Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2008) 2778–2793
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

On properties of predictors derived with a two-step bootstrap model averaging approach—A simulation study in the linear regression model

Anika Buchholz, Norbert Holländer[1], Willi Sauerbrei*

Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meier-Strasse 26, 79104 Freiburg, Germany

Received 12 July 2006; received in revised form 12 October 2007; accepted 12 October 2007
Available online 17 October 2007

Abstract

In many applications of model selection there is a large number of explanatory variables and thus a large set of candidate models. Selecting one single model for further inference ignores model selection uncertainty. Often several models fit the data equally well. However, these models may differ in terms of the variables included and might lead to different predictions. To account for model selection uncertainty, model averaging procedures have been proposed. Recently, an extended two-step bootstrap model averaging approach has been proposed. The first step of this approach is a screening step. It aims to eliminate variables with negligible effect on the outcome. In the second step the remaining variables are considered in bootstrap model averaging. A large simulation study is performed to compare the MSE and coverage rate of models derived with bootstrap model averaging, the full model, backward elimination using Akaike and Bayes information criterion and the model with the highest selection probability in bootstrap samples. In a data example, these approaches are also compared with Bayesian model averaging. Finally, some recommendations for the development of predictive models are given.

Keywords: Bootstrap; Model averaging; Model selection uncertainty; Linear regression; Variable screening

1. Introduction

The identification of factors and models in predicting an outcome is of major interest in many areas of application. Often a large number of potential explanatory variables are collected, leading to a large set of candidate models, from which usually one single model is chosen for prediction. Here we consider in- or exclusion of candidate variables in a linear regression model as the only model building task. The model space would be extensively enlarged if further issues, such as determination of a functional form for a continuous variable or another type of regression model, would be considered. When a model is constructed from 15 variables, $2^{15} = 32\,768$ model combinations are possible. Ultimately, only one model will be selected. It is well known that often several models fit the data equally well, but may differ substantially in terms of the variables included and might lead to different predictions for individual observations

* Corresponding author. Tel.: +49 761 2036669; fax: +49 761 2035002.
E-mail address: wfs@imbi.uni-freiburg.de (W. Sauerbrei).
[1] New affiliation: Novartis Pharma AG, 4057 Basel, Switzerland.

0167-9473/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.csda.2007.10.007

(Miller, 2002). Efficient algorithms are available (Hofmann et al., 2007), but ignoring model selection uncertainty may lead to biased parameter estimates and underestimation of variance (Draper, 1995; Chatfield, 1995).

To account for model selection uncertainty, model averaging (MA) procedures have been proposed. The MA estimate is obtained as a weighted average of a set of estimated predictors obtained under different models. Advantages of MA are stressed in many papers, but usually the evidence is restricted to case studies (Hoeting et al., 1999; Volinsky et al., 1997; Augustin et al., 2005) or some analytical results for restricted problems, such as a very small set of candidate models, independence between predictors, assuming a local asymptotic framework (Buckland et al., 1997; Candolo et al., 2003; Yuan and Yang, 2005; Hjort and Claeskens, 2003).

Over the past few years, a lot of work has been done in a Bayesian model averaging (BMA) framework (Hoeting et al., 1999; Raftery et al., 1997). An alternative to BMA is bootstrap model averaging (bootstrapMA), first proposed by Buckland et al. (1997) but modified by Augustin et al. (2005) to include a variable screening step prior to bootstrap model averaging in order to identify and eliminate variables with no or a negligible effect on the outcome. This results in a much smaller class of candidate models for the MA step. For problems with a larger number of variables (say more than 10) the importance of a screening step seems to be well accepted. With BMA, Occam's window is usually used (Hoeting et al., 1999). Burnham and Anderson (2002) argue for a selection of models based on subject matter knowledge and Yuan and Yang (2005) propose in their ARMS algorithm to keep the top m models (in examples they use $m = 40$) based on Akaike information criterion (AIC) or Bayes information criterion (BIC) in one part of the data.

In contrast to other screening approaches which eliminate models not strongly supported by the data, we eliminate variables not strongly supported by the data. The main reason is the increase of potential future use of our models, which means that it is not required to collect all variables in a new data set. A simulation study showed that our screening step reduces the number of variables and correspondingly the number of candidate models, without eliminating models strongly supported by the data (Sauerbrei et al., 2006). The first promising results of bootstrapMA could be shown in a small simulation study (Holländer et al., 2006). Here we will present details of the design and the simulation study will be substantially extended. Using MSE and coverage rate as criteria, we will compare the predictive performance of models derived with bootstrapMA to the full model, backward elimination (BE) using AIC and BIC and the model with the highest selection probability in bootstrap samples.

In a study on school children, the aim is to predict forced vital capacity (FVC) from 24 variables. In this example we will compare results from bootstrapMA with the others and also with BMA (Hoeting et al., 1999). In all methods we restrict ourselves to only fitting linear terms in the models and in- or exclusion of variables only.

We will introduce the model building approaches as well as the assessment of predictive ability in Section 2. In Section 3, we describe the design of our simulation study and present the results in Section 4. Section 5 gives an example for further illustration and comparisons. In Section 6, we will discuss the results and give some recommendations for the development of predictive models.

2. Methods

The linear regression model is defined by

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_P X_P + \varepsilon = \boldsymbol{X\beta} + \varepsilon, \tag{1}$$

where $Y = (Y_1, \ldots, Y_N)'$ denotes the response vector with N being the number of observations, $X_1, \ldots, X_P$ are the vectors of the explanatory variables and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)'$ is the error vector. The error terms ε_i are assumed to be uncorrelated and normally distributed, i.e. $\varepsilon \sim \mathrm{N}(0, \sigma_\varepsilon^2 I)$. Thus, it follows that $(Y \mid \boldsymbol{X})$ is normally distributed with mean $\mu = \boldsymbol{X\beta}$ and covariance matrix $\sigma_\varepsilon^2 I$. In this paper, we focus on the prediction of the outcome and its variance.

In many applications, the number of variables, P, that are considered as potential influence factors is large, but in a multivariable context, only a few of them have an effect on the outcome Y. A common approach to data analysis is variable selection to determine a 'best' model and to make inference as if the selected model was prespecified (Burnham and Anderson, 2002; Miller, 2002). In this paper, we apply BE using AIC and BIC. These BE procedures and the corresponding results are labelled by BE_{AIC} and BE_{BIC}, respectively. With BE_{AIC} we usually select larger models, whereas BE_{BIC} puts more penalty on variables, resulting in sparser models (for more details see Burnham and Anderson, 2004; Sauerbrei et al., 2006). We also consider the full model containing all P variables. In our simulation study, we use the true model as reference.