KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Missing covariates in logistic regression, estimation and distribution selection

Fabrizio Consentino and Gerda Claeskens

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0810

Missing Covariates in Logistic Regression, Estimation and Distribution Selection

Fabrizio Consentino and Gerda Claeskens

K.U. Leuven

ORSTAT and Leuven Statistics Research Center

Naamsestraat 69, 3000 Leuven, Belgium

Fabrizio.Consentino@econ.kuleuven.be; Gerda.Claeskens@econ.kuleuven.be

May 4, 2009

Abstract

We derive explicit formulae for estimation in logistic regression models where some of the covariates are missing. Our approach allows for modeling the distribution of the missing covariates either as a multivariate normal or multivariate t-distribution. A main advantage of this method is that it is fast and does not require the use of iterative procedures. A model selection method is derived which allows to choose amongst these distributions. In addition we consider versions of AIC that are based on the EM algorithm and on multiple imputation methods that have a wide applicability to model selection in likelihood models in general.

Keywords: Akaike information criterion, likelihood model, logistic regression, missing covariates, model selection, multiple imputation, t-distribution.

1 Introduction

The problem of missing data is quite common in statistical analysis, affecting most scientific fields. The main issue consists in the difficulty of dealing with the missingness. In the literature, there are different approaches; the simplest (though most naive) method is that of using the complete cases only by discarding all items with missing observations from the dataset. Then two important problems arise: first of all, information is lost, since the original sample size is reduced, which in some cases can be significantly high. Second, if the missingness depends on the data the results may be biased, depending on the missingness mechanism (Lipsitz et al., 1998; Little and Rubin, 2002).

The dataset considered for discussion is the European Values Study (EVS), obtained from the study catalogue ZACAT, a social science data portal from the University of Cologne. The EVS is carried out under the responsibility of the European Values Study Foundation and it represents a large-scale, cross-national and longitudinal survey research program; its scope is to explore important social value patterns in order to analyze similarities and differences in Europe. Representative national samples were interviewed using uniformly structured questionnaires to enable generalization and comparison in 33 European countries and were drawn from the population of citizens over 18 years of age. The data are based on the third wave analysis of $1999-2001$. We focus on the data related to Belgium (Flemish, French and Brussels communities) and take the following variables into account. The considered dataset has 1603 observations and 6 variables. The outcome variable is binary where $Y = 0$ indicates that the person is not satisfied with his/her job hours, while $Y = 1$ indicates satisfaction with job hours; it is completely observed. Other variables are x_1: the age when the education was completed; x_2: gender, using 1 for male and 2 for female; x_3: job payment, using 1 if they are paid and 2 if not; x_4: education levels, using 1 for primary education, 2 for secondary and 3 for post-secondary; x_5: socio-economic status, using 1 for upper-class, 2 for middle-class and 3 for manual workers. Variable x_1 contains missing values for 56 out of the 1603 cases,

the other variables do not contain missing observations. An additional difficulty is that x_1 seems to come from a distribution with heavier tails than the normal distribution.

There is an extensive literature on maximum likelihood estimation in the context of missing observations. One of the most popular tools is the expectation-maximization (EM) algorithm introduced by Dempster et al. (1977). The EM algorithm provides an efficient way of estimation in incomplete data problems, because it relates maximum likelihood estimation of the incomplete data to maximum likelihood estimation based on the complete data. This is for example used in the method of weights (Ibrahim, 1990) that is used to fit logistic regression models with missing covariates (see also Ibrahim et al., 1999a,b). The addition of the Gibbs sampler and an adaptive rejection algorithm for sampling from the distribution of the missing data given the observed data makes the estimation method computationally intensive, since it requires (often many) iterations. Since our intention is to perform model selection with missing covariate data, complexity and computation time are highly important since possibly many models will be fit to the data.

In this paper we illustrate the use of the EM algorithm for model selection and we provide alternatives to its use. First we look at the special case of logistic regression models. Here it is possible to perform fast, non-iterative calculations. For more general situations, we explain how to use model selection in combination with multiple imputation.

For logistic regression models with incomplete data Gao and Hui (1997), building further on Blackhurst and Schluchter (1989), proposed a maximum likelihood estimation approach where no iterations are needed. They assume that there is either a single normally distributed covariate that contains missing values, or more than one missing covariate with a monotone pattern. Since in data sets the covariates often might have heavier tailed distributions, we extend this estimation approach to univariate and multivariate t-distributions. The method does not require any iteration, which speeds up the estimation process. Then we turn to the issue of model selection. While the obtained likelihood can be used for variable selection, we

work out the case where the focus is mainly on the selection of the better distribution for the missing covariates. We obtain a fast model selection method to choose between normal and t-distributed errors with certain degrees of freedom. This fast method is restricted in its application to the logistic regression models as described in Section 2. For model selection in more general situations, we rely on the EM algorithm, see Section 3.1, or alternatively on multiple imputation methods, see the extension in Section 6.2.

The paper is structured as follows. Section 2 deals with estimation for t-distributed missing covariates, while Section 3 explains the construction of an AIC-type criterion to select amongst these distributions. Simulation results and the analysis of a data example are contained in Sections 4 and 5, respectively. In Section 6 we mention possible extensions to other distributions and we give an AIC for general purpose model selection for multiply imputed data.

2 Estimation with multiple missing covariates

We propose first an extension to the method of Gao and Hui (1997) by releasing the normality assumption of the error term and allowing for a multivariate t-distribution when either one or more covariates under monotone missingness contain missing values.

2.1 Multivariate t-distributed missing covariates

We introduce some notation. We consider a response variable Y that is binary and fully observed, while some of the explanatory variables $X_1, \ldots, X_p$ contain missing values. Let $\boldsymbol{X}$ be the design matrix of regression variables, of dimension $n \times p$, while $\boldsymbol{Y}$ represents the vector of response values of length n; the vectors $(Y_i, X_{i1}, \ldots, X_{ip})$ for $i = 1, \ldots, n$ are independent. Because of the missing observations, the design matrix can be split in two parts, $\boldsymbol{X} = (\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}})$; $\boldsymbol{X}_{\text{obs}}$ represents the part of the design matrix $\boldsymbol{X}$ with covariates

that are completely observed and $\boldsymbol{X}_{\text{mis}}$ is the subset of $\boldsymbol{X}$ with explanatory variables which have at least one value not observed, stressing that $\boldsymbol{X}_{\text{mis}}$ can be a matrix, containing more than one variable. We assume that $\boldsymbol{X}_{\text{mis}}$ follows a monotone pattern, as defined by Little and Rubin (2002), meaning that the missing variables are rearranged and ordered in a way that the first 'block' has more observations than the second one, that, in turn, has more observations than the following one, with the last 'block' the one with the fewest observed values. In particular $\boldsymbol{X}_{\text{mis}}$ can be seen as a columnwise partitioned matrix of covariates, where each 'block' contains either univariate or multivariate covariates, depending on the monotone pattern. The missing covariates are modeled constructing J conditional regression models, with a t-distribution for the error terms, where J denotes the number of 'blocks'. The models are built for the least observed missing $\boldsymbol{X}_{\text{mis}}^{(j)}$ in the jth block, given the more observed $\boldsymbol{X}_{\text{mis}}^{1,\dots,j-1}$ in blocks 1 to $j-1$, and so on. In this way the conditional distribution to construct depends on the missingness pattern of the explanatory variables. We want to stress that the conditional models could be either univariate if the j-th 'block' is formed by one variable or multivariate if the j-th 'block' contains more than one variable. We denote by d_j the number of variables in block j. If the 'block' is multivariate, the number of missing values is the same for each variable of the 'block'. The ith rows of $\boldsymbol{X}_{\text{obs}}$ and $\boldsymbol{X}_{\text{mis}}$ are denoted by, respectively, $\boldsymbol{X}_{\text{obs},i}$ and $\boldsymbol{X}_{\text{mis},i}$.

It is necessary to make assumptions on the missing-data mechanism because this is crucial to understand the relationship between the missing and observed data values. A missing-data indicator matrix $\boldsymbol{R}$ can be built from the missing variable, with (i,j)th element $R_{ij}=1$ if X_{ij} is observed and $R_{ij}=0$ if X_{ij} is missing. The missing data indicator is an important part of the model, depending on the missingness mechanism. In this paper we focus on the 'missing at random' (MAR) assumption which assumes that the probability of the missing-data indicator $\boldsymbol{R}$ depends only on the observed part of the data $\boldsymbol{X}$, but not on the missing part, following the definition introduced by Little and Rubin (2002). With

this assumption and the additional assumption that the parameters indexing the model of interest are distinct from those indexing the missingness model, the missing data mechanism is said to be ignorable. This implies that it does not need to be modeled as part of the parameter estimation process.

We now introduce the estimation method. For a logistic regression model, and using Bayes' formula repeatedly, it is readily obtained that

$$\begin{aligned}\operatorname{logit} P(Y_i = 1|\boldsymbol{X}_{\text{obs},i}, \boldsymbol{X}_{\text{mis},i}^{1,\ldots,J}) = \sum_{j=2}^{J} \log \frac{f(\boldsymbol{X}_{\text{mis},i}^{(j)}|\boldsymbol{X}_{\text{obs},i}, \boldsymbol{X}_{\text{mis},i}^{1,\ldots,j-1}, Y_i = 1)}{f(\boldsymbol{X}_{\text{mis},i}^{(j)}|\boldsymbol{X}_{\text{obs},i}, \boldsymbol{X}_{\text{mis},i}^{1,\ldots,j-1}, Y_i = 0)} \\ + \log \frac{f(\boldsymbol{X}_{\text{mis},i}^{(1)}|\boldsymbol{X}_{\text{obs},i}, Y_i = 1)}{f(\boldsymbol{X}_{\text{mis},i}^{(1)}|\boldsymbol{X}_{\text{obs},i}, Y_i = 0)} + \operatorname{logit} P(Y_i = 1|\boldsymbol{X}_{\text{obs},i}). \qquad (1)\end{aligned}$$

We denote

$$\operatorname{logit} P(Y_i = 1|\boldsymbol{X}_{\text{obs},i}, \boldsymbol{X}_{\text{mis},i}) = \alpha_0 + \boldsymbol{X}_{\text{obs},i}\boldsymbol{\alpha}_1 + \boldsymbol{X}_{\text{mis},i}\boldsymbol{\alpha}_2, \qquad (2)$$

where $\boldsymbol{\alpha}_1$ represents the parameters corresponding to the fully observed explanatory variables and $\boldsymbol{\alpha}_2$ stands for the parameters associated to the covariates with missing observations. For the jth block, containing d_j variables with one or more missing observation, we construct a d_j-variate model

$$\boldsymbol{X}_{\text{mis},i}^{(j)} = \boldsymbol{\gamma}_0^{(j)} + Y_i\boldsymbol{\gamma}_1^{(j)} + \boldsymbol{X}_{\text{obs},i}\boldsymbol{\gamma}_2^{(j)} + \boldsymbol{X}_{\text{mis},i}^{1,\ldots,j-1}\boldsymbol{\gamma}_{\text{mis}}^{1,\ldots,j-1} + \boldsymbol{\epsilon}_{t,j} \qquad (3)$$

with $\boldsymbol{\epsilon}_{t,j} \sim t_{d_j}(\nu, \boldsymbol{\Sigma}_j)$ a d_j-variate t-distribution with ν degrees of freedom and variance matrix $\boldsymbol{\Sigma}_j$. The coefficient matrix $\boldsymbol{\gamma}_{\text{mis}}^{1,\ldots,j-1}$ has d_j columns and the number of rows is equal to the number of variables in blocks 1 to $j-1$, that is, to $\sum_{k=1}^{j-1} d_k$. Further, we model

$$\operatorname{logit} P(Y_i = 1|\boldsymbol{X}_{\text{obs},i}) = \beta_0 + \boldsymbol{X}_{\text{obs},i}\boldsymbol{\beta}_1. \qquad (4)$$

For easier programming and representing the results in a compact way, we define a matrix $\boldsymbol{\gamma}_{\text{mis}}$ containing all such coefficients $\boldsymbol{\gamma}_{\text{mis}}^{1,\ldots,j-1}$. Its construction is easiest explained through an example. Suppose that in total we have seven covariates with missing values

where the monotone pattern defines the dimension of each 'block' by the following vector $d = (2, 3, 1, 1)$. This means that the first 'block' of $\boldsymbol{X}_{\text{mis}}$ has two variables, the second has 3 and the other two have only one variable. Thus two 'blocks' are univariate, one is bivariate and another is trivariate. The number of rows of the $\boldsymbol{\gamma}_{\text{mis}}$ matrix is equal to $\sum_{j=2}^{J} d_j$, reflecting that the first variable $\boldsymbol{X}_{\text{mis}}^1$ is only regressed on Y and $\boldsymbol{X}_{\text{obs}}$. The number of columns is equal to $\sum_{j=1}^{J-1} d_j$. Starting from a matrix filled with zeros, we insert the parameter for the missing covariates. We insert the matrix $(\boldsymbol{\gamma}_{\text{mis}}^{(2)})^t$ in the first d_2 rows, and first d_1 columns. In our example this means inserting 6 parameters in rows (1,2,3) and columns (1,2). Then, from the first row available, that is row $d_2 + 1$, and from the first column, we insert the matrix $(\boldsymbol{\gamma}_{\text{mis}}^{(3)})^t$. In our example 5 parameters are inserted in row (4) and columns (1,2,3,4,5). This is repeatedly done for each block $j = 1, \ldots, J$. We further define $\boldsymbol{\Sigma}$=blockdiag$\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_J\}$, $\widetilde{\boldsymbol{\Sigma}}$=blockdiag$\{\boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_J\}$ and likewise, for $k = 0, 1, 2$, $\boldsymbol{\gamma}_k = (\boldsymbol{\gamma}_k^{(1)}, \ldots, \boldsymbol{\gamma}_k^{(J)})^t$, $\widetilde{\boldsymbol{\gamma}}_1 = (\boldsymbol{\gamma}_1^{(2)}, \ldots, \boldsymbol{\gamma}_1^{(J)})^t$, and finally $\boldsymbol{\nu}$ =diag$_{j=1,\ldots,J}\{(\nu + d_j)/\nu \mathbf{1}_{d_j}\}$ and $\widetilde{\boldsymbol{\nu}}$ =diag$_{j=2,\ldots,J}\{(\nu + d_j)/\nu \mathbf{1}_{d_j}\}$, with $\mathbf{1}_{d_j}$ a vector of all ones of length d_j.

For a fully comprehension of the matrix building we add the final $\boldsymbol{\gamma}_{\text{mis}}$ and $\widetilde{\boldsymbol{\Sigma}}$;

$$\boldsymbol{\gamma}_{\text{mis}} = \begin{pmatrix} \gamma_{13}^{\text{mis}} & \gamma_{23}^{\text{mis}} & 0 & 0 & 0 & 0 \\ \gamma_{14}^{\text{mis}} & \gamma_{24}^{\text{mis}} & 0 & 0 & 0 & 0 \\ \gamma_{15}^{\text{mis}} & \gamma_{25}^{\text{mis}} & 0 & 0 & 0 & 0 \\ \gamma_{16}^{\text{mis}} & \gamma_{26}^{\text{mis}} & \gamma_{36}^{\text{mis}} & \gamma_{46}^{\text{mis}} & \gamma_{56}^{\text{mis}} & 0 \\ \gamma_{17}^{\text{mis}} & \gamma_{27}^{\text{mis}} & \gamma_{37}^{\text{mis}} & \gamma_{47}^{\text{mis}} & \gamma_{57}^{\text{mis}} & \gamma_{67}^{\text{mis}} \end{pmatrix} \quad \widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \sigma_{33} & \sigma_{34}^{\text{mis}} & \sigma_{35}^{\text{mis}} & 0 & 0 \\ \sigma_{43} & \sigma_{44}^{\text{mis}} & \sigma_{45}^{\text{mis}} & 0 & 0 \\ \sigma_{53} & \sigma_{54}^{\text{mis}} & \sigma_{55}^{\text{mis}} & 0 & 0 \\ 0 & 0 & 0 & \sigma_{66}^{\text{mis}} & 0 \\ 0 & 0 & 0 & 0 & \sigma_{77}^{\text{mis}} \end{pmatrix}.$$

In Section 3, we discuss how to select the best degrees of freedom of the error distribution for the data at hand. The main aim of this work is to exploit the properties of the multivariate t-distribution. This distribution is appealing in statistical analysis, as an alternative choice for the multivariate normal, particularly because its tails are heavier. It is often used from a robustness point of view to take outlying observations into account. For more details on its construction and implementation see Kotz and Nadarajah (2004). For particular application

to missing data problems, we refer to Liu (1995) and Liu and Rubin (1995).

We then arrive at the following result.

Proposition 1. *Consider a logistic regression model with a monotone pattern of missingness resulting in J blocks of covariates with missing observations where the number of covariates in each block is given by the vector $d = (d_1, \ldots, d_J)$. When the variables from each of those blocks are modelled conditionally on the observed variables and the variables in the previous blocks by means of a $t_{d_j}(\nu)$ distribution (for block j), the coefficients of the logistic regression model (2) are approximated by*

$$\begin{aligned}
\alpha_0 &\approx \beta_0 - \boldsymbol{\gamma}_1^t \boldsymbol{\nu} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1/2), \\
\boldsymbol{\alpha}_1 &\approx \boldsymbol{\beta}_1 - \boldsymbol{\gamma}_1^t \boldsymbol{\nu} \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}_2, \\
\boldsymbol{\alpha}_2 &\approx \boldsymbol{\gamma}_1^t \boldsymbol{\nu} \boldsymbol{\Sigma}^{-1} - \widetilde{\boldsymbol{\gamma}}_1^t \widetilde{\boldsymbol{\nu}} \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\gamma}_{\text{mis}},
\end{aligned}$$

in terms of the coefficients of model (3) for $\boldsymbol{X}_{\text{mis}}$ given $\boldsymbol{X}_{\text{obs}}$ and Y, and of model (4) for Y given $\boldsymbol{X}_{\text{obs}}$.

Proof. From equations (1), (3) and the error distribution assumption, for $j = 1, \ldots, J$

$$\begin{aligned}
f(\boldsymbol{X}_{\text{mis},i}^{(j)} | \boldsymbol{X}_{\text{obs},i}, \boldsymbol{X}_{\text{mis},i}^{1,\ldots,j-1}, Y_i) &= \frac{\nu^{\nu/2} \Gamma\left(\frac{\nu+d_j}{2}\right)}{(\pi)^{d_j/2} \Gamma\left(\frac{\nu}{2}\right) |\boldsymbol{\Sigma}_j|^{1/2}} \\
&\quad \cdot \left\{ \nu + (\boldsymbol{X}_{\text{mis},i}^{(j)} - \boldsymbol{\mu}_{i,j})^t \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{X}_{\text{mis},i}^{(j)} - \boldsymbol{\mu}_{i,j}) \right\}^{-\left(\frac{\nu+d_j}{2}\right)}
\end{aligned}$$

with $\boldsymbol{\mu}_{i,j} = \boldsymbol{\gamma}_0^{(j)} + \boldsymbol{\gamma}_1^{(j)} \cdot I(Y_i = 1) + \boldsymbol{X}_{\text{obs},i} \boldsymbol{\gamma}_2^{(j)} + \boldsymbol{X}_{\text{mis},i}^{1,\ldots,j-1} \boldsymbol{\gamma}_{\text{mis}}^{1,\ldots,j-1}$, and ν the degree of freedom of the t-distribution. Considering a Taylor approximation around $x = 0$, $\log(\nu + x) \approx \log(\nu) + \left(\frac{x}{\nu}\right)$, the leading term of the approximation to $\text{logit} P(Y_i = 1 | \boldsymbol{X}_{\text{obs},i}, \boldsymbol{X}_{\text{mis},i})$ in equation (1) can be written as,

$$\begin{aligned}
&- \sum_{j=1}^{J} \left(\frac{\nu + d_j}{2} \right) \frac{1}{\nu} \left\{ (\boldsymbol{X}_{\text{mis},i}^{(j)} - \boldsymbol{\mu}_{i,j}^{[1]})^t \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{X}_{\text{mis},i}^{(j)} - \boldsymbol{\mu}_{i,j}^{[1]}) \right. \\
&\quad \left. - (\boldsymbol{X}_{\text{mis},i}^{(j)} - \boldsymbol{\mu}_{i,j}^{[0]})^t \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{X}_{\text{mis},i}^{(j)} - \boldsymbol{\mu}_{i,j}^{[0]}) \right\} + \beta_0 + \boldsymbol{X}_{\text{obs},i} \boldsymbol{\beta}_1, \qquad (5)
\end{aligned}$$

with, when $Y_i = 0$, $\boldsymbol{\mu}_{i,j}^{[0]} = \boldsymbol{\gamma}_0^{(j)} + \boldsymbol{X}_{\text{obs},i}\boldsymbol{\gamma}_2^{(j)} + \boldsymbol{X}_{\text{mis},i}^{1,\ldots,j-1}\boldsymbol{\gamma}_{\text{mis}}^{1,\ldots,j-1}$ for $j = 2,\ldots,J$ and $\boldsymbol{\mu}_{i,1}^{[0]} = \boldsymbol{\gamma}_0^{(1)} + \boldsymbol{X}_{\text{obs},i}\boldsymbol{\gamma}_2^{(1)}$, and when $Y_i = 1$, define $\boldsymbol{\mu}_{i,j}^{[1]} = \boldsymbol{\mu}_{i,j}^{[0]} + \boldsymbol{\gamma}_1^{(j)}$ for $j = 1,\ldots,J$. Equation (5) can be simplified by using that

$$\begin{aligned}
&(\boldsymbol{X}_{\text{mis},i}^{(j)} - \boldsymbol{\mu}_{i,j}^{[1]})^t\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{X}_{\text{mis},i}^{(j)} - \boldsymbol{\mu}_{i,j}^{[1]}) \\
&\quad = \quad (\boldsymbol{X}_{\text{mis},i}^{(j)} - \boldsymbol{\mu}_{i,j}^{[0]})^t\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{X}_{\text{mis},i}^{(j)} - \boldsymbol{\mu}_{i,j}^{[0]}) - 2(\boldsymbol{\gamma}_1^{(j)})^t\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{X}_{\text{mis},i}^{(j)} - \boldsymbol{\mu}_{i,j}^{[0]}) + (\boldsymbol{\gamma}_1^{(j)})^t\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\gamma}_1^{(j)}. \quad (6)
\end{aligned}$$

Inserting equation (6) in equation (1) we obtain that

$$\begin{aligned}
\text{logit}P(Y_i \;=\; 1|\boldsymbol{X}_{\text{obs},i}, \boldsymbol{X}_{\text{mis},i}) \approx \beta_0 &- \sum_{j=1}^{J}\left(\frac{\nu + d_j}{\nu}\right)(\boldsymbol{\gamma}_1^{(j)})^t\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\gamma}_0^{(j)} - \frac{1}{2}\boldsymbol{\gamma}_1^{(j)}) \\
&+ \left\{\boldsymbol{\beta}_1 - \sum_{j=1}^{J}\left(\frac{\nu + d_j}{\nu}\right)(\boldsymbol{\gamma}_1^{(j)})^t\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\gamma}_2^{(j)}\right\}\boldsymbol{X}_{\text{obs},i} \qquad (7) \\
&+ \sum_{j=1}^{J}\left(\frac{\nu + d_j}{\nu}\right)(\boldsymbol{\gamma}_1^{(j)})^t\boldsymbol{\Sigma}_j^{-1}\boldsymbol{X}_{\text{mis},i}^{(j)} - \sum_{j=2}^{J}\left(\frac{\nu + d_j}{\nu}\right)(\boldsymbol{\gamma}_1^{(j)})^t\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\gamma}_{\text{mis},i}^{1,\ldots,j-1}\boldsymbol{X}_{\text{mis}}^{1,\ldots,j-1}.
\end{aligned}$$

Equating the corresponding coefficients of equation (7) to those of equation (2) proves the stated result. □

The special case of a single variable with missing observations and a univariate t-distribution for the error terms, is a direct extension of the univariate normal results.

Corollary 1. *For the logistic regression model with a single univariate $t_1(\nu)$-distributed covariate with missing observations, the coefficients of the logistic regression model (2) are approximated by*

$$\begin{aligned}
\alpha_0 &\approx \beta_0 - \left(\frac{\nu+1}{\nu}\right)\frac{1}{\sigma_{\epsilon_t}^2}(\gamma_0\gamma_1 + \gamma_1^2/2), \\
\boldsymbol{\alpha}_1 &\approx \boldsymbol{\beta}_1 - \left(\frac{\nu+1}{\nu}\right)\frac{1}{\sigma_{\epsilon_t}^2}\gamma_1\boldsymbol{\gamma}_2, \\
\alpha_2 &\approx \left(\frac{\nu+1}{\nu}\right)\frac{\gamma_1}{\sigma_{\epsilon_t}^2},
\end{aligned}$$

in terms of the coefficients of model (3) with $d_1 = 1$ for X_{mis} given $\boldsymbol{X}_{\text{obs}}$ and Y, and of model (4) for Y given $\boldsymbol{X}_{\text{obs}}$.

The fit of the model with missing covariates consists of three steps: (i) the missing covariates $\boldsymbol{X}_{\text{mis}}$ are fitted using a complete cases linear regression model, (ii) formula (4) is fitted using a classical logistic regression model, and (iii) the estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$ obtained from the two steps above are combined to obtain the $\hat{\boldsymbol{\alpha}}$ values corresponding to model (2).

The limiting case with degrees of freedom ν tending to infinity results in a multivariate normal distribution for the covariates with missing values. For that case we obtain the following result, see also Gao and Hui (1997) for the case of all $d_j = 1$.

Proposition 2. *For the logistic regression model with multivariate normal missing covariates for each block as defined by the missingness pattern, the coefficients of the logistic regression model (2) are equal to*

$$\begin{aligned} \alpha_0 &= \beta_0 - \boldsymbol{\gamma}_1^t \boldsymbol{\Sigma}^{-1}(2\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1), \\ \boldsymbol{\alpha}_1 &= \boldsymbol{\beta}_1 - \boldsymbol{\gamma}_1^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}_2, \\ \boldsymbol{\alpha}_2 &= \boldsymbol{\gamma}_1^t \boldsymbol{\Sigma}^{-1} - \widetilde{\boldsymbol{\gamma}}_1^t \widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\gamma}_{\text{mis}}, \end{aligned}$$

in terms of the coefficients $\boldsymbol{\gamma}$ for the model for $\boldsymbol{X}_{\text{mis}}$ given $\boldsymbol{X}_{\text{obs}}$ and Y, and of model (4) for Y given $\boldsymbol{X}_{\text{obs}}$.

3 Selection of the distribution

Section 2 focussed on the estimation of the parameters of interest when missing covariates are present in a logistic regression context. Because of the random nature of the missing covariates $\boldsymbol{X}_{\text{mis}}$, modeling them is an important issue; in fact, choosing the right distribution will help to get better results in the estimation process (see also the simulation study). For this reason a natural and direct question arises about the possibility to choose, given different distributions, the one that is modeling the data in a better way. Another problem arises because model selection, with the usual criteria which are mostly likelihood-based, fails in

a missing data context. Probably the most well-known criterion is the Akaike information criterion (AIC) (Akaike, 1973), which essentially is a penalized log-likelihood function. The criterion gives a balance between the goodness-of-fit, represented by the log-likelihood, and the complexity of the model, represented by the penalty term. In incomplete data problems, however, the log-likelihood is not available. In the context of incomplete data different variations of the AIC criterion have been proposed; Shimodaira (1994) proposed the selection method through the predictive divergence for indirect observation models (PDIO); Cavanaugh and Shumway (1998) suggested a variation of the former criterion using the likelihood of the incomplete data as goodness of fit of the criterion (AICcd). Hens et al. (2006) considered a modification of the AIC by weighting the complete cases by their inverse selection probabilities, dealing with the missingness mechanism as a nuisance. All the last three methods focussed on the missing response variable. Claeskens and Consentino (2008) proposed a variation of the AIC for missing covariates; the criterion is based on the EM algorithm and the method of weights of Ibrahim (1990) in order to get information on the model fitting minimizing the EM algorithm itself. In this section we exploit first the complete utilization of the criteria proposed by Claeskens and Consentino (2008) for deciding the 'best' distribution for describing the missing covariates. Second, we propose a different approach for the same purpose, using the estimation method described in Section 2, with the advantage of having fast computational speed.

Define $f(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\theta})$ the joint density function for the full set of data $(\boldsymbol{Y}, \boldsymbol{X})$, with $\boldsymbol{\theta}$ the unknown parameter vector. The joint distribution of $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ is specified by the conditional distribution of $(\boldsymbol{Y}_i|\boldsymbol{X}_i)$ and the marginal distribution of $(\boldsymbol{X}_i)$. The complete data density can be modeled as

$$f_\theta = f(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\theta}) = f(\boldsymbol{Y}|\boldsymbol{X}; \boldsymbol{\beta}) f(\boldsymbol{X}; \boldsymbol{\alpha}) \tag{8}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$, and the two parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are distinct. The distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$ can be modeled using the class of generalized linear models which belongs to

the class of the exponential family. If the missing covariates are continuous we can model the marginal distribution of $\boldsymbol{X}_i$ using a normal distribution or a Student's t-distribution for robust statistical reasons (see Liu, 1995; Liu and Rubin, 1995).

3.1 Model selection via the EM algorithm's Q function

Claeskens and Consentino (2008) derived a version of Akaike's (1973) information criterion that is suitable for use with missing covariate information. Starting from the Kullback-Leibler distance, used for measuring the distance between the true data generating density and f_θ, the model density used for describing the data $(\boldsymbol{Y}, \boldsymbol{X})$, they derive a new criterion

$$\text{AIC} = -2\ Q(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) + 2\ p_\theta, \tag{9}$$

with $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n} \int w_i \log f(y_i, x_i; \boldsymbol{\theta})\, dx_{\text{mis},i}$, $w_i = f(x_{\text{mis},i}|x_{\text{obs},i}, y_i; \boldsymbol{\theta}^{(k)})$ and $p_\theta = \text{length}(\boldsymbol{\theta})$. The model with the smallest value is selected. The Q function is an estimator of the Kullback-Leibler distance and represents the goodness-of-fit part in the criterion. The classical AIC can not be used because the density f_θ can not be evaluated at $\boldsymbol{Y}, \boldsymbol{X}$, due to the presence of missing covariate data. The weights are defined via the density function (or probability mass function for categorial covariates) of the covariates with missing observations, given the observed data. Because of the factorization in (8), the Q function is written as a sum of two terms

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = Q^{(1)}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)}) + Q^{(2)}(\boldsymbol{\alpha}|\boldsymbol{\theta}^{(k)}). \tag{10}$$

In Claeskens and Consentino (2008) the main attention was restricted to the use of the $Q^{(1)}$ function only, leading to their TIC_1 and AIC_1. The reason for this was due to the problem of a direct comparison for the second component for different models, since not all considered models contained all covariates with missing observations.

In this paper the 'full' function Q is used for the model selection purpose; in this way both the part on the regression relationship between the response Y given the covariates

$\boldsymbol{X}$ and the part with the specification of the model for the missing covariates are included, hence the criteria take the complexity of the missingness modeling into account. The idea is to also use the $Q^{(2)}$ function for model selection, but in a different way than the classical one, by using that function for deciding which distribution describes the missing covariates in a better way. This particular regression is done by regressing the missing covariates, used as response variables, on the fully observed variables, used as covariates. The first choice for describing continuous variables is the normal distribution, which is mathematical tractable. A valid alternative, though, is the Student t-distribution, allowing for robust statistical inference. Specifying two or more different density functions for the missing covariates leads to the question to decide which one is more feasible. The $Q^{(2)}$ function is able to provide an answer. To make a decision on the 'best' distribution $f(\boldsymbol{X};\boldsymbol{\alpha})$ for the missing covariates, a criterion such as AIC can provide guidance, possibly accompanied by a sensitivity analysis. The "full" AIC and TIC are particularly useful in the situation that one has decided upon a structure of the regression model and wishes to compare different models for $\boldsymbol{X}_{\text{mis}}|\boldsymbol{X}_{\text{obs}}$. Due to the presence of incomplete data the estimation of the 'full' Q function is carried out using the EM algorithm and the method of weights of Ibrahim (1990). For continuous missing covariates a Monte Carlo EM algorithm is used for evaluating Q, using the Gibbs sampler along with the adaptive rejection algorithm of Gilks and Wild (1992), in order to sample from $f(x_{\text{mis},i}|x_{\text{obs},i},y_i;\boldsymbol{\theta}^{(k)})$. The proposed information criteria do not depend on the particular way of computation and alternatives may be used. The criteria use the EM algorithm without additional computational effort.

Note that an AIC based on the Q function in (10) is able to deal with standard variable selection questions as well, in addition to the application to distribution selection as we apply it to in this paper.

However, its main and not negligible disadvantage is that the estimation of the Q function is computationally intensive and can be quite time consuming, especially in a bivariate (or

higher dimensional) situation. This motivated us to explore an alternative and fast method for distribution selection.

3.2 Non-iterative distribution selection

The purpose of this section is to use the method introduced in Section 2 for estimating the parameters and employ the corresponding maximized likelihood for model selection purposes. The log-likelihood function is based on the logistic regression model in (1). Since the conditional model of Y given $\boldsymbol{X}_{\text{obs}}$ will be the same for different distribution specifications of $\boldsymbol{X}_{\text{mis}}$ given $\boldsymbol{X}_{\text{obs}}$, we can ignore this part. Hence, for selecting the distribution we can restrict attention to the use of the model for $\boldsymbol{X}_{\text{mis}}$ given Y and $\boldsymbol{X}_{\text{obs}}$, still assuming the missingness at random assumption. The corresponding AIC is

$$\text{AIC} = -2 \ \log f(\boldsymbol{X}_{\text{mis}}|\boldsymbol{X}_{\text{obs}}, Y) + 2 \ p_\gamma, \tag{11}$$

with p_γ the number of parameters in the model. The smallest obtained value of this AIC indicates the best distribution for modeling the data. The simulation study shows that this approach works well to identify an error distribution for $\boldsymbol{X}_{\text{mis}}$.

Note that the complete data density $f(\boldsymbol{X}, \boldsymbol{Y})$ as in (8) can be used to build an AIC for broader use. Indeed, using the likelihood obtained from model (1) AIC can be used for variable selection as well.

4 Numerical results: simulation study

An extensive simulation study is performed in order to assess the validity of the method introduced in the former sections. Two different scenarios are taken into account: the first scenario is based on the presence of a univariate missing covariate, while the second one is dealing with the multivariate covariates setting, focussing, particularly, on the presence of bivariate missing covariates. The missing covariates in both scenarios are continuous.

4.1 Estimation with a univariate missing covariate

We simulate a logistic regression model as in (2). The vector of covariates for the ith observation is given by $(1, x_{i1}, \ldots, x_{i5})$, with corresponding coefficients $\boldsymbol{\alpha}^t = (\alpha_0, \ldots, \alpha_5)$. The true values chosen for the coefficients are $\boldsymbol{\alpha}^t = (1, 0, 0, -1, -1, 1)$. In this scenario only the first covariate vector $\boldsymbol{X}_1 = (x_{11}, \ldots, x_{n1})^t$ contains missing values, while all the other variables are fully observed. The fully observed covariates are generated independently from a standard normal distribution. The missing covariate is generated under the MAR assumption as $X_{mis} = \gamma_0 + Y\gamma_1 + \boldsymbol{X}_{obs}\boldsymbol{\gamma}_2 + \epsilon$. The true coefficients are $\boldsymbol{\gamma}^t = (1, 0, 0, -1, 2, -1)$ in the univariate missing covariate setting. Data are simulated using, for the error terms, either a normal distribution or a t-distribution, with one of four different degrees of freedom df= $(5, 7, 15, 50)$. Furthermore, in order to test which distribution fits the data in a better way, the missing covariate x_{i1} is generated in two ways, using a normal distribution and a t-distribution with the same degree of freedom as above and depending only on the fully observed variables $(x_{i2}, \ldots, x_{i5})$. Independent standard normal errors u_i are generated, and a data value x_{i1} is set to be missing, or in other words, $R_{1i} = 0$ when $a_1(x_{i2} - \overline{x}_{2.}) + \ldots + a_4(x_{i5} - \overline{x}_{5.}) + a_5(Y_i - \overline{Y}) + u_i \leq z_\alpha$ and $R_{1i} = 1$ otherwise, from which the distribution of $\boldsymbol{R}$ conditional on $\boldsymbol{X}_{\text{obs}}$ and Y can be deduced. We used the following notation: $\overline{x}_{k.}$ is the sample mean of $\boldsymbol{x}_k$, $\overline{Y}$ is the sample mean of Y_i, and z_α is the α-quantile of $N(0, (\sum_{i=1}^5 a_i^2 + 1))$ with α the chosen percentage of missingness. The coefficient vector is set equal to $\boldsymbol{a} = (2, 4, 3, 1, 3)$. We used four different sample sizes $n = 50$, 100, 200 and 500 and three different percentages of missingness $5\%, 15\%$ and 30%. For each setting we run $N = 2000$ simulations. Since the simulation study is quite extensive, we selected relevant parts of the output to discuss.

In Table 1 partial results of the simulation are displayed; namely those for sample size equal to 100 and 30% of missingness. From this table we can extract some useful comments. First of all, independently of the simulated data, the averaged estimates are close to the true values. Estimates obtained with a t-distribution are very close to the true parameter and

Table 1: Estimation with one covariate with 30% missingness and four fully observed covariates. The table shows the simulated mean values of the estimates and the mean squared error (in parenthesis) for $n = 100$, and 2000 simulation runs. The true value of the parameters is $\boldsymbol{\alpha} = (1, 0, 0, -1, -1, 1)$.

Simulated data	Fitted data	α_0	α_1	α_2	α_3	α_4	α_5
Normal	Normal	1.108	−0.000	0.014	−1.099	−1.119	1.100
		(0.228)	(0.093)	(0.090)	(0.246)	(0.509)	(0.231)
	t_{50}	1.111	−0.001	0.014	−1.101	−1.122	1.102
		(0.267)	(0.124)	(0.091)	(0.285)	(0.653)	(0.267)
	t_{15}	1.114	−0.001	0.014	−1.105	−1.125	1.105
		(0.323)	(0.159)	(0.093)	(0.341)	(0.859)	(0.320)
	t_7	1.121	−0.001	0.014	−1.111	−1.130	1.110
		(0.449)	(0.228)	(0.096)	(0.467)	(1.329)	(0.440)
	t_5	1.126	−0.001	0.014	−1.117	−1.135	1.115
		(0.581)	(0.293)	(0.099)	(0.598)	(1.823)	(0.566)
	CC	1.146	−0.002	0.022	−1.140	−1.158	1.145
		(0.373)	(0.125)	(0.141)	(0.412)	(0.743)	(0.348)
t_{50}	Normal	1.092	0.005	−0.005	−1.121	−1.122	1.113
		(0.230)	(0.095)	(0.093)	(0.241)	(0.529)	(0.250)
	t_{50}	1.094	0.005	−0.005	−1.123	−1.126	1.116
		(0.271)	(0.127)	(0.094)	(0.280)	(0.681)	(0.290)
	t_{15}	1.097	0.006	−0.005	−1.126	−1.132	1.121
		(0.329)	(0.164)	(0.095)	(0.338)	(0.903)	(0.349)
	t_7	1.103	0.007	−0.005	−1.132	−1.142	1.129
		(0.462)	(0.238)	(0.099)	(0.470)	(1.408)	(0.482)
	t_5	1.107	0.008	−0.005	−1.137	−1.151	1.136
		(0.600)	(0.307)	(0.102)	(0.608)	(1.938)	(0.621)
	CC	1.118	0.007	−0.001	−1.154	−1.158	1.167
		(0.364)	(0.130)	(0.139)	(0.394)	(0.790)	(0.386)
t_7	Normal	1.108	0.000	−0.005	−1.098	−1.102	1.105
		(0.193)	(0.068)	(0.089)	(0.205)	(0.418)	(0.207)
	t_{50}	1.111	0.001	−0.005	−1.101	−1.106	1.108
		(0.224)	(0.094)	(0.090)	(0.236)	(0.538)	(0.239)
	t_{15}	1.115	0.001	−0.005	−1.104	−1.111	1.113
		(0.274)	(0.126)	(0.092)	(0.286)	(0.727)	(0.290)
	t_7	1.122	0.002	−0.005	−1.111	−1.120	1.121
		(0.388)	(0.192)	(0.096)	(0.400)	(1.166)	(0.407)
	t_5	1.128	0.003	−0.005	−1.116	−1.128	1.127
		(0.509)	(0.254)	(0.100)	(0.520)	(1.631)	(0.530)
	CC	1.150	0.001	−0.002	−1.144	−1.144	1.154
		(0.345)	(0.094)	(0.142)	(0.366)	(0.607)	(0.323)

to the one estimated under normality, meaning that the linear approximation used in the construction of the coefficients (see the proof of Proposition 1) is working properly. Higher order approximations might improve the estimation results.

The estimated values when fitted with normal and t_{50} distributions are very similar, due to the propriety that increasing the degree of freedom in a t-distribution will make the distribution tend to a normal one; but even when the normality assumption is released and the degree of freedom ν is small, the estimation is working in a good way. The complete case estimates are slightly larger than the ones fitted under normal and t-distributions, while the mean squared error is slighter higher than normal and t_{50} estimates. The smallest mean squared error is usually observed when using a normal distribution for fitting the data, with the t_{50} very close. The main point, nevertheless, is that the missingness needs to be taken into account, and that the t-distribution may give more robust results than the normal distribution when the data are heavier tailed.

As stated before, the method was applied with different sample sizes and percentages of missingness. In order to avoid too many tables, a brief summary is presented. When $n = 50$ the results both for the estimates and the mean squared errors are not that good, especially for the complete case methods, while the estimates under normal and t distribution are less biased; this is due to the fact that the sample size is too small. In fact, just increasing the sample size to 100 (see Table 1) yields significant improvement. When the sample size grows to either 200 or 500 the results are very good for all the distributions used for fitting the missing covariate, both for bias and mean squared error, independently of the percentage of missingness.

4.2 AIC distribution selection with a univariate missing covariate

Table 2 displays the result of the distribution selection. We use the Akaike information criterion to investigate which distribution is modeling the data better. The results are quite good, already for the smallest sample size, and improving with sample size. Especially for the larger sample sizes, the AIC is selecting, with higher frequency, the model fitted with the true distribution and this is valid with all the considered percentages of missingness.

Table 2: Distribution selection by AIC when there is one covariate with missing observations and four fully observed covariates. The table shows the simulated percentage of times that the AIC selected a certain model, for different true models, percentages of missingness and sample sizes, for 2000 simulation runs.

Sample Size	Simulated data	Distribution selection					Distribution selection				
		missingness= 5%					missingness= 30%				
		Norm	t_{50}	t_{15}	t_7	t_5	Norm	t_{50}	t_{15}	t_7	t_5
50	Norm	0.065	0.697	0.112	0.078	0.048	0.000	0.780	0.096	0.059	0.065
	t_{50}	0.052	0.681	0.127	0.078	0.061	0.000	0.748	0.101	0.073	0.078
	t_{15}	0.034	0.589	0.147	0.117	0.113	0.000	0.673	0.108	0.093	0.126
	t_7	0.016	0.432	0.162	0.162	0.228	0.000	0.542	0.126	0.120	0.211
	t_5	0.008	0.331	0.144	0.155	0.362	0.000	0.438	0.118	0.128	0.316
100	Norm	0.468	0.314	0.141	0.065	0.013	0.351	0.434	0.126	0.061	0.027
	t_{50}	0.401	0.317	0.182	0.078	0.021	0.296	0.429	0.150	0.082	0.042
	t_{15}	0.255	0.295	0.227	0.146	0.076	0.199	0.378	0.205	0.124	0.095
	t_7	0.105	0.165	0.218	0.249	0.261	0.096	0.249	0.186	0.216	0.252
	t_5	0.050	0.106	0.144	0.261	0.440	0.050	0.168	0.151	0.224	0.406
200	Norm	0.595	0.222	0.160	0.021	0.002	0.566	0.237	0.157	0.035	0.004
	t_{50}	0.478	0.251	0.215	0.051	0.004	0.463	0.254	0.208	0.068	0.007
	t_{15}	0.234	0.233	0.330	0.163	0.041	0.264	0.231	0.284	0.171	0.050
	t_7	0.060	0.094	0.230	0.358	0.257	0.089	0.117	0.226	0.311	0.257
	t_5	0.007	0.029	0.124	0.287	0.552	0.024	0.055	0.145	0.265	0.512
500	Norm	0.631	0.259	0.108	0.002	0.000	0.608	0.251	0.135	0.006	0.000
	t_{50}	0.436	0.333	0.221	0.010	0.000	0.450	0.286	0.240	0.024	0.000
	t_{15}	0.103	0.222	0.534	0.139	0.002	0.141	0.209	0.484	0.158	0.007
	t_7	0.004	0.015	0.213	0.558	0.209	0.011	0.029	0.239	0.489	0.231
	t_5	0.000	0.000	0.029	0.287	0.683	0.000	0.003	0.052	0.312	0.631

For instance, in the setting $n = 500$ and 5% missingness, when the data are simulated from a normal distribution, the criterion selects 63% of times the normal model for fitting, in addition to 26% the t_{50} distribution, which is hard to distinguish from the normal. If the data are simulated from a t_5 distribution, then, according to the AIC, we should model the data using a t_5 or t_7 (which are hard to distinguish) in 91.7% of times. When the sample size is small, fitting the data with either normal or t_{50} seems quite a reasonable decision. Furthermore, when the percentage of missingness increases, there is reduction in performance for the smallest sample size due to the difficulty of dealing with missing data,

Table 3: Estimation with one covariate with 30% missingness and four fully observed covariates. The table shows the simulated mean values of the estimates and the mean squared error (in parenthesis) for $n = 100$, and 2000 simulation runs. For each simulation run, the estimates have been computed in the model selected by the AIC. The true value of the parameters is $\boldsymbol{\alpha} = (1, 0, 0, -1, -1, 1)$.

Simulated data	α_0	α_1	α_2	α_3	α_4	α_5
Normal	1.112	−0.001	0.014	−1.102	−1.121	1.103
	(0.290)	(0.135)	(0.092)	(0.307)	(0.736)	(0.291)
t_{50}	1.097	0.004	−0.005	−1.126	−1.125	1.116
	(0.313)	(0.150)	(0.095)	(0.320)	(0.841)	(0.331)
t_{15}	1.094	0.009	−0.008	−1.121	−1.146	1.136
	(0.299)	(0.149)	(0.096)	(0.320)	(0.892)	(0.356)
t_7	1.116	0.004	−0.005	−1.105	−1.123	1.120
	(0.346)	(0.168)	(0.096)	(0.359)	(1.027)	(0.366)
t_5	1.096	0.023	0.010	−1.098	−1.181	1.149
	(0.379)	(0.186)	(0.099)	(0.396)	(1.173)	(0.426)
CC	1.133	0.006	0.007	−1.145	−1.163	1.162
	(0.350)	(0.108)	(0.142)	(0.382)	(0.692)	(0.349)

while the method remains to perform well for the larger sample sizes.

For comparison we also computed the values shown in Table 3 where we show the average of the estimates and their mean squared errors over the simulation runs where we now each time compute the estimate in the model that is selected by AIC. This gives us a table that should be compared to Table 1. For the frequencies of the selected models we refer to Table 2. We observe that the mean squared error values over the 2000 simulation runs are still comparable to those of Table 1, showing that the model selection method behaves well and is not inflating the variances.

4.3 Estimation with multiple missing covariates

We consider the same simulation settings as described in Section 4.1, with the difference that now $\boldsymbol{X}_{\text{mis}}$ contains two components. In this scenario the first two covariates vectors (x_{i1}, x_{i2}) contain missing values, while all the other variables are fully observed. The two covariates with missing observations are generated under the MAR assumption as $X^{(1)}_{\text{mis}} = \gamma_{0,1} + Y\gamma_{1,1} + \boldsymbol{X}_{\text{obs}}\boldsymbol{\gamma}_{2,1} + \boldsymbol{\epsilon}_1$ and $X^{(2)}_{\text{mis}} = \gamma_{0,2} + Y\gamma_{1,2} + \boldsymbol{X}_{\text{obs}}\boldsymbol{\gamma}_{2,2} + \boldsymbol{X}^{(1)}_{\text{mis}}\gamma^{(1)}_{\text{mis}} + \boldsymbol{\epsilon}_2$. The true coefficients are $\boldsymbol{\alpha}^t = (1, 0, 0, -1, -1, 1)$, $\boldsymbol{\gamma}^t_k = (1, 0, 0, -1, 2)$ for $k = 1, 2$ and $\gamma^{(1)}_{\text{mis}} = -2$. Data are simulated using, for the error terms, either a normal distribution or a t-distribution, with one of four different degrees of freedom df$= (5, 7, 15, 50)$. The covariates with missing observations (x_{i1}, x_{i2}) are generated using either a multivariate normal distribution $N_2(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = \sigma^2 I_2$ a 2×2 covariance matrix or a multivariate t-distribution with the same degree of freedom as above and depending only on the fully observed variables $(x_{i3}, \ldots, x_{i5})$ and Y, and for x_{i2} also on x_{i1}. Independent standard normal errors u_{ik} are generated, and a data value x_{ik} is set to be missing, that is, $R_{1i} = 1$ when $a_1(x_{i3} - \overline{x}_{3.}) + \ldots + a_3(x_{i5} - \overline{x}_{5.}) + a_4(Y_i - \overline{Y}) + u_i \leq z_\alpha$ and, conditional on x_{i1} being missing, $R_{2i} = 1$ when $a_1(x_{i3} - \overline{x}_{3.}) + \ldots + a_3(x_{i5} - \overline{x}_{5.}) + a_4(Y_i - \overline{Y}) + u_i \leq z_\alpha$ with $(a_1 \ldots, a_4) = (2, 4, 3, 1)$, and z_α is the α-quantile of $N(0, (\sum_{i=1}^4 a_i^2 + 1))$ with α the chosen percentage of missingness. We used four different sample sizes $n = 50, 100, 200$ and 500; and three different choices of percentages of missingness $(5\%, 5\%)$, $(15\%, 5\%)$ and $(30\%, 5\%)$. For each setting we run $N = 2000$ simulations. Since the simulation study is, again, quite extensive, we selected relevant parts of the output to discuss.

Table 4 displays partial results of the simulation, for the multivariate missing covariates, namely those for sample size equal to 50 and 15% and 5% of missingness for respectively X_1 and X_2. First, when the data are fitted using the complete case method (CC), the result are biased and inefficient, independent of the simulated data. The estimates fitted under normal and t-distribution a slightly biased, even though the estimates for the missing

Table 4: Estimation with two covariates with 15% and 5% missingness and three fully observed covariates. The table shows the simulated mean values of the estimates and the mean squared error (in parenthesis) for $n = 50$, and 2000 simulation runs. The true value equals $\boldsymbol{\alpha} = (1, 0, 0, -1, -1, 1)$.

Simulated data	Fitted data	α_0	α_1	α_2	α_3	α_4	α_5
Normal	Normal	1.256	−0.020	−0.001	−1.270	−1.266	1.270
		(0.927)	(1.037)	(0.033)	(0.577)	(0.946)	(2.164)
	t_{50}	1.277	−0.026	−0.001	−1.285	−1.284	1.294
		(1.244)	(1.632)	(0.049)	(0.602)	(1.264)	(3.333)
	t_{15}	1.304	−0.032	−0.001	−1.304	−1.308	1.323
		(1.741)	(2.265)	(0.067)	(0.638)	(1.762)	(5.178)
	t_7	1.358	−0.042	−0.001	−1.343	−1.355	1.382
		(3.085)	(3.648)	(0.109)	(0.723)	(3.102)	(10.207)
	t_5	1.407	−0.050	−0.001	−1.379	−1.399	1.436
		(4.719)	(5.030)	(0.150)	(0.818)	(4.731)	(16.368)
	CC	2.251	3.956	2.016	−5.315	0.511	0.323
		(240.730)	(35201.180)	(8535.211)	(24912.710)	(15856.910)	(10220.380)
t_{50}	Normal	1.217	−0.019	−0.003	−1.242	−1.264	1.249
		(0.833)	(1.042)	(0.034)	(0.499)	(0.922)	(2.160)
	t_{50}	1.234	−0.023	−0.003	−1.258	−1.282	1.269
		(1.150)	(1.644)	(0.049)	(0.527)	(1.233)	(3.369)
	t_{15}	1.256	−0.028	−0.004	−1.279	−1.305	1.295
		(1.654)	(2.293)	(0.068)	(0.567)	(1.730)	(5.299)
	t_7	1.301	−0.037	−0.006	−1.319	−1.352	1.347
		(3.031)	(3.719)	(0.109)	(0.663)	(3.085)	(10.584)
	t_5	1.342	−0.045	−0.007	−1.356	−1.394	1.394
		(4.723)	(5.152)	(0.151)	(0.769)	(4.749)	(17.096)
	CC	2.579	−0.236	−0.164	−2.835	−2.882	2.774
		(383.546)	(540.275)	(83.913)	(471.495)	(698.678)	(1389.983)
t_7	Normal	1.217	−0.015	−0.001	−1.250	−1.267	1.284
		(0.691)	(0.784)	(0.024)	(0.511)	(0.865)	(1.765)
	t_{50}	1.235	−0.020	−0.002	−1.266	−1.285	1.305
		(0.922)	(1.250)	(0.036)	(0.539)	(1.118)	(2.674)
	t_{15}	1.258	−0.023	−0.002	−1.286	−1.310	1.333
		(1.299)	(1.773)	(0.052)	(0.580)	(1.527)	(4.160)
	t_7	1.304	−0.030	−0.002	−1.327	−1.359	1.389
		(2.335)	(2.930)	(0.086)	(0.679)	(2.643)	(8.257)
	t_5	1.346	−0.036	−0.002	−1.364	−1.404	1.440
		(3.611)	(4.093)	(0.122)	(0.787)	(4.010)	(13.318)
	CC	6.230	−4.613	−2.419	−3.856	−9.392	8.087
		(14735.990)	(25129.420)	(7089.444)	(2229.608)	(75593.980)	(28524.250)

covariates are close to the true values. The mean squared errors are small under normal and t_{50} distributions, while under complete cases we obtain very high values. Increasing the sample size to 100 or more leads to a significant improvement. The estimation method seems

to perform in a good way even when the degrees of freedom of the t-distribution is small. However, for purposes of distribution selection, this accuracy of the estimated coefficients is already sufficient, see Section 4.4 for the corresponding simulation results.

Table 5: Estimation with two covariates with 30% and 5% missingness and three fully observed covariates. The table shows the simulated mean values of the estimates and the mean squared error (in parenthesis) for $n = 50$, and 2000 simulation runs. The true value equals $\boldsymbol{\alpha} = (1, 0, 0, -1, -1, 1)$.

Simulated data	Fitted data	α_0	α_1	α_2	α_3	α_4	α_5
Normal	Normal	1.263	−0.012	−0.004	−1.267	−1.257	1.276
		(1.029)	(1.371)	(0.034)	(0.558)	(0.999)	(2.574)
	t_{50}	1.285	−0.016	−0.005	−1.286	−1.277	1.298
		(1.490)	(2.324)	(0.049)	(0.592)	(1.453)	(4.306)
	t_{15}	1.312	−0.019	−0.005	−1.309	−1.301	1.324
		(2.158)	(3.230)	(0.068)	(0.639)	(2.109)	(6.826)
	t_7	1.366	−0.025	−0.006	−1.354	−1.347	1.376
		(3.978)	(5.223)	(0.109)	(0.753)	(3.896)	(13.747)
	t_5	1.415	−0.030	−0.007	−1.396	−1.390	1.423
		(6.211)	(7.225)	(0.151)	(0.882)	(6.087)	(22.285)
	CC	20.124	−3.572	−0.335	−19.565	−20.342	21.890
		(163715.500)	(38681.030)	(1404.872)	(78111.100)	(166855.600)	(200484.700)
t_{50}	Normal	1.251	−0.036	−0.002	−1.243	−1.279	1.273
		(1.013)	(1.308)	(0.034)	(0.548)	(1.120)	(2.631)
	t_{50}	1.278	−0.047	−0.003	−1.262	−1.304	1.306
		(1.498)	(2.217)	(0.049)	(0.585)	(1.592)	(4.403)
	t_{15}	1.310	−0.057	−0.003	−1.284	−1.335	1.345
		(2.201)	(3.089)	(0.069)	(0.636)	(2.282)	(6.982)
	t_7	1.374	−0.075	−0.004	−1.329	−1.395	1.422
		(4.118)	(5.011)	(0.111)	(0.758)	(4.160)	(14.062)
	t_5	1.433	−0.090	−0.005	−1.370	−1.451	1.493
		(6.471)	(6.946)	(0.154)	(0.895)	(6.465)	(22.796)
t_7	Normal	1.235	−0.020	−0.003	−1.246	−1.277	1.291
		(0.815)	(1.027)	(0.024)	(0.514)	(1.000)	(2.124)
	t_{50}	1.258	−0.026	−0.003	−1.266	−1.302	1.315
		(1.184)	(1.753)	(0.035)	(0.556)	(1.396)	(3.484)
	t_{15}	1.286	−0.031	−0.004	−1.291	−1.332	1.345
		(1.731)	(2.474)	(0.051)	(0.616)	(1.981)	(5.507)
	t_7	1.341	−0.039	−0.005	−1.341	−1.392	1.404
		(3.238)	(4.074)	(0.085)	(0.763)	(3.585)	(11.114)
	t_5	1.391	−0.045	−0.006	−1.386	−1.446	1.458
		(5.103)	(5.694)	(0.119)	(0.930)	(5.563)	(18.077)
	CC	6.767	2.086	0.948	−7.786	−6.991	5.652
		(6285.348)	(4290.447)	(986.993)	(6762.889)	(8500.068)	(7838.147)

Table 5 confirms the results, with a worsening for the complete cases method when the

Table 6: Estimation with two covariates with missingness and four fully observed covariates. The table shows the simulated mean values of the estimates and the mean squared error (in parenthesis) for $n = 50$, and 2000 simulation runs. For each simulation run, the estimates have been computed in the model selected by the AIC. The true value of the parameters is $\boldsymbol{\alpha} = (1, 0, 0, -1, -1, 1)$.

Simulated data	α_0	α_1	α_2	α_3	α_4	α_5
15% and 5% missingness for x_1, x_2						
Normal	1.293	−0.032	−0.001	−1.294	−1.298	1.316
	(1.579)	(2.019)	(0.059)	(0.622)	(1.603)	(4.601)
t_{50}	1.258	−0.036	−0.003	−1.269	−1.308	1.305
	(1.557)	(2.115)	(0.059)	(0.553)	(1.632)	(4.871)
t_{15}	1.228	0.037	−0.004	−1.280	−1.258	1.221
	(1.438)	(1.968)	(0.057)	(0.586)	(1.456)	(4.563)
t_7	1.262	−0.024	−0.002	−1.294	−1.313	1.338
	(1.545)	(2.040)	(0.057)	(0.603)	(1.757)	(5.079)
t_5	1.307	−0.037	0.001	−1.310	−1.360	1.382
	(1.867)	(2.049)	(0.062)	(0.611)	(2.059)	(5.998)
CC	3.353	−0.481	−0.245	−3.450	−3.699	3.664
	(3429.5)	(12380.6)	(3179.4)	(5711.2)	(19066.5)	(8906.8)
30% and 5% missingness for x_1, x_2						
Normal	1.305	−0.021	−0.004	−1.299	−1.295	1.323
	(2.115)	(3.030)	(0.060)	(0.624)	(2.065)	(6.654)
t_{50}	1.314	−0.064	−0.001	−1.278	−1.342	1.356
	(2.247)	(3.002)	(0.062)	(0.630)	(2.358)	(7.014)
t_{15}	1.262	0.030	−0.004	−1.285	−1.266	1.257
	(2.115)	(2.796)	(0.062)	(0.640)	(2.273)	(6.772)
t_7	1.293	−0.030	−0.003	−1.299	−1.339	1.353
	(2.249)	(2.907)	(0.053)	(0.671)	(2.552)	(7.166)
t_5	1.317	−0.011	−0.003	−1.317	−1.364	1.358
	(2.239)	(2.701)	(0.059)	(0.650)	(2.465)	(7.131)
CC	∞	∞	∞	∞	∞	∞
	∞	∞	∞	∞	∞	∞

percentage of missingness is increased. For example the estimates, when data are simulated from a t_{50} are very large (not shown). Again increasing the sample size leads to an improvement for the estimates and the mean squared error.

In Table 6 we show the average of the estimates and their mean squared errors over the simulation runs where we now each time compute the estimate in the model that is selected by AIC. These numbers are to be compared with those in Tables 4 and 5. For the frequencies of the selected models we refer to Table 7. Here again we can observe that the

model selection method behaves well and is not inflating the variances.

4.4 AIC distribution selection with multivariate missing covariates

Table 7: Distribution selection by AIC when two covariates with missing observations are modeled and four fully observed covariates. The table shows the simulated percentage of times that the AIC selected a certain model, for different true models, percentages of missingness and sample sizes, over 2000 simulation runs.

Sample Size	Simulated data	Distribution selection					Distribution selection				
		missingness= (5%, 5%)					missingness= (30%, 5%)				
		Norm	t_{50}	t_{15}	t_7	t_5	Norm	t_{50}	t_{15}	t_7	t_5
50	Norm	0.026	0.760	0.121	0.064	0.028	0.000	0.789	0.101	0.070	0.040
	t_{50}	0.019	0.772	0.114	0.058	0.037	0.000	0.779	0.102	0.067	0.051
	t_{15}	0.014	0.692	0.151	0.090	0.052	0.000	0.710	0.128	0.092	0.070
	t_7	0.006	0.523	0.186	0.158	0.126	0.000	0.604	0.162	0.119	0.115
	t_5	0.002	0.394	0.194	0.203	0.208	0.000	0.490	0.171	0.153	0.186
100	Norm	0.461	0.374	0.118	0.044	0.002	0.331	0.472	0.133	0.054	0.010
	t_{50}	0.394	0.383	0.160	0.050	0.014	0.299	0.470	0.148	0.062	0.020
	t_{15}	0.244	0.385	0.243	0.108	0.019	0.199	0.447	0.215	0.111	0.028
	t_7	0.095	0.254	0.287	0.256	0.108	0.080	0.318	0.257	0.228	0.117
	t_5	0.030	0.150	0.247	0.318	0.254	0.034	0.217	0.236	0.270	0.243
200	Norm	0.594	0.267	0.129	0.011	0.000	0.570	0.279	0.132	0.019	0.000
	t_{50}	0.496	0.301	0.181	0.020	0.001	0.483	0.290	0.188	0.037	0.002
	t_{15}	0.265	0.310	0.340	0.082	0.004	0.272	0.295	0.335	0.087	0.010
	t_7	0.060	0.123	0.364	0.369	0.084	0.070	0.153	0.362	0.316	0.097
	t_5	0.013	0.040	0.205	0.443	0.299	0.020	0.072	0.234	0.399	0.276
500	Norm	0.678	0.270	0.051	0.000	0.000	0.675	0.250	0.073	0.001	0.000
	t_{50}	0.485	0.381	0.132	0.002	0.000	0.508	0.323	0.166	0.002	0.000
	t_{15}	0.144	0.338	0.486	0.032	0.000	0.197	0.322	0.429	0.051	0.001
	t_7	0.002	0.029	0.429	0.514	0.025	0.006	0.066	0.415	0.471	0.040
	t_5	0.000	0.002	0.073	0.608	0.318	0.000	0.005	0.110	0.566	0.318

Table 7 contains the results for the distribution selection in a multivariate context. We use the AIC to decide on the best distribution of the missing covariates. The results are comparable to the case with one missing covariate. When the sample size is 50 and data are simulated using normal, t_{50} or t_{15} distributions, the t_{50}-distribution fits the missing covariates

in a better way, while for data coming from either t_7 or t_5, there is not a clear choice. When the sample size grows, there is a significant improvement in the association between simulated and fitted data, numerically showing that the method is able to catch the best model for the data. For example, with $n = 500$ and there is 5% missingness for both incompletely observed variables, when the data are simulated from a t_5 distribution, 60.8% of the models are fitted using a t_7 and almost 32% using t_5 (which are hard to distinguish). On the contrary, if data come from a t_{50}-distribution 38.1% of the model are fitted using a t_{50} and 48.5% using a normal distribution (which are again quite similar and hard to distinguish). Furthermore, when the percentage of missingness increases, there is some reduction in performance in each setting due to the difficulty of dealing with missing data, even though the performance of the criterion remains valid. We conclude that as a distribution selection method, the AIC based on the non-iterative method performs well and is able to distinguish normal data from low-degree t-distributed data in the presence of (multiple) covariates with missing observations. This is valid even for small sample sizes. The linear approximation that is used to obtain the results for the t-distributions seems sufficient for distribution selection purposes.

5 Data analysis

The European Values Study (EVS) represents a large-scale, cross-national and longitudinal survey research program, we focus on the data related to Belgium. The outcome variable is a binary variable on people between the ages 18 and 65, where $Y = 0$ indicates if the person is not satisfied with his/her job hours, while $Y = 1$ indicates their satisfaction with job hours. The surveyed people consider the satisfaction with job hours as an important aspect of their general work satisfaction. Other variables are x_1: the age when the education was completed; x_2: gender, x_3: job payment, x_4: education level, x_5: socio-economic status. The original dataset contains 1912 observations, while the considered dataset has 1603 observations, corresponding to the subset with completely observed values on Y and variables x_2–x_5.

Variable x_1 contains missing values for 56 out of the 1603 cases; the missingness mechanism assumed is 'missing at random' (MAR). Since the response variable is binary, we model the data with a logistic regression, using equation (2), without removing cases with missing observations on variable x_1.

We performed the Jarque-Bera test for normality on variable x_1. The test is a goodness-of-fit measure of departure from normality, based on the sample kurtosis and skewness. The observed value of the test statistic is 179.11 on two degrees of freedom, resulting in a p-value of $2.2 \cdot 10^{-16}$, indicating a clear rejection of the null hypothesis of normality. While the median of age at the completion of education is 19, with a third quartile at 22, there are observed values of x_1 as large as 35. We use the model selection method AIC to investigate whether a t-distribution would give a better fit than a normal distribution.

As a comparison, we also applied the iterative method using the EM algorithm (see Section 3). Table 8 shows the results for the AIC obtained with the two different methods, for degrees of freedom equal to $(5, 15, 50)$. For both methods, the smallest value of the AIC corresponds to fitting a t_5-distribution, which can be considered as the most suitable distribution for fitting the distribution of x_1, the covariate with missing observations. We want to stress that the two methods lead to the same conclusion regarding the choice of x_1's distribution, but with hugely different computation times needed. The non-iterative methods returns results immediately, while the Q-function method employs much more time. For fitting the missing covariate's distribution with the normal distribution, the method employs about 22 minutes, while for a t-distribution, the time needed to fit just a single model varies between 14 hours to almost 22 hours for fitting the missing covariate's distribution as a t_5. Clearly, the non-iterative method is much more convenient for deciding how to fit the missing covariate's distribution.

After the distribution for the missing covariate is chosen, we can fit the data. The complete cases estimates are $\widehat{\boldsymbol{\alpha}} = (0.123, -0.036, 0.127, -0.076, 0.094, 0.227, 0.0473, 0.144)$.

Table 8: Results of distribution selection for the ESV data. The table displays, for each method, the four results depending on the distribution of the covariate with missing values, as well as the corresponding computation times.

Method	Missing Covariate Models	AIC	Goodness of fit	penalty term	Timing
			$Q^{(2)}$		
Q-function	Normal	7658.384	3820.192	9	21'42"
	t_{50}	7580.776	3781.388	9	13h59'00"
	t_{15}	7471.422	3726.711	9	17h55'45"
	t_5	7403.142	3692.571	9	21h39'55"
			LogLik		
Non iterative	Normal	7389.142	3685.571	9	< 1"
	t_{50}	7317.908	3649.954	9	< 1"
	t_{15}	7220.962	3601.481	9	< 1"
	t_5	7125.912	3553.956	9	< 1"

The results for the α parameters, with the missing covariate fitted with a t_5 distribution are: $(1.072, -0.079, 0.084, -0.051, 0.290, 0.640, 0.056, 0.098)$. The parameters can be interpreted as follows; taking the exponential of the α parameters we obtain the odds ratio: (2.921, 0.924, 1.088, 0.950, 1.336, 1.896, 1.057, 1.103). The odds ratio of age when the education was completed is 0.924, showing that the higher is the age for completing the education the less is the satisfaction with job hours. The odds ratio for sex is 1.088, showing a better satisfaction for male, while for the job payment a better satisfaction for the job hours is higher for people that do not get paid. Furthermore, the higher is the education and socio-economic status levels, the better is the satisfaction of the job hours. For each parameter we estimate the 90% confidence interval, see Table 9, using 1000 bootstrap replications.

Table 9: Parameter estimates and 90% bootstrap confidence interval of the parameters of interest based on 1000 bootstrap replications.

Parameters	Estimates	Confidence interval
intercept	1.072	(−0.332, 2.535)
x_1 age	−0.079	(−0.150, −0.008)
x_2 gender	0.084	(−0.098, 0.257)
x_3 job payment	−0.051	(−0.234, 0.124)
x_4 education, secondary	0.290	(−0.103, 0.667)
x_4 education, post-secondary	0.640	(−0.074, 1.304)
x_5 status, middle class	0.056	(−0.195, 0.305)
x_5 status, manual worker	0.098	(−0.163, 0.349)

6 Discussion and extensions

The distribution selection method in this paper works particularly well for the logistic regression model. While the results for the multivariate normal distribution to model the missing covariates as function of the observed variables led to exact expressions, those of the multivariate t-distributions were first order approximations. This approximation has been numerically shown to be precise enough to lead to accurate distribution selection results. A major advantage of the proposed method is that it leads to fast results (in contrast to the application of the EM algorithm). For estimation purposes, the approximation works better for larger degrees of freedom. Higher order approximations could be tried if this approach is needed for estimation with a small degree of freedom. Another strategy for estimation with a low degree of freedom t-distribution could be to first select the distribution with the fast method, and once decided on the distribution and its degrees of freedom, then only for that model apply an alternative estimation method.

R code to perform calculations as presented in this paper is available from the website `http://www.econ.kuleuven.be/gerda.claeskens/public/papers/AICmissing.html`.

6.1 Application to other distributions

For other types of distributions, a similar line of arguments can be constructed, though the particular type of simplification to get expressions for the coefficients no longer applies. As an example we give the case of a Poisson distributed outcome variable Y where we model $\log E(Y|\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}}) = \alpha_0 + \boldsymbol{X}_{\text{obs}}\boldsymbol{\alpha}_1 + \boldsymbol{X}_{\text{mis}}\boldsymbol{\alpha}_2$, where again the covariates $\boldsymbol{X}_{\text{mis}}$ contain one or more missing observations and Y and $\boldsymbol{X}_{\text{obs}}$ are fully observed. Writing $\log E(Y|\boldsymbol{X}_{\text{obs}}) = \boldsymbol{\beta}_0 + \boldsymbol{X}_{\text{obs}}\boldsymbol{\beta}_1$ and modelling $\boldsymbol{X}_{\text{mis}} = \boldsymbol{\gamma}_0 + Y\gamma_1 + \boldsymbol{X}_{\text{obs}}\gamma_2 + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon}$ having some multivariate distribution to be selected, leads to

$$\begin{aligned} \log E(Y|\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}}) &= \log\left\{\textstyle\sum_{k=0}^{\infty} k f(\boldsymbol{X}_{\text{mis}}|Y=k, \boldsymbol{X}_{\text{obs}}) P(Y=k|\boldsymbol{X}_{\text{obs}})\right\} \\ &\quad - \log\left\{\textstyle\sum_{k=0}^{\infty} f(\boldsymbol{X}_{\text{mis}}|Y=k, \boldsymbol{X}_{\text{obs}}) P(Y=k|\boldsymbol{X}_{\text{obs}})\right\}. \end{aligned} \tag{12}$$

For the case of a multivariate normal distribution for $\boldsymbol{X}_{\text{mis}}$, this further gives for the first term in (12)

$$\log\left\{A \sum_{k=0}^{\infty} \frac{k}{k!} \exp(ak^2 + bk)\right\},$$

with $a = -\frac{1}{2}k^2\boldsymbol{\gamma}_1^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}_1$; $b = \boldsymbol{\gamma}_1^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{\text{mis}} - \boldsymbol{\gamma}_0 - \boldsymbol{X}_{\text{obs}}\boldsymbol{\gamma}_2) + \beta_0 + \boldsymbol{X}_{\text{obs}}\boldsymbol{\beta}_1$,

$$A = \phi_q(\boldsymbol{X}_{\text{mis}}, \boldsymbol{\mu} = \boldsymbol{\gamma}_0 + \boldsymbol{X}_{\text{obs}}\boldsymbol{\gamma}_2, \boldsymbol{\Sigma}) \exp\{-\exp(\beta_0 + \boldsymbol{X}_{\text{obs}}\boldsymbol{\beta}_1)\},$$

and $\phi_q(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ the density of a q-variate normal with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. A similar calculation results for the second term in (12). Since this expression is highly non-linear in $\boldsymbol{X}_{\text{obs}}$ and $\boldsymbol{X}_{\text{mis}}$, numerical optimization methods need to be used to find (approximations for) the coefficients $\alpha_0, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$. A similar non-linearity issue arises for other distribution specifications to model $\boldsymbol{X}_{\text{mis}}$, for example for categorical covariates with missing values. While the same reasoning could be applied, numerical methods would need to be used to identify approximate values for the coefficients.

6.2 Model selection via AIC for multiply imputed data

We discussed a handling of general model selection for missing data via the EM algorithm (see Section 3.1) and by means of a non-iterative method for the specific setting of logistic regression models with a monotone pattern of missingness (Section 3.2). We here extend the model selection mechanism to handle imputation methods, which are generally applicable. The general philosophy is to impute for missing values and to analyze the resulting imputed sets of data via standard analysis methods. It is well-known that to account for imputation variability, multiple imputations should be used. For model selection this creates a new problem. It is straightforward to apply any traditional variable selection criterion to the separate imputed sets of data. But how should they be combined? Yang et al. (2005) work in a Bayesian setting and average the posterior probabilities over the imputed data sets. Schomaker et al. (2007) compute an "averaged" dataset that consists of the average of each data value after imputation, to which the classical AIC can be applied. An alternative suggestion is to compute the classical AIC for each imputed dataset and then compute the average of the AIC values to select the best model. Consentino and Claeskens (2009) derive an expression for the AIC through the connection with hypothesis testing, that is valid in combination with multiple imputation. This is the construction of the AIC that we describe below.

Multiple imputation for a model S leads to m different datasets, each with its own maximized log likelihood function. Denote by S_0 the smallest model under consideration. Meng and Rubin (1992) combine m separate likelihood ratio values (one for each imputed dataset) into one single test statistic with an approximate F-distribution. This idea of combining test statistics over different imputed datasets, builds on an earlier combined testing procedure using Wald statistics, see Li et al. (1991). We denote by $\mathcal{L}_{S,j}$ the log-likelihood ratio statistic for testing model S_0 versus model S, for the jth imputed set of data, with $j = 1, \ldots, m$. The average of these test statistics is $\bar{\mathcal{L}}_{S,\bullet} = \frac{1}{m} \sum_{j=1}^{m} \mathcal{L}_{S,j}$. The parameter estimator in model S

for the jth imputed data set is denoted by $\hat{\theta}_S^{(j)}$, and the average of the m estimators by $\bar{\theta}_S$. A 'log likelihood ratio' value $\tilde{\mathcal{L}}_{S,j}(\bar{\theta}_S)$ can also be defined for each of the m imputed data sets, where instead of using the estimator $\hat{\theta}_S^{(j)}$ we fill in the average parameter value $\bar{\theta}_S$. Their average is denoted by $\widetilde{\mathcal{L}}_{S,\bullet} = \frac{1}{m}\sum_{j=1}^{m} \tilde{\mathcal{L}}_{S,j}$. Also, denote the number of parameters in model S by $|S|$, and the difference in numbers of parameters of the two models by $p_S = |S| - |S_0|$. Using Meng and Rubin (1992), the test statistic for testing model S_0 versus model S is

$$\frac{D_S}{p_S} = \frac{\widetilde{\mathcal{L}}_{S,\bullet}}{p_S\{1 + \frac{m+1}{p_S(m-1)}(\bar{\mathcal{L}}_{S,\bullet} - \widetilde{\mathcal{L}}_{S,\bullet})\}}, \tag{13}$$

with an approximate F distribution with degrees of freedom p_S and ν where

$$\nu = \begin{cases} 4 + (t-4)\{1 + (1-2t^{-1})D^{-1}\}^2 & \text{if } t = p_S(m-1) > 4 \\ t(1+p_S^{-1})(1+D^{-1})^2/2 & \text{otherwise,} \end{cases}$$

with $D = \frac{m+1}{p_S(m-1)}(\bar{\mathcal{L}}_{S,\bullet} - \widetilde{\mathcal{L}}_{S,\bullet})$. Consentino and Claeskens (2009) define the AIC difference for model S compared to model S_0 as

$$\text{aic}(S, S_0) = -D_S + 2\,p_S. \tag{14}$$

Note that there is a constant 2 absorbed in the notation for the log likelihood ratio statistics. Model selection proceeds by computing these AIC differences for all models S under consideration. The model with the smallest AIC difference is considered the best one.

Criterion (14) is generally applicable for use with multiple imputation for likelihood models. In particular, it may be applied with imputations obtained via the method of multiple imputation via chained equations (Raghunathan et al., 2001). This technique draws values from (Bayesian) predictive distributions, allowing for other models than the logistic one.

Acknowledgements

The authors wish to thank all reviewers of this paper for their constructive comments and questions. This research is supported in part by the Fund for Scientific Research Flanders

(G0542.06).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Blackhurst, D. W. and Schluchter, M. D. (1989). Logistic regression with a partially observed covariate. *Communication in Statistics - Simulation*, 18(1):163–177.

Cavanaugh, J. E. and Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference*, 67:45–65.

Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, 64(4):1062–1069.

Consentino, F. and Claeskens, G. (2009). Order selection tests with multiply-imputed data. FBE Research Report KBI 0905, K.U.Leuven.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from imcomplete data via em algorithm. *Journal of the Royal Society, Series B*, 39(1):1–22.

Gao, S. and Hui, S. L. (1997). Logistic regression models with missing covariate values for complex survey data. *Statistics in Medicine*, 16:2419–2428.

Gilks, W. R. and Wild, P. (1992). Adaptive rejction sampling for Gibbs sampling. *Applied Statistics*, 41:337–348.

Hens, N., Aerts, M., and Molenberghs, G. (2006). Model selection for incomplete and design-based samples. *Statistics in Medicine*, 25(14):2502–2520.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85:765–769.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999a). Monte carlo EM for missing covariates in parametric regression models. *Biometrics*, 55:591–596.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M.-H. (1999b). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society, Series B*, 61(1):173–190.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *J. Amer. Statist. Assoc.*, 86(416):1065–1073.

Lipsitz, Stuart, R., Parzen, M., and Ewell, M. (1998). Inference using conditional logistic regression with missing covariates. *Biometrics*, 54:295–303.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.

Liu, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of Multivariate Analysis*, 53:139–158.

Liu, C. and Rubin, D. B. (1995). ML estimation of the multivariate t distribution with unknown degrees of freedom. *Statistica Sinica*, 5:19–39.

Meng, X.-L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1):103–111.

Raghunathan, T. E., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values. *Survey Methodology*, 27(1):85–95.

Schomaker, M., Heumann, C., and Toutenburg, H. (2007). New approaches for model selection under missing data. Technical report, Department of Statistics, LudwigMaximiliansUniversitäat München.

Shimodaira, H. (1994). A new criterion for selecting models from partially observed data. In Cheeseman, P. and Oldford, R. W., editors, *Selecting models from data: artificial intelligence and statistics IV*, pages 21–29. Springer, New York.

Yang, X., Belin, T. R., and Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, 61(2):498–506.