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.

fdsf