Journal of Econometrics 184 (2015) 452–463

Contents lists available at ScienceDirect

Journal of Econometrics

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

Model averaging estimation of generalized linear models with imputed covariates

Valentino Dardanoni[a], Giuseppe De Luca[a], Salvatore Modica[a], Franco Peracchi[b,*]

[a] University of Palermo, Italy

[b] University of Rome Tor Vergata and Einaudi Institute for Economics and Finance (EIEF), Italy

ARTICLE INFO

Article history:
Received 3 July 2013
Received in revised form
19 March 2014
Accepted 15 June 2014
Available online 24 June 2014

JEL classification:
C11
C25
C35
C81

Keywords:
Model averaging
Bayesian averaging of maximum likelihood estimators
Generalized linear models
Missing covariates
Generalized missing-indicator method
SHARE

ABSTRACT

We address the problem of estimating generalized linear models when some covariate values are missing but imputations are available to fill-in the missing values. This situation generates a bias-precision trade-off in the estimation of the model parameters. Extending the generalized missing-indicator method proposed by Dardanoni et al. (2011) for linear regression, we handle this trade-off as a problem of model uncertainty using Bayesian averaging of classical maximum likelihood estimators (BAML). We also propose a block model averaging strategy that incorporates information on the missing-data patterns and is computationally simple. An empirical application illustrates our approach.

1. Introduction

In this paper we address the problem of estimating generalized linear models (GLMs) when the outcome of interest is always observed, some covariate values are missing, and 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 item nonresponse. The focus of this paper is on how to make use of the available imputations, not on methods to impute the missing values.

Two standard approaches to the problem of missing covariate values are complete-case analysis and the fill-in approach. The first drops all the observations with missing values ignoring the imputations altogether, while the second fills-in the missing 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. When the complete cases are few the loss of precision may be substantial, 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), that is, the imputation model is more restrictive than the model used to analyze the filled-in data. Validity of the assumptions behind the fill-in approach is often taken for granted, so this bias-precision trade-off is usually ignored. However, when imputations are provided by an external source, the congeniality assumption may fail because the two models are based on different parametric assumptions or they condition on different sets of covariates. The estimates from the fill-in approach may therefore be inconsistent, especially in the case of nonlinear estimators.

Using the generalized missing-indicator approach originally proposed for linear regression by Dardanoni et al. (2011), we transform the bias-precision trade-off between complete-case analysis and the fill-in approach into a problem of model uncertainty regarding which covariates should be dropped from an augmented GLM, or 'grand model', which includes two subsets of regressors:

[*] Corresponding author. Tel.: +39 06 7259 5934; fax: +39 06 2040 219.
E-mail address: franco.peracchi@uniroma2.it (F. Peracchi).

http://dx.doi.org/10.1016/j.jeconom.2014.06.002
0304-4076/© 2014 Elsevier B.V. All rights reserved.

the focus covariates, corresponding to the observed or imputed covariates, and a set of auxiliary regressors consisting of binary indicators for the various missing-data patterns and 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 approach correspond to two extreme specifications of the grand model. Complete-case analysis corresponds to using an unrestricted specification, while the fill-in approach corresponds to using a restricted specification that includes only the focus regressors. Instead of focusing on these extreme specifications of the grand model, we consider Bayesian averaging of classical maximum likelihood estimators (BAML) that takes into account all the intermediate specifications obtained by dropping from the grand model alternative subsets of auxiliary regressors associated with the various missing-data patterns. 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 choices of using either the complete-case or the fill-in approach are still available, but neither is likely 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 extending the generalized missing-indicator method to the wide class of GLMs, we depart from Dardanoni et al. (2011) in three important respects. First, we propose a new block model averaging 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 models and ordered, multinomial or conditional logit and probit models. Third, we investigate the robustness of our block-BAML procedure to the choice of priors by considering two families of prior distributions: the calibrated information criteria priors introduced by Clyde (2000), which use approximations based on the Laplace method for integrals to calibrate posterior model probabilities to classical model section criteria, and the conjugate priors for GLMs introduced by Chen and Ibrahim (2003), which allow to directly estimate posterior model probabilities using a computationally simple Markov chain Monte Carlo algorithm.

In our empirical illustration we analyze how cognitive functioning varies with physical health and socio-economic status using data from the fourth wave of the Survey on Health, Aging and Retirement in Europe (SHARE). Like for other household surveys, sensitive variables such as household income, household net worth, and other objective health measures are affected by substantial item nonresponse. Using the imputations contained in the public-use SHARE data, we investigate the bias-precision trade-off arising from different approaches for dealing with the problem of imputed covariates in GLMs. Further, we employ multiple imputation methods to account for the additional sampling uncertainty due to the imputation of missing covariate values.

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 our BAML procedure. 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 a realization of a random vector $\boldsymbol{Y} = (Y_1, \dots, Y_N)$, whose components are independently distributed random variables with mean μ_n and finite nonzero variance σ_n^2.[1] We assume that the distribution of any component Y_n of $\boldsymbol{Y}$ belongs to the one-parameter linear exponential family with density function of the form

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

where γ_n is a scalar parameter called the canonical parameter, $b(\cdot)$ is a known, strictly convex and twice differentiable function, and $c(\cdot)$ is a known function.[2] 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). Different choices of the functions $b(\cdot)$ and $c(\cdot)$ result in different distributions within this family. For example, letting $b(\gamma_n) = \gamma_n^2/2$ and $c(y) = -1/2[y^2 + \ln(2\pi)]$ gives the density of a normal distribution with mean γ_n and unit variance, while letting $b(\gamma_n) = \exp(\gamma_n)$ and $c(y) = -\ln(y!)$ gives the density of a Poisson distribution with intensity parameter equal to $\exp(\gamma_n)$.

In a GLM the dependence of Y_n on a vector of covariates X_n (assumed to include a constant term) is modeled by assuming that there exists a continuously differentiable and invertible function $h(\cdot)$, known as 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 $\eta_n = X_n^\top \beta$ 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 with nth row equal to $X_n^\top$.

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

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

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

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

with $\upsilon(X_n^\top \beta) = h'(X_n^\top \beta)/b''(\gamma_n(\beta))$. Provided the assumed model is correctly specified, and the mild regularity conditions in Fahrmeir and Kaufmann (1985) hold, $\widehat{\beta}$ is unique, consistent, and asymptotically normal with asymptotic variance equal to the inverse of the Fisher information matrix. The fact that β enters the likelihood equations only through the linear predictor $\eta_n = X_n^\top \beta$ is the key property of GLMs that drives our main result in Theorem 1. If $b'(\cdot) = h(\cdot)$ (the "canonical link" case), then $\gamma_n(\beta) = X_n^\top \beta$ and the likelihood equations simplify considerably because $\upsilon(X_n^\top \beta) = 1$ for all n. An example is the Gaussian model with identity link $h(X_n^\top \beta) = X_n^\top \beta$, where the likelihood equations reduce to the familiar normal equations for OLS.

In this paper we depart from the standard GLM setup by allowing some covariate values to be missing. We also assume that imputations, as provided by an external source (typically the producers of the dataset), are available to fill-in the missing covariate values. Since the constant term is always observed, the number of possible missing-data patterns is equal to 2^{K-1}. Not all the possible

[1] Vectors are always column vectors, and boldface denotes vector and matrices of sample observations or of functions of sample observations.

[2] In the original formulation of Nelder and Wedderburn (1972), the density in Eq. (1) includes an additional dispersion parameter which, without loss of generality, we set equal to one.

patterns need be present in the data, so we index by $j = 0, \dots, J$ the patterns that are present, with $j = 0$ corresponding to the subsample with complete data and $J \le 2^{K-1} - 1$. We assume that the jth subsample contains N_j observations, K_j observed (non missing) covariates and $K - K_j$ missing covariates. By definition, $\sum_{j=0}^{J} N_j = N, K_0 = K$, and $1 \le K_j \le K$ for $j = 1, \dots, J$. For each missing-data pattern, let $\boldsymbol{Y}_j$ be the $N_j \times 1$ vector of observations on the outcome 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 missing we define the $N \times K$ missing indicator matrix $\boldsymbol{M}$, with (n, k)th element equal to one if the kth covariate is missing for the nth observation, and to zero otherwise. Finally, for each subsample $j = 1, \dots, J$ with missing covariates we denote by $\boldsymbol{W}_j$ the $N_j \times K$ matrix containing the values of the K_j observed covariates and the imputed values of the $K - K_j$ 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 the 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 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 $\widehat{\beta}_0$ 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 identifiability 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 implies that the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is the same in subsamples with and without missing covariates. Given the true value of the covariates, the pattern of missing data can then be ignored when predicting $\boldsymbol{Y}$. Notice that this conditional independence assumption 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. Indeed, MAR requires the missing-data process to be independent of the missing covariates given the observed data (Rubin, 1976, Seaman et al., 2013). For example, suppose that health is the outcome of interest and income is a covariate subject to missing data problems. If missing income depends on true income but not on health, then conditional independence is satisfied but MAR is not, while if missing income depends on health but not on true income then MAR is satisfied but conditional independence is not. Thus Assumption 2 is neither stronger nor weaker than MAR.

However, even when Assumptions 1 and 2 hold, the severe loss of precision that complete-case entails when the fraction of missing data is substantial cannot be ignored.

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 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 $\boldsymbol{W}$.

In addition to the assumption that the population model is correctly specified and identifiable, the validity of this approach requires two conditions. The first is that the model used to create the imputations is correctly specified, including the assumptions on the posited missing-data mechanism. The second is that the imputation model and the GLM for the filled-in data $[\boldsymbol{Y}, \boldsymbol{W}]$ are 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. Uncongeniality may occur, for instance, when the model of interest and the imputation model are based either on different parametric assumptions or on different sets of explanatory variables. When these two conditions hold, the fill-in ML estimator $\widehat{\beta}_F$ is asymptotically equivalent to the missing-free data ML estimator $\widehat{\beta}$ introduced in Section 2. Further, as shown in Appendix, $\widehat{\beta}_F$ is asymptotically more precise than the complete-case ML estimator $\widehat{\beta}_0$ introduced in Section 3.1. 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, $\widehat{\beta}_F$ may be expected to have higher precision than $\widehat{\beta}_0$ 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 $\widehat{\beta}_F$ is likely to be biased and inconsistent because it ignores the fact that the imputations are not the same as the missing covariate values.

An additional issue with the fill-in approach is how to account for the additional variability induced by the imputation process when assessing the precision of $\widehat{\beta}_F$. As illustrated in our empirical application, this problem can be easily handled by applying the combination rules of Rubin (1987) to multiple imputations of the missing covariate values.

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 we define the $N \times JK$ matrix

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

The statistical model for the full sample is an augmented GLM where the conditional density of $\boldsymbol{Y}$ given $\boldsymbol{W}$ and $\boldsymbol{Z}$ is assumed to belong to the exponential family (1) with linear predictor equal to $\boldsymbol{\eta} = \boldsymbol{W}\beta + \boldsymbol{Z}\delta$, where $\delta = (\delta_1^\top, \dots, \delta_J^\top)^\top$ is a JK-dimensional parameter vector. In the terminology of Danilov and Magnus (2004), the columns of $\boldsymbol{W}$ represent our focus regressors, while the columns of $\boldsymbol{Z}$ represent our auxiliary regressors. Similarly, the components of β are called focus parameters, and the components of δ auxiliary parameters. Following Dardanoni et al. (2011), we shall refer to this augmented GLM as the grand model.

4.1. Equivalence theorem

The following theorem extends to GLMs the result obtained by Dardanoni et al. (2011) for linear regression models.

Theorem 1. *Let $\widetilde{\beta}$ and $\widetilde{\delta}_j$, $j = 1, \dots, J$, denote the ML estimators of β and δ_j in the grand model with linear predictor equal to $\boldsymbol{\eta} = \boldsymbol{W}\beta + \boldsymbol{Z}\delta$. Then, for any set of imputations, $\widetilde{\beta} = \widehat{\beta}_0$.*

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 estimate $\widehat{\beta}_0$ 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 of the form $v(X_n^\top \beta)\,[Y_n - h(X_n^\top \beta)]$. Similarly, the ML estimates $(\widetilde{\beta}, \widetilde{\delta})$ in the grand model with linear predictor equal to $\boldsymbol{W}\beta + \boldsymbol{Z}\delta$ solve the system of $(1+J)K$ likelihood equations

$$\begin{aligned} &\boldsymbol{X}_0^\top \boldsymbol{U}_0(\beta) + \boldsymbol{W}_*^\top \boldsymbol{U}_*(\beta, \delta) = 0, \\ &\boldsymbol{Z}_*^\top \boldsymbol{U}_*(\beta, \delta) = 0, \end{aligned} \tag{2}$$

where $\boldsymbol{U}_*(\beta, \delta)$ is the $(N - N_0) \times 1$ vector with generic element of the form $v(W_n^\top \beta + Z_n^\top \delta)\,[Y_n - h(W_n^\top \beta + Z_n^\top \delta)]$. Since $\boldsymbol{Z}_*$ is a block-diagonal matrix, the last JK equations in (2) imply that

$$\boldsymbol{W}_j^\top \boldsymbol{U}_j(\beta, \delta_j) = 0, \quad j = 1, \dots, J,$$

where $\boldsymbol{U}_j(\beta, \delta_j)$ is the $N_j \times 1$ vector obtained by selecting from $\boldsymbol{U}_*(\beta, \delta)$ the observations in the jth subsample, and $\boldsymbol{W}_*^\top \boldsymbol{U}_*(\beta, \delta) = \sum_{j=1}^J \boldsymbol{W}_j^\top \boldsymbol{U}_j(\beta, \delta_j) = 0$. The ML estimate $\widetilde{\beta}$ then solves $\boldsymbol{X}_0^\top \boldsymbol{U}_0(\beta) = 0$, so $\widetilde{\beta} = \widehat{\beta}_0$. □

Theorem 1 shows that the unrestricted ML estimate of β in the grand model is numerically the same as the complete-case ML estimate. To interpret the unrestricted ML estimate $\widetilde{\delta}_j$, let $\boldsymbol{\eta}_j = \boldsymbol{W}\beta_j$ denote the linear predictor in the GLM for the jth sample with missing covariates and let $\widehat{\beta}_j$ denote the ML estimate of β_j in the jth subsample. Under Assumptions 1 and 2, $\beta_0 = \beta$ but in general $\beta_j \neq \beta$ for $j = 1, \dots, J$. The next corollary shows that $\widetilde{\delta}_j$ coincides numerically with the difference between $\widehat{\beta}_j$ and the complete-case ML estimate.

Corollary. *For any set of imputations, $\widetilde{\delta}_j = \widehat{\beta}_j - \widehat{\beta}_0$, $j = 1, \dots, J$, where $\widehat{\beta}_j$ is the ML estimator of β_j in the jth missing-data pattern.*

Proof. The ML estimate $\widehat{\beta}_j$ in the jth subsample with missing covariates satisfies $\boldsymbol{W}_j^\top \boldsymbol{V}_j(\widehat{\beta}_j) = 0$, where $\boldsymbol{V}_j(\beta)$ is the $N_j \times 1$ vector with generic element of the form $v(W_n^\top \beta)\,[Y_n - h(W_n^\top \beta)]$. Because the linear predictor for the jth subsample in the grand model is of the form $\boldsymbol{\eta}_j = \boldsymbol{W}_j(\beta + \delta_j)$, the ML estimate $\widetilde{\delta}_j$ in the grand model must satisfy

$$0 = \boldsymbol{W}_j^\top \boldsymbol{U}_j(\widetilde{\beta}, \widetilde{\delta}_j) = \boldsymbol{W}_j^\top \boldsymbol{V}_j(\widetilde{\beta} + \widetilde{\delta}_j), \quad j = 1, \dots, J.$$

Thus $\widetilde{\beta} + \widetilde{\delta}_j = \widehat{\beta}_j$. Since $\widetilde{\beta} = \widehat{\beta}_0$, it follows that $\widetilde{\delta}_j = \widehat{\beta}_j - \widehat{\beta}_0$. □

Notice that the fill-in estimator of β coincides with the restricted ML estimator when all elements of δ in the grand model are set to zero. If the population model is correctly specified and the imputations are valid, then this estimator is asymptotically more precise than the complete-case estimator (Appendix). However, if the population model is misspecified or the imputations are not valid, then the fill-in estimator is inconsistent. The generalized missing-indicator approach handles this trade-off between bias and precision 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 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 an improved estimator of β.

4.2. A dual result

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. Our 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. This result, which may be regarded as the dual of Theorem 1, provides a justification for the block model averaging approach presented in Section 5.4.

Given a collection $\mathcal{J}$ of subsamples with missing covariates, let

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

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

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. The 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 of the form $U_n^+(\beta) = v(W_n^{+\top}\beta)[Y_n^+ - h(W_n^{+\top}\beta)]$. Also let $\boldsymbol{U}^-(\beta, \delta^-)$ be the vector of dimension $N - (N_0 + \sum_{j \in \mathcal{J}} N_j)$ with generic element of the form $v(W_n^\top \beta + Z_n^{-\top}\delta^-)\,[Y_n - h(W_n^\top \beta + Z_n^{-\top}\delta^-)]$, 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. □

Thus, constraining δ_j to zero in 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. So each δ_j controls for a particular subsample with missing covariates, a feature that we exploit in our 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 non negligible effects on the statistical properties of the resulting estimates.

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

Suppose that the model space $\mathcal{M}$ includes R possible GLMs, that is, $\mathcal{M} = \{M_1, \dots, M_R\}$. 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 $0 \le P_r \le JK$ 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. Our model averaging estimates of β and δ are of the form

$$\widehat{\beta} = \sum_{r=1}^{R} \lambda_r \widehat{\beta}_r, \qquad \widehat{\delta} = \sum_{r=1}^{R} \lambda_r S_r \widehat{\delta}_r, \tag{3}$$

where the λ_r are non-negative weights that add up to one, the $\widehat{\beta}_r$ and $\widehat{\delta}_r$ are the ML 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-vectors of conditional estimate $\widehat{\delta}_r$ into JK-vectors by setting to zero the elements of δ which are excluded from the rth model.

5.1. Bayesian averaging of ML estimators

As pointed out by Magnus and De Luca (2014), the parameters of each model can be estimated from either a frequentist or a Bayesian perspective. Also, one can choose the weights from a frequentist or a Bayesian perspective. This gives rise to four types of model averaging. In the spirit of the BACE (Bayesian Averaging of Classical Estimates) approach of Sala-i-Martin et al. (2004) and the WALS (Weighted-Average Least Squares) approach of Magnus et al. (2010), our model averaging approach is based on Bayesian averaging of classical ML estimators (BAML). The parameters of each model are estimated by ML, hence under a classical frequentist perspective, while the weighting scheme is developed under a Bayesian perspective using posterior model probabilities $\pi_r(\boldsymbol{Y})$ that reflect our confidence in the ML estimates based on prior beliefs and the observed data. Thus, the weights used in our model averaging estimates of β and δ are

$$\lambda_r = \pi_r(\boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid M_r)\,\pi_r}{\sum_{r=1}^{R} p(\boldsymbol{Y} \mid M_r)\,\pi_r}, \quad r = 1, \dots, R, \tag{4}$$

where π_r is the prior probability of the rth model,

$$p(\boldsymbol{Y} \mid M_r) = \int p(\boldsymbol{Y} \mid \theta_r, M_r)\,\pi(\theta_r \mid M_r)\,d\theta_r \tag{5}$$

is its marginal likelihood, $\theta_r = (\beta, \delta_r)$ is the vector of its parameters, $p(\boldsymbol{Y} \mid \theta_r, M_r)$ is its sample likelihood, and $\pi(\theta_r \mid M_r)$ is the prior density of θ_r under the rth model. Notice that the conditional ML estimates $\widehat{\beta}_r$ and $\widehat{\delta}_r$ are approximately equal to the posterior means of β and δ under the rth model when the sample likelihood is unimodal, approximately symmetric and dominates the prior, either because the sample size is large or because the prior is uninformative. Under these assumptions, the model averaging estimates in (3) can be interpreted as the posterior means of β and δ given the data and all the models in the model space, and therefore coincide with those obtained under a Bayesian model averaging (BMA) approach.

The posterior variance–covariance matrix of β and δ consists of the following blocks (Raftery, 1993; Draper, 1995)

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

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

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

The posterior variances of β and δ involve 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 our model averaging estimator incorporates the uncertainty due to both parameter estimation and model selection.

The choice between alternative BAML estimates depends on the strategies used to handle a number of methodological and computational problems arising in the development of its Bayesian weighting scheme. The main problems are: (i) how to specify the prior probabilities π_r of the various models, (ii) how to specify the prior distribution $\pi(\theta_r \mid M_r)$ for the parameters of each model, (iii) how to evaluate the integrals in (5), which in the context of GLMs do not usually have closed form solutions, and (iv) how to compute the model averaging estimates in (3) 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 likelihood for the various models, not on the prior weight assigned to each of them.

As for problem (ii), we consider two families of prior distributions over the parameters in the rth model. The first is the family of calibrated information criteria (CIC) prior distributions introduced by Clyde (2000), which are uninformative priors derived from the following modification of Jeffrey's prior (Jeffreys, 1961)

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

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 calibrating the posterior model probabilities to classical model selection criteria like the Akaike Information Criterion (AIC; Akaike, 1978), the Bayesian Information Criterion (BIC; Schwarz, 1978), or the Risk Inflation Criterion (RIC; Foster and George, 1994). The use of model averaging estimators 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.

The second is the family of conjugate priors for GLMs proposed by Chen and Ibrahim (2003). The conjugate prior for the parameters of the rth model is proportional to

$$\mathcal{L}(\theta_r \mid M_r) = \exp\left[\bar{a}(\bar{Y}^\top\gamma(\theta_r) - \iota_N^\top b(\gamma(\theta_r)))\right],$$

where $\bar{Y}$ is an N-dimensional vector of prior parameters which can be viewed as the prior predictions for the marginal means of $\boldsymbol{Y}$ at $\boldsymbol{W}$ and $\boldsymbol{Z}$, $\bar{a} > 0$ is a scalar prior parameter which can be interpreted as a precision parameter that quantifies the strength of our prior belief in $\bar{Y}$, $\gamma(\theta_r)$ is the N-dimensional vector of canonical parameters in model M_r, and ι_N is the N-dimensional vector of ones. This family of priors is attractive because the resulting posterior distribution is proportional to

$$\mathcal{L}(\theta_r \mid \boldsymbol{Y}, M_r) = \exp\left[(\boldsymbol{Y} + \bar{a}\bar{Y})^\top \gamma(\theta_r) - (1+\bar{a})\iota_N^\top b(\gamma(\theta_r))\right].$$

When $\bar{a} \to 0$, this posterior reduces to the sample likelihood for the rth model.

5.3. Marginal likelihood

For CIC priors, we use approximations 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. On the basis of this approximation, Clyde (2000) shows that the posterior probability of model M_r is approximately

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

where D_r is the deviance of model M_r (namely -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. 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 $\log c = 2 \log JK$ 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 are attractive for sensitivity analysis in BAML estimation.

For conjugate priors, the marginal likelihood for model M_r satisfies

$$p(\boldsymbol{Y} \mid M_r) = \int p(\boldsymbol{Y} \mid \theta_r, M_r)\, \pi(\theta_r \mid M_r)\, \mathrm{d}\theta_r \propto \frac{C_r(\boldsymbol{Y})}{\bar{C}_r},$$

where $C_r(\boldsymbol{Y}) = \int \mathcal{L}(\theta_r \mid \boldsymbol{Y}, M_r)\, \mathrm{d}\theta_r$ and $\bar{C}_r = \int \mathcal{L}(\theta_r \mid M_r)\, \mathrm{d}\theta_r$ are the posterior and the prior normalization constants respectively. Because these normalization constants cannot be evaluated analytically, we consider the Markov Chain Monte Carlo (MCMC) method developed by Chen et al. (2008). This method is computationally convenient as it requires drawing only two MCMC samples: one from the posterior distribution and one from the prior distribution of θ_R under the unrestricted model M_R. By the results in Chen et al. (2008), the ratio of the posterior normalization constants in models M_r and M_R can be written

$$C_{r,R}(\boldsymbol{Y}) = \frac{C_r(\boldsymbol{Y})}{C_R(\boldsymbol{Y})} = \mathbb{E}\left[\left.\frac{\mathcal{L}(\theta_r \mid \boldsymbol{Y}, M_r)\, w(\theta_{-r} \mid \theta_r, M_R)}{\mathcal{L}(\theta_R \mid \boldsymbol{Y}, M_R)}\right| \boldsymbol{Y}\right],$$

where the expectation is taken with respect to the posterior distribution of θ_R under model M_R, θ_{-r} is a $(JK - P_r)$-dimensional vector of parameters obtained by deleting θ_r from θ_R, and $w(\theta_{-r} \mid \theta_r, M_R)$ is the conditional posterior density of θ_{-r} given θ_r under model M_R. Given a MCMC sample $\{\theta_R^s = \left(\theta_r^s, \theta_{-r}^s\right), s = 1, \dots, S\}$ from the posterior $\pi(\theta_R \mid \boldsymbol{Y}, M_R)$, under appropriate regularity conditions (e.g. ergodicity), $C_{r,R}(\boldsymbol{Y})$ can be consistently estimated by

$$\widehat{C}_{r,R} = \frac{1}{S}\sum_{s=1}^{S} \frac{\mathcal{L}(\theta_r^s \mid \boldsymbol{Y}, M_r)\, w(\theta_{-r}^s \mid \theta_r^s, M_R)}{\mathcal{L}(\theta_R^s \mid \boldsymbol{Y}, M_R)}.$$

Although the conditional density $w(\theta_{-r} \mid \theta_r, M_R)$ is generally not available in closed form, it can be approximated using the asymptotically normal approximation to the joint posterior of $\theta_R = (\theta_r, \theta_{-r})$ given by Chen (1985). The ratio $\bar{C}_{r,R}$ of the prior normalization constants can be estimated in a similar fashion using a MCMC sample from the prior $\pi(\theta_R \mid M_R)$. Given an estimate $\widetilde{C}_{r,R}$ of $\bar{C}_{r,R}$, posterior model probabilities can be estimated by

$$\widehat{\pi}_r = \frac{\widehat{C}_{r,R}/\widetilde{C}_{r,R}}{\sum_{r=1}^{R} \widehat{C}_{r,R}/\widetilde{C}_{r,R}},$$

where $\widehat{C}_{r,R}/\widetilde{C}_{r,R}$ is an estimate of the Bayes factor for model M_r against model M_R.

5.4. Block-BAML

Our last issue is how to handle the case when the number of candidate models in the model space $\mathcal{M}$ is large. With K covariates (including the constant term) and J subsamples with missing 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, Theorem 2 justifies confining attention to the J blocks of auxiliary variables associated with the various missing-data patterns, where the K auxiliary variables in each block capture the asymptotic bias of the fill-in estimator of β due to the imputation of the missing covariate values.

From the computation viewpoint, this block-BAML procedure has the important advantage of reducing the dimension of the model space from 2^{KJ} to 2^J. In applications where J does not exceed 20, one may then proceed by directly exploring all models. When J is large, our block-BAML 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 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 be achieved by stochastic search strategies based on MCMC methods, which allow exploring a considerably larger subset of models and provide 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 Sections 4 and 5 extend to settings where there is more than one outcome of interest and the nth component Y_n of $\boldsymbol{Y}$ is a Q-dimensional vector whose distribution is assumed to belong to the multivariate exponential family. This setup covers ordered, multinomial or conditional logit and probit models where the outcome can take $Q+1$ possible values corresponding to $Q+1$ mutually exclusive categories. The expression for the density of Y_n is now

$$f(y; \gamma_n) = \exp\left[\gamma_n^\top y - b(\gamma_n) + c(y)\right], \tag{6}$$

where γ_n is a Q-dimensional vector of canonical parameters, 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 $X_n^\top \beta_q$, 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 $Q \times Q$ identity matrix and $\otimes$ is Kronecker's product. The dependence of Y_n on the covariates is again modeled by assuming that there exists an inverse link 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 sample log-likelihood for the missing-free data is now

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

where the vector $\gamma_n(\beta)$ solves $b'(\gamma) = h((I_Q \otimes X_n^\top)\beta)$, and the missing-free data ML estimator $\widehat{\beta}$ 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 $\left[b''(\gamma_n(\beta)) \right]^{-1} h'((I_Q \otimes X_n^\top)\beta)$. The conditions for uniqueness, consistency and asymptotic normality of $\widehat{\beta}$ 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 enters the likelihood equations only through the linear predictor $\eta_n = (I_Q \otimes X_n^\top)\beta$ is all we need in order to adapt the proofs of the theorems in Section 4 to this case. To see this, it is enough to write the grand model as a 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-BAML 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. Our data are from release 1.1.1 of the fourth wave of the Survey of Health, Ageing and Retirement in Europe (SHARE), a multidisciplinary and cross-national household panel survey which covers about 58,500 individuals aged 50+, plus their spouses irrespective of age, in 16 European countries. To reduce the impact of cross-country differences in the fraction of the population living in institutions as opposed to households, we confine attention to people between 50 and 80 years of age.

To measure of cognitive ability we focus on the test of verbal fluency that consists of counting how many distinct members of the animal kingdom the respondent can name in one minute. The test outcome is an integer variable ranging from 0 to 100, which we model through a Poisson process estimated separately for four broad European regions: North (Denmark, the Netherlands, Sweden), West (Austria, Belgium, France, Germany, Switzerland), East (Czech Republic, Estonia, Hungary, Poland, Slovenia) and South (Italy, Portugal, Spain). Our covariates include self-reported measures of physical health (number of limitations in the 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 outcome and the covariates are presented in Table 1, separately by region.

Table 1
Descriptive statistics for the outcome and the covariates by region.

Region	Variable	Median	Mean	St.dev.	Min	Max
North	Fluency	22.0	22.6	6.8	0.0	58.0
	ADL	0.0	0.1	0.5	0.0	6.0
	Chronic	1.0	1.3	1.3	0.0	8.0
	Grip strength	35.0	36.9	12.0	4.0	80.0
	Age	64.0	64.5	7.8	50.0	80.0
	Male	0.0	0.5	0.5	0.0	1.0
	Education	1.0	0.7	0.5	0.0	1.0
	Income	1.6	1.9	1.3	0.0	13.9
	Net worth	1.7	2.7	4.3	−2.3	91.7
	Complete obs.					1 278
	Imputed obs.					5 841
West	Fluency	20.0	21.1	7.4	0.0	100.0
	ADL	0.0	0.1	0.6	0.0	6.0
	Chronic	1.0	1.5	1.5	0.0	11.0
	Grip strength	33.0	35.2	11.7	2.0	99.0
	Age	63.0	63.7	8.2	50.0	80.0
	Male	0.0	0.5	0.5	0.0	1.0
	Education	1.0	0.7	0.5	0.0	1.0
	Income	1.6	2.3	2.8	0.0	98.1
	Net worth	1.9	3.1	5.5	−5.0	197.2
	Complete obs.					4 697
	Imputed obs.					17 966
East	Fluency	21.0	21.1	7.5	0.0	93.0
	ADL	0.0	0.2	0.7	0.0	6.0
	Chronic	2.0	1.9	1.6	0.0	10.0
	Grip strength	33.0	34.5	12.0	2.0	99.0
	Age	64.0	64.2	8.1	50.0	80.0
	Male	0.0	0.4	0.5	0.0	1.0
	Education	1.0	0.7	0.5	0.0	1.0
	Income	0.8	3.1	6.2	0.0	216.6
	Net worth	1.0	6.4	20.9	−16.0	483.0
	Complete obs.					3 525
	Imputed obs.					17 443
South	Fluency	14.0	15.0	6.3	0.0	99.0
	ADL	0.0	0.2	0.8	0.0	6.0
	Chronic	1.0	1.7	1.5	0.0	10.0
	Grip strength	30.0	31.7	11.3	1.0	92.0
	Age	64.0	64.7	8.2	50.0	80.0
	Male	0.0	0.5	0.5	0.0	1.0
	Education	0.0	0.3	0.5	0.0	1.0
	Income	0.7	1.0	2.4	0.0	161.7
	Net worth	1.6	2.4	3.9	−3.6	152.1
	Complete obs.					2 074
	Imputed obs.					7 634

Notes: Fluency is the score in the verbal fluency test; ADL is the number of limitations in the activities of daily living; chronic is the number of chronic conditions; grip strength is the score in the grip strength test; age is the respondents' age in years; male is an indicator equal to one for males and to zero for females; education is an indicator equal to one for higher educational attainments and to zero otherwise; income is PPP-adjusted per-capita household income in units of 10,000 Euro; net worth is PPP-adjusted per-capita household net worth in units of 100,000 Euro.

Hand-grip strength, per-capita household income and household net worth are affected by substantial item nonresponse. The item nonresponse rates on these three covariates are respectively equal to 5%, 37% and 68%. In total, complete-case analysis would drop 76% of the sample. The number of subsamples with miss-

Table 2
Estimated coefficients and standard errors (in parentheses) of Poisson regression models for fluency by region.

Region	Variable	CC	FI	Block-BAML				
				CIC priors		CNJ priors		
				AIC	BIC	$\bar{\alpha} = 0.1$	$\bar{\alpha} = 0.01$	$\bar{\alpha} = 0.001$
North	ADL	−0.0634	−0.0476	−0.0648	−0.0475	−0.0663	−0.0677	−0.0712
		(0.0149)	(0.0064)	(0.0137)	(0.0065)	(0.0122)	(0.0111)	(0.0172)
	Chronic	−0.0046	−0.0084	−0.0022	−0.0084	0.0001	−0.0019	−0.0016
		(0.0047)	(0.0022)	(0.0053)	(0.0022)	(0.0045)	(0.0040)	(0.0054)
	Grip strength	0.0028	0.0038	0.0027	0.0038	0.0026	0.0027	0.0031
		(0.0008)	(0.0004)	(0.0008)	(0.0004)	(0.0007)	(0.0007)	(0.0008)
	Age	−0.0073	−0.0063	−0.0077	−0.0063	−0.0081	−0.0081	−0.0079
		(0.0008)	(0.0004)	(0.0009)	(0.0004)	(0.0008)	(0.0007)	(0.0010)
	Male	−0.0762	−0.0883	−0.0689	−0.0883	−0.0615	−0.0622	−0.0700
		(0.0197)	(0.0095)	(0.0203)	(0.0095)	(0.0179)	(0.0171)	(0.0186)
	Education	0.0925	0.1263	0.0945	0.1263	0.0967	0.0984	0.1109
		(0.0142)	(0.0063)	(0.0135)	(0.0063)	(0.0124)	(0.0119)	(0.0178)
	Income	0.0155	0.0086	0.0135	0.0086	0.0103	0.0054	0.0072
		(0.0073)	(0.0030)	(0.0069)	(0.0030)	(0.0066)	(0.0053)	(0.0061)
	Net worth	0.0053	0.0049	0.0061	0.0049	0.0072	0.0080	0.0074
		(0.0022)	(0.0014)	(0.0022)	(0.0014)	(0.0019)	(0.0018)	(0.0027)
	Constant	3.1213	3.0673	3.1169	3.0674	3.1121	3.1081	3.0907
		(0.0143)	(0.0062)	(0.0142)	(0.0062)	(0.0127)	(0.0120)	(0.0197)
West	ADL	−0.0530	−0.0530	−0.0461	−0.0420	−0.0437	−0.0428	−0.0424
		(0.0069)	(0.0033)	(0.0076)	(0.0037)	(0.0060)	(0.0041)	(0.0038)
	Chronic	0.0090	0.0078	0.0076	0.0074	0.0071	0.0067	0.0071
		(0.0023)	(0.0012)	(0.0021)	(0.0013)	(0.0017)	(0.0014)	(0.0013)
	Grip strength	0.0065	0.0059	0.0065	0.0062	0.0065	0.0063	0.0062
		(0.0004)	(0.0002)	(0.0004)	(0.0002)	(0.0004)	(0.0003)	(0.0002)
	Age	−0.0045	−0.0051	−0.0046	−0.0049	−0.0047	−0.0048	−0.0049
		(0.0004)	(0.0002)	(0.0004)	(0.0002)	(0.0003)	(0.0003)	(0.0002)
	Male	−0.1482	−0.1308	−0.1453	−0.1362	−0.1434	−0.1385	−0.1366
		(0.0099)	(0.0054)	(0.0087)	(0.0053)	(0.0082)	(0.0070)	(0.0055)
	Education	0.1993	0.1907	0.1966	0.1860	0.1947	0.1899	0.1871
		(0.0074)	(0.0039)	(0.0069)	(0.0040)	(0.0069)	(0.0056)	(0.0042)
	Income	0.0034	0.0042	0.0046	0.0038	0.0049	0.0042	0.0038
		(0.0014)	(0.0007)	(0.0016)	(0.0008)	(0.0014)	(0.0012)	(0.0009)
	Net worth	0.0037	0.0019	0.0030	0.0019	0.0026	0.0023	0.0019
		(0.0010)	(0.0005)	(0.0010)	(0.0005)	(0.0009)	(0.0008)	(0.0006)
	Constant	2.9668	2.9539	2.9641	2.9638	2.9635	2.9654	2.9650
		(0.0072)	(0.0038)	(0.0066)	(0.0039)	(0.0062)	(0.0047)	(0.0041)
East	ADL	−0.0423	−0.0498	−0.0423	−0.0463	−0.0423	−0.0423	−0.0433
		(0.0070)	(0.0027)	(0.0070)	(0.0091)	(0.0070)	(0.0070)	(0.0063)
	Chronic	−0.0083	−0.0043	−0.0083	−0.0114	−0.0083	−0.0084	−0.0114
		(0.0027)	(0.0012)	(0.0027)	(0.0024)	(0.0027)	(0.0027)	(0.0024)
	Grip strength	0.0074	0.0063	0.0074	0.0074	0.0074	0.0074	0.0074
		(0.0005)	(0.0003)	(0.0005)	(0.0004)	(0.0005)	(0.0005)	(0.0004)
	Age	−0.0084	−0.0074	−0.0084	−0.0075	−0.0084	−0.0084	−0.0075
		(0.0005)	(0.0002)	(0.0005)	(0.0005)	(0.0005)	(0.0006)	(0.0005)
	Male	−0.1318	−0.1280	−0.1318	−0.1326	−0.1318	−0.1318	−0.1324
		(0.0111)	(0.0055)	(0.0111)	(0.0099)	(0.0111)	(0.0110)	(0.0099)
	Education	0.1478	0.1382	0.1478	0.1453	0.1478	0.1477	0.1451
		(0.0081)	(0.0037)	(0.0081)	(0.0072)	(0.0081)	(0.0081)	(0.0072)
	Income	0.0089	0.0040	0.0089	0.0089	0.0089	0.0089	0.0092
		(0.0011)	(0.0006)	(0.0011)	(0.0017)	(0.0011)	(0.0011)	(0.0012)
	Net worth	−0.0003	0.0004	−0.0003	−0.0003	−0.0003	−0.0003	−0.0003
		(0.0003)	(0.0001)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
	Constant	2.9553	2.9918	2.9553	2.9633	2.9553	2.9555	2.9642
		(0.0076)	(0.0038)	(0.0076)	(0.0071)	(0.0076)	(0.0077)	(0.0068)

(*continued on next page*)

ing covariates is $J = 2^3 - 1 = 7$, so our model space consists of $R = 2^7 = 128$ models for each region. The public-use SHARE data include multiple imputations of income and net worth, which are constructed using five independent replicates of the fully conditional specification method of van Buuren et al. (2006). In our analysis, validity of these imputations may be questioned because verbal fluency 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. We produce our own multiple imputations for the missing values on hand grip strength using a simple hot-deck procedure.

The estimates of the focus parameters in the Poisson models for verbal fluency are presented in Table 2.[3] For each European region, we compare estimated coefficients and standard errors for the complete-case ML estimator (CC), the fill-in ML estimator (FI), and the block-BAML estimators based on CIC and conjugate priors. In the fill-in and the generalized missing-data approaches, estimated coefficients and standard errors resulting from the five multiple imputed datasets are combined using the formulas in Rubin (1987). For block-BAML estimates with CIC priors, we only

[3] Results for the auxiliary regressors are omitted to save space but are available upon request.

Table 2 (*continued*)

Region	Variable	CC	FI	Block-BAML				
				CIC priors		CNJ priors		
				AIC	BIC	$\bar{\alpha} = 0.1$	$\bar{\alpha} = 0.01$	$\bar{\alpha} = 0.001$
South	ADL	−0.0482	−0.0656	−0.0482	−0.0580	−0.0482	−0.0540	−0.0560
		(0.0095)	(0.0045)	(0.0095)	(0.0066)	(0.0095)	(0.0073)	(0.0064)
	Chronic	0.0118	0.0058	0.0118	0.0080	0.0118	0.0084	0.0079
		(0.0040)	(0.0022)	(0.0040)	(0.0027)	(0.0040)	(0.0033)	(0.0027)
	Grip strength	0.0027	0.0052	0.0027	0.0039	0.0027	0.0038	0.0040
		(0.0007)	(0.0004)	(0.0007)	(0.0005)	(0.0007)	(0.0007)	(0.0005)
	Age	−0.0082	−0.0078	−0.0082	−0.0080	−0.0082	−0.0079	−0.0079
		(0.0008)	(0.0004)	(0.0008)	(0.0005)	(0.0008)	(0.0006)	(0.0005)
	Male	0.0292	−0.0230	0.0291	−0.0054	0.0290	−0.0000	−0.0054
		(0.0164)	(0.0088)	(0.0164)	(0.0110)	(0.0165)	(0.0178)	(0.0110)
	Education	0.0839	0.1302	0.0839	0.0941	0.0839	0.0918	0.0935
		(0.0140)	(0.0071)	(0.0140)	(0.0090)	(0.0140)	(0.0105)	(0.0090)
	Income	0.1045	0.0087	0.1045	0.0850	0.1044	0.0885	0.0855
		(0.0080)	(0.0037)	(0.0080)	(0.0050)	(0.0081)	(0.0091)	(0.0049)
	Net worth	0.0023	0.0061	0.0023	0.0046	0.0024	0.0042	0.0045
		(0.0012)	(0.0007)	(0.0012)	(0.0009)	(0.0012)	(0.0013)	(0.0010)
	Constant	2.6589	2.6669	2.6589	2.6674	2.6589	2.6668	2.6680
		(0.0098)	(0.0052)	(0.0098)	(0.0066)	(0.0098)	(0.0080)	(0.0066)

Notes: The ML estimates from complete-case analysis are denoted by CC, from the fill-in approach by FI. The block-BAML estimates are based on the family of calibrated information criteria (CIC) priors and the family of conjugate (CNJ) priors. The standard errors of the fill-in ML estimates and the posterior standard deviations of the block-BAML estimates are computed by multiple imputation methods to account for the additional sampling variability due to imputation of missing covariate values. Results for the auxiliary regressors are omitted to save space.

report the results obtained using AIC and BIC because RIC leads to a penalty for complexity which is very similar to that used in BIC. For block-BAML estimates with conjugate priors, we set all elements of the vector $\bar{Y}$ of prior parameters equal to the sample mean of verbal fluency in the second wave of SHARE and consider three different choices of the prior parameter $\bar{a}$, namely 0.10, 0.01, and 0.001. These prior specifications imply that all regression coefficients have a zero prior mode except the constant term, the prior mode of which is instead equal to the logarithm of the marginal mean of the outcome in the second wave. This choice is attractive because the prior prediction for verbal fluency does not depend on the value of the covariates for a given individual. Further, as $\bar{a}$ decreases, we can assess how our block-BAML estimates changes as the prior become less informative. For conjugate priors, posterior model probabilities are always estimated through the MCMC algorithm discussed in Section 5.3 using a sample of $S = 20{,}000$ draws, after a "burn-in sample" of 10,000 draws, from the prior and the posterior under the unrestricted model.

Interpretation of the standard errors differs depending on the estimation strategy. For the complete-case approach, they can be interpreted as classical standard errors that ignore the additional sampling variability induced by both the imputation process and the model selection step. For the fill-in approach, the standard errors take into account the sampling variability induced by the imputation process but not the sampling variability induced by the model selection step. Finally, 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 given the multiple imputed data. By construction, they take explicitly into account the sampling variability due to both the imputation process and the model selection step.

Our results show little differences in the sign of the estimated associations across regions and estimation methods. Verbal fluency is typically higher for women than for men, is negatively related to age, and is positively related to self-reported and objective physical health measures and to variables typically associated with higher socio-economic status. In some regions, the size of the coefficients and the standard errors are however subject to non-negligible differences across estimation methods. 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. For example, these two approaches lead to sign changes for the estimated coefficients on net worth in the Eastern region and for the male dummy in the Southern region. For the coefficient on income in the Southern region, we obtain a complete-case ML estimate of 0.105 with a standard error of 0.008 and a fill-in ML estimate of 0.009 with a standard error of 0.004. As shown in Table 3, the resulting marginal effect of income on the expected value of the verbal fluency score (evaluated at the means of all covariates in the complete-case sample) is equal to 1.504 with a standard error of 0.116 for complete-case analysis and 0.125 with a standard error of 0.053 for the fill-in approach. For the fill-in ML estimates, the average percentage increase of standard errors caused by imputations is found to be large for the coefficients on income (98%) and net worth (65%). Despite this additional source of sampling variability, standard errors for the complete-case ML estimates are always much larger than those obtained for the fill-in ML estimates.

To facilitate the interpretation of our block-BAML estimates, we report in Table 4 the posterior probabilities of the top two models under the various prior specifications. AIC priors assign large posterior inclusion probabilities to almost all blocks of auxiliary covariates and thus leads to block-BAML estimates that are very close to the complete-case ML estimates. BIC priors support instead more parsimonious models by assigning large posterior model probabilities to either the restricted fill-in model or some intermediate model between the complete-case and fill-in model specifications. Block-BAML estimates with conjugate priors are somewhat in between the estimates resulting from AIC and BIC priors. Consistently with previous findings by Chen et al. (2008), conjugate priors with larger values of $\bar{a}$ tend to favor less parsimonious models but, as $\bar{a}$ decreases and priors become less informative, more parsimonious models receive higher posterior probabilities. For both families of priors, posterior model probabilities are typically concentrated at a few models. This suggests that uncertainty due to the model selection step is limited. Posterior standard deviations of block-BAML estimates are therefore similar to classical standard errors of ML estimates in the models with the highest posterior probabilities.

Our results cast some doubts about the validity of the SHARE imputations when studying cognitive functioning. This issue appears to be particularly important for countries belonging to the Eastern and the Southern regions, where discrepancies between complete-case and fill-in ML estimates are substantial and the

Table 3
Estimated marginal effects on the expected value of fluency and standard errors (in parentheses) by region and estimation method.

Region	Variable	CC	FI	Block-BAML				
				CIC priors		CNJ priors		
				AIC	BIC	$\bar{\alpha} = 0.1$	$\bar{\alpha} = 0.01$	$\bar{\alpha} = 0.001$
North	ADL	−1.4319	−1.0215	−1.4584	−1.0196	−1.4869	−1.5121	−1.5674
		(0.3355)	(0.1376)	(0.3076)	(0.1382)	(0.2722)	(0.2467)	(0.3968)
	Chronic	−0.1037	−0.1807	−0.0490	−0.1811	0.0014	−0.0418	−0.0336
		(0.1058)	(0.0480)	(0.1186)	(0.0481)	(0.1011)	(0.0891)	(0.1176)
	Grip strength	0.0631	0.0814	0.0600	0.0814	0.0573	0.0613	0.0682
		(0.0193)	(0.0091)	(0.0185)	(0.0091)	(0.0171)	(0.0167)	(0.0184)
	Age	−0.1653	−0.1355	−0.1733	−0.1354	−0.1816	−0.1818	−0.1731
		(0.0184)	(0.0087)	(0.0195)	(0.0088)	(0.0169)	(0.0157)	(0.0243)
	Male	−1.6577	−1.8134	−1.4991	−1.8131	−1.3363	−1.3460	−1.4828
		(0.4269)	(0.1948)	(0.4405)	(0.1949)	(0.3878)	(0.3690)	(0.3840)
	Education	2.1892	2.8902	2.2306	2.8900	2.2760	2.3088	2.5716
		(0.3286)	(0.1415)	(0.3114)	(0.1415)	(0.2861)	(0.2733)	(0.3876)
	Income	0.3503	0.1854	0.3045	0.1854	0.2300	0.1204	0.1564
		(0.1648)	(0.0635)	(0.1568)	(0.0635)	(0.1474)	(0.1174)	(0.1317)
	Net worth	0.1197	0.1055	0.1381	0.1055	0.1605	0.1780	0.1626
		(0.0494)	(0.0297)	(0.0499)	(0.0297)	(0.0434)	(0.0400)	(0.0608)
West	ADL	−1.0384	−1.0219	−0.9017	−0.8194	−0.8536	−0.8358	−0.8288
		(0.1345)	(0.0644)	(0.1508)	(0.0719)	(0.1176)	(0.0807)	(0.0746)
	Chronic	0.1768	0.1511	0.1486	0.1441	0.1377	0.1310	0.1378
		(0.0448)	(0.0237)	(0.0417)	(0.0246)	(0.0339)	(0.0269)	(0.0258)
	Grip strength	0.1274	0.1138	0.1277	0.1208	0.1271	0.1231	0.1213
		(0.0090)	(0.0049)	(0.0078)	(0.0048)	(0.0078)	(0.0063)	(0.0049)
	Age	−0.0885	−0.0992	−0.0905	−0.0955	−0.0917	−0.0943	−0.0952
		(0.0086)	(0.0043)	(0.0071)	(0.0045)	(0.0062)	(0.0052)	(0.0046)
	Male	−2.6973	−2.3645	−2.6423	−2.4833	−2.6078	−2.5269	−2.4933
		(0.1793)	(0.0963)	(0.1549)	(0.0969)	(0.1438)	(0.1224)	(0.0994)
	Education	4.3197	4.0500	4.2449	3.9885	4.1978	4.0867	4.0174
		(0.1573)	(0.0806)	(0.1496)	(0.0835)	(0.1486)	(0.1217)	(0.0907)
	Income	0.0658	0.0817	0.0896	0.0736	0.0964	0.0811	0.0744
		(0.0283)	(0.0143)	(0.0309)	(0.0155)	(0.0277)	(0.0228)	(0.0173)
	Net worth	0.0729	0.0360	0.0589	0.0371	0.0515	0.0452	0.0378
		(0.0187)	(0.0099)	(0.0194)	(0.0105)	(0.0178)	(0.0151)	(0.0114)
East	ADL	−0.8088	−0.9853	−0.8088	−0.8925	−0.8088	−0.8096	−0.8368
		(0.1344)	(0.0531)	(0.1344)	(0.1728)	(0.1344)	(0.1342)	(0.1215)
	Chronic	−0.1583	−0.0851	−0.1583	−0.2193	−0.1583	−0.1600	−0.2195
		(0.0514)	(0.0230)	(0.0514)	(0.0459)	(0.0514)	(0.0523)	(0.0465)
	Age	−0.1612	−0.1456	−0.1612	−0.1447	−0.1612	−0.1608	−0.1448
		(0.0103)	(0.0046)	(0.0103)	(0.0091)	(0.0103)	(0.0106)	(0.0093)
	Male	−2.3623	−2.3751	−2.3623	−2.3969	−2.3623	−2.3632	−2.3965
		(0.1964)	(0.1014)	(0.1964)	(0.1768)	(0.1964)	(0.1961)	(0.1779)
	Education	3.0457	2.9304	3.0457	3.0172	3.0457	3.0450	3.0170
		(0.1649)	(0.0782)	(0.1649)	(0.1480)	(0.1649)	(0.1646)	(0.1481)
	Grip strength	0.1414	0.1244	0.1414	0.1429	0.1414	0.1415	0.1439
		(0.0099)	(0.0052)	(0.0099)	(0.0091)	(0.0099)	(0.0099)	(0.0089)
	Income	0.1711	0.0796	0.1711	0.1724	0.1711	0.1711	0.1786
		(0.0219)	(0.0116)	(0.0219)	(0.0332)	(0.0219)	(0.0218)	(0.0234)
	Net worth	−0.0049	0.0085	−0.0049	−0.0051	−0.0049	−0.0049	−0.0056
		(0.0059)	(0.0022)	(0.0059)	(0.0061)	(0.0059)	(0.0059)	(0.0059)
South	ADL	−0.6940	−0.9419	−0.6940	−0.8393	−0.6945	−0.7820	−0.8112
		(0.1368)	(0.0649)	(0.1368)	(0.0945)	(0.1368)	(0.1053)	(0.0924)
	Chronic	0.1701	0.0836	0.1701	0.1152	0.1698	0.1218	0.1138
		(0.0578)	(0.0313)	(0.0578)	(0.0387)	(0.0578)	(0.0477)	(0.0389)
	Age	−0.1176	−0.1115	−0.1176	−0.1157	−0.1176	−0.1148	−0.1147
		(0.0116)	(0.0059)	(0.0116)	(0.0077)	(0.0116)	(0.0085)	(0.0077)
	Male	0.4258	−0.3262	0.4257	−0.0785	0.4234	0.0006	−0.0779
		(0.2394)	(0.1249)	(0.2395)	(0.1582)	(0.2416)	(0.2593)	(0.1594)
	Education	1.2595	1.9974	1.2595	1.4274	1.2602	1.3918	1.4194
		(0.2146)	(0.1134)	(0.2146)	(0.1395)	(0.2145)	(0.1648)	(0.1399)
	Grip strength	0.0391	0.0754	0.0391	0.0569	0.0392	0.0549	0.0575
		(0.0108)	(0.0061)	(0.0108)	(0.0076)	(0.0108)	(0.0108)	(0.0076)
	Income	1.5037	0.1247	1.5037	1.2296	1.5025	1.2812	1.2385
		(0.1164)	(0.0531)	(0.1165)	(0.0724)	(0.1176)	(0.1289)	(0.0721)
	Net worth	0.0337	0.0883	0.0337	0.0671	0.0338	0.0602	0.0656
		(0.0174)	(0.0096)	(0.0174)	(0.0136)	(0.0175)	(0.0190)	(0.0139)

Notes: The marginal effects are evaluated at the mean value of all the covariates in the complete-case sample. The block-BAML estimates of the marginal effects are computed as weighted averages of the conditional marginal effects under each model with weights equal to the posterior model probabilities. The standard errors of the conditional marginal effects are computed by the delta-method. The standard errors of the fill-in ML estimates and the posterior standard deviations of the block-BAML estimates are computed by multiple imputation methods to account for the additional sampling variability due to imputation of the missing covariate values. Results for the auxiliary regressors are omitted to save space.

Table 4
Posterior probabilities for the top two models by region and block-BAML approach.

Region	Block	Top	Blocks of auxiliary covariates							$\pi_r(\boldsymbol{Y})$
	BAML	Models	M_1	M_2	M_3	M_4	M_5	M_6	M_7	
North	AIC	1	1	0	1	1	1	1	1	0.5409
		2	1	1	1	1	1	1	1	0.4364
	BIC	1	0	0	0	0	0	0	0	0.9989
		2	0	0	0	0	0	0	1	0.0011
	CNJ($\bar{a} = 0.1$)	1	1	0	1	1	1	1	1	0.7963
		2	1	0	1	1	0	1	1	0.0929
	CNJ($\bar{a} = 0.01$)	1	1	0	1	1	0	0	1	0.9961
		2	1	0	1	0	0	0	1	0.0025
	CNJ($\bar{a} = 0.001$)	1	0	0	1	0	0	0	0	0.9023
		2	1	0	1	0	0	0	0	0.0756
West	AIC	1	0	1	1	1	1	1	1	0.5644
		2	1	1	1	1	1	1	1	0.3506
	BIC	1	0	0	0	1	1	0	1	1.0000
		2	0	0	0	1	1	1	1	0.0000
	CNJ($\bar{a} = 0.1$)	1	0	1	1	1	1	1	1	0.7590
		2	1	1	0	1	1	1	1	0.1218
	CNJ($\bar{a} = 0.01$)	1	0	1	0	1	1	1	1	0.9960
		2	0	1	1	1	1	1	1	0.0033
	CNJ($\bar{a} = 0.001$)	1	0	0	0	1	1	0	1	0.6612
		2	0	1	0	1	1	0	1	0.2673
East	AIC	1	1	1	1	1	1	1	1	1.0000
		2	1	0	1	1	1	1	1	0.0000
	BIC	1	1	0	1	1	1	1	1	0.9942
		2	1	0	1	0	1	1	1	0.0055
	CNJ($\bar{a} = 0.1$)	1	1	1	1	1	1	1	1	1.0000
		2	1	0	1	1	1	1	1	0.0000
	CNJ($\bar{a} = 0.01$)	1	1	1	1	1	1	1	1	0.9496
		2	1	0	1	1	1	1	1	0.0504
	CNJ($\bar{a} = 0.001$)	1	1	0	1	1	1	1	1	0.9996
		2	1	1	1	1	1	1	1	0.0004
South	AIC	1	1	1	1	1	1	1	1	0.9993
		2	0	1	1	1	1	1	1	0.0007
	BIC	1	0	1	1	1	1	0	1	0.9960
		2	0	1	1	1	1	1	1	0.0040
	CNJ($\bar{a} = 0.1$)	1	1	1	1	1	1	1	1	0.9967
		2	0	1	1	1	1	1	1	0.0033
	CNJ($\bar{a} = 0.01$)	1	0	1	1	1	1	1	1	0.8668
		2	1	1	1	1	1	1	1	0.1332
	CNJ($\bar{a} = 0.001$)	1	0	1	1	1	1	1	1	0.6307
		2	0	1	1	1	1	0	1	0.3693

Notes: The $M_j, j = 1, \ldots, 7$, are indicators for the missing-data patterns in grip strength, income and net worth. If "1" denotes missing and "0" denotes observed, then $M_1 = (0, 0, 1)$, $M_2 = (0, 1, 0)$, $M_3 = (0, 1, 1)$, $M_4 = (1, 0, 0)$, $M_5 = (1, 0, 1)$, $M_6 = (1, 1, 0)$, and $M_7 = (1, 1, 1)$.

fill-in approach unambiguously receives little support from our set of block-BAML estimates, whatever the chosen prior distribution.

The main lessons we draw from this exercise are the following. First, the loss of precision from using only the complete cases is substantial. Hence the need to somehow exploit the incomplete observations. On the other hand, taking straight the fill-in approach can be quite misleading. Hence routes such as our BAML procedure are highly desirable. Finally, it is noteworthy that although all models are equally likely *a priori*, our block-BAML procedure leads to a posterior distribution concentrated at a few models.

8. Conclusions

This paper considers the problem of estimating GLMs 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 their validity for granted and should explicitly consider the trade-off between bias and precision involved in their use. Our approach reformulates this trade-off as a problem of model uncertainty, which can be handled very naturally through Bayesian averaging of classical ML estimators. The particular structure of this problem allows us to adopt a block model averaging strategy that is straightforward and makes it possible to explore all the relevant submodels.

Our empirical application shows that inference based on standard approaches to missing covariates and on our generalized missing-data approach may be substantially different.

In future work we plan to use our approach to formally test the validity of imputations given the specific GLM of interest.

Acknowledgments

We thank Francesco Bartolucci and Jan Magnus for very useful discussions. We also thank an Associate Editor and two anonymous referees for detailed comments that greatly helped improve the paper. Valentino Dardanoni and Franco Peracchi acknowledge financial support from the MIUR PRIN 2010T8XAXB, Salvatore Modica from the MIUR PRIN 20103S5RN3.

Appendix. Asymptotic properties of complete-case and fill-in ML estimators

From Section 4.1, the complete-case ML estimator $\widehat{\beta}_0$ coincides with the ML estimator of β in the grand model with linear predictor $\boldsymbol{\eta} = \boldsymbol{W}\beta + \boldsymbol{Z}\delta$. This ML estimator, denoted by $\widetilde{\theta} = (\widetilde{\beta}, \widetilde{\delta})$, converges in probability to the true population value $\theta^0 = (\beta^0, \delta^0)$ which solves the equation system

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

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 $\widetilde{\beta}$ is the top-left block of the inverse of $\mathcal{I}^0$, it follows that

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

On the other hand, the fill-in ML estimator $\widehat{\beta}_F$ solves the equation system

$$\frac{1}{N}\sum_{n=1}^{N} s_\beta(\beta, 0; W_n, Z_n) = 0. \qquad (7)$$

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

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

which does not generally coincide with the true population value β^0. A first-order Taylor expansion of (7) around the pseudo-true value β^* gives

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

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

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

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

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

a positive definite matrix. Therefore,

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

When the imputations are valid, the restriction that $\delta = 0$ is valid. So, the fill-in ML estimator $\widehat{\beta}_F$ is consistent and asymptotically more precise than the complete-case ML estimator $\widehat{\beta}_0$, that is $AV(\widehat{\beta}_F)^{-1} - AV(\widehat{\beta}_0)^{-1} \geq 0$.

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

$$AV(\widehat{\beta}_F)^{-1} - AV(\widehat{\beta}_0)^{-1} = \mathcal{I}^0_{\beta\delta}\mathcal{I}^{0-1}_{\delta\delta}\mathcal{I}^0_{\delta\beta},$$

which is a nonnegative definite matrix.

References

Akaike, H., 1978. A new look at the Bayes procedure. Biometrika 65, 53–59.
Burnham, K.P., Anderson, D.R. (Eds.), 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach, second ed. Springer-Verlag, New York.
Chen, M.H., 1985. On asymptotic normality of limiting density functions with Bayesian implications. J. R. Stat. Soc. Ser. B 47, 540–546.
Chen, M.H., Huang, L., Ibrahim, J.G., Kim, S., 2008. Bayesian variable selection and computation for generalized linear models with conjugate priors. Bayesian Anal. 3, 585–614.
Chen, M.-H., Ibrahim, J.G., 2003. Conjugate priors for generalized linear models. Statist. Sinica 13, 461–476.
Clyde, M.A., 2000. Model uncertainty and health effect studies for particular matter. Environmetrics 11, 745–763.
Clyde, M.A., George, E.I., 2004. Model uncertainty. Statist. Sci. 19, 81–94.
Danilov, D., Magnus, J.R., 2004. On the harm that ignoring pretesting can cause. J. Econometrics 122, 27–46.
Dardanoni, V., Modica, S., Peracchi, F., 2011. Regression with imputed covariates: a generalized missing-indicator approach. J. Econometrics 162, 362–368.
Draper, D., 1995. Assessment and propagation of model uncertainty. J. R. Stat. Soc. Ser. B 57, 45–97.
Fahrmeir, L., Kaufmann, H., 1985. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Ann. Statist. 13, 342–368.
Foster, D.P., George, E.I., 1994. The risk information criterion for multiple regression. Ann. Statist. 22, 1947–1975.
Furnival, G.M., Wilson, R.W., 1974. Regression by leaps and bounds. Technometrics 16, 499–511.
Gourieroux, C., Monfort, A., Renault, E., Trognon, A., 1987. Generalized residuals. J. Econometrics 34, 5–32.
Han, C., Carlin, B.P., 2001. Markov chain Monte Carlo methods for computing Bayes factors. J. Amer. Statist. Assoc. 96, 1122–1132.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statist. Sci. 14, 382–401.
Jeffreys, H., 1961. Theory of Probability, third ed. Oxford Unversity Press, Oxford, United Kingdom.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90, 773–795.
Madigan, D., Raftery, A.E., 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Amer. Statist. Assoc. 89, 1535–1546.
Magnus, J.R., De Luca, G., 2014. Weighted-average least squares (WALS): a survey. Mimeo.
Magnus, J.R., Durbin, J., 1999. Estimation of regression coefficients of interest when other regression coefficients are of no interest. Econometrica 67, 639–643.
Magnus, J.R., Powell, O., Prüfer, P., 2010. A comparison of two averaging techniques with an application to growth empirics. J. Econometrics 154, 139–153.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, London.
Meng, X.L., 1994. Multiple-imputation inferences with uncongenial sources of input. Statist. Sci. 9, 538–558.
Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. J. R. Stat. Soc. Ser. A 135, 370–384.
Raftery, A.E., 1993. Bayesian model selection in structural equation models. In: Bollen, K., Long, J. (Eds.), Testing Structural Equation Models. Sage, Newbury Park, CA, pp. 163–180.
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.
Sala-i-Martin, X., Doppelhofer, G., Miller, R.I., 2004. Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. Amer. Econ. Rev. 94, 813–835.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.
Seaman, S., Galati, J., Jackson, D., Carlin, J., 2013. What is meant by "missing at random?" Statist. Sci. 28, 257–268.
Tierney, L., Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities. J. Amer. Statist. Assoc. 81, 82–86.
van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M., Rubin, D.B., 2006. Fully conditional specification in multivariate imputation. J. Stat. Comput. Simul. 76, 1049–1064.
Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data, second ed. MIT Press, Cambridge, Massachusetts.