# Frequentist Model Averaging with Missing Observations

Michael Schomaker*, Alan T.K. Wan†, Christian Heumann*

**Abstract**

Model averaging or combining is often considered as an alternative to model selection. Frequentist model averaging (FMA) is considered extensively and strategies for the application of FMA methods in the presence of missing data based on two distinct approaches are presented. The first approach combines estimates from a set of appropriate models which are weighted by scores of a missing data adjusted criterion developed in the recent literature of model selection. The second approach averages over the estimates of a set of models with weights based on conventional model selection criteria but with the missing data replaced by imputed values prior to estimating the models. For this purpose three easy-to-use imputation methods that have been programmed in currently available statistical software are considered, and a simple recursive algorithm is further adapted to implement a generalized regression imputation in a way such that the missing values are predicted successively. The latter algorithm is found to be quite useful when one is confronted with two or more missing values simultaneously in a given row of observations. Focusing on a binary logistic regression model, the properties of the FMA estimators resulting from these strategies are explored by means of a Monte-Carlo study. The results show that in many situations, averaging after imputation is preferred to averaging using weights that adjust for the missing data, and model average estimators often provide better estimates than those resulting from any single model. As an illustration, the proposed methods are applied to a dataset from a study of Duchenne muscular dystrophy detection.

## 1 Introduction

It is well-known that model selection introduces additional uncertainty into the process of statistical modeling. Properties of estimators and tests subsequent to model selection depend on the way the model has been selected in addition to the stochastic nature of the chosen model. However, researchers engaged in applied statistical work usually only take into account the latter, and report estimates obtained from the chosen model as if they were unconditional estimates even though they are actually conditional. The consequences may be very serious - for example, what we believe to be a 95% confidence interval may actually be a 75% interval; a hypothesis tested at the nominal 5% level may in fact have been tested at a much higher level. Readers interested in post-model selection inference may consult work by Leeb and Pötscher (2003, 2005, 2006, 2008) and the references therein.

It has been argued by many that a simple way to overcome the under-reporting problem of model selection is by model averaging; rather than attaching to a single 'winning' model, a model average estimator weighs across many potential models. Model averaging techniques from a Bayesian perspective

*Ludwig Maximilian University of Munich, Department of Statistics, Akademiestr. 1, 80799 München, Germany
†City University of Hongkong, Department of Management Sciences, 83 Tat Chee Avenue, Kowloon, Hongkong

have been developed since the late 1970's, but were not widely used until recent advances in computing power facilitated their practical usage. A useful overview of this literature is given in Hoeting et al. (1999). Contributions from a frequentist perspective have been fewer, but recent studies of Buckland et al. (1997), Yang (2001, 2003), Hjort and Claeskens (2003, 2006), Yuan and Yang (2005), Leung and Barron (2006), Hansen (2007, 2008) and Buchholz et al. (2008) have made some important progress. Unlike the Bayesian approach for which prior probabilities for the potential models have to be specified, and computer intensive methods such as Markov Chain Monte Carlo are required for computing the posterior distribution, frequentist model averaging (FMA) can be implemented without much difficulty or protracted computations. One requirement of frequentist model averaging is the specification of model weights. Buckland et al. (1997) suggested exponential AIC weights, Hansen (2007, 2008) suggested weights based on minimizing a Mallows criterion, while Claeskens and Hjort (2008) proposed selecting model weights based on the Focused Information Criterion advocated in Claeskens and Hjort (2003).

This paper addresses frequentist model averaging when observations are partially missing. The situation of missing data is very common in many fields of statistics and there is a large collection of literature dealing with missing data from both practical and inference perspectives. When data are missing and the focus is on model selection, various studies have found that the use of conventional model selection criteria can lead to choices of inferior models. See Cavanaugh and Shumway (1998), Hens et al. (2006) and Claeskens and Consentino (2008) who studied the robustness of the AIC in the face of missing data. These authors also suggested modifications of the AIC that are applicable to different missing data circumstances and investigated the properties of the modified criteria. Extending the application of these criteria from model selection to model averaging is of both theoretical and practical interest in light of the increased attention model averaging has received in the recent literature. Here, we consider a model average estimation scheme that uses weights based on the weighted AIC (abbreviated as $\mathrm{AIC_W}$ hereafter) derived in Hens et al. (2006). This modification of the AIC is based on reweighing the complete observations by their inverse selection probabilities. The $\mathrm{AIC_W}$ has been shown to work quite well under a variety of simulation settings, although it has one obvious drawback in that it requires the estimation of selection probabilities.

When confronted with missing data, a common approach is to fill in the missing data by replacement values before analyzing the resultant complete data. This added dimension makes it necessary to consider the alternative scenario where model averaging is preceded by imputation of the missing data. We consider here three different imputation techniques, namely, the $k$-nearest neighbor (kNN) imputation (Chen and Shao (2000)), an EM algorithm-based imputation developed by King et al. (2001) and a generalization of the well-known regression imputation (Little and Rubin (2002)). With the kNN imputation method, a missing value is replaced by the mean of its $k$ nearest neighbors. King et al.'s (2001) method for imputation uses a bootstrap EM algorithm. Regression imputation, as is well-known, fills in the missing values as predictions from a regression. These are just some of the many approaches for handling missing data, but we draw on them here because they all have a very specific and attractive feature, namely, the computational algorithms for implementing these methods either have been programmed into existing software packages or can be easily programmed; the kNN imputation algorithm is available in the EMV package (Gottardo (2008)) for the $R$ programming environment; the Amelia II package for $R$ (Honaker et al. (2008)) is specifically designed to implement King et al.'s (2001) method; and steps for regression imputation can be easily programmed. As such, the findings of this

paper should appeal to practitioners who are more likely to use an imputation method included in the toolbox than a sophisticated method that requires considerable expertise and computational algorithms not readily available. We further adapt a simple and easy-to-use recursive algorithm to implement a generalized version of regression imputation in a way such that the missing values are predicted successively. We find this algorithm quite useful when one is confronted with two or more missing values simultaneously in a given row of observations. Readers interested in other more sophisticated imputation methods, many of which are in fact optimized for their particular applications, may consult Zhou et al. (2008) and the references therein.

In light of possible strategies discussed above, this paper seeks answers to the following important questions: 1) In the context of model combining, are we better or worse off by adopting a weighting mechanism based on modified criteria such as the $\mathrm{AIC_W}$, or by a procedure whereby the missing values are first replaced by imputed values before implementing model combining? 2) How do the properties of these model average estimators compared with their model selection counterparts? The nature of these questions precludes analytical answers, and in this paper we attempt to provide some insights by means of a Monte-Carlo study with attention restricted to a binary logistic regression model. As such, the approach adopted is quite modest and the nature of our investigation is preliminary, but even within this simple framework our results provide a good deal of insights into the aforementioned questions and are of direct relevance to applied statisticians engaged in empirical work. There inevitably remain many opportunities to extend and improve the current analysis, and it is hoped that this paper will pave the way for further investigations in this area.

The paper proceeds with a description of the model framework and the FMA estimator in Section 2. Section 3 is devoted to a discussion of missing data adjusted model selection criteria and imputation methods. We also introduce a recursive algorithm for handling two or more missing values simultaneously in conjunction with regression imputation. Section 4 deals with the design of the Monte-Carlo study and reports results of this study. Our experimental designs assume that the data are missing at random (MAR), that is, the probability of missingness depends only on the observed data and not the missing data. A real-data example from a study on Duchenne muscular dystrophy detection is presented in Section 5 followed by concluding remarks in Section 6.

## 2 Model framework and the FMA estimator

Consider the following "incomplete" $n \times (p+1)$ data matrix

$$D_* = \begin{pmatrix} y_1 & x_{11} & \dots & \dots & x_{1p} \\ \vdots & \vdots & * & & \vdots \\ \vdots & \vdots & & & * \\ \vdots & \vdots & & * & \vdots \\ y_n & x_{n1} & \dots & \dots & x_{np} \end{pmatrix}.$$

Denote by $y = (y_1, \dots, y_n)^t$ an $n \times 1$ vector of response values in a parametric regression setup, $X_j = (x_{1j}, \dots, x_{nj})^t$ an $n \times 1$ vector of values of the $j^{th}$ covariate, $j = 1, \dots, p$, and $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, an $1 \times p$ vector containing the $i^{th}$ observation of each of the $p$ covariates, $i = 1, \dots, n$. Some of the covariates $X_1, \dots, X_p$ contain missing observations while the response values $y_i$'s are fully observed.

Denote by $\mathcal{X}^*$ the set of covariates that contain missing values and $D_*^c \subseteq D_*$ the subsample of complete cases (CC) that contain only the fully observed rows.

Let $y_1, \ldots, y_n$ be independently distributed and denote by $f(y|\mathbf{x}_i; \beta, \sigma)$ the conditional density of $y|\mathbf{x}_i$, where $\beta$ is an unknown parameter vector and $\sigma$ is a scale parameter. Several recent studies (e.g., Hjort and Claeskens (2003)) adopt the framework where $\beta$ is partitioned into two parts, $\beta = (\theta, \gamma)$, such that $\theta$ and $\gamma$ contain the coefficients of the respective sets of covariates that we certainly want to include in the model and those that may potentially be included in the model. In many situations $\theta$ includes only the intercept term. For consistency one may also adopt the same framework here but the results to be reported do not depend on it. Now, let $\mathcal{M} = \{M_1, \ldots, M_k\}$ denote the set of all candidate models for describing $y$ based on varying combinations of $X_j$'s. A model selection procedure is one that singles out a "winning" model from the set $\mathcal{M}$ on the basis of a criterion $\Gamma$ (e.g., the AIC (Akaike (1973))). Typically, all subsequent estimation and inference are then conducted within this single chosen model. Model averaging, on the other hand, compromises across some or all of these candidate models. This results in the compromise estimator

$$\hat{\hat{\beta}} = \sum_{\kappa=1}^{k} w_\kappa \hat{\beta}_\kappa \tag{1}$$

based on the estimators $\hat{\beta}_\kappa$ for each of the candidate models belonging in $\mathcal{M}$. This compromise estimator may be called a FMA estimator if $\beta$ in each model is estimated by a frequentist principle such as maximum likelihood. It is readily seen that the frequentist model selection (FMS) estimator is a special case of $\hat{\hat{\beta}}$ by assigning a value of 1 to a particular $w_\kappa$ and 0 to all other $w_\kappa$'s. In regards to the weight choice in $\hat{\hat{\beta}}$, Buckland et al. (1997) proposed the following exponential AIC weights:

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

where $\mathrm{AIC}_\kappa$ is the AIC value of model $M_\kappa \in \mathcal{M}$. One may also construct weights based on values of the generalized cross validation (Golub et al. (1979)) or Focused Information Criterion (Claeskens and Hjort (2003)) scores or by minimizing a Mallows criterion as suggested in recent studies by Hansen (2007, 2008). An important limitation of Hansen's approach, however, is that the optimality properties regarding the Mallows criterion apply only in the context of linear regression and not elsewhere.

# 3 Missing data adjusted criteria and missing values imputation

## 3.1 Averaging based on missing data adjusted criteria

When observations are partially missing, it is seductive for the investigator to base his/her analysis only on the complete cases contained in the subsample $D_*^c$. Hens et al. (2006) showed that the naive use of AIC on the complete cases can lead to the selection of rather poor models and proposed a modification of the AIC that explicitly accounts for the fact that data are missing. One purpose of the present paper is to extend the application of Hens et al.'s (2006) criterion to model averaging. In this section we

briefly review Hens et al.'s (2006) approach. Consider $D_*$ and let

$$\delta_i = \begin{cases} 1, & \text{if the } i^{th} \text{ row } (y_i, \mathbf{x}_i) \text{ is observed completely} \\ 0, & \text{if the } i^{th} \text{ row } (y_i, \mathbf{x}_i) \text{ is not observed completely} \end{cases}$$

Further, denote by $\pi_i = P(\delta_i = 1)$ the probability of completely observing the $i^{th}$ row. Let the weight for the $i^{th}$ observation be

$$\tilde{w}_i = \frac{\delta_i}{\pi_i} \ . \tag{3}$$

Then the weighted AIC may be defined as

$$\text{AIC}_\text{W} \quad = \quad -2 \left\{ \sum_{i=1}^{n} \tilde{w}_i \, L_i(\hat{\beta}_W, \hat{\sigma}_W | \text{data}) \right\} + 2K \, , \tag{4}$$

where $L(\cdot)$ is the likelihood function, $\hat{\beta}_W$ and $\hat{\sigma}_W$ are the weighted maximum likelihood estimators and $K$ is the number of parameters to be estimated. Essentially the $\text{AIC}_\text{W}$ approach weighs each observation by its inverse selection probability. One may estimate $\tilde{w}_i$ nonparametrically using, for example, a generalized additive model with the smoothing parameter chosen through generalized cross validation. It is worth noting that weighted model selection criteria have also found applications in the literature of robust statistics. Interested readers may consult Ronchetti (1997), Agostinelli (2002) and the references therein. Now, Buckland et al.'s (1997) exponential AIC weight for weighting models as shown in (2) may be rephrased to fit the present framework, leading to

$$w_\kappa^{(A)} \quad = \quad \frac{\exp(-\frac{1}{2}\text{AIC}_{\text{W},\kappa})}{\sum_{\kappa=1}^{k} \exp(-\frac{1}{2}\text{AIC}_{\text{W},\kappa})} \, , \tag{5}$$

where $\text{AIC}_{\text{W},\kappa}$ is the $\text{AIC}_\text{W}$ of model $M_\kappa \in \mathcal{M} = \{M_1, \ldots, M_k\}$.

Other approaches for adjusting the AIC when data are incomplete include the approach proposed by Cavanaugh and Shumway (1998) and the EM-based AIC developed in Claeskens and Consentino (2008). Alternative FMA weights may be constructed by replacing $\text{AIC}_\text{W}$ in (5) by the model scores corresponding to these alternative criteria.

## 3.2 Averaging after imputation

An alternative to using values of $\text{AIC}_\text{W}$ or other missing data adjusted criteria as weights is to weigh the models based on scores of conventional model selection criteria but with the missing data replaced by imputed values prior to estimating the models. There is an extensive literature dealing with missing data imputation; in many cases, the procedures developed are mathematically intricate and computationally demanding. Here we limit our attention to some computationally simple and frequently used imputation methods, namely, regression imputation, kNN imputation and an EM algorithm-based imputation developed by King et al. (2001). As noted already, these methods all have the attractive feature of either having been programmed into existing software or can be programmed easily.

**(i)** Regression imputation refers to the type of methods whereby missing values are replaced by predicted values from a regression of the missing item on the items observed. Consider a missing value

$x_{ij} \in D_*$. In the simplest case one can use the subsample $D_*^c$ to fit the linear regression model

$$X_j = \theta + \gamma_0 y + \sum_{\substack{l=1 \\ l \neq j}}^{p} \gamma_l X_l + \epsilon, \quad \epsilon \sim N(0, \sigma^2). \tag{6}$$

Then, based on the parameter estimates $(\hat{\theta}, \hat{\gamma}_0, \ldots, \hat{\gamma}_{j-1}, \hat{\gamma}_{j+1}, \ldots, \hat{\gamma}_p)$ we can impute a value $\tilde{x}_{ij}$ for $x_{ij}$ via

$$\tilde{x}_{ij} = \hat{\theta} + \hat{\gamma}_0 y_i + \sum_{\substack{l=1 \\ l \neq j}}^{p} \hat{\gamma}_l x_{il}. \tag{7}$$

This very basic idea has two shortcomings; the first being that a dataset might include missing categorial, binary or count data such that imputation by linear regression would not be appropriate. Second, and more important, there is no prescribed modification to the method when one is confronted with two or more missing values simultaneously in a given row of $D_*$. One way to get around the first problem is to use an alternative modeling approach - for example, a generalized additive model (GAM). The second problem may be circumvented by a simple recursive algorithm that imputes the missing values successively. We describe the algorithm here in the context of estimation based on GAMs, and call our algorithm GAM based recursive imputation (GAMRI) which proceeds according to the following steps:

**Step 1**: Let $s$ be an integer and begin by setting $s = 1$.
**Step 2**: Denote by $D_s$ the subsample containing the rows with exactly $s$ missing values.
**Step 3**: Consider $X_j \in \mathcal{X}^*$ with the least amount of missing data in $D_s$.
**Step 4**: For a given missing value $x_{ij} \in X_j$, $i = 1, \ldots, n$, and $x_{ij} \in D_s$,

**a)** fit the GAM,

$$\eta(X_j) = \theta + f_0(y) + \sum_{\substack{l=1 \\ l \neq j,\, l \notin \Phi_l}}^{p} f_l(X_l) + \epsilon,$$

using the fully observed data in the set $D_*^c$, where $\Phi_l = \{l : x_{il} \text{ is missing}\}$ and $\eta(\cdot)$ is a proper link function (note: if $x_{ij}$ and $x_{kj}$ are both missing and $i < k$, then the imputation of $x_{ij}$ takes precedence over that of $x_{kj}$).

**b)** impute a new value $\tilde{x}_{ij}$ for $x_{ij}$ based on the estimated GAM

$$\eta(\tilde{x}_{ij}) = \hat{\theta} + \widehat{f}_o(y_i) + \sum_{\substack{l=1 \\ l \neq j,\, l \notin \Phi_l}}^{p} \widehat{f}_l(x_{il}),$$

where $\hat{\theta}$, $\hat{f}_o$ and $\hat{f}_l$ are estimates of $\theta$, $f_o$ and $f_l$.

**c)** Update $D_s$, $D_*^c$ and $\mathcal{X}^*$ by treating the imputed value as if it were an actual observation.

**Step 5**: Repeat from Step 3 until $D_s = \emptyset$.
**Step 6**: Set $s$ to $s + 1$.
**Step 7**: Repeat from Step 2 until $s$ has reached the maximum number of missing values in any row.
**Step 8**: Repeat Steps 1–7 until $\mathcal{X}^* = \emptyset$.

It is clear from steps 4a and 4b that the imputation model uses covariates with fully observed values. Moreover, step 4c of the algorithm updates the set of covariates that contain missing values and the subsample of complete cases that contain the fully observed rows after each imputation step; that is, after each imputation step the imputed value is considered to be part of the CC and used for fitting the next imputation model. Note that if $\mathbf{x}_i$, the vector containing the $i^{th}$ observation of each of the individual covariates, has two or more missing values, the algorithm first imputes only one of the missing values in the row and imputes the remaining missing values in the same row at subsequent stages as determined by the successive updating of $D_s$ in step 4c of the algorithm. For purposes of exposition consider the data from Table 1, where the variables $y$, $X_1$ and $X_2$ are continuous, while $X_3$ and $X_4$ contain count data and binary data respectively. Altogether five observations, namely, $x_{41}$, $x_{23}$, $x_{24}$, $x_{34}$ and $x_{54}$, (each represented by a "$*$" in Table 1) are missing from the dataset.

Table 1: Data for illustrating the GAMRI-algorithm

| $i$ | $y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|---|
| 1 | 1.2 | 24.2 | -8.2 | 7 | 0 |
| 2 | 2.3 | 23.4 | -3.2 | $*$ | $*$ |
| 3 | 2.2 | 30.0 | -0.1 | 9 | $*$ |
| 4 | 2.3 | $*$ | -4.1 | 10 | 1 |
| 5 | 2.6 | 22.0 | -0.8 | 7 | $*$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 100 | 1.7 | 26.4 | -2.9 | 7 | 1 |

According to steps 1 and 2, the algorithm begins by setting $s = 1$ and selecting the subsample $D_1$ that contains only the rows with one missing value (i.e., $D_1$ includes the $3^{rd}$, $4^{th}$ and $5^{th}$ row in Table 1). Now, according to step 3, the algorithm initially imputes the missing observation belonging to the variable with the least amount of missing data in $D_1$. In the current context, $X_1$ contains only one missing value whereas $X_4$ has two missing values in $D_1$, so the algorithm first imputes $x_{41}$ by fitting the model $\eta(X_1) = \theta + f_0(y) + f_2(X_2) + f_3(X_3) + f_4(X_4) + \epsilon$ based on the complete cases $D_*^c$ and the Gaussian link function as described in steps 4a and 4b. Now, by step 4c, the CC are updated and the imputed value $\tilde{x}_{41}$ is now treated as part of the CC. By Step 5, the algorithm next imputes $x_{34}$. Based on the GAM $\eta(X_4) = \theta + f_0(y) + f_1(X_1) + f_2(X_2) + f_3(X_3) + \epsilon$ with the logit link function to account for binary data we obtain an imputed value $\tilde{x}_{34}$. The algorithm then updates $D_s$, $D_*^c$ and $\mathcal{X}^*$ and proceeds with the imputation of $x_{54}$ as per the manner described for imputing $x_{34}$. With the missing values of $x_{41}$, $x_{34}$ and $x_{54}$ all replaced by imputed values, the subset $D_1$ reduces to an empty set. Thus, we set $s = 2$ (step 6), consider the subset $D_2$ which contains only the $2^{nd}$ row in Table 1 and proceed with the imputation of $x_{23}$, which entails fitting the model $\eta(X_3) = \theta + f_0(y) + f_1(X_1) + f_2(X_2) + \epsilon$ (without $X_4$, since $X_{24}$ is missing) with the log link function to account for the count data. The algorithm then includes the newly imputed value $\tilde{x}_{23}$ in the CC and updates $D_2$, which by now is an empty set since

there is only one missing value left, namely, $x_{24}$. As $s$ has reached the maximum number of missing value in any row, by step 8, the algorithm returns to step 1, where again $s$ is set to 1 and the algorithm imputes $x_{24}$ using a GAM with $y$, $X_1$, $X_2$, $X_3$ as covariates and the logit link function.

The R-package 'mgcv' (Wood (2006)) provides an excellent basis for the estimation of GAMs and the implementation of the GAMRI algorithm. However, to reduce computational burden one may consider using a generalized linear model (GLM) instead of a GAM. In this case the steps for the above algorithm remain the same except that a GLM instead of a GAM is used as the basis for analysis. We call this algorithm GLM based recursive imputation (GLMRI).

**(ii)** The kNN procedure (Chen and Shao (2000)) is another procedure to be considered as part of the proposed averaging after imputation strategy. The idea of this procedure is simple - based on the Euclidian distance or sample correlation one chooses $k$ rows that are nearest to the row that contains missing values; these $k$ rows must not contain any missing observation; the missing values in the row under consideration are then replaced by the average of the observations in these neighboring $k$ rows. The $R$ program for the kNN procedure is available in the EMV package (Gottardo (2008)).

**(iii)** The last imputation procedure to be considered is a new, bootstrap based version of the EM algorithm introduced recently by King et al. (2001). This method transforms the data in such a way that they can be treated as multivariate normal, i.e., $\tilde{D}_* \sim N(\mu, \Sigma)$, where $\tilde{D}_*$ is the sample of the transformed data. Then, the well-known EM algorithm is applied to bootstrap samples of the data $\tilde{D}_*$ in order to produce point estimates of $\mu$ and $\Sigma$ and use them as the basis for imputation in the original data. Finally, the data are retransformed to their original scale. The Amelia II package (Honaker et al. (2008)) for $R$ is specifically designed to implement this method and allows for multiple imputation. The implemented algorithm is usually fast, robust and reliable.

**Remark:** In general, the averaging after imputation scheme also allows for multiple imputation. Consider $M$ imputed sets of data. Then for each dataset $D^{(m)}$, $m = 1, \ldots, M$, one may construct some FMA estimator $\hat{\hat{\theta}}^{(m)}$ to obtain an "average" FMA estimator $\hat{\hat{\theta}}_{\mathrm{MI}} = 1/M \sum_{m=1}^{M} \hat{\hat{\theta}}^{(m)}$. Multiple imputations can be carried out by model based approaches such as the bootstrap EM-approach described in **(iii)** or by extending the GLMRI algorithm in **(i)**. The latter may be realized by adding an additional stochastic error term to the imputation model in step 4b and/or by not explicitly taking the fits $\hat{\gamma}_l$'s, but by random draws from their estimated distribution (see, for example, Schafer (1997) for details in the context of linear regression models). However, it is unclear as to what properties the resultant estimator possesses. Nevertheless, this warrants further research that goes beyond the scope of the current paper.

# 4 A Monte-Carlo study

As with all procedures developed it is desirable to investigate the properties of the procedures based on a finite amount of data. We do so here by means of a Monte-Carlo study with attention confined to the binary logistic regression model:

$$p_i = P(y_i = 1|\mathbf{x}_i) \quad = \quad F(\mu)\,,$$

where $y$ is a binary outcome, $\mu = \theta + \mathbf{x}_i^t\gamma$, $F(.) = 1/\{1 + \exp(-.)\}$, $\theta$ is the intercept coefficient, $\gamma$ is a vector of coefficients corresponding to the covariate vector $\mathbf{x}_i$.

Observations for the covariates are generated by the following distributions: $X_1 \sim \text{Exp}(0.2)$, $X_2 \sim \text{N}(0,1)$, $X_3 \sim \text{Unif}(0,10)$, $X_4 \sim \text{N}(10,2)$, and $X_5 \sim \text{Bin}(1,0.65)$. To model the dependency between the covariates we use a Clayton Copula with a copula parameter of 1.25 which indicates medium correlation among the covariates. The R-package 'copula' (Yan (2007)) provides an efficient tool to utilize this approach. Samples of 550 observations are generated with the first $n = 500$ observations used as training data and the remaining $n_{test} = 50$ observations as test data. Throughout the study we assume the data are MAR, which means the missingness mechanism depends only on the observed values. Although being more restrictive than non-ignorable missingness, MAR is also justified in many practical situations and there is a large collection of literature that adopts this missing data mechanism as a baseline for analysis. See Lu and Copas (2004) and Zhou et al. (2008) for recent examples. Our Monte-Carlo study considers the following FMA estimators:

1) **CC-FMA** estimator - the FMA estimator based on data of the complete cases in the subsample $D_*^c$, using exponential AIC weights as in (2);
2) **GAMRI-FMA**, **GLMRI-FMA**, **kNN-FMA** and **Amelia-FMA** estimators - FMA estimators based on the averaging after imputation principle using, respectively, the GAMRI, GLMRI, kNN and Amelia II imputation methods described in Section 3.2; these FMA estimators use exponential AIC weights as in (2);
3) **AIC$_{\text{W}}$-FMA** estimator - the FMA estimator that uses weights based on the AIC$_{\text{W}}$ missing data adjusted criterion defined in (5).

To facilitate comparability with model selection, we also consider the FMS counterparts of all of the FMA estimators described above. For each FMA estimator the corresponding FMS estimator is a special case of (1) by setting the $w_\kappa$ for the model with the best AIC score to 1 and all other $w_\kappa$'s to 0. In each instance the experiment is replicated $\mathcal{R} = 500$ times, with the same 500 samples used across all estimators to facilitate direct comparisons. The reported sampling properties include the mean square error (MSE) and mean absolute error (MAE) of estimators for estimating $\beta = (\theta, \gamma^t)^t$, the MSE and MAE of estimators for $\mu$ in the test sample and the error rate, defined as the probability of misclassifying an observation $y$ in the test sample; our classification of $y$ is based on a simple rule that $\hat{y}_i = 1$ if $\hat{\mu}_i \geq 0$ and $\hat{y}_i = 0$ if $\hat{\mu}_i < 0$, where $\hat{y}$ and $\hat{\mu}$ are respectively estimates of $y$ and $\mu$. These criteria may be written as follows:

$$L_1 = \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \sum_{j=1}^{p} (\hat{\beta}_j - \beta_j)^2 \right\}, \tag{8}$$

$$L_2 = \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \sum_{j=1}^{p} |\hat{\beta}_j - \beta_j| \right\}, \tag{9}$$

$$L_3 = \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\hat{\mu}_i^{(r)} - \mu_i^{(r)})^2 \right\}, \tag{10}$$

$$L_4 = \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |\hat{\mu}_i^{(r)} - \mu_i^{(r)}| \right\}, \tag{11}$$

and

$$L_5 = \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left\{ \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} I(\hat{y}_i^{(r)} \neq y_i^{(r)}) \right\}, \tag{12}$$

where $\hat{\beta}_j$, $j = 1, \ldots, p$, is the $j^{th}$ element of $\hat{\beta} = (\hat{\theta}, \hat{\gamma}^t)^t$, an estimator of $\beta$, $\beta_j$ is defined analogously, $\hat{\mu}_i^{(r)} = \hat{\theta} + \mathbf{x}_i^t \hat{\gamma}$ is an estimator of $\mu_i^{(r)}$, the true value of $\mu_i$ in the $r^{th}$ simulation run, and $n_{\text{test}}$ is the number of observations in the test sample.

**Experiment 1**. Our first experiment is based on the following parameter values of the coefficients: $\theta = 18$ and $\gamma = (0.3, 10, 0, -2.5, 0.25)$. Insofar, we consider realizations of the binary outcome variable depending on two strong factors $X_2$ and $X_4$ and two mild factors $X_1$ and $X_5$ via $y_i \sim \text{Bin}(1, p_i)$, $p_i = F(\theta + \mathbf{x}_i^t \gamma + \epsilon)$ and $\epsilon \sim N(0, 2)$. Observations of $X_2$, $X_4$ and $X_5$ in the generated samples are made missing at random (MAR) via the missing probability functions $\pi_{X_2}(X_1)$, $\pi_{X_4}(X_3)$ and $\pi_{X_5}(X_1)$, respectively:

$$\begin{aligned} \pi_{X_2}(X_1) &= 1 - \{0.005 \cdot (X_1 - 13)^2 + 1\}^{-1}, \quad \pi_{X_4}(X_3) = 1 - \{0.006 \cdot X_3^2 + 1\}^{-1}, \\ \pi_{X_5}(X_1) &= 1 - \{1 + \exp(1 - 0.35 \cdot (X_1 + 2))\}^{-1}. \end{aligned}$$

This results in about 30% missing values for $X_2$, 29% for $X_4$ and 14% for $X_5$. The FMA estimators we consider weigh over the following models:

$M_1$: $\ln \frac{p_i}{1-p_i} = \theta + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3} + \gamma_4 X_{i4} + \gamma_5 X_{i5}$

$M_2$: $\ln \frac{p_i}{1-p_i} = \theta + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_4 X_{i4} + \gamma_5 X_{i5}$

$M_3$: $\ln \frac{p_i}{1-p_i} = \theta + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_4 X_{i4}$

$M_4$: $\ln \frac{p_i}{1-p_i} = \theta + \gamma_2 X_{i2} + \gamma_4 X_{i4} + \gamma_5 X_{i5}$

$M_5$: $\ln \frac{p_i}{1-p_i} = \theta + \gamma_2 X_{i2} + \gamma_3 X_{i3} + \gamma_4 X_{i4}$

$M_6$: $\ln \frac{p_i}{1-p_i} = \theta + \gamma_2 X_{i2} + \gamma_4 X_{i4}$

Table 2 reports the results under the current setting. It is apparent that in this experiment averaging after imputation using GAMRI or GLMRI is strongly favored relative to all other averaging approaches.

Regardless of the criterion of comparison, the GAMRI-FMA and GLMRI-FMA estimators are uniformly superior, and often by a large margin, to all other FMA estimators under consideration. Overall, the differences between the GAMRI-FMA and GLMRI-FMA estimators are quite small across all five performance criteria, with the GLMRI approach being a marginally better choice than GAMRI. The FMA estimator that uses the missing data adjusted criterion, $\text{AIC}_\text{W}$, performs markedly worse than the two regression imputation based FMA's. There is generally not much difference in the performance of the $\text{AIC}_\text{W}$-FMA estimator and the FMA that uses the complete data only. Interestingly, the remark regarding the superiority of the imputation based FMA estimators relative to the $\text{AIC}_\text{W}$-FMA estimator would be substantially altered, and generally reversed, if comparison has been made between the latter estimator and the imputation based FMA estimators that use the kNN or Amelia II methods. The kNN-FMA estimator is habitually the worst performer followed by the Amelia-FMA estimator. These inference comparisons among the estimators generally hold also for the model selection estimators, with

Table 2: Monte-Carlo results for Experiment 1. The table presents the losses $L_1$-$L_5$ for different FMA- and FMS-estimators: 1) the estimator based on the complete cases and exponential AIC weights, 2) the estimators based on the averaging after imputation principle using the GAMRI, GLMRI, kNN and Amelia II imputation methods and exponential AIC weights, and 3) the estimator that uses the weights based on $\text{AIC}_\text{W}$ as defined in (5).

| | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ |
|---|---|---|---|---|---|
| | Frequentist Model Averaging (FMA) | | | | |
| 1) CC | 37.2660 | 8.6294 | 55.6963 | 6.3516 | 0.2633 |
| 2) GAMRI | 12.2505 | 5.1376 | 18.0860 | 3.3074 | 0.1158 |
| GLMRI | 11.1168 | 4.9395 | 16.2934 | 3.2109 | 0.1102 |
| kNN | 145.7839 | 18.2297 | 227.1061 | 14.0487 | 0.7236 |
| Amelia | 80.7745 | 13.2958 | 120.2725 | 9.9809 | 0.4960 |
| 3) $\text{AIC}_\text{W}$ | 36.2549 | 8.6658 | 53.0029 | 6.1309 | 0.2474 |
| | Frequentist Model Selection (FMS) | | | | |
| 1) CC | 36.8990 | 8.5521 | 54.4415 | 6.2660 | 0.2588 |
| 2) GAMRI | 12.2569 | 5.1205 | 17.6145 | 3.2520 | 0.1136 |
| GLMRI | 11.5614 | 5.0041 | 16.4461 | 3.1990 | 0.1076 |
| kNN | 145.0149 | 18.2209 | 226.1434 | 14.0171 | 0.7235 |
| Amelia | 80.6955 | 13.3465 | 119.9025 | 9.9553 | 0.4954 |
| 3) $\text{AIC}_\text{W}$ | 36.3040 | 8.7655 | 53.0203 | 6.1598 | 0.2488 |

the GLMRI-FMS estimator yielding the best estimates followed closely by the GAMRI-FMS estimator across all criteria. The results in Table 2 also indicate that when comparing model combining with model selection, model combining is often better but sometimes worse than model selection, although the improvement in accuracy offered by combining models over selecting a model is typically indiscernible, *ceteris paribus*.

**Experiment 2**. The basic setting of this experiment is the same as that of the last experiment except for a variation in the the amount of missing values. We now consider the missing probability functions

$$\pi_{X_2}(X_1) = 1 - \{1 + \exp(1 - 0.6 \cdot (X_1 + 2))\}^{-1},\ \pi_{X_5}(X_3) = 1 - \{0.006 \cdot X_3^2 + 1\}^{-1}$$

and turn values of $X_2$ missing via $\pi_{X_2}(X_1)$ and values of $X_5$ missing via $\pi_{X_5}(X_3)$. This results in about 15% of missing values for $X_2$ and 13% for $X_5$. Values of $X_4$ are assumed to be fully observed. The results for this experiment are shown in Table 3.

Qualitatively, the results are very similar to those of the last experiment but there are also differences. As far as the ranking of the estimators is concerned, all of the general comments made under Experiment 1 continue to apply in broad terms. An exception is that the GLMRI-FMA estimator does not uniformly dominate the GAMRI-FMA estimator across all evaluation criteria under the current setting; the latter estimator is found to have a slight edge over the former in terms of $L_5$, the error rate. It is evident that for both model combining and model selection, these two regression imputation based approaches continue to offer substantial advantages over all other approaches. On the other hand, contrary to the results of the last experiment, the FMA and FMS estimators based on the Amelia II imputation method were found to provide slightly more accurate estimates than do the corresponding

Table 3: Monte-Carlo results for Experiment 2. The table presents the losses $L_1$-$L_5$ for different FMA- and FMS-estimators: 1) the estimator based on the complete cases and exponential AIC weights, 2) the estimators based on the averaging after imputation principle using the GAMRI, GLMRI, kNN and Amelia II imputation methods and exponential AIC weights, and 3) the estimator that uses the weights based on $\text{AIC}_\text{W}$ as defined in (5).

| | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ |
|---|---|---|---|---|---|
| | Frequentist Model Averaging (FMA) | | | | |
| 1) CC | 48.3463 | 9.7190 | 74.1622 | 7.7545 | 0.3096 |
| 2) GAMRI | 26.5638 | 7.2170 | 44.4934 | 5.6660 | 0.1872 |
| GLMRI | 26.1480 | 6.9933 | 42.6427 | 5.4362 | 0.1940 |
| kNN | 71.1478 | 12.6733 | 126.3754 | 10.1841 | 0.4698 |
| Amelia | 44.6938 | 9.6504 | 71.0355 | 7.4703 | 0.3166 |
| 3) $\text{AIC}_\text{W}$ | 42.3117 | 9.1562 | 65.0313 | 7.1745 | 0.2762 |
| | Frequentist Model Selection (FMS) | | | | |
| 1) CC | 49.2513 | 9.9589 | 75.7272 | 7.8353 | 0.3090 |
| 2) GAMRI | 27.2321 | 7.4196 | 45.7310 | 5.7447 | 0.2042 |
| GLMRI | 26.9721 | 7.2563 | 44.1511 | 5.5489 | 0.2024 |
| kNN | 71.8526 | 12.8576 | 127.8945 | 10.2463 | 0.4854 |
| Amelia | 45.4797 | 9.8019 | 72.4092 | 7.5481 | 0.3154 |
| 3) $\text{AIC}_\text{W}$ | 45.1303 | 9.6245 | 68.9807 | 7.4717 | 0.2836 |

CC estimators in most situations. Under the current setting the $\text{AIC}_\text{W}$ based estimators have better performance than the estimators based on complete cases only, while the kNN based estimators remain the least preferred in both the categories of model combining and model selection. Other things being equal, model combining appears to offer some slight gains over model selection in terms of all evaluation criteria.

**Experiment 3**. This experiment is based on a modification of the original Experiment 1 through the introduction of a new covariate $X_2^2$ and the following alternative values for the coefficients: $\theta = -7$ and $\gamma = (0, -2, -2, 0, 0, 1.8)$, where 1.8, the last value in $\gamma$, is the coefficient corresponding to $X_2^2$. The missing probability functions are the same as in Experiment 1, and we consider the following seven candidate models:

$$M_1\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_2 X_{i2} + \gamma_3 X_{i3} + \gamma_6 X_{i2}^2 \qquad M_5\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_2 X_{i2} + \gamma_3 X_{i3}$$
$$M_2\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3} \qquad M_6\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_2 X_{i2}$$
$$M_3\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_2 X_{i2} + \gamma_3 X_{i3} + \gamma_4 X_{i4} \qquad M_7\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_3 X_{i3}$$
$$M_4\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_2 X_{i2} + \gamma_3 X_{i3} + \gamma_5 X_{i5}$$

Results for this experiment, as displayed in Table 4, alter the conclusions reached under the two previous experiments to some extent. First, it is found that under the current setting, performance differences among estimators are far less pronounced than in the previous settings. Another noticeable difference between results reported in Table 4 and those reported for the previous settings is that the two regression imputation approaches GAMRI and GLMRI are no longer the uniformly preferred strategies under Experiment 3. Whether used in model averaging or selection, GAMRI and GLMRI are dominated

by the Amelia II based estimators in terms of performance evaluation criteria $L_1$ and $L_2$, and by the kNN and CC based estimators in terms of $L_2$. Interestingly, while the kNN approach is rated relatively favorably in terms of criterion $L_2$, it produces the worst estimates among all strategies considered in terms of $L_3$, $L_4$ and $L_5$ for both model averaging and selection. Also, the $\text{AIC}_\text{W}$-FMA (FMS) estimator is always dominated by the FMA (FMS) estimators based on one of the imputation approaches. In the current setting, model selection estimators always yield slightly less accurate estimates than do their respective model averaging counterparts, *ceteris paribus*.

Table 4: Monte-Carlo results for Experiment 3. The table presents the losses $L_1$-$L_5$ for different FMA- and FMS-estimators: 1) the estimator based on the complete cases and exponential AIC weights, 2) the estimators based on the averaging after imputation principle using the GAMRI, GLMRI, kNN and Amelia II imputation methods and exponential AIC weights, and 3) the estimator that uses the weights based on $\text{AIC}_\text{W}$ as defined in (5).

| | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ |
|---|---|---|---|---|---|
| | Frequentist Model Averaging (FMA) | | | | |
| 1) CC | 8.0273 | 4.1146 | 12.8297 | 2.8611 | 0.2148 |
| 2) GAMRI | 7.3940 | 4.3102 | 11.0110 | 2.5597 | 0.1716 |
| GLMRI | 7.1925 | 4.4497 | 11.1249 | 2.5646 | 0.1682 |
| kNN | 7.6102 | 4.2021 | 14.4072 | 3.1124 | 0.2240 |
| Amelia | 5.7252 | 4.1982 | 12.0097 | 2.8081 | 0.1838 |
| 3) $\text{AIC}_\text{W}$ | 8.9510 | 4.5134 | 12.7359 | 2.8820 | 0.2066 |
| | Frequentist Model Selection (FMS) | | | | |
| 1) CC | 8.7400 | 4.1969 | 13.5728 | 2.9661 | 0.2200 |
| 2) GAMRI | 7.8313 | 4.4091 | 11.3237 | 2.6002 | 0.1744 |
| GLMRI | 7.9062 | 4.5975 | 11.7556 | 2.6593 | 0.1722 |
| kNN | 7.9200 | 4.2112 | 14.7734 | 3.1520 | 0.2258 |
| Amelia | 6.3072 | 4.3393 | 12.4974 | 2.8668 | 0.1886 |
| 3) $\text{AIC}_\text{W}$ | 9.1307 | 4.5586 | 12.8334 | 2.8926 | 0.2060 |

# 5 An illustrative example

In this section we perform model averaging and selection with the methods described in the previous paragraphs for a dataset concerning the detection of Duchenne Muscular Dystrophy (DMD) carriers in a sample of female patients. These data are taken from Andrews and Herzberg (1985) and the same data were used in studies by Tibsharani and Hinton (1998) and Zhou et al. (2008). DMD is a genetically transmitted disease that passes from a mother to her children. In general, only male offsprings can be afflicted by DMD, while female offsprings can be carriers but do not show any apparent symptoms of the disease. DMD carriers tend to exhibit elevated levels of certain serum enzymes or proteins. The data consist of 209 records of female patients, among whom 75 are carriers and 134 are non-carriers. The response variable $y$ is a 0/1 variable where 1 indicates that the patient is a DMD carrier. The covariates comprise $AGE$, the age of the patient, and the serum markers creatine kinase, hemopexin, pyruvate kinase, and lactate dehydroginase, denoted as $CK$, $H$, $PK$ and $LD$ respectively. The serum markers $CK$ and $H$ may be measured rather inexpensively from frozen serum, whereas $PK$ and $LD$

require fresh serum. Observations are missing for 7 values of $LD$ and 8 values of $PK$. Here, the MAR assumption appears reasonable, because the lack of data is caused by the design of the study.

The purpose of the present investigation is to devise a model that appropriately describes the risk of a female patient's being a DMD-carrier and allows predictions for future patients based on their levels of the serum markers. We consider the following five competing logistic regression models:

$$M_1\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_1\text{AGE} + \gamma_2\text{CK} + \gamma_3\text{H} + \gamma_4\text{PK} + \gamma_5\text{LD} + \gamma_6\text{CK} \times \text{H} + \gamma_7\text{PK} \times \text{LD}$$
$$M_2\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_1\text{AGE} + \gamma_2\text{CK} + \gamma_3\text{H} + \gamma_4\text{PK} + \gamma_5\text{LD} + \gamma_6\text{CK} \times \text{H}$$
$$M_3\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_1\text{AGE} + \gamma_2\text{CK} + \gamma_3\text{H} + \gamma_4\text{PK} + \gamma_5\text{LD} + \gamma_7\text{PK} \times \text{LD}$$
$$M_4\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_1\text{AGE} + \gamma_2\text{CK} + \gamma_3\text{H} + \gamma_4\text{PK} + \gamma_5\text{LD}$$
$$M_5\text{: } \ln \tfrac{p_i}{1-p_i} = \theta + \gamma_1\text{AGE} + \gamma_2\text{CK} + \gamma_3\text{H}$$

Maximum likelihood estimates of the coefficients of the model selected by AIC are shown in the row labeled "original data" in the lower panel of Table 5. Note that these estimates are based on the 194 fully observed records out of a total of 209 records. The AIC selects the model $M_2$ that includes all regressors except the interaction term $PK \times LD$. The corresponding FMA estimates based on exponential AIC weights as in (2), are given in the top row of the upper panel of the same table. We observe that based on the complete cases of the original data, estimates produced by model selection and model averaging have the same signs and, except for the coefficient of the intercept term, are very close in magnitude. To increase the realism of the study, we increase the amount of missing data of $LD$ and $PK$ in the dataset via the probability function

$$\pi_{LD,PK}(y) \quad = \quad \{3 + \exp(\tilde{y} + 0.5(\text{sign}(\tilde{y}) + 1) + \epsilon)\}^{-1}\,,$$

where $\tilde{y} = y - 0.5$ and $\epsilon \sim N(0, 0.5)$. This results in 22% of missing values for $PK$ and about 23% for $LD$. We then estimate the coefficients based on the FMA and FMS strategies described in the previous section. The standard errors of the estimates are calculated via the following approximation formula given in Buckland et al. (1997)

$$s.e.(\hat{\gamma}_{j,\text{FMA}}) \quad = \quad \sum_{\kappa=1}^{k} w_\kappa \sqrt{Var(\hat{\gamma}_{j,\text{FMS}(\kappa)}) + (\hat{\gamma}_{j,\text{FMS}(\kappa)} - \hat{\gamma}_{j,\text{FMA}})^2}$$

where $\hat{\gamma}_{j,\text{FMA}}$ is the FMA estimator of $\gamma_j$ and $\hat{\gamma}_{j,\text{FMS}(\kappa)}$ is the corresponding FMS estimator based on the $\kappa^{th}$ model. As Hjort and Claeskens (2003) remarked, this formula is not 100% correct, but it is easy to program and gives reasonably accurate standard errors that account for the additional uncertainty due to model selection. The estimation results are reported in Table 5.

Table 5 reveals some interesting disagreements in the results produced by the different methods, most notably in the signs of the coefficient estimates of some of the serum markers. Except for the coefficient of the intercept term, the CC and $\text{AIC}_W$ based FMA and FMS estimators produce estimates that possess the same signs as the corresponding estimates based on the original data, while such is not the case for the GAMRI, GLMRI, kNN and Amelia based estimators. The standard errors of the FMA estimators are invariably larger than those of the corresponding FMS estimators, as is expected because standard errors reported for the FMS estimators are really conditional errors, whereas FMA standard errors are unconditional and account for the uncertainty due to model selection.

Table 5: FMA- and FMS- coefficient estimates (standard errors in brackets) for the illustrative example.

| | Intercept | | AGE | | CK | | H | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Frequentist Model Averaging (FMA) | | | | | |
| Original Data | −9.097 | (8.342) | 0.162 | (0.047) | −0.243 | (0.175) | −0.061 | (0.089) | ··· |
| 1) CC | 1.854 | (9.320) | 0.171 | (0.066) | −0.662 | (0.272) | −0.203 | (0.123) | ··· |
| 2) GAMRI | −13.827 | (3.312) | 0.175 | (0.046) | 0.024 | (0.027) | −0.006 | (0.024) | ··· |
| GLMRI | −13.411 | (3.332) | 0.173 | (0.046) | 0.021 | (0.030) | −0.008 | (0.026) | ··· |
| kNN | −13.057 | (3.269) | 0.173 | (0.043) | 0.021 | (0.026) | −0.007 | (0.023) | ··· |
| Amelia | −13.712 | (3.099) | 0.181 | (0.045) | 0.032 | (0.027) | 0.000 | (0.024) | ··· |
| 3) $\mathrm{AIC_W}$ | 2.527 | (9.306) | 0.173 | (0.066) | −0.679 | (0.267) | −0.211 | (0.120) | ··· |
| | | | | Frequentist Model Selection (FMS) | | | | | |
| Original Data | −5.132 | (6.708) | 0.166 | (0.047) | −0.297 | (0.160) | −0.088 | (0.081) | ··· |
| 1) CC | 1.193 | (7.572) | 0.173 | (0.066) | −0.657 | (0.245) | −0.202 | (0.112) | ··· |
| 2) GAMRI | −13.943 | (2.361) | 0.173 | (0.045) | 0.035 | (0.013) | 0.002 | (0.014) | ··· |
| GLMRI | −13.492 | (2.276) | 0.170 | (0.045) | 0.035 | (0.012) | 0.002 | (0.015) | ··· |
| kNN | −13.082 | (2.212) | 0.170 | (0.042) | 0.031 | (0.012) | 0.001 | (0.014) | ··· |
| Amelia | −14.396 | (2.432) | 0.178 | (0.044) | 0.043 | (0.012) | 0.008 | (0.014) | ··· |
| 3) $\mathrm{AIC_W}$ | 1.193 | (7.572) | 0.173 | (0.066) | −0.657 | (0.245) | −0.202 | (0.112) | ··· |

| | PK | | LD | | PK×LD | | CK×H | |
|---|---|---|---|---|---|---|---|---|
| | | | | Frequentist Model Averaging (FMA) | | | | |
| Original Data | ··· | 0.227 | (0.177) | 0.020 | (0.014) | −0.001 | (0.001) | 0.003 | (0.002) |
| 1) CC | ··· | 0.045 | (0.191) | 0.026 | (0.016) | 0.000 | (0.001) | 0.009 | (0.003) |
| 2) GAMRI | ··· | 0.188 | (0.123) | 0.019 | (0.010) | 0.000 | (0.000) | 0.000 | (0.000) |
| GLMRI | ··· | 0.177 | (0.124) | 0.020 | (0.010) | 0.000 | (0.000) | 0.000 | (0.000) |
| kNN | ··· | 0.176 | (0.122) | 0.019 | (0.010) | 0.000 | (0.001) | 0.000 | (0.000) |
| Amelia | ··· | 0.115 | (0.075) | 0.015 | (0.007) | 0.000 | (0.000) | 0.000 | (0.000) |
| 3) $\mathrm{AIC_W}$ | ··· | 0.037 | (0.207) | 0.026 | (0.016) | 0.000 | (0.001) | 0.009 | (0.003) |
| | | | | Frequentist Model Selection (FMS) | | | | | |
| Original Data | ··· | 0.115 | (0.048) | 0.011 | (0.006) | 0.000 | (0.000) | 0.004 | (0.002) |
| 1) CC | ··· | 0.085 | (0.105) | 0.028 | (0.012) | 0.000 | (0.000) | 0.009 | (0.003) |
| 2) GAMRI | ··· | 0.155 | (0.068) | 0.017 | (0.006) | 0.000 | (0.000) | 0.000 | (0.000) |
| GLMRI | ··· | 0.133 | (0.064) | 0.017 | (0.006) | 0.000 | (0.000) | 0.000 | (0.000) |
| kNN | ··· | 0.138 | (0.052) | 0.016 | (0.005) | 0.000 | (0.000) | 0.000 | (0.000) |
| Amelia | ··· | 0.120 | (0.046) | 0.016 | (0.005) | 0.000 | (0.000) | 0.000 | (0.000) |
| 3) $\mathrm{AIC_W}$ | ··· | 0.085 | (0.105) | 0.028 | (0.012) | 0.000 | (0.000) | 0.009 | (0.003) |

Note that there is no unitary set of data that allows for a fair comparison of all FMA and FMS estimators in terms of the estimated error rate because the CC and $\mathrm{AIC}_W$ based estimators are based on the dataset $D_*^c$, while the imputation based estimators are based on the imputed dataset. Thus, to compare these estimators, we consider the squared error loss

$$L_6 = \sum_{j=1}^{7} (\hat{\gamma}_{j,\mathrm{od}} - \hat{\gamma}_{j,\mathrm{FM}})^2 , \tag{13}$$

where $\hat{\gamma}_{j,\mathrm{od}}$ is the maximum likelihood estimate of $\gamma_j$ based on the complete cases of the original data, and $\hat{\gamma}_{j,\mathrm{FM}}$ is the corresponding estimate based on a given FMA or FMS strategy that accounts for the missing observations. Table 6 presents a comparison of the various strategies based on this criterion.

Table 6: Comparison of the FMA- and FMS-estimator performance based on $L_6$

| | CC | GAMRI | GLMRI | kNN | Amelia | $\mathrm{AIC}_W$ |
|---|---|---|---|---|---|---|
| FMA | 0.2285 | 0.0761 | 0.0752 | 0.0756 | 0.0925 | 0.2483 |
| FMS | 0.1437 | 0.1200 | 0.1186 | 0.1164 | 0.1250 | 0.1437 |

It is clear that the four imputation based strategies have the edge over the CC and $\mathrm{AIC}_W$ in both model combining and model selection. In other words, the imputation based FMA and FMS methods produce closer estimates to the maximum likelihood estimates than do the corresponding CC and $\mathrm{AIC}_W$ methods. Interestingly, with the imputation based methods, we are gaining accuracy by combining models, whereas with the CC and $\mathrm{AIC}_W$ based strategies we are better off with selecting a single model.

# 6 Concluding remarks

The main aim of the current paper is to explore the following questions: 1) When observations are partially missing, are we better or worse off by adopting a model selection criterion that explicitly adjusts for data missingness, or by a procedure which replaces the missing data by imputed values prior to model selection or averaging? 2) Is model averaging necessarily superior to selecting a single model when observations are missing? In regards to the second question, we have found, based on the Monte-Carlo settings and dimensions of performance considered above, that there generally can be some small gains in estimation efficiency by adopting model combining rather than selecting a single model. Whether these small efficiency gains are enough to justify the additional computational efforts involved in model averaging is a highly subjective matter. As for the first question, the answer is not as clear cut since it depends on the choice of the imputation method as well as the data. Indeed, one of the notable features of our investigation is the extent to which the imputation mechanism affects the results. As noted already, the GAMRI and GLMRI based FMA and FMS estimators generally perform well relative to the corresponding estimators that adopt the criterion $\mathrm{AIC_W}$, but the performance of the kNN and Amelia II imputation methods based estimators can vary considerably across the experimental settings and performance criteria. The $\mathrm{AIC_W}$-FMA and FMS estimators generally yield more accurate estimates than do the corresponding complete cases estimators, though often the gains in efficiency are not as significant as one would have hoped for.

It is worth mentioning that as with all Monte Carlo experiments, the results we have reported are tentative, and care must be exercised in attempting to generalize our conclusions to cases other than those investigated here. We have limited attention to only very few of the possible model average estimators that could be considered in this context, and it would be interesting to broaden the analysis to include the FMA and FMS estimators based on, for example, the EM-based AIC developed in Claeskens and Consentino (2008), as well as model frameworks other than logistic regression. Our study has been limited further in the range of the imputation methods for missing data that we have

investigated. Mindful of the intended audience (mainly practitioners), our approach concentrates on some of the most computationally convenient imputation methods. It will be interesting to include other more sophisticated imputation techniques (e.g., Zhou et al. (2008)) in the investigation. Despite these limitations, this study has offered some interesting insights into some very practical questions on model averaging that certainly warrant further studies.

# References


Agostinelli, C., 2002. Robust model selection in regression via weighted likelihood methdology. Statistics and Probability Letters 64, 583–639.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. Proceeding of the Second International Symposiumon Information Theory Budapest, 267–281.

Andrews, D. F., Herzberg, A. M., 1985. Data: A collection of problems from many fields for the student and research worker. Springer, New York.

Buchholz, A., Holländer, N., Sauerbrei, W., 2008. On properties of predictors derived with a two-step bootstrap model averaging approach - a simulation study in the linear regression model. Computational Statistics and Data Analysis 52, 2778–2793.

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.

Chen, J., Shao, J., 2000. Nearest neighbor imputation for survey data. Journal of Official Statistics 16, 113–131.

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.

Claeskens, G., Hjort, N. L., 2008. Minimising average risk in regression models. Econometric Theory 24, 493–527.

Golub, G. H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21, 215–223.

Gottardo, R., 2008. EMV: Estimation of missing values for a data matrix. R package version 1.3.1.

Hansen, B. E., 2007. Least squares model averaging. Econometrica 75, 1175–1189.

Hansen, B. E., 2008. Least squares forecast averaging. Journal of Econometrics 146, 342–350.

Hens, N., Aerts, M., G., M., 2006. Model selection for incomplete and design based samples. Statistics in Medicine 25, 2502–2520.

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., Blackwell, M., 2008. Amelia 2: A program for missing data. R Package version 1.1–33, `http://gking.harvard.edu/amelia`.

King, G., Honaker, J., Joseph, A., Scheve, K., 2001. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. American Political Science Review 95, 49–69.

Leeb, H., Pötscher, B. M., 2003. The finite sample distribution of post-model-selection estimators and uniform versus non-uniform approximations. Econometric Theory 19, 100–142.

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.

Leung, G., Barron, A. R., 2006. Information theory and mixing least squares regressions. IEEE Transaction on Information Theory 52, 3396–3410.

Little, R., Rubin, D., 2002. Statistical analysis with missing data. Wiley, New York.

Lu, G., Copas, J. B., 2004. Missing at random, likelihood ignorability and model completeness. Annals of Statistics 32, 754–765.

Ronchetti, E., 1997. Robustness aspects of model choice. Statistica Sinica 7, 327–338.

Schafer, J. L., 1997. Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

Tibsharani, R., Hinton, G., 1998. Coaching variables for regression and classification. Statistics and Computing 8, 25–33.

Wood, S. N., 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Yan, J., 2007. Enjoy the joy of copulas: with package copula. Journal of Statistical Software 21, 1–21.

Yang, Y., 2001. Adaptive regression by mixing. Journal of the American Statistical Association 96, 574–586.

Yang, Y., 2003. Regression with multiple candidate models: selecting or mixing. Statistica Sinica 13, 783–809.

Yuan, Z., Yang, Y., 2005. Combining linear regression models: when and how. Journal of the American Statistical Association 100, 1202–1214.

Zhou, Y., Wan, A. T. K., Wang, X., 2008. Estimating equations inference with missing data. Journal of the American Statistical Association 103, 1187–1199.