Variable selection based on Allen's PRESS-Statistic

Schütz, Thiel

6 7 2021

Introduction

Within this report variable selection based on **Allen's PRESS-statistic** is presented. Variable selection itself is a popular topic now more than ever. Due to increasing machine resources such variable selection algorithms, which are mostly already existing for a long time, are now possible to execute even on larger scale. The desire to find relations between variables based on algorithms has been present for some time too. It comes from the idea that selection based on algorithms is more objective and therefore in some way better. Another reason may be that one is getting the feeling that no connections were overlooked with such algorithms. It has already been found that variable selection based on algorithms is not automatically the best solution, nevertheless it can help to find a *good* set of variables to describe data.

Although the basic procedures like **backward** or **forward elimination** exist for some time, there is still research done to improve or investigate these algorithms. Another very common approach is the **best subset selection** which is often based on **AIC, BIC, R-squared**, … Especially for this way of selection increasing machine resources are important. The procedure is as follows:

1. Define …

- the criteria on which the selection process is based on.
- the set of variables selection should be done on.
- the range of subset sizes you want to find best subsets on (optional).

2. The algorithm finds the *best* subset of variables for each subset size according to the defined criteria.

The selection process shows why the computational effort should not be underestimated. For a working set of 10 variables, 2^{10} models have to be calculated in order to find the *best* according to a given criteria.

Before referring to **Allen's PRESS-statistic** we want to give a short overview of other common, already named, criteria:

- **AIC** (Akaike's information criteria): $AIC = -2l(\hat{\theta}_{ML}) + 2k$ with k denoting the number of explanatory variables in the model and $l(\hat{\theta}_{ML})$ describing the log-likelihood of the vector θ which includes k coefficients that result from the maximum-likelihood-estimation. In the case of linear models the AIC can be calculated directly as $AIC = n \cdot ln(\sigma^2) + 2k$. [1]
- **BIC** (Bayesian information criteria): $BIC = -2l(\hat{\theta}_{ML}) + k \cdot ln(n)$. AIC and BIC differ only regarding the penalty term. While AIC always adds $2k$, BIC considers sample size n. It can be concluded that for $n > 7$ ($ln(8) > 2$), BIC is larger than AIC. [1]
- **evtl. noch weiteres Maß**

The idea behind these measures is to reward a good model fit and punish for model complexity. The difference to the **PRESS-statistc** is, that a good model fit is assessed different. Instead of maximum-likelihood-estimation, the goodness of prediction for every point in the data set is assessed in the following way:

$$PRESS = \sum_{i=1}^{n} (Y - \hat{Y}_i)^2$$

with $\hat{Y}_i$ describing $E(Y_i)$ excluding the i-th observation. In other words: For every point i in the data set, the goodness of prediction of a model which is based on the remaining $n - 1$ data points is calculated (as squared difference between the actual and predicted value). The sum of these predicted errors is then describing the PRESS-statistic which stands for **PRE**diction **S**um of **S**quares.

Sources

1. Sachs, L., & Hedderich, J. (2006). Angewandte Statistik: Methodensammlung mit R. Springer-Verlag.
2.