Estatística e Investigación de Operacións
Pontevedra, 3–4–5 de novembro de 2011

A forward stepwise algorithm to select variables in regression context

M. Sestelo, N. M. Villanueva and J. Roca-Pardiñas
Department of Statistics and O. R, University of Vigo
sestelo@uvigo.es, nmvillanueva@uvigo.es, roca@uvigo.es

ABSTRACT

In multiple regression models with p variables is important to determine the best subset or subsets of q ($q \leq p$) predictors which will establish the model or models with the best discrimination capacity. This problem is particularly important where p is high and/or where there are mutually redundant predictors. With this study, we present a new approach to this problem, where we will try to predict a new emission episode of SO_2, but focusing our attention in the importance to know the best combinations of time instants to obtain the best prediction. The proposed method is a new forward stepwise-based selection procedure that selects a model containing a subset of variables (or time instants) according to an optimal criteria (determination coefficient) and taking into account the computational cost. Additionally, bootstrap resampling techniques are used to implement tests capable of detecting whether significant effects of the unselected variables are present in the model.

Keywords: variable selection, forward stepwise, R^2, bootstrap

1. INTRODUCTION

The combustion of fuel oil or coal release to the atmosphere sulphur dioxide in different quantities. An emission episode is said to occur when the series of bi-hourly means of SO_2 is greater than a defined level. Therefore, both economical as environmentally, it will be interest to be able to predict, with sufficient time for effective countermeasures to be taken, when the legal limit will be exceeded.

In this work, we will try to predict a new emission episode, but focusing our attention in the importance to know the best combinations of time instants to obtain the best prediction. For this reason, the selection of the optimal subset of variables can be a good approach to this subject. The proposed procedure selects a model containing a subset of variables (or time instants), according to an optimal criteria (determination coefficient). This procedure is based on (Roca-Pardiñas et al., 2009), where selection of variables is based on searching through all the possible subsets. If the number of variables p is high, this selection procedure may have a computacional cost that is too high (for instance, if $p = 20$, the number of estimated models has reached 1048575), and the problem becomes intractable. With this paper we propose an adaptation of the previous method, a new forward stepwise-based selection procedure that greatly decreases the computacional cost.

Once we have selected the best subset of variables and eliminated the remainder from the model, it will be interest to know the number of covariates to be included. Hence, bootstrap resampling techniques are used to implemented tests capable of detecting whether significant effects of the unselected variables are present in the model.

Accordingly, the goal of this work was to provide a new methodology for detecting optimal combinations of variables in the regression context. For this purpose, a computational algorithm was developed and implement, based on forward selection methods and bootstrap techniques.

2. BOOTSTRAP-BASED TESTS FOR COVARIATE SELECTION

In this section, we propose a procedure that will help to determine the number of variables that have to be included into the model. Given a set of p initial variables, $X_1,\ X_2,\ \ldots,\ X_p$, for a given size k, being $\sigma^2(k)$ the residual variance obtained with the best subset of k variables in the sense that

$$\sigma^2\left(k\right) = \min_{1 \leq j_1 < j_2 < \ldots < j_k \leq p} \sigma^2_{j_1,\ldots,j_k}$$

being $\sigma^2_{j_1,\ldots,j_k}$ the σ^2 obtained by the model

$$Y = \beta_0 + \beta_{j_1} X_{j_1} + \beta_{j_2} X_{j_2}, \ldots, \beta_{j_k} X_{j_k} + \varepsilon \tag{1}$$

Given a subset of size q, considerations will be given to a test for the null hypothesis

$$H_0\left(q\right) : \sigma^2\left(q\right) = \min_{r>q} \sigma^2\left(r\right)$$

vs. the general hypothesis

$$H_1 : \sigma^2\left(q\right) > \min_{r>q} \sigma^2\left(r\right)$$

To test H_0 we propose the use of the test statistic

$$\widehat{T} = \frac{\hat{\sigma}^2\left(q\right)}{\min\limits_{r>q} \hat{\sigma}^2\left(r\right)}$$

with

$$\hat{\sigma}^2\left(k\right) = \min_{1 \leq j_1 < j_2 < \ldots < j_k \leq p} \hat{\sigma}^2_{j_1,\ldots,j_k}$$

where $\hat{\sigma}^2_{j_1,\ldots,j_k}$ is the $\hat{\sigma}^2$ obtained from the model in (1).

Applying this test to $q = 1, \ldots, p-1$ may be an important step in a covariate selection procedure. If H_0 is not rejected, only the subset of the covariables $X_{j_1}, \ldots, X_{j_q}$ will be retained, and the remainder variables will be eliminated from the model. In all other cases, the test will be repeated with $q+1$ variables until the null hypothesis is not rejected.

The bootstrap methods (Efron, 1979; Härdle and Mammen, 1993; Kauermann and Opsomer, 2003) was used to approximate the critical values T^{α}. Given the sample $\{\mathbf{X}_i, Y_i\}_{i=1}^{n}$ of $(\mathbf{X}, Y)$, the steps are as follows:

Step 1: Obtain $\widehat{T}$ from the sample size as we explained above.

Step 2: For $b = 1$ to B (e.g. B=1000), simulate the bootstrap sample $\left\{\left(\mathbf{X}_i, Y_i^{*b}\right)\right\}_{i=1}^{n}$ by randomly sampling the n items from the original data set with replacement (that is, each individual value (X_i, Y_i) has a probability $n^{-}1$ of occurring), and obtain the bootstrap estimates $\hat{T}^{*b}$.

The test rule based on T is given by rejecting the null hypothesis if $\hat{T}^{\alpha} < 1$, where T^{p} is the empirical $p-$percentile of values T^{*b} $(b = 1, \ldots, B)$ obtained before.

3. APPLICATION EXAMPLE

The methodology development in this paper was tested on the prediction of atmospheric SO_2 pollution incidents discussed previously. One of the problems that arises is to decide which temporal instants $(X_{t-1}, X_{t-2}, \ldots, X_{t-n})$ are relevant for prediction purpose, since inclusion of all the times X_l may well degrade the overall performance of the prediction model.

Figure 1 shows the R^2 values for each subset size q. Also, we represent a black line that joints the models with the best values of R^2. These models appear in the Table 1. As we can see, the R^2 increases as the number of variables included in the model rises up to a given q. Furthermore, it should be noted that, as q increases, the R^2 tends to stabilize.

It is possible to deduce that there is an optimal intermediate point between the number of variables that enter to form part of the model (preferably low) and the R^2 value (preferably high). To delimit this point, the above described test for the null hypothesis $H_0(q)$ was applied for each q (Table 1). The deduction to be drawn from the results obtained is that, for a 5 percent significance level, the null hypothesis is rejected with $q = 1$ and accepted thereafter. From this results, we can conclude that the best temporal instants for prediction of an emission episode would be X_{18} and X_{16}.

Table 1: R^2 obtained with each selected model of size q

q	R^2	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
1	0.87002																		x
2	0.93316																x		x
3	0.93452					x												x	x
4	0.93550		x	x														x	x
5	0.93590	x			x		x											x	x
6	0.93660	x			x		x										x	x	x
7	0.93686	x			x		x	x									x	x	x
8	0.93734	x			x		x			x	x			x				x	x
9	0.93758	x			x		x			x	x		x				x	x	x
10	0.93788	x			x		x	x		x	x		x				x	x	x
11	0.93803	x			x		x	x	x	x	x		x				x	x	x
12	0.93819	x			x		x	x	x	x	x			x		x	x	x	x
13	0.93821	x		x	x		x	x	x	x	x			x		x	x	x	x
14	0.93823	x	x	x	x		x	x	x	x	x			x		x	x	x	x
15	0.93824	x	x	x	x		x	x	x	x	x	x		x		x	x	x	x
16	0.93825	x	x	x	x	x	x	x	x	x	x	x		x		x	x	x	x
17	0.93825	x	x	x	x		x	x	x	x	x	x		x	x	x	x	x	x
18	0.93825	x	x	x	x		x	x	x	x	x	x	x	x		x	x	x	x

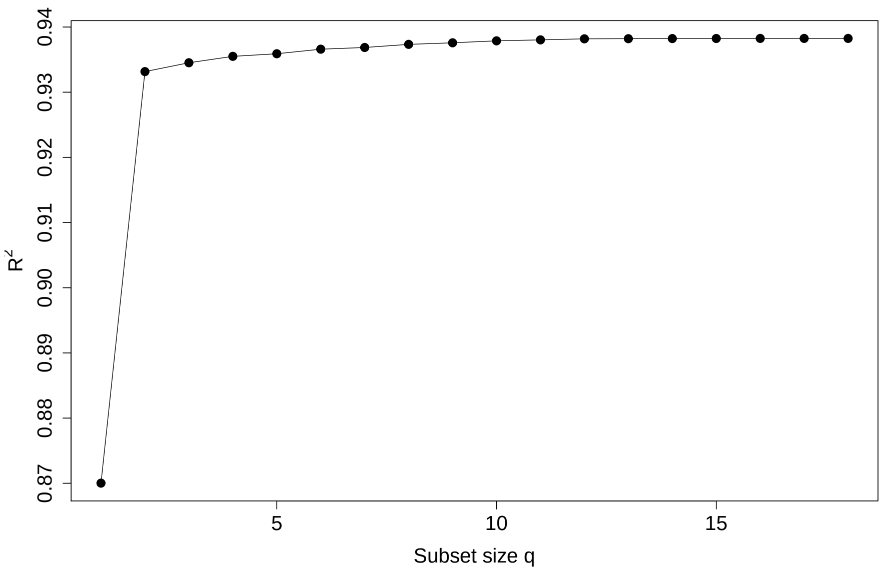

Figure 1: For each subset size, R^2 obtained for the best model.

4. CONCLUSION

In this paper, we present a new procedure of selecting variables in regression models based on bootstrap techniques that determines the optimal number and identifies them according to an optimal criteria.

REFERENCES

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26.

Härdle, W. and Mammen, E. (1993). Testing parametric versus nonparametric regression. *Annals of Statistics*, 21:1926–1947.

Kauermann, G. and Opsomer, J. (2003). Local Likelihood Estimation in Generalized Additive Models. *Scandinavian Journal of Statistics*, 30:317–337.

Roca-Pardiñas, J., Cadarso-Suárez, C., Tahoces, P. G., and Lado, M. J. (2009). Selecting variables in non-parametric regression models for binary response. an application to the computerized detection of breast cancer. *Statistics in Medicine*, 28(2):240–259.