EIEF

E I N A U D I   I N S T I T U T E   F O R   E C O N O M I C S   A N D   F I N A N C E

Bayesian Model Averaging for Generalized Linear Models with Missing Covariates

by

Valentino Dardanoni

(University of Palermo)

Giuseppe De Luca

(University of Palermo)

Salvatore Modica

(University of Palermo)

Franco Peracchi

(University of Rome "Tor Vergata" and EIEF)

BAYESIAN MODEL AVERAGING FOR GENERALIZED LINEAR MODELS WITH MISSING COVARIATES

VALENTINO DARDANONI, GIUSEPPE DE LUCA, SALVATORE MODICA, AND FRANCO PERACCHI

ABSTRACT. We address the problem of estimating generalized linear models (GLMs) when the outcome of interest is always observed, the values of some covariates are missing for some observations, but imputations are available to fill-in the missing values. Under certain conditions on the missing-data mechanism and the imputation model, this situation generates a trade-off between bias and precision in the estimation of the parameters of interest. The complete cases are often too few, so precision is lost, but just filling-in the missing values with the imputations may lead to bias when the imputation model is either incorrectly specified or uncongenial. Following the generalized missing-indicator approach originally proposed by Dardanoni et al. (2011) for linear regression models, we characterize this bias-precision trade-off in terms of model uncertainty regarding which covariates should be dropped from an augmented GLM for the full sample of observed and imputed data. This formulation is attractive because model uncertainty can then be handled very naturally through Bayesian model averaging (BMA). In addition to applying the generalized missing-indicator method to the wider class of GLMs, we make two extensions. First, we propose a block-BMA strategy that incorporates information on the available missing-data patterns and has the advantage of being computationally simple. Second, we allow the observed outcome to be multivariate, thus covering the case of seemingly unrelated regression equations models, and ordered, multinomial or conditional logit and probit models. Our approach is illustrated through an empirical application using the first wave of the Survey on Health, Aging and Retirement in Europe (SHARE).

Keywords: Bayesian model averaging; generalized linear models; missing covariates; generalized missing-indicator method; SHARE.

JEL classification numbers: C11; C25; C35; C81.

1. Introduction

In this paper we address the problem of estimating generalized linear models (GLMs) in the empirically relevant case when the outcome of interest is always observed, the values of some covariates are missing for some observations, but imputations are available to fill-in the missing values. This situation is becoming quite common, as public-use data files increasingly include imputations of key variables affected by missing data problems, and specialized software for carrying out imputations directly is also becoming increasingly available.

Two standard approaches to the problem of missing covariate values are complete-case analysis and the fill-in approach. The first approach drops all the observations with missing covariate values ignoring the imputations altogether, while the second approach fills in the missing covariate values with the available imputations without distinguishing between observed and imputed values. Under certain conditions on the missing-data mechanism and the imputation model, the choice between these two approaches generates a trade-off between bias and precision in the estimation of the parameters of interest. The complete cases are often too few, so precision is lost, but just filling in the missing values with the imputations may lead to bias when the imputation model is either incorrectly specified or uncongenial in the sense of Meng (1994).

Following the generalized missing-indicator approach originally proposed for linear regression models by Dardanoni et al. (2011), we transform this bias-precision trade-off into a problem of uncertainty about which regressors should be dropped from an augmented GLM, or 'grand model', which includes two subsets of regressors: the focus regressors, corresponding to the observed or imputed covariates, and a set of auxiliary regressors consisting of binary indicators for the various missing-data patterns plus their interactions with the focus regressors. Our formulation of the bias-precision trade-off in terms of model uncertainty exploits the fact that complete-case analysis and the fill-in correspond to using two extreme specifications of the grand model. Complete-case analysis corresponds to using a fully unrestricted specification, while the fill-in approach corresponds to using a restricted specification that includes only the focus regressors.

A natural way of handling model uncertainty is Bayesian model averaging (BMA). In our context, this implies considering all the intermediate specifications obtained by dropping from the grand model alternative subsets of auxiliary regressors. In this way, we avoid restricting attention to the complete cases but, at the same time, we exploit the available imputations in

a sensible way by allowing the imputation model to be incorrectly specified or uncongenial with the GLM of interest. The extreme choice of using the complete-case or the fill-in approach is still available, but it is unlikely to emerge as the best one since all the intermediate models in the expanded model space carry information about the parameters of interest.

In addition to applying the generalized missing-indicator method to the wider class of GLMs, we make two important extensions. First, we propose a block-BMA strategy that incorporates the information on the available patterns of missing data while being computationally simple. Second, we allow the observed outcome to be multivariate, thus covering the case of seemingly unrelated regression equations (SURE) models and ordered, multinomial or conditional logit and probit models. STATA commands which implement our procedure are available upon request.

To illustrate our methods, we analyze how cognitive functioning varies with physical health and socio-economic status using data from the first (2004) wave of the Survey on Health, Aging and Retirement in Europe (SHARE), a multi-purpose cross-national household panel carried out in 11 continental European countries. Like other household surveys, sensitive variables such as household income, household net worth, and other objective health measures are affected by substantial item nonresponse.

The remainder of the paper is organized as follows. Section 2 presents our statistical framework. Section 3 discusses complete-case analysis and the fill-in approach. Section 4 describes the generalized missing-indicator method. Section 5 discusses how to apply BMA to deal with model uncertainty in the context of the grand model. Section 6 extends our results to the case of multivariate outcomes. Section 7 presents an empirical application. Finally, Section 8 offers some conclusions.

2. STATISTICAL FRAMEWORK

We represent the available set of N observations on an outcome of interest as the realization of a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$ whose components are independently distributed random variables with mean μ_n and finite nonzero variance σ_n^2, $n = 1, \ldots, N$. We assume that the distribution of each component of $\boldsymbol{Y}$ belongs to the (one-parameter) linear exponential family with density function of the form

$$f(y; \gamma_n) = \exp\left[\gamma_n\, y - b(\gamma_n) + c(y)\right], \qquad n = 1, \ldots, N, \tag{1}$$

where γ_n is a scalar parameter that may vary across observations depending on a K-dimensional vector of covariates X_n (assumed to always include a constant term), $b(\cdot)$ is a known, strictly convex and twice differentiable function, and $c(\cdot)$ is a known function. In the original formulation of Nelder and Wedderburn (1972), the density function (1) includes an additional dispersion parameter which, without loss of generality, we set equal to one. By the properties of the linear exponential family, the mean and variance of Y_n are equal to $\mu_n = b'(\gamma_n)$ and $\sigma_n^2 = b''(\gamma_n)$ respectively (McCullagh and Nelder 1989).

In a GLM, the dependence of Y_n on the vector of covariates X_n is modelled by assuming that there exists a continuously differentiable and invertible function $h(\cdot)$, sometimes called the inverse link, such that the mean of Y_n is equal to $\mu_n = h(X_n^\top \beta)$ for a unique value of the K-dimensional parameter vector β. The linear combination $\eta_n = \beta^\top X_n$ is called the linear predictor associated with the nth observation. Collecting together the linear predictors associated with the sample observations gives the N-dimensional vector $\boldsymbol{\eta} = \boldsymbol{X}\beta$, where $\boldsymbol{X}$ is the $N \times K$ matrix of observations on the covariates.

In the absence of missing data, the classical approach to estimating β is maximum likelihood (ML). The sample log-likelihood for the missing-free data is

$$L(\beta) = c + \sum\nolimits_{n=1}^{N} \left[\gamma_n(\beta) Y_n - b\left(\gamma_n(\beta)\right) \right],$$

where $\gamma_n(\beta)$ is the unique root of the equation $b'(\gamma) = h(\beta^\top X_n)$, and the missing-free data ML estimator of β is obtained by solving the system of K likelihood equations

$$0 = L'(\beta) = \sum\nolimits_{n=1}^{N} v(\beta^\top X_n) \left[Y_n - h(\beta^\top X_n) \right] X_n, \tag{2}$$

where $v(\beta^\top X_n) = h'(\beta^\top X_n)/b''(\gamma_n(\beta))$. Notice that β enters the above equations only through the linear predictor $\eta_n = \beta^\top X_n$. This is the property that drives our main result in Theorem 1 below. If $b'(\cdot) = h(\cdot)$ (the "canonical link" case), then $\gamma_n(\beta) = \beta^\top X_n$ and the likelihood equations (2) simplify considerably because $v(\beta^\top X_n) = 1$ for all n. When the GLM is correctly specified, and the mild regularity conditions in Fahrmeir and Kaufmann (1985) hold, the missing-free data ML estimator of β is unique, consistent, and asymptotically normal with asymptotic variance equal to the inverse of the Fisher information matrix.

In this paper we depart from the standard GLM setup in two ways. First, we assume that some covariate values are missing for some observations. Second, we assume that imputations are available to fill-in the missing covariate values. Since the constant term is always observed,

the number of possible subsamples with missing covariates is equal to $2^{K-1} - 1$ so, including the subsample with complete data, the number of missing-data patterns is equal to 2^{K-1}. Because a particular data set need not to contain all the possible patterns, we index by $j = 0, \ldots, J$ the patterns that are present in the data, with $j = 0$ corresponding to the subsample with complete data and $J \le 2^{K-1} - 1$.

Let N_j be the number of observations, K_j the number of observed (non missing) covariates and $\bar{K}_j = K - K_j$ the number of missing covariates in the jth missing data pattern. By definition, $\sum_{j=0}^{J} N_j = N$, $K_0 = K$, and $1 \le K_j \le K$ for $j = 1, \ldots, J$. For each missing-data pattern, let $\boldsymbol{Y}_j$ be the $N_j \times 1$ vector of observations on the outcome of interest and let $\boldsymbol{X}_j$ be the $N_j \times K$ matrix containing the values of the covariates, which could be either observed or missing. Clearly, $\boldsymbol{X}_0$ is always observed. To keep track of which covariate values are observed and which are missing, we define the $N \times K$ missing indicator matrix $\boldsymbol{M}$, whose (n, k)th element is equal to one if the kth covariate is missing for the nth observation, and is equal to zero otherwise. Finally, for each subsample $j = 1, \ldots, J$ with missing covariates, we let $\boldsymbol{W}_j$ be the $N_j \times K$ matrix containing the values of the K_j observed covariates and the imputed values of the $\bar{K}_j$ missing covariates. We shall refer to $\boldsymbol{W}_j$ as the filled-in design matrix for the jth subsample.

3. Complete-case analysis and the fill-in approach

This section discusses two standard approaches to the problem of missing covariate values, namely complete-case analysis and the fill-in approach.

3.1. **Complete-case analysis.** This approach amounts to estimating a GLM on the subsample $[\boldsymbol{X}_0, \boldsymbol{Y}_0]$ without missing covariates ignoring the imputations altogether. Complete-case analysis is a useful benchmark because it gives a consistent ML estimator of β under the following two assumptions (Wooldridge 2010, p. 798).

Assumption 1. *The Fisher information matrix for the subsample with complete data is positive definite with probability approaching one as $N \to \infty$.*

Assumption 2. *$\boldsymbol{Y}$ and $\boldsymbol{M}$ are independent conditionally on $\boldsymbol{X}$.*

Assumption 1 guarantees that the model parameters are identified using only the information in the subsample with complete data. Because the function $b(\cdot)$ is strictly convex, this assumption holds if the matrix $N^{-1} \boldsymbol{X}_0^\top \boldsymbol{X}_0$ converges in probability to a positive definite matrix as $N \to \infty$.

Assumption 2 is an ignorability assumption on the missing-data process. It implies that the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and $\boldsymbol{M}$ is the same as the distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ or, equivalently, the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is the same in subsamples with and without missing covariates. Given the true values of the covariates, the pattern of missing data can then be ignored when predicting $\boldsymbol{Y}$. Notice that Assumption 2 is stronger than the conditional mean independence assumption needed to ensure unbiasedness of the complete-case OLS estimator of β in classical linear regression models, but is weaker than the missing completely at random (MCAR) assumption which instead requires that the distribution of $\boldsymbol{M}$ does not depend on $\boldsymbol{Y}$ and $\boldsymbol{X}$. Also notice that Assumption 2 is not the same as the standard missing at random (MAR) assumption, usually imposed when imputing missing values, which requires the missing-data process to be independent of the missing covariates given the observed data (Rubin 1976). For example, suppose that health is the outcome of interest and the only covariate, household income, is subject to missing data problems. If missing income depends on true income but not on health, then Assumption 2 is satisfied but MAR is not, while if missing income depends on health but not on true income then MAR is satisfied but Assumption 2 is not. Thus, Assumption 2 is neither stronger nor weaker than MAR.

Although the asymptotic results implied by Assumptions 1 and 2 provide the main justification for complete-case analysis, one cannot ignore the severe loss of precision that may result from this approach when the fraction of missing data is not small.

3.2. **Fill-in approach.** Reordering the observations by stacking on top of each other the $J + 1$ available missing-data patterns gives

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_0 \\ \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_J \end{bmatrix}, \qquad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{X}_0 \\ \boldsymbol{W}_1 \\ \vdots \\ \boldsymbol{W}_J \end{bmatrix},$$

where the $N \times K$ matrix $\boldsymbol{W}$ is called the filled-in design matrix for the whole sample. The fill-in approach consists of estimating a GLM for $\boldsymbol{Y}$ replacing $\boldsymbol{X}$ by $\boldsymbol{W}$.

The validity of this approach requires two conditions: (i) the model used to create the imputations must be correctly specified (including the assumptions on the posited missing-data mechanism); and (ii) the imputation model and the GLM for the filled-in data $[\boldsymbol{Y}, \boldsymbol{W}]$ must be congenial in the sense of Meng (1994), i.e., the imputation model cannot be more restrictive than the model used to analyze the filled-in data). We say that the fill-in approach is valid when these

two conditions hold, in which case the fill-in ML estimator of β is asymptotically equivalent to the missing-free data ML estimator introduced in Section 2. Further, as shown in Appendix A, this estimator is asymptotically more precise than the complete-case ML estimator introduced in Section 3.1.

In general, little can be said about the finite sample properties of these two ML estimators. Since the number of unknown parameters is the same in the complete-case and the fill-in approaches, but the number of observations is greater in the latter, the fill-in ML estimator is expected to have higher precision than the complete-case ML estimator provided that the additional sampling variability induced by imputation is small. On the other hand, if the imputation model is not correctly specified or is not congenial, then the fill-in estimator is likely to be biased and inconsistent because it ignores the fact that the imputations are not the same as the missing covariate values.

We would like to stress the importance of the assumption that the imputation model is congenial. If the model of interest and the imputation model are uncongenial, because they are based either on different parametric assumptions or on different sets of explanatory variables, then the fill-in approach may lead to inconsistent estimates. This is especially true in the case of nonlinear estimators, such as ML estimators for GLMs. We refer to Nicoletti and Peracchi (2006) for a simple test of congeniality.

An additional issue with the fill-in estimator is how to account for the additional variability induced by the imputation process when assessing the precision of this estimator, a problem that we ignore throughout this paper because it is easily handled by multiple imputation methods (Rubin 1987; Dardanoni et al. 2012).

4. THE GENERALIZED MISSING-INDICATOR APPROACH

The key idea of this approach is to augment the set of K observed or imputed covariates in the filled-in design matrix $\boldsymbol{W}$ with a set of JK additional regressors corresponding to binary indicators for the subsamples with missing covariate values and their interactions with the regressors in $\boldsymbol{W}$. Thus define the $N \times JK$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} 0 & \cdots & 0 \\ \boldsymbol{W}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{W}_J \end{bmatrix}.$$

Our statistical model for the full sample is an augmented GLM where the conditional density of $\boldsymbol{Y}$ given $\boldsymbol{W}$ and $\boldsymbol{Z}$ belongs to the exponential family (1) with linear predictor $\boldsymbol{\eta} = \boldsymbol{W}\beta + \boldsymbol{Z}\delta$. The columns of $\boldsymbol{W}$ represent, in the terminology of Danilov and Magnus (2004), our focus regressors, while the columns of $\boldsymbol{Z}$ represent the auxiliary regressors. Following Dardanoni et al. (2011), we shall refer to this augmented GLM as the *grand model*. As shown in Appendix B, the JK-dimensional vector of auxiliary parameters δ can be interpreted as the asymptotic bias of the fill-in estimator of the focus parameter β when the imputations are not valid.

4.1. **Equivalence theorem.** The following theorem extends to GLMs the main result in Dardanoni et al. (2011).

Theorem 1. *For any set of imputations, the ML estimator of β in the grand model with linear predictor $\boldsymbol{\eta} = \boldsymbol{W}\beta + \boldsymbol{Z}\delta$ is equal to the complete-case ML estimator of β.*

Proof. Let

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_0 \\ \boldsymbol{Y}_* \end{bmatrix}, \qquad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{X}_0 \\ \boldsymbol{W}_* \end{bmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix} 0 \\ \boldsymbol{Z}_* \end{bmatrix},$$

where

$$\boldsymbol{Y}_* = \begin{bmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_J \end{bmatrix}, \qquad \boldsymbol{W}_* = \begin{bmatrix} \boldsymbol{W}_1 \\ \vdots \\ \boldsymbol{W}_J \end{bmatrix}, \qquad \boldsymbol{Z}_* = \begin{bmatrix} \boldsymbol{W}_1 & & \\ & \ddots & \\ & & \boldsymbol{W}_J \end{bmatrix}.$$

The complete-case ML estimator $\widehat{\beta}$ of β solves the system of K likelihood equations

$$\boldsymbol{X}_0^\top \boldsymbol{U}_0(\beta) = 0,$$

where $\boldsymbol{U}_0(\beta)$ is the $N_0 \times 1$ vector of generalized residuals (Gourieroux et al. 1987) with generic element equal to $v(\beta^\top X_n)\left[Y_n - h(\beta^\top X_n)\right]$. On the other hand, the ML estimator $(\widetilde{\beta}, \widetilde{\delta})$ of (β, δ) in the grand model with linear predictor equal to $\boldsymbol{W}\beta + \boldsymbol{Z}\delta$ solves the system of $K(1 + J)$ likelihood equations

$$\boldsymbol{X}_0^\top \boldsymbol{U}_0(\beta) + \boldsymbol{W}_*^\top \boldsymbol{U}_*(\beta, \delta) = 0,$$
$$\boldsymbol{Z}_*^\top \boldsymbol{U}_*(\beta, \delta) = 0, \tag{3}$$

where $\boldsymbol{U}_*(\beta, \delta)$ is the $(N - N_0) \times 1$ vector with generic element equal to $v(\beta^\top W_n + \delta^\top Z_n)\left[Y_n - h(\beta^\top W_n + \delta^\top Z_n)\right]$. Since $\boldsymbol{Z}_*$ is a block-diagonal matrix, the last JK equations in (3) imply that

$$\boldsymbol{W}_*^\top \boldsymbol{U}_*(\beta, \delta) = 0.$$

The ML estimator $\widetilde{\beta}$ then solves $\boldsymbol{X}_0^\top \boldsymbol{U}_0(\beta) = 0$, so it must coincide with the complete-case estimator $\widehat{\beta}$. $\qquad\qquad\square$

Theorem 1 shows that the complete-case ML estimator of β coincides with the unrestricted ML estimator of β in the grand model that places no restrictions on the vector δ of auxiliary parameters. The intuition for this result is that estimating the grand model is essentially equivalent to jointly estimating a set of unrestricted GLMs, one for each missing-data pattern. The model for the subsample without missing covariates determines the estimate of β, which coincides by construction with the complete-case ML estimate and is therefore consistent under the assumptions in Section 3.1. Given the estimate of β, the various elements of δ are estimated from the models for the subsamples with imputed covariates.

The fill-in estimator of β coincides instead with the restricted ML estimator when all elements of δ in the grand model are set to zero. If the imputations are valid (that is, $\delta = 0$), then the fill-in estimator is asymptotically more precise than the complete-case estimator (see Appendix A). However, if the imputations are not valid (that is, $\delta \neq 0$), then the fill-in estimator is inconsistent. Thus, testing the hypothesis that $\delta = 0$ provides an alternative to the test for the validity of imputations proposed by Nicoletti and Peracchi (2006).

The aim of the generalized missing-indicator approach is to handle the trade-off between bias and precision in the estimation of β by considering all intermediate models obtained from the grand model by setting to zero arbitrary subsets of elements in δ. This strategy has two advantages. First, the original bias-precision trade-off is transformed into a problem of uncertainty about a subset of covariates of the grand model, for which a variety of alternative strategies are available. Second, instead of focusing on two extreme specifications of the grand model, any intermediate model in the expanded model space may now play a role in constructing the best available estimator of β in the asymptotic mean squared error sense.

4.2. **A dual result.** We now prove a result that may be regarded as a dual of Theorem 1. This result provides the justification for the block-BMA approach presented in Section 5.4.

Theorem 1 says that the complete-case approach is equivalent, as far as estimation of β is concerned, to using the grand model that includes all the observations (observed or imputed) and all the auxiliary regressors. The next theorem shows that, more generally, pooling together the subsample with complete data and arbitrary subsamples with missing covariates is equivalent to removing blocks of auxiliary regressors from the grand model.

Given a collection $\mathcal{J}$ of subsamples with missing covariates, let $\boldsymbol{Y}^+$ be the subvector of $\boldsymbol{Y}$ obtained by stacking $\boldsymbol{Y}_0$ and all $\boldsymbol{Y}_j$ such that $j \in \mathcal{J}$, and $\boldsymbol{Y}^-$ be the subvector consisting of the

remaining rows of $\boldsymbol{Y}$; let $\boldsymbol{W}^+$ be the submatrix of $\boldsymbol{W}$ obtained by stacking $\boldsymbol{X}_0$ and all $\boldsymbol{W}_j$ such that $j \in \mathcal{J}$, and $\boldsymbol{W}^-$ be the submatrix consisting of the remaining rows of $\boldsymbol{W}$; and let $\boldsymbol{Z}_*^-$ be the submatrix obtained by dropping from $\boldsymbol{Z}_*$ the rows and columns containing the elements of $\boldsymbol{W}^+$. Finally, let

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}^+ \\ \boldsymbol{Y}^- \end{bmatrix}, \quad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}^+ \\ \boldsymbol{W}^- \end{bmatrix}, \quad \boldsymbol{Z}^- = \begin{bmatrix} 0 \\ \boldsymbol{Z}_*^- \end{bmatrix}.$$

Theorem 1 can now be restated as follows: If $\mathcal{J}$ is the empty set, then the ML estimates of β in the GLM for $[\boldsymbol{Y}^+, \boldsymbol{W}^+]$ and in the GLM for $[\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z}^-]$ coincide. Next theorem shows that this is actually true if $\mathcal{J}$ is any collection of subsamples with missing covariates.

Theorem 2. *For any collection $\mathcal{J}$ of subsamples with missing covariates, the ML estimates of β in the GLM for $[\boldsymbol{Y}^+, \boldsymbol{W}^+]$ and in the GLM for $[\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z}^-]$ coincide.*

Proof. Let $\boldsymbol{U}^+(\beta)$ be the vector of dimension $N_0 + \sum_{j \in \mathcal{J}} N_j$ with generic element $U_n^+(\beta) = v(\beta^\top W_n^+)[Y_n^+ - h(\beta^\top W_n^+)]$. Also let $\boldsymbol{U}^-(\beta, \delta^-)$ be the vector of dimension $N - (N_0 + \sum_{j \in \mathcal{J}} N_j)$ with generic element $v(\beta^\top W_n + \delta^{-\top} Z_n^-)\,[Y_n - h(\beta^\top W_n + \delta^{-\top} Z_n^-)]$, where δ^- denotes the subvector of δ obtained by deleting the coefficients associated with the $\boldsymbol{W}_j$, $j \in \mathcal{J}$. The proof of the theorem follows immediately from the proof of Theorem 1 after replacing $\boldsymbol{X}_0$, $\boldsymbol{U}_0(\beta)$, $\boldsymbol{W}_*$, $\boldsymbol{U}_*(\beta, \delta)$, and $\boldsymbol{Z}_*$ with $\boldsymbol{W}^+$, $\boldsymbol{U}^+(\beta)$, $\boldsymbol{W}^-$, $\boldsymbol{U}^-(\beta, \delta^-)$, and $\boldsymbol{Z}_*^-$ respectively. $\square$

Thus, removing the vector δ_j from the grand model gives the same estimate of β that would be obtained from the model without auxiliary regressors when using only the complete data and the jth subsample with missing covariates. Therefore each δ_j controls for a particular subsample with missing covariates, a feature that we exploit in our Bayesian model averaging procedure in Section 5.4.

5. ESTIMATION UNDER MODEL UNCERTAINTY

Model uncertainty can be handled by either model selection or model averaging. In model selection one first selects the best model in the available model space and then estimates β conditional on the selected model. A problem with this approach is pre-testing. As shown by Magnus and Durbin (1999), Burnham and Anderson (2002) and Danilov and Magnus (2004), the initial model selection step matters and is likely to have nonnegligible effects on the statistical properties of the resulting estimates.

Model averaging provides a more coherent approach to inference because it takes explicitly into account uncertainty due to both the estimation and the model selection steps. In this case, the parameters of interest are first estimated conditional on each model in the model space, then an unconditional estimate is computed using a weighted average of these conditional estimates.

Suppose that the model space $\mathcal{M}$ includes R possible GLMs, that is, $\mathcal{M} = \{M_1, \ldots, M_R\}$. The rth model M_r is obtained by including in the linear predictor the K focus regressors in $\boldsymbol{W}$ and only a subset of $0 \leq P_r \leq JK$ auxiliary regressors in $\boldsymbol{Z}$. Thus, the linear predictor for the rth model is equal to $\boldsymbol{\eta_r} = \boldsymbol{W}\beta + \boldsymbol{Z}_r\delta_r$, where $\boldsymbol{Z}_r$ is the matrix containing the N observations on the included subset of P_r auxiliary regressors and δ_r is the corresponding vector of coefficients. The model averaging estimates of β and δ are of the form

$$
\begin{aligned}
\widehat{\beta} &= \sum_{r=1}^{R} \lambda_r \widehat{\beta}_r, \\
\widehat{\delta} &= \sum_{r=1}^{R} \lambda_r S_r \widehat{\delta}_r,
\end{aligned}
\tag{4}
$$

where the λ_r are non-negative weights that add up to one, the $\widehat{\beta}_r$ and $\widehat{\delta}_r$ are the estimates of β and δ_r under the rth model, and the S_r are $JK \times P_r$ selection matrices that transform the P_r-dimensional vectors of conditional estimate $\widehat{\delta}_r$ into JK-dimensional vectors by setting to zero the elements of δ which are excluded from the rth model.

5.1. **Bayesian model averaging.** In Bayesian model averaging (BMA) the conditional estimates $\widehat{\beta}_r$ and $\widehat{\delta}_r$ are weighted by the posterior probability of the rth model to reflect our confidence in that model based on prior beliefs and the observed data. Thus

$$
\lambda_r = p(M_r \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid M_r)\, p(M_r)}{\sum_{r=1}^{R} p(\boldsymbol{Y} \mid M_r)\, p(M_r)}, \qquad r = 1, \ldots, R,
\tag{5}
$$

where $p(M_r)$ is the prior probability of model M_r,

$$
p(\boldsymbol{Y} \mid M_r) = \int p(\boldsymbol{Y} \mid \theta_r, M_r)\, p(\theta_r \mid M_r)\, d\theta_r
\tag{6}
$$

is the marginal likelihood of the rth model, $\theta_r = (\beta, \delta_r)$ the vector of its parameters, $p(\boldsymbol{Y} \mid \theta_r, M_r)$ its likelihood, and $p(\theta_r \mid M_r)$ the prior density of θ_r under the rth model. In this setting the model averaging estimates in (4) can be interpreted as the posterior means of the distribution of β and δ. The posterior variance-covariance matrix consists of the following blocks (Raftery

1993; Draper 1995)

$$\text{Var}(\widehat{\beta} \mid \boldsymbol{Y}) = \sum_{r=1}^{R} \lambda_r \left[\text{Var}\left(\widehat{\beta}_r \mid \boldsymbol{Y}, M_r\right) + \widehat{\beta}_r \widehat{\beta}_r^{\top} \right] - \widehat{\beta}\widehat{\beta}^{\top},$$

$$\text{Var}(\widehat{\delta} \mid \boldsymbol{Y}) = \sum_{r=1}^{R} \lambda_r S_r \left[\text{Var}\left(\widehat{\delta}_r \mid \boldsymbol{Y}, M_r\right) + \widehat{\delta}_r \widehat{\delta}_r^{\top} \right] S_r^{\top} - \widetilde{\delta}\widetilde{\delta}^{\top},$$

$$\text{Cov}(\widehat{\beta}, \widehat{\delta} \mid \boldsymbol{Y}) = \sum_{r=1}^{R} \lambda_r \left[\text{Cov}\left(\widehat{\beta}_r, \widehat{\delta}_r \mid \boldsymbol{Y}, M_r\right) + \widehat{\beta}_r \widehat{\delta}_r^{\top} \right] S_r^{\top} - \widehat{\beta}\widetilde{\delta}^{\top}.$$

The posterior variances $\text{Var}(\widehat{\beta} \mid \boldsymbol{Y})$ and $\text{Var}(\widehat{\delta} \mid \boldsymbol{Y})$ consist of two components: the weighted average of the conditional variances in each model and the weighted variance of the conditional estimates across models. Thus, unlike pretest estimators, the posterior variance of the BMA estimator incorporates the uncertainty due to both parameter estimation and model selection.

The choice between alternative BMA estimates depends on the strategies used to handle a number of methodological and computational problem. The main problems are: (i) the specification of the prior probabilities $p(M_r)$ of the various models, (ii) the specification of the prior distribution $p(\theta_r \mid M_r)$ for the parameters of each model, (iii) the procedure to evaluate the integrals in (6), which do not usually have closed form solutions in the context of GLMs, and (iv) the procedure to compute the posterior model probabilities in (5) when exploring all models is infeasible due to the large dimension of the model space.

5.2. **Choice of priors.** As for problem (i), the assumption that all models are equally likely a priori is a reasonable neutral choice when there is little prior information about the relative plausibility of the models considered (Hoeting et al. 1999). This choice, which corresponds to assuming a uniform prior distribution on the model space, implies that the posterior model probabilities depend only on the marginal likelihoods of the various models but not on the prior weight assigned to each of them.

As for problem (ii), our choice of prior distributions over the parameters in the rth model is the family of calibrated information criteria (CIC) prior distributions introduced by Clyde (2000). This is a family of uninformative priors derived from the following modification of Jeffrey's prior (Jeffreys 1961)

$$p(\theta_r \mid M_r) = (2\pi)^{-d_r/2} \left| \frac{1}{c} \mathcal{I}(\widehat{\theta}_r) \right|^{1/2},$$

where $d_r = K + P_r$ is the number of parameters in the rth model, $\mathcal{I}(\widehat{\theta}_r)$ is the observed Fisher information for the rth model evaluated at the ML estimate $\widehat{\theta}_r$, and c is a hyperparameter which allows one to calibrate the posterior model probabilities to classical model selection criteria like the Akaike Information Criterion (AIC; Akaike 1978), the Bayesian Information Criterion (BIC;

Schwarz 1996), or the Risk Inflation Criterion (RIC; Foster and George 1994). The use of BMA with a weighting scheme based on BIC was originally suggested by Raftery (1996), who showed that BIC is an approximation to twice the logarithm of the Bayes factor for model M_r against the restricted model with $\delta = 0$. Clyde's formulation of the CIC prior is attractive because it provides a general Bayesian justification for the entire family of model selection criteria.

5.3. **Evaluating marginal likelihoods.** As for problem (iii), the integrals in (6) may be difficult to evaluate, except for linear regression models where closed form solutions are sometimes available (see e.g. Magnus et al. 2010), so approximations to the marginal likelihood of each model are typically used. Alternatively, the problem may be circumvented by directly estimating the posterior probability of each model using Markov chain Monte Carlo (MCMC) methods.

For regular statistical models approximations to the marginal likelihoods may be obtained by the Laplace method for integrals (Tierney and Kadane 1986). As suggested by Kass and Raftery (1995), this method is reasonably accurate when the sample size is greater than 20 times the number of covariates. Moreover, its use has been justified by several authors (Raftery 1995, 1996; Hoeting et al. 1999; Clyde 2000; Volinsky and Raftery 2000; Clyde and George 2004). On the basis of this approximation, Clyde (2000) shows that the posterior probability of model M_r is

$$p(M_r \mid \boldsymbol{Y}) \simeq \frac{\exp\left[1/2 \left(D_r - d_r \, \log c\right)\right]}{\sum_{h=1}^{R} \exp\left[1/2 \left(D_h - d_h \, \log c\right)\right]},$$

where D_r is the deviance of model M_r (-2 times the log-likelihood ratio between model M_r and the restricted model with $\delta = 0$). Hence, under CIC priors, the logarithm of the posterior probability of each model is approximately proportional to its deviance minus a penalty for complexity, which depends on the hyperparameter c. The posterior model probabilities can be calibrated to classical model selection criteria by setting $\log c = 2$ for AIC, $\log c = \log N$ for BIC, and $\log c = 2 \log P_r$ for RIC. Although debate over the choice of an optimal model-selection criterion is still open, AIC and BIC are known to be two extreme strategies which tend to favor, respectively, more and less complicated model structures. From this view point, CIC priors represent an attractive family of prior distributions for sensitivity analysis in BMA estimation.

5.4. **Block BMA.** Our last issue is how to handle the case when the number of candidate models in the model space $\mathcal{M}$ is large. If there are J subsamples with missing covariates and K covariates, the number of models obtained by dropping alternative subsets of auxiliary regressors is $R = 2^{JK}$. Even for moderate values of J and K, exploring all these models is unfeasible.

However, in the context of our generalized missing-indicator approach there are both theoretical and computational reasons to confine model uncertainty to the J blocks of auxiliary variables associated with the various missing-data patterns.

From a theoretical view point, the K auxiliary variables in a given block capture the asymptotic bias of the fill-in estimator of β due to the imputation of the missing covariate values. Thus, instead of considering the separate contribution of the JK auxiliary variables in $\boldsymbol{Z}$, we may just consider the separate contribution of the J blocks of auxiliary variables corresponding to each of the subsamples with missing covariate values.

From a computation view point, a great advantage of our block-BMA procedure is its simplicity, as the dimension of the model space reduces to $R = 2^J$. Thus, in typical applications where J does not exceed 20, one may proceed by directly exploring all models. When J is large, our block-BMA procedure may be combined with some deterministic or stochastic search method over the space of 2^J models. For example, deterministic search strategies such as the Occam's window of Madigan and Raftery (1994) and the leaps and bounds algorithm of Furnival and Wilson (1974) may be used for moderately sized problems where J does not exceed 30. For larger problems, these methods can be too expensive computationally or may not explore a large enough region of the model space leading to poor predictive performances (Hoeting et al. 1999). More accurate results can then be achieved by stochastic search strategies based on MCMC methods, which allow exploring a considerably larger subset of models and proivide direct estimates of the posterior model probabilities using the proportion of times the Markov chain visits each model. We refer to Han and Carlin (2001) and Clyde and George (2004) for a review of the methodological and computational issues arising with the various MCMC methods.

6. The multivariate case

The results of Section 4 extend to settings where there is more than one outcome of interest and the nth observation Y_n is a Q-dimensional vector whose distribution is assumed to belong to the multivariate exponential family. These include models for ordered or unordered multinomial outcomes where the outcome of interest can take $Q + 1$ possible values corresponding to $Q + 1$ mutually exclusive categories, leading to ordered, multinomial or conditional logit and probit regressions.

The expression for the density of Y_n is now

$$f(y; \gamma_n) = \exp \left[\gamma_n^\top y - b(\gamma_n) + c(y) \right], \tag{7}$$

where γ_n is a Q-dimensional vector that may depend on covariates, and $b(\cdot)$ and $c(\cdot)$ are known functions which satisfy the regularity conditions in Fahrmeir and Kaufmann (1985). The mean and variance of Y_n are equal to $\mu_n = b'(\gamma_n)$ and $\Sigma_n = b''(\gamma_n)$ respectively, where $b'(\cdot)$ is the Q-dimensional gradient vector and $b''(\cdot)$ is the $Q \times Q$ Hessian matrix of $b(\cdot)$.

Given a K-dimensional vector of covariates X_n, the linear predictor associated with the qth component of Y_n is $\beta_q^\top X_n$, with $\beta_q \in \mathbb{R}^K$. Stacking all the β_q into the QK-dimensional vector $\beta = (\beta_1^\top, \ldots, \beta_Q^\top)^\top$, the linear predictor associated with Y_n is the Q-dimensional vector $\eta_n = (I_Q \otimes X_n^\top)\beta$, where I_Q is the $Q \times Q$ unit matrix and $\otimes$ is Kronecker's product. As before, the dependence of Y_n on the covariates is modelled by assuming that there exists a continuously differentiable and invertible function $h \colon \mathbb{R}^Q \to \mathbb{R}^Q$ such that the mean of Y_n is equal to $\mu_n = h((I_Q \otimes X_n^\top)\beta)$ for a unique value of β.

The missing-free data now consist of the Q-dimensional vectors Y_n of observations on the outcomes of interest and the K-dimensional vectors X_n of covariates, with $n = 1, \ldots, N$, and the sample log-likelihood is

$$L(\beta) = c + \sum_{n=1}^{N} \left[\gamma_n(\beta)^\top Y_n - b(\gamma_n(\beta)) \right],$$

where the vector $\gamma_n(\beta)$ solves $b'(\gamma) = h((I_Q \otimes X_n^\top)\beta)$. The missing-free data ML estimator of β is obtained by solving the QK likelihood equations

$$0 = L'(\beta) = \sum_{n=1}^{N} (I_Q \otimes X_n) V_n(\beta) \left[Y_n - h \left((I_Q \otimes X_n^\top)\beta \right) \right],$$

where $V_n(\beta)$ is the transpose of the $Q \times Q$ matrix $[b''(\gamma_n(\beta))]^{-1} h'((I_Q \otimes X_n^\top)\beta)$. The conditions for uniqueness, consistency and asymptotic normality of this estimator are as before (Fahrmeir and Kaufmann 1985).

With missing covariates, we consider a grand model that now includes, in addition to the filled-in design matrix, a set of JK auxiliary regressors for each of the Q equations corresponding to the individual components of Y_n. The property that the vector β of parameters of interest enters the likelihood equations only through the linear predictor $\eta_n = (I_Q \otimes X_n^\top)\beta$ is all we need to adapt the proofs of the equivalence theorems in Section 4 to this case. To see this, is enough to write the grand model as an augmented GLM with linear predictor equal to NQ-dimensional

vector $\boldsymbol{\eta} = \boldsymbol{W}\beta + \boldsymbol{Z}\delta$, where

$$\boldsymbol{W} = \begin{bmatrix} I_Q \otimes W_1^\top \\ \vdots \\ I_Q \otimes W_N^\top \end{bmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix} I_Q \otimes Z_1^\top \\ \vdots \\ I_Q \otimes Z_N^\top \end{bmatrix},$$

and $\delta = (\delta_1^\top, \ldots, \delta_q^\top)^\top$ is a QJK-dimensional vector of auxiliary parameters. As before, our block-BMA procedure considers all intermediate models obtained from the grand model by simultaneously restricting arbitrary blocks of K elements in δ_q to be equal to zero for all q. The dimension of the model space is again $R = 2^J$.

7. Empirical application

In this section we use data on the elderly European population to investigate how cognitive functioning varies with physical health and socio-economic status. As argued by Mazzonna and Peracchi (2012), cognitive functioning is fundamental for decision making, for it influences individuals' ability to process information and to make the right choices.

Our data are from release 2.4.0 of the first wave of the Survey of Health, Ageing and Retirement in Europe (SHARE), a multidisciplinary and cross-national household panel survey which provides information on cognitive abilities, physical health, socio-economic status, and social networks for nationally representative samples of people aged 50 or more, plus their spouses irrespective of age, in the participating countries. The data can be freely downloaded from the SHARE web site (`http://www.share-project.org`). We refer to Börsch-Supan et al. (2005) for detailed information on survey design, target population, response rates and other methodological issues.

The first wave, conducted in 2004–05, covers about 28,500 individuals in 11 European countries (Austria, Belgium, Denmark, France, Germany, Greece, Italy, the Netherlands, Spain, Sweden and Switzerland). To reduce the impact of cross-country differences in the fraction of the institutionalized population we confine attention to people between 50 and 80 years of age.

The measures of cognitive ability available in SHARE are the outcomes of simple tests of orientation in time, memory, verbal fluency and numeracy. Here we consider two dimensions of cognitive functioning: verbal fluency and numeracy. The test of verbal fluency consists of counting how many distinct members of the animal kingdom the respondent can name in one minute, and the test outcome is an integer variable ranging from 0 to 90. The test of numeracy consists instead of four possible questions involving simple arithmetical calculations based on

real life situations, and the test outcome is an integer ranging from 1 (no correct answer) to 5 (correct answer to the most difficult question).

Our covariates include self-reported measures of physical health (number of limitations with activities of daily living and number of chronic diseases), an objective measure of physical health (hand grip strength), and a number of socio-economic variables (age, gender, an indicator for educational attainments, per-capita household income and household net worth). To ensure cross-country comparability, the information on educational attainments has been recoded using the 1997 International Standard Classification of Education (ISCED-97), while per-capita household income and household net worth have been adjusted for differences in purchasing power across countries. Summary statistics for the outcomes and the covariates are presented in Table 1, separately for three regions: North (Denmark, the Netherlands, Sweden), Center (Austria, Belgium, France, Germany, Switzerland) and South (Greece, Italy, Spain).

Among the selected covariates, hand grip strength, per-capita household income and household net worth are affected by substantial item nonresponse. The item nonresponse rate for hand grip strength is equal to 6 percent. Missing data occur either because respondents are excluded from the grip strength test in case of swelling, inflammation, severe pain, recent injury or surgery to both hands in the last 6 months, or because the measurements obtained during the test are considered as unreliable. The item nonresponse rates for household income and household net worth are much higher, and are equal to 62 and 64 percent, respectively. The substantial amount of item nonresponse on these two variables reflects three problems. First, they are not directly reported by the respondents but obtained by aggregating a large number of income and wealth components. Second, information about incomes, assets, mortgages and other debts are asked through open-ended and retrospective questions that are sensitive and difficult to answer. Third, according to SHARE fieldwork rules, a household with two spouses is considered as interviewed if at least one of them agrees to participate. If the other does not, then household income and household net worth must be imputed because the individual components are missing for the nonresponding spouse. In total, complete-case analysis would drop 83 percent of the sample.

To deal with the potential selectivity effects generated by missing data the public-use SHARE data include imputations of key variables. As described in Christelis (2011), these imputations are constructed using the multivariate iterative procedure of van Buuren et al. (2006), which attempts to preserve the correlation structure of the imputed data. In our analysis, validity of

the SHARE imputations for income and net worth may be questioned because verbal fluency, number of chronic diseases and hand grip strength are not among the explanatory variables used by the SHARE imputation model. Thus, even when correctly specified, the imputation model is likely to be uncongenial with the models of interest as they are based on different sets of explanatory variables. Finally, we produce our own imputations for the missing values on hand grip strength using a simple hot-deck procedure.

Given the high level of cross-country comparability of SHARE, we pool data within each region and estimate a Poisson regression model for verbal fluency and an ordered probit model for numeracy, separately by region. The Poisson model for verbal fluency is an example of univariate model with canonical inverse link functions, while the ordered probit model for numeracy is an example of multivariate model with non-canonical inverse link function. The number of missing-data patterns is $J = 7$ in both models. Thus, our block-BMA procedure requires to consider $R = 2^7 = 128$ models for each outcome and region.

The estimates for the two models, obtained from STATA routines developed on purpose, are presented in Tables 2 and 3. For each outcome and region we compare estimated coefficients and standard errors for the complete-case ML estimator, the fill-in ML estimator, and the block-BMA estimators based on AIC, BIC and RIC priors, respectively. Notice that interpretation of the standard errors differs depending on the estimation strategy. For the complete-case and fill-in approaches, they can be interpreted as classical standard errors that ignore the additional sampling variability induced by the model selection step. For the generalized missing-data approach, the standard errors have the usual Bayesian interpretation of measuring the spread of the posterior distribution of the parameters of interest given the data. As discussed in Section 5, these standard errors take model uncertainty explicitly into account by construction.

Our results show little differences in the sign of the estimated associations across cognitive dimensions, regions and estimation methods. In particular, we find that verbal fluency and numeracy are negatively related to age, and positively related to self-reported and objective physical health measures and to variables typically associated with higher socio-economic status. On average, verbal fluency is higher for women than for men, while the opposite is true for numeracy. The size of the coefficients and the standard errors are instead subject to non-negligible differences across estimation method. Complete-case and fill-in ML estimates tend

to be different, and one can notice the substantial loss of precision resulting from complete-case analysis. As expected, block-BMA estimates based on AIC priors are closer to the less parsimonious complete-case model, while those based on BIC and RIC priors are closer to the more parsimonious fill-in model. The differences are particularly striking in the Poisson models for fluency, especially for the Central and the Southern regions, possibly because this outcome is not included in the set of explanatory variables used by the SHARE imputation model. Standard errors of block-BMA estimators are often greater than those obtained with the restricted fill-in ML estimator which ignores uncertainty due to the model selection step.

Overall, these results cast some doubt on the validity of the SHARE imputations when studying cognitive functioning. The issue appears to be particularly important for verbal fluency and for countries belonging to the Central and the Southern regions, where discrepancies between standard approaches to the problem of missing covariate values and our generalized missing-data approach are substantial independently of the chosen prior distribution.

8. Conclusions

This paper considers the problem of estimating GLMs in the empirically relevant case when the values of some covariates are missing for some observations but imputations are available to fill-in the missing values. Although using imputed covariates is quite common, researchers should not take the validity of imputations for granted and should explicitly consider the trade-off between bias and precision involved in their use. We address this problem by extending to the class of GLMs the generalized missing-indicator approach, originally proposed for linear regression by Dardanoni et al. (2011). Our approach reformulates the trade-off between bias and precision as a problem of model uncertainty, which can be handled very naturally through Bayesian model averaging. The particular structure of this problem allows us to adopt a block-BMA strategy that is straightforward and makes it possible to explore all the relevant submodels.

An empirical application using the first wave of the Survey on Health, Aging and Retirement in Europe (SHARE) illustrates the practical use of our approach for GLMs with both univariate and multivariate observed outcomes. Our results show that inference based on standard approaches to missing covariates and on our generalized missing-data approach may be substantially different.

REFERENCES

Akaike, H. (1978), 'A new look at the Bayes procedure', *Biometrika* **65**, 53–59.

Börsch-Supan, A., A. Brugiavini, H. Jürges, J. Mackenbach, J. Siegrist and G. Weber, eds (2005), *Health, Ageing and Retirement in Europe: First Results from the Survey of Health, Ageing and Retirement in Europe*, MEA, Mannheim, DE.

Burnham, K.P. and D.R. Anderson, eds (2002), *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*, 2nd edn, Springer-Verlag, New York.

Christelis, D. (2011), 'Imputation of missing data in waves 1 and 2 of SHARE', *SHARE Working Paper Series* pp. 1–50.

Clyde, M.A. (2000), 'Model uncertainty and health effect studies for particular matter', *Environmetrics* **11**, 745–763.

Clyde, M.A. and E.I. George (2004), 'Model uncertainty', *Statistical Science* **19**, 81–94.

Danilov, D. and J.R. Magnus (2004), 'On the harm that ignoring pretesting can cause', *Journal of Econometrics* **122**, 27–46.

Dardanoni, V., G. De Luca, S. Modica and F. Peracchi (2012), 'A generalized missing-indicator approach to regression with imputed covariates', *Stata Journal* **12**, 575–604.

Dardanoni, V., S. Modica and F. Peracchi (2011), 'Regression with imputed covariates: A generalized missing-indicator approach', *Journal of Econometrics* **162**, 362–368.

Draper, D. (1995), 'Assessment and propagation of model uncertainty', *Journal of the Royal Statistical Society (Series B)* **57**, 45–97.

Fahrmeir, L. and H. Kaufmann (1985), 'Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models', *Annals of Statistics* **13**, 342–368.

Foster, D.P. and E.I. George (1994), 'The risk information criterion for multiple regression', *Annals of Statistics* **22**, 1947–1975.

Furnival, G.M. and R.W. Wilson (1974), 'Regression by leaps and bounds', *Technometrics* **16**, 499–511.

Gourieroux, C., A Monfort, E. Renault and A. Trognon (1987), 'Generalized residuals', *Journal of Econometrics* **34**, 5–32.

Han, C. and B.P. Carlin (2001), 'Markov chain Monte Carlo methods for computing Bayes factors', *Journal of the American Statistical Association* **96**, 1122–1132.

Hoeting, J.A., D. Madigan, A.E. Raftery and C.T. Volinsky (1999), 'Bayesian model averaging: a tutorial', *Statistical Science* **14**, 382–401.

Jeffreys, H. (1961), *Theory of Probability*, 3rd edn, Oxford Unversity Press, Oxford, United Kingdom.

Kass, R.E. and A.E. Raftery (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**, 773–795.

Madigan, D. and A.E. Raftery (1994), 'Model selection and accounting for model uncertainty in graphical models using Occam's window', *Journal of the American Statistical Association* **89**, 1535–1546.

Magnus, J. R., O. Powell and P. Prüfer (2010), 'A comparison of two averaging techniques with an application to growth empirics', *Journal of Econometrics* **154**, 139–153.

Magnus, J.R. and J. Durbin (1999), 'Estimation of regression coefficients of interest when other regression coefficients are of no interest', *Econometrica* **67**, 639–643.

Mazzonna, F. and F. Peracchi (2012), 'Aging, cognitive abilities and retirement', *European Economic Review* **56**, 691–710.

McCullagh, P. and J.A. Nelder (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

Meng, X.L. (1994), 'Multiple-imputation inferences with uncongenial sources of input', *Statistical Science* **9**, 538–558.

Nelder, J.A. and R.W.M. Wedderburn (1972), 'Generalized linear models', *Journal of Royal Statistical Society (Series A)* **135**, 370–384.

Nicoletti, C. and F. Peracchi (2006), 'The effects of income imputation on microanalyses: evidence from the Europen Community Household Panel', *Journal of Royal Statistical Society (Series A)* **169**, 625–646.

Raftery, A.E. (1993), Bayesian model selection in structural equation models, *in* K.Bollen and J.Long, eds, 'Testing Structural Equation Models', Sage, Newbury Park, CA, pp. 163–180.

Raftery, A.E. (1995), 'Bayesian model selection in social research', *Sociological Methodology* **25**, 111–163.

Raftery, A.E. (1996), 'Approximate Bayes factors and accounting for model uncertainty in generalized linear models', *Biometrika* **83**, 251–266.

Rubin, D.B. (1976), 'Inference and missing data', *Biometrika* **63**, 581–592.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Schwarz, G. (1996), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464.

Tierney, L. and J.B. Kadane (1986), 'Accurate approximations for posterior moments and marginal densities', *Journal of the American Statistical Association* **81**, 82–86.

van Buuren, S., J. P. L. Brand, C. G. M. Groothuis-Oudshoorn and D. B. Rubin (2006), 'Fully conditional specification in multivariate imputation', *Journal of Statistical Computation and Simulation* **76**, 1049–1064.

Volinsky, C.T. and A.E. Raftery (2000), 'Bayesian information criterion for censored survival models', *Biometrics* **56**, 256–262.

Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edn, MIT Press, Cambridge, Massachusetts.

TABLE 1. Descriptive statistics for the outcomes and the covariates by region (PPP-adjusted per-capita household income is in 10,000 Euro and household net worth is in 100,000 Euro).

Region	Variable	Median	Mean	SD	Min	Max
North	fluency	23.0	22.9	7.0	.0	53.0
	numeracy	4.0	3.6	1.0	1.0	5.0
	adl	.0	.1	.5	.0	6.0
	chronic	1.0	1.5	1.4	.0	9.0
	grip strength	35.0	36.6	12.5	4.0	75.0
	age	62.0	62.8	8.4	50.0	80.0
	male	.0	.5	.5	.0	1.0
	education	1.0	.6	.5	.0	1.0
	income	2.1	2.5	1.8	.0	47.5
	net worth	1.2	2.3	4.9	-20.3	83.6
	Complete obs.					813
	Imputed obs.					4,088
Center	fluency	20.0	20.3	7.1	.0	90.0
	numeracy	4.0	3.5	1.1	1.0	5.0
	adl	.0	.1	.6	.0	6.0
	chronic	1.0	1.4	1.3	.0	12.0
	grip strength	35.0	36.1	12.2	1.0	92.0
	age	62.0	63.1	8.4	50.0	81.0
	male	.0	.5	.5	.0	1.0
	education	1.0	.6	.5	.0	1.0
	income	1.7	2.4	2.6	.0	44.7
	net worth	1.9	3.5	8.9	-20.7	291.1
	Complete obs.					2,575
	Imputed obs.					13,635
South	fluency	14.0	14.6	5.5	.0	88.0
	numeracy	3.0	3.0	1.1	1.0	5.0
	adl	.0	.2	.6	.0	6.0
	chronic	1.0	1.6	1.5	.0	10.0
	grip strength	30.0	31.7	11.7	3.0	80.0
	age	63.0	63.3	8.4	50.0	81.0
	male	.0	.5	.5	.0	1.0
	education	.0	.3	.4	.0	1.0
	income	.9	1.3	1.4	.0	27.5
	net worth	1.5	3.1	9.5	-22.5	260.9
	Complete obs.					1,754
	Imputed obs.					6,745

TABLE 2. Estimated coefficients and standard errors (in parentheses) of Poisson regression models for fluency by region. Results for the constant term and the auxiliary regressors are omitted to save space.

Region	Variable	CC	FI	Block BMA		
				AIC	BIC	RIC
North	adl	-.0625	-.0525	-.0564	-.0525	-.0499
		(.0147)	(.0072)	(.0134)	(.0072)	(.0073)
	chronic	-.0117	-.0162	-.0115	-.0162	-.0160
		(.0058)	(.0026)	(.0058)	(.0026)	(.0026)
	grip strength	.0051	.0039	.0045	.0039	.0040
		(.0010)	(.0004)	(.0011)	(.0004)	(.0005)
	age	-.0055	-.0049	-.0056	-.0049	-.0049
		(.0011)	(.0005)	(.0011)	(.0005)	(.0005)
	male	-.0873	-.0817	-.0750	-.0817	-.0829
		(.0238)	(.0108)	(.0235)	(.0108)	(.0108)
	education	.0406	.0794	.0389	.0794	.0780
		(.0159)	(.0070)	(.0160)	(.0070)	(.0071)
	income	.0287	.0124	.0286	.0124	.0123
		(.0054)	(.0017)	(.0062)	(.0017)	(.0017)
	net worth	.0061	.0032	.0060	.0032	.0032
		(.0023)	(.0006)	(.0024)	(.0006)	(.0006)
Center	adl	-.0325	-.0561	-.0400	-.0611	-.0591
		(.0097)	(.0038)	(.0094)	(.0056)	(.0071)
	chronic	-.0029	-.0031	-.0024	-.0042	-.0037
		(.0036)	(.0016)	(.0030)	(.0023)	(.0025)
	grip strength	.0043	.0043	.0044	.0046	.0046
		(.0006)	(.0003)	(.0005)	(.0004)	(.0006)
	age	-.0054	-.0058	-.0056	-.0055	-.0054
		(.0006)	(.0003)	(.0005)	(.0004)	(.0006)
	male	-.0718	-.0770	-.0770	-.0846	-.0848
		(.0134)	(.0058)	(.0116)	(.0092)	(.0115)
	education	.1388	.1574	.1438	.1540	.1527
		(.0095)	(.0041)	(.0086)	(.0060)	(.0064)
	income	.0092	.0073	.0101	.0114	.0118
		(.0021)	(.0007)	(.0018)	(.0022)	(.0022)
	net worth	-.0006	.0005	.0005	.0007	.0011
		(.0009)	(.0002)	(.0010)	(.0004)	(.0009)
South	adl	-.0567	-.0567	-.0604	-.0526	-.0526
		(.0131)	(.0060)	(.0122)	(.0073)	(.0073)
	chronic	.0073	.0128	.0068	.0011	.0011
		(.0049)	(.0024)	(.0048)	(.0036)	(.0036)
	grip strength	.0063	.0067	.0060	.0059	.0059
		(.0008)	(.0004)	(.0008)	(.0006)	(.0006)
	age	-.0069	-.0071	-.0072	-.0071	-.0071
		(.0009)	(.0005)	(.0009)	(.0007)	(.0007)
	male	-.0488	-.0559	-.0441	-.0423	-.0423
		(.0185)	(.0091)	(.0183)	(.0135)	(.0135)
	education	.1651	.1574	.1678	.1869	.1869
		(.0158)	(.0075)	(.0157)	(.0118)	(.0118)
	income	.0105	.0059	.0095	.0005	.0005
		(.0058)	(.0021)	(.0058)	(.0041)	(.0041)
	net worth	.0048	.0008	.0049	.0008	.0008
		(.0017)	(.0003)	(.0017)	(.0006)	(.0006)

TABLE 3. Estimated coefficients and standard errors (in parentheses) of ordered probit models for numeracy by region. Results for the thresholds of the outcome and the auxiliary regressors are omitted to save space.

Region	Variable	CC	FI	Block BMA		
				AIC	BIC	RIC
North	adl	-.1373	-.1138	-.1099	-.1138	-.1156
		(.0631)	(.0317)	(.0512)	(.0317)	(.0373)
	chronic	-.0472	-.0327	-.0351	-.0327	-.0277
		(.0282)	(.0130)	(.0204)	(.0130)	(.0160)
	grip strength	.0143	.0107	.0108	.0107	.0109
		(.0051)	(.0023)	(.0037)	(.0023)	(.0029)
	age	-.0068	-.0061	-.0097	-.0061	-.0107
		(.0053)	(.0023)	(.0038)	(.0023)	(.0029)
	male	.1458	.2071	.1907	.2071	.1903
		(.1196)	(.0544)	(.0841)	(.0544)	(.0682)
	education	.1782	.4042	.2985	.4042	.3438
		(.0801)	(.0357)	(.0662)	(.0357)	(.0445)
	income	.1055	.0472	.0688	.0472	.0447
		(.0322)	(.0096)	(.0318)	(.0096)	(.0121)
	net worth	.0518	.0161	.0414	.0161	.0403
		(.0150)	(.0036)	(.0122)	(.0037)	(.0065)
Center	adl	-.0386	-.1113	-.0411	-.1113	-.0959
		(.0414)	(.0160)	(.0345)	(.0160)	(.0169)
	chronic	-.0281	-.0297	-.0406	-.0297	-.0312
		(.0169)	(.0073)	(.0164)	(.0073)	(.0074)
	grip strength	.0138	.0137	.0131	.0137	.0141
		(.0028)	(.0012)	(.0024)	(.0012)	(.0012)
	age	-.0084	-.0078	-.0061	-.0078	-.0076
		(.0029)	(.0012)	(.0028)	(.0012)	(.0012)
	male	.1293	.1257	.1298	.1257	.1235
		(.0637)	(.0273)	(.0526)	(.0273)	(.0278)
	education	.5447	.6315	.5549	.6315	.6316
		(.0446)	(.0196)	(.0406)	(.0196)	(.0198)
	income	.0446	.0390	.0491	.0390	.0399
		(.0107)	(.0037)	(.0095)	(.0037)	(.0038)
	net worth	.0072	.0029	.0073	.0029	.0035
		(.0044)	(.0010)	(.0035)	(.0010)	(.0011)
South	adl	-.0179	-.0972	-.0522	-.0972	-.0972
		(.0453)	(.0213)	(.0460)	(.0213)	(.0213)
	chronic	-.0358	-.0562	-.0426	-.0562	-.0562
		(.0188)	(.0097)	(.0163)	(.0097)	(.0097)
	grip strength	.0177	.0185	.0171	.0185	.0185
		(.0033)	(.0017)	(.0029)	(.0017)	(.0017)
	age	-.0152	-.0180	-.0172	-.0180	-.0180
		(.0036)	(.0018)	(.0031)	(.0018)	(.0018)
	male	.1730	.2063	.2037	.2063	.2063
		(.0720)	(.0365)	(.0599)	(.0365)	(.0365)
	education	.9448	.9430	.9334	.9430	.9430
		(.0678)	(.0328)	(.0574)	(.0328)	(.0328)
	income	.0248	.0250	.0306	.0250	.0250
		(.0241)	(.0093)	(.0210)	(.0093)	(.0093)
	net worth	.0007	.0019	.0022	.0019	.0019
		(.0068)	(.0014)	(.0042)	(.0014)	(.0014)

APPENDIX A. ASYMPTOTIC PROPERTIES OF COMPLETE-CASE AND FILL-IN ML ESTIMATORS

From Section 4.1, the complete-case ML estimator coincides with the ML estimator of β in the grand model with linear predictor $\boldsymbol{\eta} = \boldsymbol{\eta} = \boldsymbol{W}\beta + \boldsymbol{Z}\delta$. The ML estimator $\widehat{\theta} = (\widehat{\beta}, \widehat{\delta})$ in this model converges in probability to the true population value $\theta^0 = (\beta^0, \delta^0)$, which solves the equation system

$$\mathrm{E}\, s_\beta(\beta, \delta; W_n, Z_n) = 0,$$

$$\mathrm{E}\, s_\delta(\beta, \delta; W_n, Z_n) = 0,$$

where s_β and s_δ denote the elements of the score vector corresponding to β and δ respectively. Further, $\sqrt{N}(\widehat{\theta} - \theta^0) \Rightarrow \mathcal{N}(0, \mathcal{I}^{0-1})$, where

$$\mathcal{I}^0 = \begin{bmatrix} \mathcal{I}^0_{\beta\beta} & \mathcal{I}^0_{\beta\delta} \\ \mathcal{I}^0_{\delta\beta} & \mathcal{I}^0_{\delta\delta} \end{bmatrix}$$

is the Fisher information matrix evaluated at θ^0. Because the asymptotic variance of $\widehat{\beta}$ is the top-left block of the inverse of $\mathcal{I}^0$, it follows that

$$\sqrt{N}(\widehat{\beta} - \beta^0) \Rightarrow \mathcal{N}(0, (\mathcal{I}^0_{\beta\beta} - \mathcal{I}^0_{\beta\delta}\mathcal{I}^{0-1}_{\delta\delta}\mathcal{I}^0_{\delta\beta})^{-1}).$$

The fill-in ML estimator $\widetilde{\beta}$ solves the equation system

$$\frac{1}{N}\sum\nolimits_{n=1}^{N} s_\beta(\beta, 0; W_n, Z_n) = 0. \tag{8}$$

Because the restriction that $\delta = 0$ may be invalid, $\widetilde{\beta}$ converges in probability to the pseudo-true value β^*, defined as the root of the equation system

$$\mathrm{E}\, s_\beta(\beta, 0; W_n, Z_n) = 0,$$

which does not generally coincide with the true population value β^0. A first-order Taylor expansion of (8) around the pseudo-true value β^* gives

$$\sqrt{N}(\widetilde{\beta} - \beta^*) = \left[-\frac{1}{N}\sum\nolimits_{n=1}^{N} S_{\beta\beta}(\beta^*, 0; W_n, Z_n)\right]^{-1} \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N} s_\beta(\beta^*, 0; W_n, Z_n) + \mathrm{o}_p(1),$$

where $S_{\beta\beta}$ denotes the Hessian of the log-likelihood with respect to β. Under the regularity conditions in Fahrmeir and Kaufmann (1985), as $N \to \infty$, the Central Limit Theorem implies that

$$\frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N} s_\beta(\beta^*, 0; W_n, Z_n) \Rightarrow \mathcal{N}(0, V^*_{\beta\beta}),$$

where $V^*_{\beta\beta} = \mathrm{Var}\, s_\beta(\beta^*, 0; W_n, Z_n)$, and the Law of Large Numbers implies that

$$\mathrm{plim}\, \frac{1}{N}\sum\nolimits_{n=1}^{N} S_{\beta\beta}(\beta^*, 0; W_n, Z_n) = H^*_{\beta\beta},$$

a positive definite matrix. Therefore,

$$\sqrt{N}(\widetilde{\beta} - \beta^0) \Rightarrow \mathcal{N}(\beta^* - \beta^0, (H^*_{\beta\beta})^{-1} V^*_{\beta\beta} (H^*_{\beta\beta})^{-1}).$$

A characterization of the asymptotic bias $\beta^* - \beta^0$ is given in Appendix B.

If the imputations are valid, then the restriction that $\delta = 0$ is valid, so the fill-in ML estimator $\widetilde{\beta}$ is consistent and asymptotically more precise than the complete-case ML estimator $\widehat{\beta}$, that is

$$AV(\widetilde{\beta})^{-1} - AV(\widehat{\beta})^{-1} \geq 0.$$

In this case, the asymptotic variance of the fill-in ML estimator is equal to the inverse of the Fisher information. Thus,

$$AV(\widetilde{\beta})^{-1} - AV(\widehat{\beta})^{-1} = \mathcal{I}^0_{\beta\delta} \mathcal{I}^{0-1}_{\delta\delta} \mathcal{I}^0_{\delta\beta},$$

which is a nonnegative definite matrix.

APPENDIX B. ASYMPTOTIC BIAS OF THE FILL-IN ML ESTIMATOR

In this appendix, we focus on a GLM with canonical inverse link function in order to characterize the asymptotic bias of the fill-in estimator of β in terms of the imputations. To keep the notation simple, we drop the subscript n and only consider the case of a constant term and two covariates X_1 and X_2. Without loss of generality, we focus on the components of the asymptotic bias arising from one of the possible missing-data pattern, namely the case when X_1 is missing, X_2 is fully observed, and the missing values of X_1 are replaced by the imputations L_1.

In this case, asymptotically, the fill-in ML estimator of β solves the following system of first-order conditions

$$\mathrm{E}\left[Y - h(\beta_0^* + \beta_1^* L_1 + \beta_2^* X_2)\right] = 0,$$

$$\mathrm{E}\left[L_1(Y - h(\beta_0^* + \beta_1^* L_1 + \beta_2^* X_2))\right] = 0,$$

$$\mathrm{E}\left[X_2(Y - h(\beta_0^* + \beta_1^* L_1 + \beta_2^* X_2))\right] = 0,$$

where the $\beta_k^* = \beta_k + \delta_k$ are the pseudo-true parameter values, the β_k are the true parameter values, and the δ_k are the asymptotic biases, $k = 0, 1, 2$. Approximating $h(\beta_0^* + \beta_1^* L_1 + \beta_2^* X_2)$ by a first-order Taylor expansion around the true linear index $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ and noticing that $\mathrm{E}[Y - h(\eta)] = \mathrm{E}[X_2(Y - h(\eta))] = 0$, we can rewrite the system of asymptotic first-order conditions as

$$\delta_0 m_0 + \beta_1(\widetilde{m}_1 - m_1) + \delta_1 \widetilde{m}_1 + \delta_2 m_2 \simeq 0,$$

$$\widetilde{m}_{1u} - \delta_0 \widetilde{m}_1 - \beta_1(\widetilde{m}_{11} - \widetilde{m}_{11}^*) - \delta_1 \widetilde{m}_{11} - \delta_2 \widetilde{m}_{12} \simeq 0,$$

$$\delta_0 m_2 + \beta_1(\widetilde{m}_{12} - m_{12}) + \delta_1 \widetilde{m}_{12} + \delta_2 m_{22} \simeq 0,$$

where $m_0 = \mathrm{E}[h'(\eta)]$, $m_k = \mathrm{E}[h'(\eta)X_k]$ $(k = 1, 2)$, and $\widetilde{m}_1 = \mathrm{E}[h'(\eta)L_1]$ are weighted first-order moments, and $\widetilde{m}_{1u} = \mathrm{E}[L_1(Y - h(\eta))]$, $\widetilde{m}_{11} = \mathrm{E}[h'(\eta)L_1^2]$, $m_{12} = \mathrm{E}[h'(\eta)X_1X_2]$, $\widetilde{m}_{12} = \mathrm{E}[h'(\eta)L_1X_2]$, $m_{22} = \mathrm{E}[h'(\eta)X_2^2]$, and $\widetilde{m}_{11}^* = \mathrm{E}[h'(\eta)X_1L_1]$ are weighted uncentered second-order moments. Solving this system with respect to the components of the asymptotic bias gives

$$\delta_0 \simeq \left\{(\widetilde{c}_1 c_{22} - c_2\widetilde{c}_{12})\widetilde{c}_{1u} - \beta_1\left[(\widetilde{c}_1 c_{22} - c_2\widetilde{c}_{12})\Delta_{11} + (\widetilde{c}_1\widetilde{c}_{12} - c_2\widetilde{c}_{11})\Delta_{12} - \widetilde{\rho}_{12}^2\Delta_1\right]\right\}/\widetilde{\rho}_{12}^2,$$

$$\delta_1 \simeq -\left[c_{22}\widetilde{c}_{1u} - \beta_1(c_{22}\Delta_{11} - \widetilde{c}_{12}\Delta_{12})\right]/\widetilde{\rho}_{12}^2, \qquad (9)$$

$$\delta_2 \simeq \left[\widetilde{c}_{12}\widetilde{c}_{1u} - \beta_1(\widetilde{c}_{12}\Delta_{11} - \widetilde{c}_{11}\Delta_{12})\right]/\widetilde{\rho}_{12}^2,$$

where $c_k = m_k/m_0$ and $\widetilde{c}_1 = \widetilde{m}_1/m_0$ are rescaled first-order moments, $\widetilde{c}_{11} = \widetilde{m}_{11}/m_0 - (\widetilde{m}_1/m_0)^2$, $c_{22} = m_{22}/m_0 - (m_2/m_0)^2$, $\widetilde{c}_{12} = \widetilde{m}_{12}/m_0 - \widetilde{m}_1 m_2/(m_0)^2$, and $\widetilde{c}_{11}^* = \widetilde{m}_{11}^*/m_0 - m_1\widetilde{m}_1/(m_0)^2$ are weighted variances and covariances, $\widetilde{c}_{1u} = \widetilde{m}_{1u}/m_0$ is the weighted covariance between L_1 and the generalized residual $U = Y - h(\eta)$, $\widetilde{\rho}_{12}^2 = \widetilde{c}_{12}^2 - \widetilde{c}_{11}c_{22}$ is the coefficient of correlation between L_1 and X_2, and $\Delta_1 = \widetilde{c}_1 - c_1$, $\Delta_{11} = \widetilde{c}_{11} - \widetilde{c}_{11}^*$ and $\Delta_{12} = \widetilde{c}_{12} - c_{12}$ are the differences between the moments of the true and the imputed covariates.

The asymptotic bias of the fill-in ML estimator of β then depends on the asymptotic correlation between the imputations and the generalized residual, and on the differences between the weighted first- and second-order moments of the true and the imputed covariates. A sufficient condition for the asymptotic bias to vanish is that the moments of the joint distribution of (Y, L_1, X_2) converge to the moments of the joint distribution of (Y, X_1, X_2). If this condition holds, then $\widetilde{m}_{1u}$, Δ_{11} and Δ_{12} converge to zero and $\widetilde{\beta}$ is a consistent estimator of β. In general, the fill-in ML estimator of β is consistent if the moments of the joint distribution of $(\boldsymbol{Y}_j, \boldsymbol{W}_j)$ converge to the moments of the joint distribution of $(\boldsymbol{Y}_j, \boldsymbol{X}_j)$, $j = 1, \ldots, J$.

For a GLM with non-canonical inverse link function, the approximation to the asymptotic bias is more complex because it involves the additional terms resulting from the first-order Taylor expansion of $v(\eta)$. The presence of these additional terms, however, does not change the sufficient condition for consistency of the fill-in ML estimator.

In the linear regression case, our approximation to the asymptotic bias of the fill-in ML estimator coincides with expressions for the asymptotic bias of the OLS estimator given in (Dardanoni et al., 2011, p. 364). Assume for simplicity that all variables have been standardized to have zero means and unit variances. Also assume, as in Dardanoni et al. (2011), that the imputations are asymptotically uncorrelated with the regression error, so $\widetilde{c}_{1u} = 0$. Then the

asymptotic biases of the OLS estimators of β_1 and β_2 are given by

$$\delta_1 = (\Gamma^1 - 1)\beta_1, \qquad \delta_2 = \Delta^1\beta_1,$$

where

$$\Gamma^1 = \frac{\widetilde{c}_{11}^* - c_{12}\widetilde{c}_{12}}{1 - \widetilde{c}_{12}^2}, \qquad \Delta^1 = \frac{c_{12} - \widetilde{c}_{11}^*\widetilde{c}_{12}}{1 - \widetilde{c}_{12}^2},$$

are the coefficients in the best linear predictor of X_1 given L_1 and X_2. In the linear regression case, where $h(\cdot)$ is the identity function, $m_0 = 1$ and the approximations in (9) are exact because no Taylor expansion is needed. In this case, after imposing the restrictions $\widetilde{c}_{1u} = 0$, $\widetilde{c}_{11} = 1$ and $c_{22} = 1$, the asymptotic bias of the ML estimators of β_1 and β_2 reduces to

$$\delta_1 = \left(\frac{\widetilde{c}_{11}^* - c_{12}\widetilde{c}_{12}}{1 - \widetilde{c}_{12}^2} - 1 \right) \beta_1, \qquad \delta_2 = \frac{c_{12} - \widetilde{c}_{11}^*\widetilde{c}_{12}}{1 - \widetilde{c}_{12}^2} \beta_1.$$