Computational Statistics and Data Analysis 71 (2014) 758–770

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Model selection and model averaging after multiple imputation

Michael Schomaker[a,*], Christian Heumann[b]

[a] University of Cape Town, Centre for Infectious Disease Epidemiology & Research, Anzio Road, Cape Town, 7925, South Africa

[b] Ludwig Maximilians University Munich, Department of Statistics, Akademiestr. 1, 80799 München, Germany

ARTICLE INFO

Article history:
Received 29 June 2012
Received in revised form 11 February 2013
Accepted 13 February 2013
Available online 26 February 2013

Keywords:
Akaike's information criterion
Bootstrap
Frequentist model averaging
Linear regression
Missing data
Survival analysis

ABSTRACT

Model selection and model averaging are two important techniques to obtain practical and useful models in applied research. However, it is now well-known that many complex issues arise, especially in the context of model selection, when the stochastic nature of the selection process is ignored and estimates, standard errors, and confidence intervals are calculated as if the selected model was known *a priori*. While model averaging aims to incorporate the uncertainty associated with the model selection process by combining estimates over a set of models, there is still some debate over appropriate interpretation and confidence interval construction. These problems become even more complex in the presence of missing data and it is currently not entirely clear how to proceed. To deal with such situations, a framework for model selection and model averaging in the context of missing data is proposed. The focus lies on multiple imputation as a strategy to deal with the missingness: a consequent combination with model averaging aims to incorporate both the uncertainty associated with the model selection and with the imputation process. Furthermore, the performance of bootstrapping as a flexible extension to our framework is evaluated. Monte Carlo simulations are used to reveal the nature of the proposed estimators in the context of the linear regression model. The practical implications of our approach are illustrated by means of a recent survival study on sputum culture conversion in pulmonary tuberculosis.

1. Introduction

Data-driven model selection is an essential part of many statistical analyses. During the last four decades an impressive range of techniques and criteria have been developed to choose a 'best' model among a set of plausible models: Among these, Akaike's Information Criterion (AIC, Akaike (1973)) and cross validation (Stone, 1974) are popular choices, especially in the context of variable selection in regression models. There are, however, numerous alternatives which are often fine-tuned for a specific purpose or model, see Rao and Wu (2001) for a comprehensive overview.

It is common practice that statistical inference is performed conditional on the selected model and all subsequent estimates are based on the assumption that the model was chosen *a priori*. This may be problematic in many situations as, in addition to the stochastic nature of the model, the model selection process is stochastic itself and naive post model selection estimators may underestimate variability, yield therefore overconfident inference and may be unstable (Chatfield, 1995; Leeb and Pötscher, 2005; Wang et al., 2009). It is often argued that model averaging can overcome this problem by combining

* Corresponding author. Tel.: +27 21 404 7737.
E-mail address: michael.schomaker@uct.ac.za (M. Schomaker).

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.02.017

estimates of many potentially good models. Model averaging designs a weighted average across a set of candidate models to obtain a robust estimator and incorporates the uncertainty associated with the model selection process into standard errors and confidence intervals. These estimators are often called 'unconditional' in the literature since inference does not rely on a single selected model (Leeb and Pötscher, 2008), but they are still conditional on the set of candidate models under consideration. The weights would be typically constructed such that the final model averaging estimator is optimal with respect to minimizing a Mallows criterion, the trace of the estimator's MSE, or other meaningful criteria (Hansen, 2007; Wan et al., 2010; Liang et al., 2011; Schomaker, 2012; Hansen and Racine, 2012); or, more often, such that 'better' models receive a higher weight whereby the quality of a model is judged upon model selection criteria such as the AIC or the FIC (Buckland et al., 1997; Hjort and Claeskens, 2003; Claeskens and Hjort, 2003; Hjort and Claeskens, 2006; Schomaker and Heumann, 2011; Zhang et al., 2012; Wang et al., 2012). Another popular weight choice would relate to approximations of the posterior probability of a model being correct, see Hoeting et al. (1999) for an overview of Bayesian model averaging; we will, however, emphasize the frequentist perspective of model averaging in this article.

Apart from the discussion on how to appropriately select or average model estimates, data analyses often suffer from incomplete data. Nowadays, a broad range of methods, including multiple imputation and weighted estimating equations, can be employed when the missingness mechanism is ignorable, i.e. if the probability that a response is missing at any occasion depends only on observed data (Little and Rubin, 2002; Horton and Kleinman, 2007). However, the literature on model selection and averaging in the presence of missing observations is surprisingly sparse given that this is a daily task for many researchers. Adjusting the AIC when confronted with incomplete observations is the most common suggestion for model selection (Shimodaira, 1994; Cavanaugh and Shumway, 1998; Hens et al., 2006; Claeskens and Consentino, 2008). Among the proposed modifications of the AIC, using inverse probability weighting (IPW) as the method of correction (AIC_w; Hens et al. (2006)) is probably the most accessible option for many applied working researchers. Other suggestions are more pragmatic such as selecting predictors only if they are contained in most imputed sets of data (Wood et al., 2008); or selecting variables based on a stacked dataset of multiply imputed datasets and apply weights to this dataset (Wood et al., 2008); or selecting variables based on averaged model selection criteria after multiple imputation (AIC, *p*-value, etc., May et al. (2010)). While the latter suggestions are certainly valuable for solving a specific practical problem they do not provide a general and overall valid framework for model selection with missing data. Moreover, they do not incorporate model selection uncertainty, i.e. by means of applying model averaging.

When considering (frequentist) model averaging in the presence of missing data, e.g. by means of implementing model averaging with AIC-based weights, Schomaker et al. (2010) suggest to either adjust the model averaging weights by using the IPW corrected criterion AIC_w from Hens et al. (2006) instead of the classical AIC, or to perform model averaging on a single imputed set of data. Nevertheless, multiple imputation (MI) probably remains the most popular option to deal with missing data in most areas of research (assuming that omitting missing data is not an acceptable strategy). Modern software packages, such as `Amelia II` in *R* (Honaker et al., 2010; Honaker and King, 2010) or Stata's `ICE` (White et al., 2011), allow us to conveniently create multiple imputations and combine results across the imputed datasets for standard modeling exercises. Not only due to its widespread use it is of great importance to understand how to appropriately combine multiple imputation with model selection. To account for both the uncertainty related to imputation and model selection, the incorporation of model averaging is another issue of great relevance.

We aim to describe how to combine model selection and model averaging with multiple imputation correctly. As we will see, it is straightforward to integrate model selection and averaging estimates into standard MI combining rules—though it is important to discuss the consequences of this. While point estimates shrink towards zero if a variable is not supported throughout imputations and candidate models, resulting standard errors will become large due to combination of both selection and imputation uncertainty.

A somewhat neglected issue of the model averaging literature, confidence interval construction, has recently attracted more attention: In the frequentist literature, Hjort and Claeskens (2003) were the first ones pointing towards the possibly asymmetric distribution of both post model selection and model averaging estimators. Their framework allows for asymmetric confidence intervals but the discussion of the consequences of this finding have then long been avoided; indeed, in the more applied model averaging literature often only point estimates and standard errors have been reported without explicitly stating the confidence interval. Recent work of Wang et al. (2012) and Wang and Zhou (forthcoming) shows that under a fair amount of models the confidence intervals suggested by Hjort and Claeskens (2003) are asymptotically equivalent to the intervals obtained from the full model indicating limited use of model averaging. While it is still been pointed out that even symmetric confidence intervals can perform well in many situations (Fletcher and Dillingham, 2011), more and more value is seen in the evaluation and modification of interval estimation (Turek and Fletcher, 2012). Given the relevance and timeliness of these discussions we find it desirable to devote some investigations to interval estimation for our estimators: In light of the additional complication introduced by missing data and the implementation of multiple imputation, it is especially useful to address these and other important questions by means of Monte Carlo studies and a motivating data example.

The paper proceeds with a detailed description of our statistical framework in Section 2. We explore the finite sample performance of the proposed estimators through a Monte Carlo study in Section 3 with the aim of revealing the nature of model selection and averaging estimators under multiple imputation. Using analyses based on an illustrative example related to a recent study on sputum culture conversion in pulmonary tuberculosis, we discuss several aspects of software implementation and further verify our findings. We conclude with an extensive discussion in Section 5.