Computational Statistics and Data Analysis 116 (2017) 19–31

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Detection of influential points as a byproduct of resampling-based variable selection procedures

Riccardo De Bin [a,b,*], Anne-Laure Boulesteix [b], Willi Sauerbrei [c]

[a] *Department of Mathematics, University of Oslo, Postboks 1053 Blindern, 0316, Oslo, Norway*
[b] *Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Germany*
[c] *Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany*

ARTICLE INFO

Article history:
Received 17 November 2016
Received in revised form 3 July 2017
Accepted 4 July 2017
Available online 20 July 2017

Keywords:
Bootstrap
Grubbs' test
Inclusion frequency
Model averaging
Outliers
Subsampling

ABSTRACT

Influential points can cause severe problems when deriving a multivariable regression model. A novel approach to check for such points is proposed, based on the variable inclusion matrix, a simple way to summarize results from resampling-based variable selection procedures. The variable inclusion matrix reports whether a variable (column) is included in a regression model fitted on a pseudo-sample (row) generated from the original data (e.g., bootstrap sample or subsample). It is used to study the variable selection stability, to derive weights for model averaged predictors and in others investigations. Concentrating on variable selection, it also allows understanding whether the presence of a specific observation has an influence on the selection of a variable. From the variable inclusion matrix, indeed, the inclusion frequency (I-frequency) of each variable can be computed only in the pseudo-samples (i.e., rows) which contain the specific observation. When the procedure is repeated for each observation, it is possible to check for influential points through the distribution of the I-frequencies, visualized in a boxplot, or through a Grubbs' test. Outlying values in the former case and significant results in the latter point to observations having an influence on the selection of a specific variable and therefore on the finally selected model. This novel approach is illustrated in two real data examples.

1. Introduction

In the construction of a statistical model, an important aspect to take into consideration is its stability. It is well known, indeed, that small perturbations in the data may lead to the selection of different models. For example, several papers show that variable selection procedures, such as backward elimination or forward selection, may provide very different sets of relevant variables, and consequently very different models, when applied to different bootstrap samples generated from the same dataset (Sauerbrei et al., 2015).

In the literature, different approaches have been proposed to handle this issue. From a variable point of view, resampling-based variable selection techniques can handle the instability issue by investigating the inclusion frequencies of the single variables (Gong, 1982; Chen and George, 1985). The idea is rather simple. Several pseudo-samples are generated via a resampling technique and a variable selection procedure is applied to select the best model in each of them. The proportion of models which contain the specific variable (inclusion frequency) is used as an indicator of the importance of the variable itself, and those variables with higher inclusion frequencies are used in the final model.

[*] Corresponding author at: Department of Mathematics, University of Oslo, Postboks 1053 Blindern, 0316, Oslo, Norway.
E-mail addresses: debin@math.uio.no (R. De Bin), boulesteix@ibe.med.uni-muenchen.de (A.-L. Boulesteix), wfs@imbi.uni-freiburg.de (W. Sauerbrei).

http://dx.doi.org/10.1016/j.csda.2017.07.001
0167-9473/© 2017 Elsevier B.V. All rights reserved.

From a model point of view, model averaging is a technique which aims to deal with model uncertainty by fitting different models on the data and then summarizing their results. For example, in linear regression, a regression coefficient is estimated as a weighted mean of the corresponding estimates computed in each model. In particular, in the resampling-based approaches the weights are obtained by generating several pseudo-samples via a resampling technique and evaluating for how many of these pseudo-samples the different models are selected by a variable selection procedure. Other kinds of weights are based on information criteria, Mallows' criterion, etc. For a review on model averaging and on the different alternatives for the computation of the weights, we refer the reader to Wang et al. (2009). That paper, in particular, considers the frequentist approach. For a review about Bayesian model averaging, a classical reference is Hoeting et al. (1999).

Both resampling-based variable selection and resampling-based weights for model averaging require the application of a variable selection technique to several pseudo-samples. The goal of this paper is to show that the information collected in this part of the analysis can be used to check for influential points, such as outliers or single observations that have a high impact on the results. It is well known that influential points can cause problems when selecting a statistical model. For example, the inclusion or exclusion of a single or a few observations can have a dramatic effect on variables selected and on the issue of selecting linear or nonlinear function for a continuous variable (Royston and Sauerbrei, 2007). The literature on influential point detection is vast, and countless approaches have been proposed. For a simple and concise overview we refer the reader to Su and Tsai (2011) and references therein.

The detection of influential points as a byproduct of model-building procedures is not new. Tsao and Ling (2012), for example, exclude from the final model fitting procedure those observations that are not included in any of the pseudo-samples that lead to good models in terms of goodness-of-fit. A similar approach is used by Sauerbrei et al. (2015), who consider the selection probabilities of some "best models" and identify as influential points those observations which are able to modify these selection probabilities. Both approaches handle the influential point detection issue from a model point of view, ignoring the effect of these observations on the single variables. In this paper we consider the problem from a variable point of view, though maintaining a multivariable approach.

Finally, we mention Atkinson and Riani (2002), who also studied the effect of influential points from a model building point of view, using a forward search procedure (Atkinson and Riani, 2000 Ch. 2). We contrast our and their approaches in Section 4.1.5.

The paper is structured as follows. Section 2 presents two datasets later used as real examples. A brief introduction to model averaging and resampling-based variable selection is presented in Section 3, together with the description of our approach. The application of the method to the data is reported in Section 4. Finally, Section 5 contains a short discussion.

2. Data

2.1. Body fat data

The estimate of the percentage of body fat is considered a good indicator to assess the health of patients (see, e.g., Myint et al., 2014). Johnson (1996) presents a dataset in which the percentage of body fat (PBF) is collected from 252 men, together with the information about 13 further quantities, namely *age*, *weight*, *height* and 10 continuous body circumference measurements that are considered variables with potential influence on PBF. The data are publicly available at http://portal.uni-freiburg.de/imbi/Royston-Sauerbrei-book/Multivariable_Model-building/downloads/datasets/edu_bodyfat_both.zip.

It is important to note that this dataset contains at least one influential point. Royston and Sauerbrei (2007), in particular, show that observation 39 highly influences the choice of the fractional polynomial function used to model the relationship between outcome and variables. Although some variables seem to have a non-linear effects on the outcome, we re-analyze this dataset under the assumption of linear effects. Non-linear effects are not that strong and this simplifying assumption seems acceptable for the main purpose of this paper.

To show the effect of observation 39 in a classical model-building procedure, we report in Table 1 the models obtained with backward elimination when this observation is included/excluded from the sample. Three common inclusion criteria are used. In this example, results are identical for BIC and $\alpha = 0.05$. As commonly seen in the literature (see, e.g., Sauerbrei et al., 2015), more variables are selected with AIC. We note that the presence/absence of observation 39 in the sample leads to substantially different models. The selections of variables *age*, *weight*, *height* and *forearm* are clearly affected.

2.2. Myeloma data

As an application of our method to a different kind of outcome, we also use a dataset with a time-to-event response. In particular, we consider a study on patients with multiple myeloma presented by Krall et al. (1975), in which the outcome is the survival time of the patients. The 16 variables are either binary or continuous. We consider the proportional hazard assumption acceptable, being this dataset analyzed several times in the literature by using the Cox model (see, e.g., Kuk, 1984; Chen and Wang, 1991). The sample size is small, consisting of 65 patients with 48 events. As for the body fat data, we use the simplifying assumption that the effect of continuous variables is linear. This dataset is also publicly available on the same website (http://.../myeloma.zip).