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 © 2013 Elsevier B.V. All rights reserved.
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.

2. Statistical framework

To facilitate our discussion of model selection and model averaging with missing observations, we first introduce some notation. Consider some data $\mathcal{D}$ consisting of independent observations $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$. Let $\mathbf{y} = (y_1, \ldots, y_n)'$ be the $n \times 1$ vector of response values, and $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})'$ the $n \times 1$ vector of measurements related to the jth covariate. The $1 \times p$ vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ contains the ith observation of each of the p covariates and $\mathbf{X} = (\mathbf{x}_1{}', \ldots, \mathbf{x}_n{}')'$ is the matrix of all covariates.

Model selection and model averaging. To relate the response with a set of regressors one typically introduces a statistical model $M_\kappa = f(\mathbf{y}|\mathbf{X}, \theta)$, $\theta \in \Theta$, $\Theta \subset \mathbb{R}^m$, $m \geq 1$. Estimation of θ depends upon the choice of this model and $\hat{\theta}$ is therefore often called a *conditional* estimator (Leeb and Pötscher, 2006). If we consider a set of candidate models, $\mathcal{M} = \{M_1, \ldots, M_k\}$, for describing $\mathbf{y}$ based on varying combinations of $\mathbf{X}_j$'s, then a model selection procedure chooses one single 'best' model out of the set $\mathcal{M}$ based on some criterion Γ. In regression analysis such a procedure would often correspond to the choice of a model which minimizes prediction error by means of cross validation ($\Gamma = \text{CV}$) or minimizes Akaike's information criterion ($\Gamma = \text{AIC}$); hence, the corresponding post model selection estimator $\hat{\theta}^*$ is

$$\hat{\theta}^* = \hat{\theta}_\kappa : \arg\min_{M_\kappa \in \mathcal{M}} \{\Gamma(\hat{\theta}_\kappa|M_\kappa)|\mathcal{D}\}, \quad \kappa = 1, \ldots, k. \tag{1}$$

However, model selection introduces additional uncertainty into the process of statistical modeling. There might be many good models to describe the data, i.e. models with a very similar prediction error, but while in some models a specific variable may be included, in others it may be not. As a result, model selection estimators are often unstable, biased, and – most importantly – underestimate the estimator's variance by neglecting the uncertainty associated with the model selection process (Chatfield, 1995; Hjort and Claeskens, 2003; Leeb and Pötscher, 2005). It is often argued that model averaging is suitable to overcome this problem. With model averaging, one calculates a weighted average from the k estimators of $\mathcal{M}$, with the perception that 'better' models should receive a higher weight:

$$\hat{\bar{\theta}} = \sum_{\kappa=1}^{k} w_\kappa \hat{\theta}_\kappa. \tag{2}$$

A popular weight choice would be based on the exponential AIC,

$$w_\kappa = \frac{\exp\left(-\frac{1}{2}\text{AIC}_\kappa\right)}{\sum_{\kappa=1}^{k} \exp\left(-\frac{1}{2}\text{AIC}_\kappa\right)}, \tag{3}$$

where AIC_κ is the AIC value related model $M_\kappa \in \mathcal{M}$ (Buckland et al., 1997). Thus, the smaller (i.e. better) the value of the AIC for a particular model, the higher the influence of this particular model on the model averaging estimator. Other choices would favor criteria such as the focused information criterion FIC (Claeskens and Hjort, 2003) for constructing weights in the spirit of (3), or would construct weights such that the final model averaging estimator is optimal with respect to minimizing a Mallows criterion or the trace of the estimator's MSE (Hansen, 2007; Wan et al., 2010; Liang et al., 2011). Nevertheless, whatever choice is made, it is important to incorporate the uncertainty of the model selection process into the final estimates and hence the variability of estimates in the different candidate models in addition to the sampling variance. Hence, Buckland et al. (1997) suggested to estimate the variance of the scalar $\hat{\bar{\theta}}_j \in \hat{\bar{\theta}}$ via

$$\widehat{\text{Var}}(\hat{\bar{\theta}}_j) = \left\{ \sum_{\kappa=1}^{k} w_\kappa \sqrt{\widehat{\text{Var}}(\hat{\theta}_{j,\kappa}|M_\kappa) + (\hat{\theta}_{j,\kappa} - \hat{\bar{\theta}}_j)^2} \right\}^2. \tag{4}$$

While (4) certainly addresses the problem of model selection uncertainty and can be implemented easily, it has also been criticized that the coverage probability of interval estimates based on (4) may be biased (Hjort and Claeskens, 2003). It is thus important to mention that some model averaging frameworks allow alternative variance estimators, for instance: weighted-average least squares estimation (Magnus et al., 2010; Heumann and Grenke, 2010; Magnus et al., 2011), Bayesian model averaging (Draper, 1995), bootstrapping (Buckland et al., 1997), or frequentist model averaging based on a local misspecification framework (Hjort and Claeskens, 2003). We aim in evaluating the performance of (4) in the context of multiple imputation in our simulation study in Section 3 and discuss implications and consequences of this approach and alternatives, such as bootstrapping, in the following sections, especially in the discussion.

Multiple imputation. It is common for many studies that some of the data $\mathcal{D}$ will be missing: individuals may not provide all the information that is required, data are not captured regularly, or study participants may simply be lost to follow-up. Here, the data consist of both observed and missing values, $\mathcal{D} = \{\mathcal{D}^{\text{obs}}, \mathcal{D}^{\text{mis}}\}$, and valid statistical inference depends on the missingness mechanism: if the data are missing at random (MAR), meaning that the probability of a value to be missing depends only on observed quantities, multiple imputation is a commonly applied method to obtain valid inference in the presence of missing observations (Little and Rubin, 2002). Here, M imputed sets of data will be generated, and the

imputations are based on draws from the predictive posterior distribution of the missing data given the observed data $p(\mathcal{D}^{\text{mis}}|\mathcal{D}^{\text{obs}}) = \int p(\mathcal{D}^{\text{mis}}|\mathcal{D}^{\text{obs}};\theta)\, p(\theta|\mathcal{D}^{\text{obs}})\, d\theta$, or an approximation thereof. It is common to speak of 'augmented data' when referring to the M imputed datasets which consist of both the observed and imputed data. There are nowadays various options to easily utilize proper imputation, typically based on either specifying a suitable multivariate distribution or conditional distributions for each individual variable (= imputation model); an excellent review on these and other aspects, including software, can be found in Horton and Kleinman (2007). After generating the augmented data, the analysis model (e.g. any regression model) will be fitted on each augmented dataset and the M results will be combined appropriately. The point estimate of θ (without considering model selection/averaging) is

$$\hat{\theta}_{\text{MI}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}^{(m)} \tag{5}$$

where $\hat{\theta}^{(m)}$ refers to the estimate of θ in the mth imputed set of data $\mathcal{D}^{(m)}$, $m = 1,\ldots,M$. Based on the average within imputation covariance $\widehat{W} = M^{-1}\sum_m \widehat{\text{Cov}}(\hat{\theta}^{(m)})$ and the between imputation covariance $\hat{B} = (M-1)^{-1}\sum_m(\hat{\theta}^{(m)} - \hat{\theta}_{\text{MI}})(\hat{\theta}^{(m)} - \hat{\theta}_{\text{MI}})'$ one obtains variance estimates via

$$\widehat{\text{Cov}}(\hat{\theta}_{\text{MI}}) = \widehat{W} + \frac{M+1}{M}\hat{B} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\text{Cov}}(\hat{\theta}^{(m)}) + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}(\hat{\theta}^{(m)} - \hat{\theta}_{\text{MI}})(\hat{\theta}^{(m)} - \hat{\theta}_{\text{MI}})'. \tag{6}$$

To construct confidence intervals for $\hat{\theta}_{\text{MI}}$ in the scalar case, it may be assumed that $\widehat{\text{Var}}(\hat{\theta}_{\text{MI}})^{-\frac{1}{2}}(\hat{\theta}_{\text{MI}} - \theta)$ follows a t_R-distribution with approximately $R = (M-1)[1 + \{M\hat{W}/(M+1)\hat{B}\}]^2$ degrees of freedom (Rubin and Schenker, 1986), though there are alternative approximations, especially for small samples, see Lipsitz et al. (2002) among others.

2.1. Model selection after multiple imputation

Estimating θ based on model selection is part of fitting the analysis model. Hence, it is obvious to (i) specify any imputation model and utilize proper multiple imputation, (ii) to analyze the augmented data by specifying the analysis model and performing model selection, and (iii) to combine the post model selection estimates obtained from the M augmented datasets by means of the multiple imputation rules. If the model selection procedure does not select a variable in each augmented set of data, but only in some, then this implies that the corresponding estimate $\hat{\theta}_j$ is sometimes $\neq 0$ and sometimes $=0$; for instance, consider a specific parameter $\theta_j \in \theta$ such as the coefficient for a specific variable in the linear regression model: if the model selection procedure does not select the aforementioned variable, then the corresponding estimate $\hat{\theta}_j$ (as well as its variance) is defined to be equivalent to 0. This yields a kind of model averaging over different datasets instead of models and less clear effects, supported only by a few augmented datasets, will be shrunk towards zero. In essence, a model selection estimator after multiple imputation can be defined as

$$\hat{\theta}^{*}_{\text{MI}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}^{*(m)} \quad \text{with } \hat{\theta}^{*(m)} = \hat{\theta}_{\kappa} : \arg\min_{M_\kappa \in \mathcal{M}}\{\Gamma(M_\kappa;\hat{\theta}_\kappa)|\mathcal{D}^{(m)}\} \tag{7}$$

being the post model selection estimator obtained from the mth augmented dataset $\mathcal{D}^{(m)}$. The corresponding covariance matrix is

$$\widehat{\text{Cov}}(\hat{\theta}^{*}_{\text{MI}}) = \frac{1}{M}\sum_{m=1}^{M}\widehat{\text{Cov}}(\hat{\theta}^{*(m)}) + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}(\hat{\theta}^{*(m)} - \hat{\theta}^{*}_{\text{MI}})(\hat{\theta}^{*(m)} - \hat{\theta}^{*}_{\text{MI}})', \tag{8}$$

where the covariance matrix of $\hat{\theta}^{*(m)}$ is obtained from the selected model in the mth augmented dataset. Note that this means that a variable will be formally 'selected' if it is selected in at least one imputed set of data, but its overall impact will depend on how often it is chosen. This may lead to different results than pragmatic approaches which may select predictors only if they are contained in most imputed sets of data; or select variables based on a stacked dataset of all imputations and apply weights to this dataset (Wood et al., 2008); or select variables based on averaged model selection criteria (AIC, p-value, etc.) which is not supported by MI literature, see also White et al. (2011, p. 389).

2.2. Model averaging after multiple imputation

Following the arguments from the above section, model averaging and multiple imputation can be combined by first calculating model averaging estimators in each augmented dataset (as part of estimating the analysis model) and then combining them by MI rules, with the aim to address both imputation and model selection uncertainty. A general model averaging estimator is then defined as

$$\hat{\bar{\theta}}_{\text{MI}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{\theta}}^{(m)} \quad \text{with } \hat{\bar{\theta}}^{(m)} = \sum_{\kappa=1}^{k} w_\kappa^{(m)}\hat{\theta}_\kappa^{(m)} \tag{9}$$

and applies to any weight choice. If the variance of the model averaging estimator is estimated via (4), the overall variance of the estimator after multiple imputation relates to

$$\widehat{\mathrm{Var}}(\hat{\bar{\theta}}_{j,\mathrm{MI}}) = \frac{1}{M}\sum_{m=1}^{M}\left\{\sum_{\kappa=1}^{k} w_\kappa \sqrt{\widehat{\mathrm{Var}}(\hat{\theta}_{j,\kappa}^{(m)}) + (\hat{\theta}_{j,\kappa}^{(m)} - \hat{\bar{\theta}}_j^{(m)})^2}\right\}^2 + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}(\hat{\bar{\theta}}_j^{(m)} - \hat{\bar{\theta}}_{j,\mathrm{MI}})^2. \tag{10}$$

Note that a reverse order of incorporating the uncertainty associated with both imputation and model selection yields to the alternative estimate $\hat{\bar{\theta}}_{\mathrm{MI,alt}} = \sum_{\kappa=1}^{k} w_\kappa^M \hat{\theta}_\kappa^M$ where first, based on the multiple imputation combining rules, estimates of the weights and candidate model parameters are obtained, $w_\kappa^M = \frac{1}{M}\sum_{m=1}^{M} w_\kappa^{(m)}$, $\theta_\kappa^M = \frac{1}{M}\sum_{m=1}^{M}\theta_\kappa^{(m)}$, and are then combined by model averaging. This leads to different results since

$$\begin{aligned}\hat{\bar{\theta}}_{\mathrm{MI,alt}} &= \sum_{\kappa=1}^{k} w_\kappa^M \hat{\theta}_\kappa^M = \frac{1}{M^2}\sum_{\kappa=1}^{k}\left\{\sum_{m=1}^{M} w_\kappa^{(m)}\sum_{m=1}^{M}\hat{\theta}_\kappa^{(m)}\right\} \\ &\neq \frac{1}{M}\sum_{m=1}^{M}\sum_{\kappa=1}^{k} w_\kappa^{(m)}\hat{\theta}_\kappa^{(m)} = \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{\theta}}_{\mathrm{imp}}^{(m)} = \hat{\bar{\theta}}_{\mathrm{MI}}\end{aligned}$$

but is not supported by MI theory and may yield less appropriate estimates.

We would like to stress that for (7)–(10) we did not distinguish between regression parameters that are related to main effects and others that are related to interaction effects: in each augmented dataset an interaction term can either be selected or omitted (depending on Γ or the model averaging weights), no matter whether main effects are included in the model or not. Of course, the interpretation of the final estimator depends on whether main effects are eventually included or not; and it would generally be possible to allow inclusion of interactions in models only in conjunction with the respective main effects for ease of interpretation, i.e. by restricting the set of candidate models $\mathcal{M}$. Typically pragmatic reasons will decide upon the final treatment of interaction terms.

2.3. Model averaging and selection after multiple imputation using bootstrapping

It is known from the literature that estimators post model selection and after model averaging not necessarily have a normal, or even symmetric distribution (Hjort and Claeskens, 2003; Leeb and Pötscher, 2005; Wang et al., 2012). Multiple imputation and missing data add another dimension of complexity and it is not expected that these conclusions change. To address and explore the issue of suitable confidence interval construction, bootstrapping may well reveal the nature of the distributions belonging to the above proposed estimators and provide an alternative estimation procedure. For combining model averaging (selection), multiple imputation, and bootstrapping in a fruitful manner and in line with the framework developed above, the algorithm listed in Table 1 can be considered:

Table 1
Model selection and model averaging after multiple imputation using bootstrapping.

(1)	Create B bootstrap samples of the original data (including missing observations)
(2)	Generate M imputed sets of data for each bootstrap sample
(3)	Calculate a model averaging (or selection) estimator for each imputed set of data in each bootstrap sample
(4)	Create a model averaging (or selection) estimator after imputation, i.e. via (9) (or (7)), for each bootstrap sample
(5)	Use the average of the B estimates calculated in step 4 as the final point estimate
(6)	Construct confidence intervals based on the percentiles of the empirical distribution produced by the B estimates of step 4

3. Simulation study

In this section we evaluate the finite sample performance of different model averaging and model selection estimators in the presence of missing observations and in the context of the linear regression model.

3.1. Setting

Generating data: In this experiment we set the sample size as $n = 500$ and consider seven variables, one outcome and six potential covariates. The observations for the covariates are generated by the following distributions: $\mathbf{X}_1 \sim N(0.5, 1)$, $\mathbf{X}_2 \sim \log N(0.5, 0.5)$, $\mathbf{X}_3 \sim \mathrm{Weibull}(1.75, 1.9)$, $\mathbf{X}_4 \sim \mathrm{Exp}(1)$, $\mathbf{X}_5 \sim \mathrm{Gamma}(0.25, 2)$, and $\mathbf{X}_6 \sim N(0.25, 1)$. To model the dependency between the covariates we use a Clayton Copula with a copula parameter of 1 which indicates medium correlation among the covariates which are 'quasi-standardized' in the sense that the variance of each variable is $\approx$1. The *R*-package `copula` (Yan, 2007) provides an efficient tool to utilize this approach. The values of the response vector

Table 2
Squared loss L_2 with respect to $\boldsymbol{\beta}$ for both model selection (MS) and model averaging (MA) estimators—using either complete case analysis (CC), multiple imputation (MI), multiple imputation combined with bootstrapping (MI Boot), or the original data without missing observation (Org).

MA Org	0.34	MS Org	0.41
MA MI	0.51	MS MI	0.55
MA MI Boot	0.51	MS MI Boot	0.52
MA CC	1.66	MS CC	1.78

are realized via draws from $\mathbf{y} \sim N(\boldsymbol{\mu}_1, \sigma_1)$ with expectation $\boldsymbol{\mu}_1 = \mathbf{X}\boldsymbol{\beta}_1$, the true parameter (including intercept) $\boldsymbol{\beta}_1 = (2.5, -3, -0.25, 0, -1.5, 0, 0.35)'$, and $\sigma_1 = \exp(1.25)$.

Missing data: Observations of $\mathbf{X}_1, \mathbf{X}_4$ and $\mathbf{X}_5$ are now assumed to have missing observations. Missing values are generated by means of the following missingness functions: $\pi_{\mathbf{X}_1}(\mathbf{y}) = 1 - \{(0.15\mathbf{y})^2 + 1\}^{-1}$, $\pi_{\mathbf{X}_4}(\mathbf{X}_2) = 1 - \{1 + 0.02\mathbf{X}_2^3\}^{-1}$, $\pi_{\mathbf{X}_5}(\mathbf{X}_3) = 1 - \{1 + \exp(1 - 2\mathbf{X}_3)\}^{-1}$. The theoretical percentage of missing values related to these functions correspond to 26.90% for $\mathbf{X}_1$, 14.61% for $\mathbf{X}_4$ and 18.24% for $\mathbf{X}_5$. Since the probability of missingness depends only on observed variables, the data are missing at random.

Estimators and model: We aim in comparing model selection (MS) and model averaging (MA) estimators in the context of the linear regression model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, that address the missing data in a different manner. All model selection estimators rely on the model choice related to $\Gamma = \text{AIC}$; all model averaging estimators are based on exponential AIC weights as introduced in (3). To deal with missing observations four strategies are considered: (i) complete case analysis, i.e. discarding observations with missing values (MA CC, MS CC); (ii) combining multiple imputation (using the *R*-package `Amelia II`; $M = 10$; Honaker and King (2010)) and model averaging/selection as described in (7)–(10) (MA MI, MS MI); (iii) combining multiple imputation and model averaging/selection as described in (7) and (9) and combining it with bootstrapping ($B = 200$) as described in Table 1 (MA MI Boot, MS MI Boot); and (iv) applying model selection/averaging onto the original data without missing observations to provide a reference measure (MA Org, MS Org).

Measures of performance: We compare the different estimates of $\boldsymbol{\beta}$ by evaluating both the estimated mean squared error (MSE) for each $\hat{\beta}_j$ and the loss function $L_2 = \mathcal{R}^{-1}\sum_r\{\sum_j(\hat{\beta}_{j,r} - \beta_{j,r})^2\}$, whereby $r = 1, \ldots, \mathcal{R}$ describes the number of simulation runs. Moreover, we compare the estimated standard errors due to (8) and (10) to the empirical standard error over all simulation runs and describe the simulated distributions of the different estimators and the implications for their respective confidence intervals.

All results are based on $\mathcal{R} = 500$ simulation runs.

3.2. Results

On average the simulation results in about 27% missing values for X_1, 15% for X_4 and 18% for X_5.

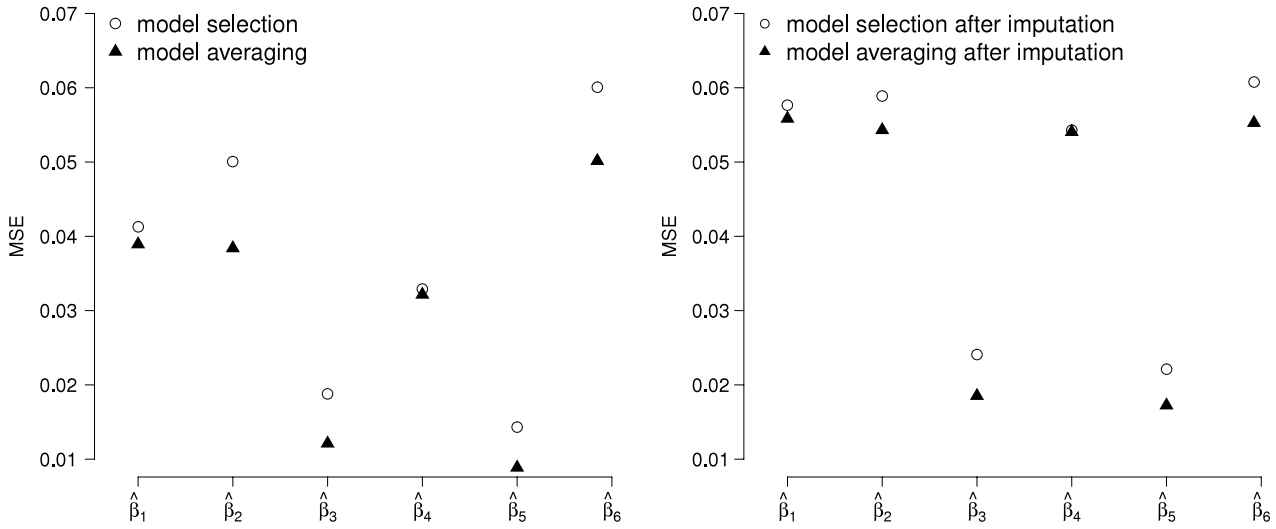
Table 2 presents the summary of each estimators' performance with respect to the loss function L_2. As expected, combining model selection and model averaging with multiple imputation generally outperforms a complete case analysis. No matter what strategy with regard to the missing data is chosen, model averaging provides slightly better estimates than model selection. There are no gains in using bootstrapping when using multiple imputation. Fig. 1 emphasizes that – for each individual parameter – the estimated MSE is lower for model averaging when compared to the AIC selected estimates; this can be observed not only for the augmented data (Fig. 1b) but also for the data without missing values (Fig. 1a).

The reason of superiority for model averaged estimates often relates to their higher stability, i.e. a decreased variance in exchange for only a small increase in bias: for instance, after multiple imputation, for β_2 the MSE (Variance, Bias^2) is 0.054 (0.027, 0.027) for model averaging and 0.059 (0.034, 0.025) for model selection; for β_6 it is 0.055 (0.044, 0.011) and 0.061 (0.054, 0.007) respectively. What we observe here reconfirms the shrinkage behavior of model averaging estimators which was already mentioned by Hansen (2007) in a different context.

A major reason for applying model averaging in favor of model selection is to incorporate model selection uncertainty into final variance estimates and confidence intervals. Table 3 compares the estimated standard errors of our estimators (averaged over all simulation runs) with the empirical standard error obtained from the variation of $\hat{\beta}_j$ over the simulation runs. For all model averaged estimates, estimated and empirical standard errors are very close, indicating that the proposed variance estimators work well in our simulation—even for a complete case analysis. On the contrary, model selection (MS Org, MS MI, MS CC) systematically underestimates the standard error; the only exception are the estimates which use model selection after multiple imputation in conjunction with bootstrapping (MS MI Boot). This suggests that bootstrapping offers an opportunity to circumvent the common problems associated with post model selection variance estimation. It is worthwhile to note that also the Bayesian model selection paradigm – putting a prior probability on each candidate model – explicitly addresses the uncertainty associated with the model selection process, though the difficulty in specifying good prior probabilities and the possibly volatile choice of the model with highest posteriori probability (zero–one loss function) point towards the usefulness of Bayesian model averaging and the reporting of posteriori model probabilities.

Figs. 2 and 3 offer insight into the behavior and distribution of the proposed estimators.

Fig. 2a depicts the simulated distribution of $\hat{\beta}_4$ for model averaging and model selection after multiple imputation (MA MI, MS MI); even though the distribution might not be exactly normal, it is rather symmetric and one might assume that



(a) Original data without missing observations.

(b) Data after multiple imputation.

Fig. 1. Estimated MSE based on the simulation study. Model selection is based on the AIC and model averaging is performed with exponential AIC weights as described in (3). For combining multiple imputation with model selection/averaging, (7) and (9) are used respectively.

Table 3
Estimated standard errors (se): (i) estimated standard error (averaged over all simulation runs, *left*); (ii) empirical standard error (based on the variability of $\hat{\beta}$ in 500 simulation runs, *right*). Multiple imputation related estimators in (i) are based on (8) and (10).

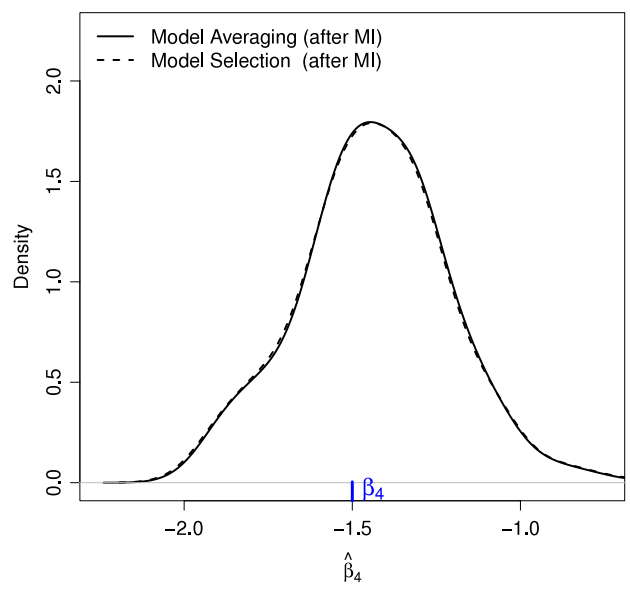
	$\widehat{\text{se}}(\beta_1)$		$\widehat{\text{se}}(\beta_2)$		$\widehat{\text{se}}(\beta_3)$		$\widehat{\text{se}}(\beta_4)$		$\widehat{\text{se}}(\beta_5)$		$\widehat{\text{se}}(\beta_6)$	
	(i)	(ii)	(i)	(ii)	(i)	(ii)	(i)	(ii)	(i)	(ii)	(i)	(ii)
MA Org	0.19	0.20	0.14	0.17	0.11	0.11	0.18	0.18	0.10	0.09	0.18	0.20
MA MI	0.23	0.23	0.16	0.16	0.13	0.14	0.22	0.22	0.14	0.13	0.20	0.21
MA MI Boot	0.23	0.23	0.18	0.18	0.17	0.16	0.22	0.22	0.16	0.15	0.21	0.20
MA CC	0.25	0.28	0.19	0.22	0.13	0.13	0.24	0.25	0.12	0.13	0.18	0.22
MS Org	0.18	0.20	0.07	0.21	0.03	0.14	0.17	0.18	0.02	0.12	0.12	0.24
MS MI	0.22	0.24	0.12	0.19	0.08	0.16	0.22	0.22	0.10	0.15	0.17	0.23
MS MI Boot	0.23	0.24	0.20	0.19	0.19	0.17	0.22	0.22	0.18	0.16	0.22	0.21
MS CC	0.24	0.29	0.07	0.27	0.03	0.16	0.23	0.25	0.03	0.16	0.08	0.27

symmetric confidence intervals, as described in Section 2.2, yield reasonable results as long as a good standard error estimate is provided (as explained in Table 2). The same applies to $\hat{\beta}_5$ as illustrated in Fig. 2b. While $\mathbf{X}_4$ is defined to have a clear negative effect on the outcome, $\mathbf{X}_5$ is defined to have none at all. When looking at the two smaller effects related to $\mathbf{X}_2$ and $\mathbf{X}_6$ (Fig. 2c and d), one can see that the distributions are now heavily skewed and non-normal. The distributions of model selection estimators have two modi, one around '0' and the other close to the true parameter, while distributions related to model averaged estimates are smoothed out. It is expected that in this setting symmetric confidence intervals will not provide accurate estimates of confidence limits. Bootstrapping can account for this, as highlighted in Fig. 3a and b for β_4 and β_6. The estimated distributions (as well as the point estimate and confidence intervals) seem to accurately describe the estimator's behavior. Combining model selection/averaging with bootstrapping may therefore help to calculate good estimates when dealing with model selection uncertainty (and missing data).

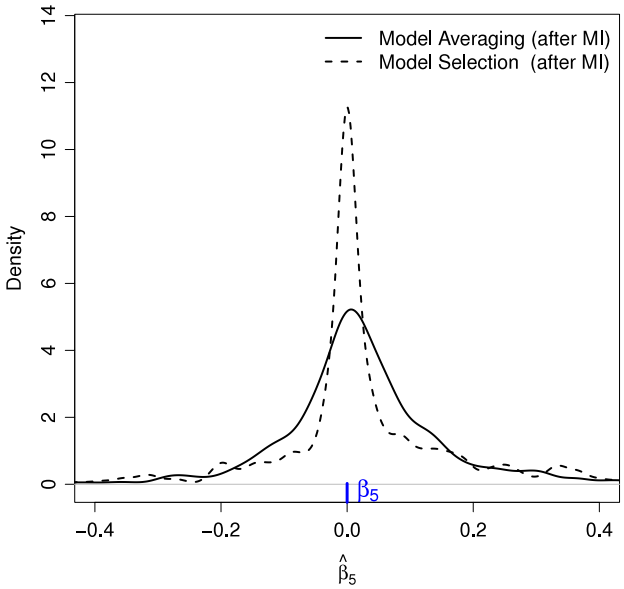
Indeed, for the clear effects (β_1, β_3, β_4, β_5) symmetric confidence intervals (for both model selection and averaging) do reasonably well in our simulation: in $\geq$93% of simulation runs both the true effect was covered and '0' was included/excluded correctly. While there is a poor coverage by means of the symmetric CIs for small effects (β_2, β_6) ($\approx$55%–76%), bootstrapping clearly helps to address this problem: for instance, for β_6 coverage of the true parameter when applying model averaging is 86% compared to 76% without bootstrapping; and in 44% of simulation runs the '0' was correctly not covered by the CI when applying bootstrapping, compared to 16% with symmetric intervals.

3.3. Sensitivity

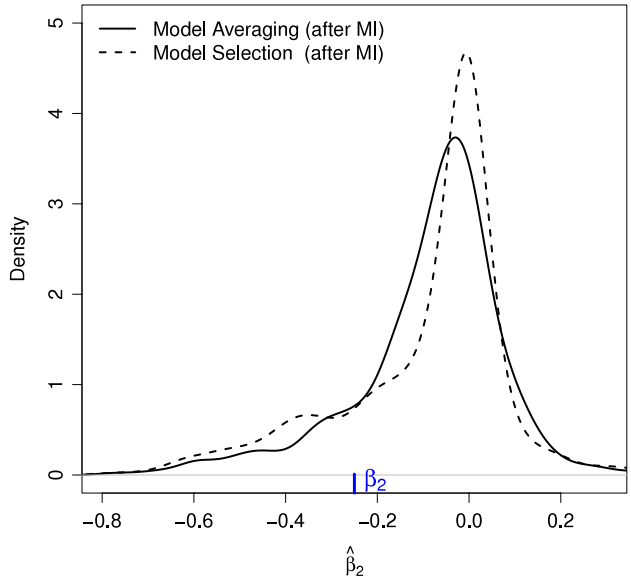
To explore the sensitivity of the reported results, we have conducted and evaluated several modifications of our simulation setting. A different number of imputed datasets ($M = 5$), another model averaging strategy (Bayesian model averaging), and a lower missingness rate (6% for $\mathbf{X_1}$, 8% for $\mathbf{X_4}$, and 8% for $\mathbf{X_5}$) led to essentially the same findings. Generating data from a general model (logistic regression, adding interaction terms, see Section 4 for survival data) yielded also very similar conclusions though there was no benefit of model averaging compared to model selection when evaluating the point



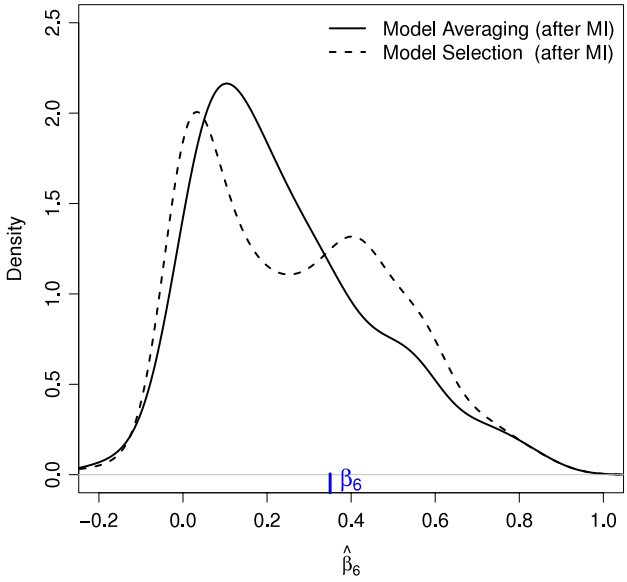
(a) Simulated distribution of $\hat{\beta}_4$ (large effect).



(b) Simulated distribution of $\hat{\beta}_5$ (no effect).



(c) Simulated distribution of $\hat{\beta}_2$ (small effect).



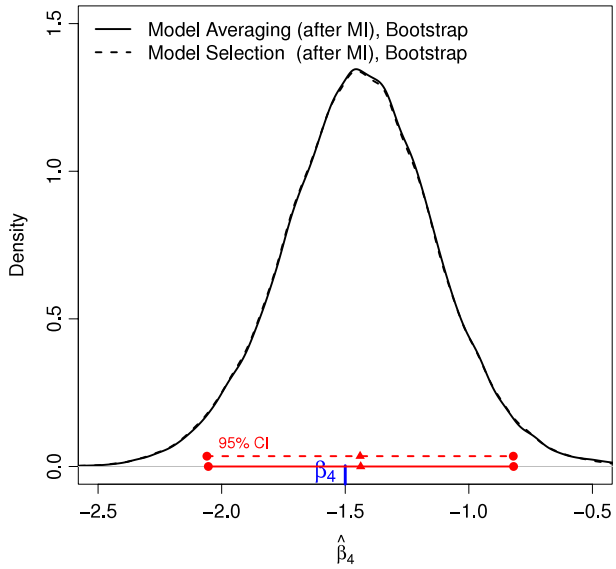
(d) Simulated distribution of $\hat{\beta}_6$ (small effect).

Fig. 2. Estimated distribution of different β_j's after multiple imputation based on the 500 simulation runs. Point estimates are based on (9) and (7).

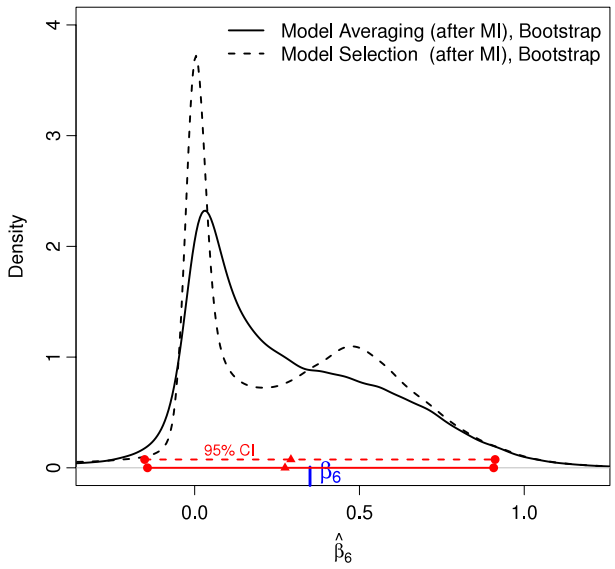
estimates. Utilizing single imputation, e.g. via `Amelia II` or GAMRI (Schomaker et al., 2010), produced variance estimates that were too small as imputation uncertainty was neglected by using only one augmented dataset.

4. Sputum culture conversion in pulmonary tuberculosis: an illustrative example

To further illustrate the implementation and behavior of the proposed estimators, also beyond the linear regression model, we focus on a recent study of Visser et al. (2012) who aim to detect predictors for delayed sputum culture conversion in pulmonary tuberculosis. Sputum culture conversion after 2 months of anti-tubercular therapy is regarded to be a good biomarker for tuberculosis cure and finding the best predictors when modeling time to culture conversion is of great importance in the fields of pharmacokinetics and infectious diseases. The authors used a Cox proportional hazards model to determine the most relevant predictors among a set of 16 pre-determined, epidemiologically and biologically potentially relevant variables. They report model selection uncertainty and missing covariate data for seven variables (up to $\approx$20%, $n = 113$). When fitting the full model with all 16 variables after multiple imputation, Visser et al. (2012) identified four important variables to predict the hazard of culture conversion. Average posterior effect probabilities obtained from Bayesian model averaging confirmed these findings, identifying lung cavities, time to culture detection and smoking as having a rather strong effect and the tuberculosis W-Beijing genotype as having a moderate effect on delayed conversion.



(a) Estimated distribution and 95% confidence interval of β_4 by using bootstrapping.



(b) Estimated distribution and 95% confidence interval of β_6 by using bootstrapping.

Fig. 3. Estimated distribution of different β_j's after multiple imputation based on 500 simulation runs. All bootstrap based estimates refer to the strategy explained in Table 1.

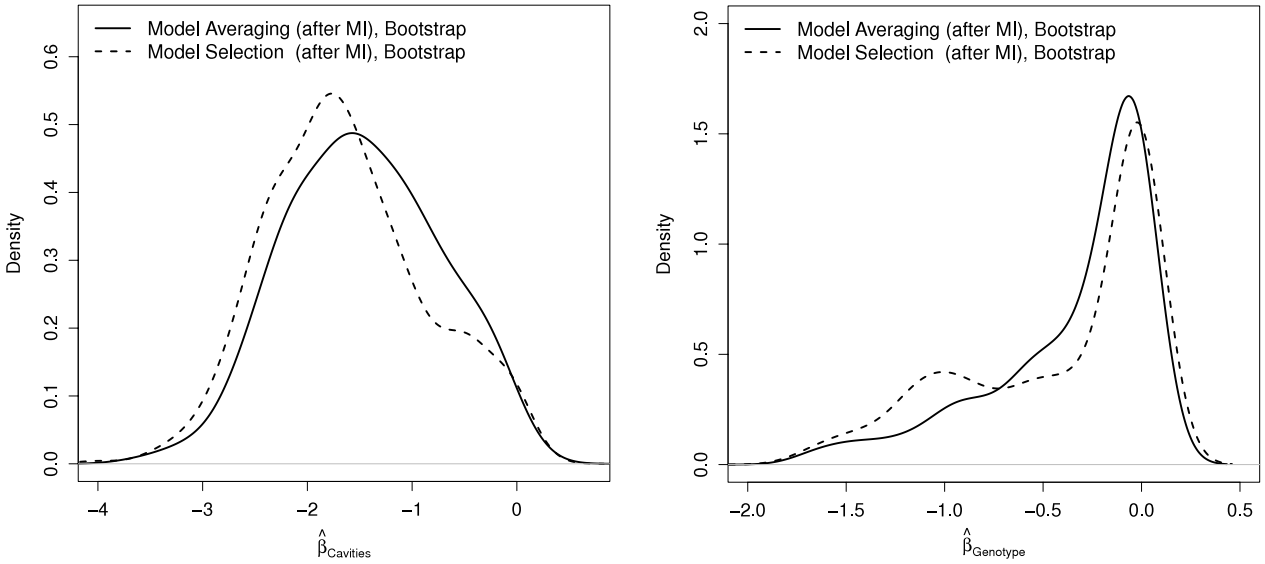
Table 4
Estimated hazard ratios and 95% confidence intervals from the model selection and model averaging estimators of the Cox model after imputation. Results of the four out of 16 most relevant variables (lung cavities, smoking status, time to culture detection in days (TTD), W-Beijing genotype) are reported.

	Model averaging after MI						Model selection after MI					
	Standard			Bootstrapping			Standard			Bootstrapping		
	HR	95% CI		HR	95% CI		HR	95% CI		HR	95% CI	
Cavities	0.18	0.06	0.56	0.23	0.06	0.89	0.17	0.06	0.48	0.20	0.05	1.00
Smoking	0.32	0.14	0.73	0.37	0.10	0.98	0.30	0.14	0.65	0.33	0.09	1.00
TTD	1.09	0.99	1.20	1.10	1.00	1.24	1.11	1.03	1.19	1.10	1.00	1.25
Genotype	0.61	0.26	1.46	0.71	0.23	1.01	0.51	0.23	1.11	0.66	0.22	1.00

We now implement frequentist model selection and model averaging estimators based on the AIC, as introduced in Section 2 and evaluated in Section 3, for this study and describe their performance. Applying model averaging in applied survival analysis has become more and more important, see also Hjort and Claeskens (2006) among others. It will be of interest whether the findings of our simulation study can also be observed in survival analysis. Table 4 reports hazard ratios and 95% confidence intervals of the four most important variables (based on 10 multiple imputations obtained from `Amelia II` and 500 bootstrap replications; the imputation model included all measured variables including the survival outcome).

All estimated hazard ratios are similar to the results obtained from Visser et al. (2012) in the full Cox model. Both the standard combination estimates related to (7)–(10) and bootstrapping support the hypothesis that lung cavities, time to culture detection and smoking status are associated with delayed sputum conversion. For these rather strong effects the standard confidence intervals are narrower than the bootstrap intervals. On the contrary, for the effect of genotype the confidence intervals with respect to the bootstrap procedure are narrower than the standard confidence intervals. Fig. 4 explains these findings by means of the bootstrap distributions of model selection and model averaging estimators for lung cavities and genotype: In line with the findings of Fig. 2 from the simulation study we observe a non-normal, but reasonably symmetric distribution for the rather large effect of lung cavities and a heavily skewed distribution towards 0 for the moderate effect of genotype. This suggests that while a straightforward combination of model selection/averaging with multiple imputation may provide good and meaningful results for many predictors, bootstrapping may indeed help to discover more moderate effects; when estimating the effect of genotype the straightforward combination of both model selection and imputation uncertainty would have led to a rather wide and conservative confidence interval (0.26; 1.46) masking the potential impact of genotype—and only the nature of the bootstrap distribution and the respective confidence interval (0.23; 1.01) would have provided stronger evidence for W-Beijing genotype as being a relevant predictor of delayed sputum culture conversion.

As expected, standard model averaging intervals are overall wider than standard model selection intervals. The shrinkage effect of model averaging as well as the often bimodal nature of post model selection distributions can also be observed in Fig. 4.



(a) Distributions for the effect of lung cavities.

(b) Distributions for the effect of W-Beijing genotype.

Fig. 4. Bootstrap distribution of model selection and model averaging estimators after multiple imputation for the effects of lung cavities and W-Beijing genotype on delayed culture conversion.

It is certainly worthwhile to add that frequentist model averaging confirms the importance of the four presented variables: indeed, the average relative importance (RI; for each variable, this is the sum of model averaging weights from the models where this variable is included) highly favors the presented four variables (RI $\geq$ 0.7) compared to the other 12 (RI $\leq$ 0.4). Applying model selection, AIC occasionally also selects other variables in a few of the augmented datasets due to its well-known property of supporting less parsimonious models than Bayesian criteria; however, averaged over all augmented datasets they do not seem to provide much additional value with hazard ratios often estimated $\approx$1 and distributions of $\hat{\beta}$ centering around 0, similar to Fig. 2b.

Another interesting remark refers to the bootstrap intervals for model selection after multiple imputation: one can see that for our example either the upper or lower confidence limit is 1 revealing that there is considerable mass around $\hat{\beta} = 0$. This makes interpretations more difficult and model averaging may be the better option for the current analysis.

5. Discussion

We have demonstrated that model averaging and model selection can easily and successfully be combined with multiple imputation. As a result, effects of variables which are not supported throughout augmented datasets and candidate models will simply be less pronounced. In our analyses, model averaging induces somewhat more stable estimates than model selection, mostly due to its inherent shrinkage properties and therefore smaller variance in exchange for some bias. Also in the context of missing observations, standard errors post model selection typically underestimate the true variability of the estimates which is not surprising and to be expected from the literature. On the contrary, we have found our model averaging estimators to produce accurate standard errors after multiple imputation for all situations under consideration. Our simulations further suggest that bootstrapping provides an attractive alternative for standard error estimation, especially for post model selection estimators. In our examples we found that if effects are either very clear or non-apparent, distributions of the proposed estimators are rather symmetric and inference works for both model selection and averaging reasonably well. For smaller effects, distributions are often highly non-normal and while the combination of standard MI and model averaging/selection rules does not always provide appropriate confidence intervals, bootstrapping helps to obtain both proper estimates and stronger evidence about the importance of variables. This may be explained by the fact that bootstrapping explicitly addresses model selection uncertainty by underscoring the different variable choice for different samples and therefore acknowledging possible bimodal or other non-normal distributions of our estimators. Interestingly, the complexity of distributions after model selection and imputation may make it difficult to discover small effects which may sometimes yield conservative conclusions in practice and therefore introduces an additional dimension to the usual overconfidence problem of model selection.

To our knowledge, this is the first comprehensive study on model selection and model averaging after multiple imputation. While our findings encourage researchers to take model selection uncertainty and missing observations into account and our framework provides a useful guideline to do so, there are some limitations one has to mention. First of all, we have restricted our studies to certain imputation procedures and selection/averaging frameworks: all estimators refer to AIC based choices and multiple imputation was solely utilized with `Amelia II`. Shifting from joint modeling approaches to fully conditional modeling, for instance via imputation by chained equations (White et al., 2011), and using

other model selection and averaging techniques, such as cross validation, shrinkage and optimal weight based averaging (Hansen, 2007; Liang et al., 2011; Schomaker, 2012; Hansen and Racine, 2012), further work will provide more evidence about the generalization of our findings.

Secondly, we emphasized the frequentist perspective on model selection and averaging. It will be interesting to extend our framework to a Bayesian perspective. Multiple imputation is naturally Bayesian and combining it with model averaging is straightforward based on the insights we have offered here. There are already very good software implementations, e.g. the package `BMA` (Raftery et al., 2011) in *R*, to effectively implement Bayesian model averaging in conjunction with imputation. The Bayesian Bootstrap (Rubin, 1981) may serve as an additional technique to well approximate the posterior distribution of (for example) the mean and an option to obtain suitable credibility intervals. Indeed, preliminary Bayesian analyses we have conducted reveal a similar nature of estimators and distributions after imputation when compared to the frequentist results.

A reflection on the principal findings of our study points towards the necessity of a more thorough discussion about the implications of our findings and their relation to existing literature. We reckon that the implementation of AIC based model selection or averaging is fairly straightforward and any researcher with the ability for utilizing multiple imputation should be able to estimate (7)–(10). In our simulation studies we found these estimators to work quite well, also the variance estimators incorporating (4). However, our investigations also show that there exist situations (i.e. when effects are rather small) where asymmetric confidence intervals are needed which is no surprise when looking at the literature: The framework of Hjort and Claeskens (2003) studies the limiting distribution and asymptotic risk properties of frequentist model average estimators based on a local misspecification framework; in this framework and under the assumption of no missing data, it was found that the asymptotic distributions of model average estimators are non-normal. The authors derived confidence interval estimates which take this matter into account. It is nowadays pointed out that the confidence intervals proposed by Hjort and Claeskens (2003) are asymptotically equivalent to the intervals obtained from the full model indicating limited use of model averaging (Wang et al., 2012; Wang and Zhou, forthcoming; Kabaila and Leeb, 2006). There are also some concerns whether a local misspecification framework addresses the variable selection problem at all (Ishwaran and Rao, 2003) and doubts whether the finite-sample distribution of model averaging estimators can be estimated well enough (Pötscher, 2006). Given these lively and ongoing discussions, as well as the corresponding debate in the post model selection literature (Leeb and Pötscher, 2005), we find that our work adds important evidence that constructing confidence intervals for model averaging or post model selection is non-trivial and dealing with missing data adds another dimension of complexity.

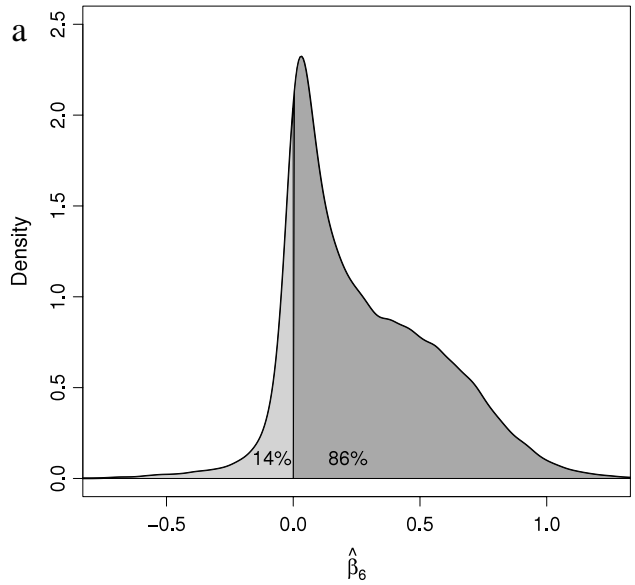
We believe that visualizing the distributions of model selection and averaging estimators by means of bootstrapping are attractive to reveal the relevance of variables in regression models, especially (but not only) in the context of missing data. Consider again the simulation study from Section 3. For both for the large effects of $\mathbf{X}_1$ and $\mathbf{X}_4$ and the variables which have no influence on the outcome ($\mathbf{X}_3, \mathbf{X}_5$) all point and interval estimates performed well when applying model selection/averaging after multiple imputation and had been interpreted as in any other statistical analysis. The evaluation of the respective bootstrap distributions would have only emphasized existing findings. However, the distributions relating to the estimators of $\mathbf{X}_2$ and $\mathbf{X}_6$ reveal possible effects that would not have been discovered otherwise and provide us with more appropriate interval estimates. It may thus be a useful qualitative approach to visualize these distributions and possibly also evaluate the area left and right of $\hat{\beta} = 0$ as indicated in Fig. 5. One can see that 86% of the mass of the bootstrap distribution of $\hat{\beta}_6$ is in the area right of zero, hence indicating a small positive effect of $\mathbf{X}_6$—while the same approach would have also given valuable insight about the effects of all other variables.

Another issue that deserves some discussion and more in-depth research is the implementation of proper multiple imputation and the consequences of specifying a wrong imputation model. We have assumed throughout the paper that data are missing at random and the implementation of multiple imputation is straightforward. However, if the missingness process is non-ignorable any valid inferential method requires careful specification of a model for the missing data mechanism, e.g. via pattern-mixture or shared-parameter models (Molenberghs and Fitzmaurice, 2009), which is difficult to realize for many analyses; thus, the use of an incorrect imputation model can cause improper imputations, biased model estimates and likewise inappropriate post model selection and model averaging estimates.

Even if data are missing at random the choice of the imputation model can affect final results: (i) if a fully conditional imputation approach is utilized, such as imputation by chained equations (White et al., 2011), convergence to the theoretical joint distribution is not always guaranteed (Drechsler and Rässler, 2008), (ii) if a joint modeling approach is taken, e.g. via `Amelia II`, the treatment of categorical variables via a multivariate normal distribution may often yield reasonable results but imputation uncertainty increases and for this case we have indeed observed quite large standard errors (8) and (10) and (iii) imputing longitudinal data is complex, very few proper approaches have been developed (Honaker and King, 2010) and it is not entirely clear how misspecification of a longitudinal imputation model may affect regression modeling. These examples demonstrate the complexity and sensitivity of analyses dealing with missing data, including model selection and model averaging after multiple imputation.

In conclusion, both model averaging and selection can easily be combined with multiple imputation. In some situations it can be of advantage to use bootstrap estimation, for example when there is only small evidence of an effect. To reveal the whole complexity of modeling uncertainty in the presence of missing data further research has to be done.

a



b

	Area:		
	Left of 0 (%)	Right of 0 (%)	
$\hat{\beta}_1$	100	0	$(\beta_1 = -3)$
$\hat{\beta}_2$	67	33	$(\beta_2 = -0.25)$
$\hat{\beta}_3$	47	53	$(\beta_3 = 0)$
$\hat{\beta}_4$	100	0	$(\beta_4 = -1.5)$
$\hat{\beta}_5$	45	55	$(\beta_5 = 0)$
$\hat{\beta}_6$	14	86	$(\beta_6 = 0.35)$

Fig. 5. (a) Bootstrap distribution of $\hat{\beta}_6$ after multiple imputation. The shaded areas illustrate the percentage of area from the distribution that is left and right of 0. (b) Percentage of area from the bootstrap distribution that is left and right of 0 for $\hat{\beta}_1, \ldots, \hat{\beta}_6$.

Acknowledgments

The authors thank Gary Maartens and Marianne Visser for providing the data for our illustrative example and sharing their insights about the meaning of the study with us. Further thanks go to the two referees and the associate editor for their valuable comments and suggestions.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle, In: Proceeding of the Second International Symposium on Information Theory Budapest, pp. 267–281.
Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53, 603–618.
Cavanaugh, J., Shumway, R., 1998. An Akaike information criterion for model selection in the presence of incomplete data. Journal of Statistical Planning and Inference 67, 45–65.
Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society A 158, 419–466.
Claeskens, G., Consentino, F., 2008. Variable selection with incomplete covariate data. Biometrics 64, 1062–1069.
Claeskens, G., Hjort, N.L., 2003. The focused information criterion (with discussion). Journal of the American Statistical Association 98, 900–916.
Draper, D., 1995. Assessment and propagation of model uncertainty. Journal of the Royal Statistical Society B 57, 45–97.
Drechsler, J., Rässler, S., 2008. Does convergence really matter? In: Shalabh, Heumann, C. (Eds.), Recent Advances in Linear Models and Related Areas. Springer, pp. 342–355.
Fletcher, D., Dillingham, P., 2011. Model-averaged confidence intervals for factorial experiments. Computational Statistics and Data Analysis 55, 3041–3048.
Hansen, B.E., 2007. Least squares model averaging. Econometrica 75, 1175–1189.
Hansen, B.E., Racine, J., 2012. Jackknife model averaging. Journal of Econometrics 167, 38–46.
Hens, N., Aerts, M., Molenberghs, G., 2006. Model selection for incomplete and design based samples. Statistics in Medicine 25, 2502–2520.
Heumann, C., Grenke, M., 2010. An efficient model averaging procedure for logistic regression models using a Bayesian estimator with Laplace prior. In: Kneib, T., Tutz, G. (Eds.), Statistical Modelling and Regression Structures. Physica, pp. 79–90.
Hjort, L., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98, 879–945.
Hjort, N.L., Claeskens, G., 2006. Focussed information criteria and model averaging for Cox's hazard regression model. Journal of the American Statistical Association 101, 1449–1464.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statistical Science 14, 382–417.
Honaker, J., King, G., 2010. What to do about missing values in time series cross-section data. American Journal of Political Science 54, 561–581.
Honaker, J., King, G., Blackwell, M., 2010. Amelia 2: a program for missing data. R Package version 1.5. http://gking.harvard.edu/amelia.
Horton, N., Kleinman, K., 2007. Much ado about nothing: a comparison of missing data methods and software to fit incomplete regression models. The American Statistician 61, 79–90.
Ishwaran, H., Rao, J., 2003. Discussion. Journal of the American Statistical Association 98, 922–925.
Kabaila, P., Leeb, H., 2006. On the large-sample minimal coverage probability of confidence intervals after model selection. Journal of the American Statistical Association 101, 619–629.
Leeb, H., Pötscher, B.M., 2005. Model selection and inference: facts and fiction. Econometric Theory 21, 21–59.
Leeb, H., Pötscher, B.M., 2006. Can one estimate the conditional distribution of post-model-selection estimators? Annals of Statistics 34, 2554–2591.
Leeb, H., Pötscher, B.M., 2008. Can one estimate the unconditional distribution of post-model-selection estimators? Econometric Theory 24, 338–376.
Liang, H., Zou, G., Wan, A., Zhang, X., 2011. Optimal weight choice for frequentist model average estimators. Journal of the American Statistical Association 106, 1053–1066.
Lipsitz, S., Parzen, M., Zhao, L., 2002. A degrees-of-freedom approximation in multiple imputation. Journal of Statistical Computation and Simulation 72, 309–318.
Little, R., Rubin, D., 2002. Statistical Analysis with Missing Data. Wiley, New York.
Magnus, J., Powell, O., Prüfer, P., 2010. A comparison of two model averaging techniques with an application to growth empirics. Journal of Econometrics 154, 139–153.
Magnus, J., Wan, A., Zhang, X., 2011. Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. Computational Statistics and Data Analysis 55, 1331–1341.

May, M., Boulle, A., Phiri, S., Messou, E., Myer, L., Wood, R., Sterne, J., Dabis, F., Egger, M., 2010. Prognosis of petients with HIV-1 infection starting therapy in sub-Saharan Africa: a collaborative analysis of scale-up programmes. Lancet 376, 449–457.
Molenberghs, G., Fitzmaurice, G., 2009. Incomplete data: introduction and overview. In: Fitzmaurice, G., Davidian, M., Verbeke, G., Molenberghs, G. (Eds.), Longitudinal Data Analysis. CRC Press, pp. 395–408.
Pötscher, B., 2006. The distribution of model averaging estimators and an impossibility result regarding its estimation. In: Ho, H., Ing, C., Lai, T. (Eds.), IMS Lecture Notes: Time Series and Related Topics, vol. 52. pp. 113–129.
Raftery, A., Hoeting, J., Volinsky, C., Painter, I., Yeung, K., 2011. BMA: Bayesian model averaging. R package version 3.14. http://CRAN.R-project.org/package=BMA.
Rao, C., Wu, Y., 2001. On model selection. IMS Lecture Notes - Monograph Series 38, 1–64.
Rubin, D., 1981. The Bayesian bootstrap. Annals of Statistics 9, 130–134.
Rubin, D., Schenker, N., 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association 81, 366–374.
Schomaker, M., 2012. Shrinkage averaging estimation. Statistical Papers 53, 1015–1034.
Schomaker, M., Heumann, C., 2011. Model averaging in factor analysis: an analysis of Olympic decathlon data. Journal of Quantitative Analysis in Sports 7 (1). Article 4.
Schomaker, M., Wan, A.T.K., Heumann, C., 2010. Frequentist model averaging with missing observations. Computational Statistics and Data Analysis 54, 3336–3347.
Shimodaira, H., 1994. A new criterion for selecting models from partially observed data. In: Cheesman, P., Oldford, R. (Eds.), Selecting Models from Data: Artificial Intelligence and Statistics, Vol. IV. Springer, pp. 21–29.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society B 36, 111–147.
Turek, D., Fletcher, D., 2012. Model-averaged wald confidence intervals. Computational Statistics and Data Analysis 56, 2809–2815.
Visser, M., Stead, M., Walzl, G., Warren, R., Schomaker, M., Grewal, H., Swart, E., Maartens, G., 2012. Baseline predictors of sputum conversion in pulmonary tuberculosis: importance of cavities, smoking, time to detection and W-Beijing genotype. PLoS ONE 7, e29588.
Wan, A.T.K., Zhang, X., Zou, G.H., 2010. Least squares model averaging by mallows criterion. Journal of Econometrics 156, 277–283.
Wang, H., Zhang, X., Zou, G., 2009. Frequentist model averaging: a review. Journal of Systems Science and Complexity 22, 732–748.
Wang, H., Zhou, S., 2012. Interval estimation by frequentist model averaging, Communications in Statistics—Theory and Methods (2013) (forthcoming).
Wang, H., Zou, G., Wan, A., 2012. Model averaging for varying-coefficient partially linear measurement error models. Electronic Journal of Statistics 6, 1017–1039.
White, I., Royston, P., Wood, A., 2011. Multiple imputation using chained equations. Statistics in Medicine 30, 377–399.
Wood, A., White, I., Royston, P., 2008. How should variable selection be performed with multiply imputed data? Statistics in Medicine 27, 3227–3246.
Yan, J., 2007. Enjoy the joy of copulas: with package copula. Journal of Statistical Software 21, 1–21.
Zhang, X., Wan, A., Zhou, S., 2012. Focused information criteria, model selection and model averaging in a tobit model with a non-zero threshold. Journal of Business and Economics Statistics 30, 132–142.