BAYESIAN MODEL AVERAGING FOR GENERALIZED LINEAR MODELS WITH MISSING COVARIATES

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

Abstract. This paper addresses the problem of estimating generalized linear models (GLMs) when the values of some covariates are missing for some observations but imputations are available to fill-in the missing values. We approach this problem using the generalized missing-indicator method applied to linear regression by Dardanoni et al. (2011, 2012). This method is attractive because it allows us to handle model uncertainty very naturally through Bayesian model averaging. In addition to applying the method to the general class of GLMs, we make two extensions. First, we allow the observed outcome to be multivariate, thus covering the case of seemingly unrelated regression equations models, ordered logit and probit models, and multinomial logit and probit models. Second, we propose a model averaging strategy that incorporates information on the available missing-data patterns and has the advantage of being computationally simple.

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

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

Date: November 15, 2012.
We thank Jan Magnus for useful discussions. Affiliations. Dardanoni, De Luca, Modica: University of Palermo; Peracchi: University of Rome Tor Vergata and EIEF.
1. Introduction

Generalized linear models (GLMs) are a widely used class of models that includes the Gaussian linear regression model and a variety of nonlinear models typically employed for categorical outcomes, such as logit, probit, and Poisson regression models. This paper addresses the problem of estimating this class of models in the empirically relevant case where 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.

We approach this problem using the generalized missing-indicator method originally applied to linear regression by Dardanoni et al. (2011, 2012). Their approach exploits the fact that complete-case analysis (which drops all observations with missing covariate values) and the fill-in approach (which uses all the observations, without distinguishing between observed and imputed values) correspond to using two extreme versions of the same model. Complete-case analysis corresponds to using a “grand model” that includes two subsets of regressors: the focus regressors consisting of the observed or imputed covariates, and a set of auxiliary regressors consisting of various missing-data indicators. The fill-in approach corresponds instead to using a restricted version of the grand model that includes only the focus regressors. The key idea of the DMP approach is to consider all the intermediate models between these two extremes by dropping alternative subsets of auxiliary regressors from the grand model. This method is attractive because it allows one to handle model uncertainty very naturally through Bayesian Model Averaging (BMA).

In addition to applying the method to the general class of GLMs, we make two extensions. First, we allow the observed outcome to be multivariate, thus covering the case of seemingly unrelated regression equations (SURE) models and ordered or multinomial logit and probit models. Second, we propose a block-BMA strategy that incorporates the information on the available patterns of missing data and has the advantage of being computationally simple.

Throughout the paper we focus on the nonlinear case. For ease of notation, results are mainly presented for the univariate case, although they all extend to the case of multivariate outcomes in a straightforward manner.
We illustrate our approach using data from the first wave of the Survey on Health, Aging and Retirement in Europe (SHARE), a multi-purpose cross-national household panel carried out in 2004–2005 in 11 continental European countries.

The remainder of the paper is organized as follows. Section 2 presents our statistical framework. Section 3 introduces our notation for missing data patterns. Section 4 discusses standard approaches to missing data. Section 5 describes our method. Section 6 discusses how to apply BMA in our case. Section 7 extends our results to the case of multivariate outcomes. Section 8 presents an empirical application. Finally, Section 9 offers some conclusions.

2. Statistical framework

GLMs assume that the conditional distribution of a scalar outcome $Y$ belongs to the (one-parameter) linear exponential family, with density functions of the form

$$f(y; \gamma) = \exp \left[ \gamma y - b(\gamma) + c(y) \right],$$

(1)

where $\gamma \in \Gamma \subset \mathbb{R}$ is the canonical parameter, $b(\cdot)$ is a known, strictly convex and twice differentiable function of $\gamma$, and $c(\cdot)$ is a known function of $y$.\(^1\) By the properties of the linear exponential family (McCullagh and Nelder 1989), the mean and variance of $Y$ are equal to $b'(\gamma)$ and $b''(\gamma)$ respectively.

To model the dependence of $Y$ on a $K$-vector of covariates $X$, it is typically assumed that there exists a continuously differentiable and invertible function $h(\cdot)$ such that, for all $x$ in the support of $X$, the conditional mean of $Y$ given $X = x$ is equal to $h(x^T \beta)$ for a unique $K$-dimensional parameter vector $\beta$. The function $h(\cdot)$ is sometimes called the inverse link function and the linear combination $x^T \beta$ is called the linear predictor.

This paper deals with the problem of how to estimate the regression parameter $\beta$ when some covariate values are missing but imputations are available to fill-in these missing values. In the absence of missing data problems, the standard approach to estimation of $\beta$ is maximum likelihood (ML). The log-likelihood for a random sample of $N$ observations $(Y_1, X_1), \ldots, (Y_N, X_N)$ is

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

\(^1\) As in the original formulation of Nelder and Wedderburn (1972), the densities in equation (1) can be generalized to include an additional dispersion parameter. Without loss of generality, we set this nuisance parameter to one as it does not affect the maximum likelihood estimator of $\gamma$ and it can be consistently estimated.
where \( \gamma_n(\beta) \) is the unique root of the equation \( b'(\gamma) = h(X_n^T \beta) \), and the full-information ML estimator of \( \beta \) is obtained by solving the system of \( K \) likelihood equations

\[
l'(\beta) = \sum_{n=1}^{N} v(X_n^T \beta) [Y_n - h(X_n^T \beta)] X_n = 0,
\]  

(2)

where \( v(X_n^T \beta) = h'(X_n^T \beta)/b''(\gamma_n(\beta)) \). Notice that \( \beta \) enters the above equations only through the linear predictor \( X_n^T \beta \); this is the property which will drive our main results. If \( b'(\cdot) = h(\cdot) \) (the “canonical link” case), then \( \gamma_n(\beta) = X_n^T \beta \) and the likelihood equations (2) simplify considerably because \( v_n(X_n^T \beta) = 1 \). Provided that the model is correctly specified and the mild regularity conditions in Fahrmeir and Kaufmann (1985) hold, the full-information ML estimator of \( \beta \) is known to be unique, consistent, and asymptotically normal with asymptotic variance equal to the inverse of the Fisher information matrix.

3. Missing-data patterns

The data are represented as an \( N \times (K + 1) \) matrix \([Y, X]\), where \( Y \) is an \( N \times 1 \) vector of observations on the outcome of interest and \( X \) is an \( N \times K \) matrix of observations on the covariates. Observations are indexed by \( n = 1, \ldots, N \), and covariates by \( k = 0, 1, \ldots, K - 1 \), with \( k = 0 \) corresponding to the constant term. We depart from the standard setup of GLMs by assuming that the elements of the matrix \( X \) can be missing for some \( n \) and \( k \). We also assume that imputations are available to fill-in the missing covariate values.

Our setup allows the missing-data patterns to be very general, as in Dardanoni et al. (2011). We define a subsample with incomplete data as a group of observations where one or more covariates are missing. Since the constant term is always observed, the number of possible subsamples with incomplete data is \( 2^{K-1} - 1 \) (only the first covariate missing, only the first and the second missing, etc.) so, including the subsample without missing 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 without missing data (which is assumed to be always available) and \( J \leq 2^{K-1} - 1 \).

We assume that the \( j \)th subsample contains \( N_j \) observations, \( K_j \) available (non missing) covariates, and \( K_j^* = K - K_j \) missing covariates. Hence, by construction, \( \sum_{j=0}^{J} N_j = N \), \( K_0 = K \), and \( 1 \leq K_j \leq K \) for \( j = 1, \ldots, J \). For each subsample, we also define the \( N_j \times 1 \) outcome vector \( y^j \), the \( N_j \times K_j \) submatrix \( X_a^j \) containing the values of the available covariates,
the $N_j \times K_{j}^*$ submatrix $X_{m_j}^j$, containing the values of the missing covariates, and the $N_j \times K$ matrix $X^j = [X_{a_j}^j, X_{m_j}^j]$. Further, to keep track of which covariate values are missing, we introduce the $N \times K$ missing indicator matrix $M$, whose $(n,k)$th element $m_{nk}$ is equal to one if the $k$th element of $X_n$ is missing, and is equal to zero otherwise.

4. Standard approaches

This section discusses the two standard approaches to the problem of missing covariate values, namely complete-case analysis (Section 4.1) and the fill-in approach (Section 4.2).

4.1. Complete-case analysis. This approach amounts to estimating a GLM on the subsample with complete data, $[Y^0, X^0]$, ignoring the imputations altogether. Complete-case analysis is a useful benchmark because it gives a consistent and asymptotically normal ML estimator of $\beta$ 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.** $M$ and $Y$ are independent conditionally on $X$.

Assumption 1 ensures 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 $N_0^{-1}X^0_0^TX^0_0$ converges in probability to a positive definite matrix as $N \to \infty$.

Assumption 2 is an ignorability assumption on the missing-data process. This assumption implies that the distribution of $Y$ given $X$ and $M$ is the same as the distribution of $Y$ given $X$ or, equivalently, the conditional distribution of $Y$ given $X$ is the same in subsamples with complete and incomplete data. Given the true values of the covariates, the pattern of missing data can then be ignored when predicting $Y$. Notice this assumption is stronger than the conditional mean independence needed to ensure unbiasedness of the complete-case OLS estimator of $\beta$ in classical linear regression models. However, as discussed in Dardanoni et al. (2012), it not the same as the standard missing-at-random (MAR) assumption which instead requires the missing-data process to be independent of the missing covariates given the observed outcome and the available covariates (Rubin 1976). For example, suppose that health is the outcome of interest and the only covariate, 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. The next section considers an alternative approach to the problem of missing covariate values which avoids dropping the incomplete observations by making use of the available imputations.

4.2. Fill-in approach. For each subsample with incomplete data, let $L^j$ be the $N_j \times K_j^*$ matrix containing the imputed values of the missing covariates. The $N_j \times K$ matrix $W^j = [X^j_0, L^j]$ is called the completed design matrix for the $j$th subsample ($j = 1, \ldots, J$). Reordering the observations by stacking on top of each other the $J + 1$ available subsamples gives

$$Y = \begin{bmatrix} Y^0 \\ Y^1 \\ \vdots \\ Y^J \end{bmatrix}, \quad W = \begin{bmatrix} X^0 \\ W^1 \\ \vdots \\ W^J \end{bmatrix},$$

where $W$ is the $N \times K$ completed design matrix for the full sample, with generic element equal to the corresponding element of $X$ if this is not missing and to its imputed value otherwise.

The fill-in approach consists of estimating the GLM for $Y$ with $X$ replaced by the completed design matrix $W$. This approach requires two conditions. First, the missing-data process must be MAR. Second, the imputation model used to create the imputations must be “congenial” in the sense of Meng (1994) (i.e. correctly specified given the true data generating process). We say that the imputations are valid when these two conditions hold. If the imputations are valid, then the fill-in ML estimator of $\beta$ is asymptotically equivalent to the full-information 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 4.1. In general, only little can be said about the finite sample properties of these ML estimators. Since the number of unknown parameters is the same in the complete-case and the fill-in approach, but the number of observations is greater in the latter, the fill-in ML estimator may be expected to have higher finite sample precision than the complete-case ML estimator provided that the additional sampling variability induced by imputation is small. On the other hand, when the MAR and congeniality assumptions do not hold, the fill-in estimator of $\beta$ 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 uncongeniality.

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 but may be handled by multiple imputation methods (Rubin 1987).

5. The generalized missing-indicator approach

We now extend the generalized missing-indicator approach introduced by Dardanoni et al. (2011) to the wider class of GLMs.

The key idea of this approach is to augment the $K$ available and imputed covariates in the completed design matrix $W$ with a set of $JK$ covariates corresponding to binary indicators for the missing-data patterns and their interactions with the covariates in $W$. Thus define the $N \times JK$ matrix of additional covariates $Z = \begin{bmatrix} 0 & \ldots & 0 \\ W^1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & W^J \end{bmatrix}$.

Our statistical model for the full sample is an augmented GLM where the conditional density of $Y$ given $W$ and $Z$ is assumed to belong to the exponential family (1) with a linear predictor of the form $\eta = W\beta + Z\delta$. By analogy with the terminology of Dardanoni et al. (2011), this augmented GLM is called the grand model. The main difference with respect to the fill-in approach is that, in addition to the vector of parameters $\beta$, the grand model now contains the $JK \times 1$ vector of nuisance parameters $\delta$ associated with $Z$. As shown in Appendix B, these nuisance parameters can be interpreted as the asymptotic bias of the fill-in estimator of $\beta$ when the imputations are
not valid, either because the missing-data process is not MAR or because the imputation model is uncongenial.

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

**Theorem 1.** For any choice of the imputation matrices \( L^1, \ldots, L^J \), the ML estimator of \( \beta \) in the grand model with linear predictor \( \eta = W \beta + Z \delta \) is numerically the same as the complete-case ML estimator of \( \beta \).

**Proof.** Let

\[
Y = \begin{bmatrix} Y^0 \cr Y^* \end{bmatrix}, \quad W = \begin{bmatrix} X^0 \cr W^* \end{bmatrix}, \quad Z = \begin{bmatrix} 0 \\ Z^* \end{bmatrix},
\]

where

\[
Y^* = \begin{bmatrix} Y^1 \\ \vdots \\ Y^J \end{bmatrix}, \quad W^* = \begin{bmatrix} W^1 \\ \vdots \\ W^J \end{bmatrix}, \quad Z^* = \begin{bmatrix} W^1 & \cdots & W^J \end{bmatrix}.
\]

The complete-case ML estimator \( \hat{\beta} \) of \( \beta \) solves the system of \( K \) likelihood equations

\[
X^0 \top U^0(\beta) = 0,
\]

where \( U^0(\beta) \) is the \( N_0 \times 1 \) vector of generalized residuals (Gourieroux et al., 1987) with generic element

\[
U^0_n(\beta) = v(X_n^\top \beta)[Y_n - h(X_n^\top \beta)].
\]

On the other hand, the ML estimator \( (\tilde{\beta}, \tilde{\delta}) \) of \( (\beta, \delta) \) in the grand model with linear predictor \( \eta = W \beta + Z \delta \) solves the system of \( K + JK \) likelihood equations

\[
X^0 \top U^0(\beta) + W^* \top U^*(\beta, \delta) = 0, \quad Z^* \top U^*(\beta, \delta) = 0,
\]

where \( U^*(\beta, \delta) \) is the \((N - N_0) \times 1\) vector with generic element

\[
U^*_n(\beta, \delta) = v(W_n^\top \beta + Z_n^\top \delta)[Y_n - h(W_n^\top \beta + Z_n^\top \delta)].
\]

Since \( Z^* \) is a block-diagonal matrix, the last \( JK \) equations in (3) imply that

\[
W^* \top U^*(\beta, \delta) = 0.
\]

The ML estimator \( \tilde{\beta} \) then solves \( X^0 \top U^0(\beta) = 0 \), so it must coincide with the complete-case estimator \( \hat{\beta} \). \( \square \)
Theorem 1 shows that the complete-case ML estimator of $\beta$ is numerically the same as the unrestricted ML estimator of $\beta$ in the grand model that places no restrictions on the vector $\delta$ of nuisance parameters. This estimator is consistent under the assumptions in Section 4.1. At the other extreme, the fill-in estimator of $\beta$ corresponds to the restricted ML estimator of $\beta$ when all elements of $\delta$ in the grand model are set to zero. If the imputations are valid (that is, $\delta = 0$), then the fill-in estimator of $\beta$ is asymptotically more precise than the complete-case estimator (see Appendix A). However, if $\delta \neq 0$ because either the missing-data mechanism is not MAR or the imputation model is uncongenial, then the fill-in estimator is inconsistent and asymptotically biased. The aim of the generalized missing-indicator approach is to handle this trade-off between asymptotic bias and asymptotic precision in the estimation of $\beta$ by considering all intermediate models obtained from the grand model setting to zero arbitrary subsets of elements in $\delta$. This strategy has two advantages. First, the original trade-off between asymptotic bias and asymptotic precision in the estimation of $\beta$ 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 finding the best available estimator of $\beta$ in the asymptotic mean squared error sense.

5.2. A dual result. We now prove a result that may be regarded as the “dual” of Theorem 1. This result provides the justification for the approach taken later on in Section 6.4.

In terms of adding observations to the subsample without missing data, Theorem 1 says that the complete-case approach is equivalent, as far as estimation of $\beta$ is concerned, to using the grand model that includes all the observations (observed or imputed) and all the auxiliary variables. The next theorem shows that, more generally, adding subsets of observations to subsample without missing data is equivalent to removing groups of auxiliary variables from the grand model.

Given a particular collection $J$ of subsamples with missing data, let $Y^+ \subset Y$ be the subvector of $Y$ obtained by keeping $Y^0$ and all the $Y^j$ such that $j \in J$, let $Y^- \subset Y$ be the subvector consisting of the remaining rows of $Y$, let $W^+ \subset W$ be the submatrix of $W$ obtained by keeping $X^0$ and all the $X^j$ such that $j \in J$, let $W^- \subset W$ be the submatrix consisting of the remaining rows of $W$, and let $Z^{++}$ be the submatrix obtained by deleting from $Z^+$ the rows and columns corresponding to
\(W^j\). Also let
\[Y = \begin{bmatrix} Y^+ \\ Y^- \end{bmatrix}, \quad W = \begin{bmatrix} W^+ \\ W^- \end{bmatrix}, \quad Z^- = \begin{bmatrix} 0 \\ Z^{**} \end{bmatrix},\]
and define \(\eta^- = W\beta + Z^-\delta^-\), where \(\delta^-\) denotes the subvector of \(\delta\) obtained by deleting the coefficients associated with the \(W^j, j \in J\).

Theorem 1 can now be restated as follows: If \(J\) is the empty set, then the ML estimates of \(\beta\) in the GLM for \([Y^+, W^+]\) and in the GLM for \([Y, W, Z^-]\) coincide. The next theorem shows that this is actually true if \(J\) is any collection of subsamples with missing data.

**Theorem 2.** For any collection \(J\) of subsamples with missing data, the ML estimates of \(\beta\) in the GLM for \([Y^+, W^+]\) and in the GLM for \([Y, W, Z^-]\) coincide.

**Proof.** Let \(U^+(\beta)\) be the vector of dimension \(N_0 + \sum_{j \in J} N_j\) with generic element
\[U^+_n(\beta) = v(W^+_n^T\beta)[Y^+_n - h(W^+_n^T\beta)].\]
Also let \(U^-(\beta, \delta^-)\) be the vector of dimension \(N - (N_0 + \sum_{j \in J} N_j)\) with generic element
\[U^-_n(\beta, \delta^-) = v(W_n^T\beta + Z_n^{-T}\delta^-)[Y_n - h(W_n^T\beta + Z_n^{-T}\delta^-)].\]
The proof of the theorem follows immediately from the proof of Theorem 1 after replacing \(X^0, U^0(\beta), W^*, U^*(\beta, \delta),\) and \(Z^*\) with \(W^+, U^+(\beta), W^-, U^-(\beta, \delta^-),\) and \(Z^{**}\), respectively. \(\square\)

Thus, removing the vector \(\delta^j\) from the grand model corresponds to estimating the model including the complete cases plus the observations in the \(j\)th group (and no auxiliary variables). So \(\delta^j\) controls separately for the corresponding \(j\) groups, which gives much flexibility in selecting the sets of \(\delta\)'s to use in Bayesian model averaging. For example, a possible choice is to consider models which include the set of observations where given sets of (“preferred”) covariates are observed; this corresponds to including several groups at a time, that is removing several \(\delta^j\)'s at a time, which may save computing time considerably.

### 6. Estimation under model uncertainty

Model uncertainty can be handled by either model selection or model averaging. Model selection involves first selecting the best model out of the available set of models and then estimating \(\beta\) conditional on the selected model. A problem with this approach is pre-testing. As shown by Magnus and Durbin (1999) 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 explicit account of the 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 includes $R$ possible candidates and let $\mathcal{M} = \{M_1, \ldots, M_R\}$ denote the set of GLMs being considered. The $r$th model $M_r$ in the model space $\mathcal{M}$ is obtained by including in the linear predictor the $K$ covariates in the completed design matrix $W$ and only a subset of $0 \leq P_r \leq JK$ auxiliary covariates in $Z$,

$$\eta_r = W\beta + Z_r\delta_r, \quad r = 1, \ldots, R,$$

where $Z_r$ is the matrix containing the $N$ observations on the included subset of $P_r$ auxiliary covariates and $\delta_r$ is the corresponding vector of auxiliary parameters. The model averaging estimates of $\beta$ and $\delta$ are of the form

$$\hat{\beta} = \sum_{r=1}^R \lambda_r \hat{\beta}_r, \quad \hat{\delta} = \sum_{r=1}^R \lambda_r S_r \hat{\delta}_r,$$

where the $\lambda_r$ are non-negative weights that add up to one, the $\hat{\beta}_r$ and $\hat{\delta}_r$ are the estimates of $\beta$ and $\delta$ under the $r$th model, and the $S_r$ are $JK \times P_r$ selection matrices defined by $S_r^T = [I_{P_r} : 0]$, or a column-permutation thereof, that transform the conditional estimates of $\delta_r$ in $JK \times 1$ vectors by setting to zero the elements of $\delta$ which are excluded from the $r$th model.

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

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

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

$$p(Y \mid M_r) = \int p(Y \mid \theta_r, M_r) p(\theta_r \mid M_r) d\theta_r$$

is the marginal likelihood of the $r$th model, $\theta_r = (\beta, \delta_r)$ is its vector of parameters, $p(Y \mid \theta_r, M_r)$ is its likelihood, and $p(\theta_r \mid M_r)$ is the prior density of $\theta_r$ under the $r$th model. In this setting, the model averaging estimates in (4) can be interpreted as the posterior means of the distribution of
The posterior variance-covariance matrix consists of the following blocks (Raftery 1993 and Draper 1995)

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

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

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

These posterior variances 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 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 issues that may arise in this estimation method. The main problems are: (i) the specification of the prior probability \(p(M_r)\) 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 approximate 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.

### 6.2. 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 prior over the parameters \(\theta_r\) in the \(r\)th model is the family of calibrated information criteria (CIC) prior distributions introduced by Clyde (2000). This is a family of uninformative prior distributions derived from the following modification of Jeffreys’s prior (Jeffreys 1961)

\[
p(\theta_r \mid M_r) = (2\pi)^{-d_r/2} \left| \frac{1}{c} \mathbf{I}(\hat{\theta}_r) \right|^{1/2},
\]

where \(d_r = K + P_r\) is the number of parameters in the \(r\)th model, \(\mathbf{I}(\hat{\theta}_r)\) is the observed Fisher information for \(M_r\) evaluated at the ML estimate of \(\theta_r\), and \(c\) is a hyperparameter which allows one to calibrate the posterior model probabilities to classical model selection criteria like BIC.
(Bayesian Information Criterion; Schwarz 1996), AIC (Akaike’s Information Criterion; Akaike 1978) and RIC (Risk Inflation Criterion; 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 of twice the log of the Bayes factor for model $M_r$ against the restricted model which excludes all auxiliary covariates. The Clyde’s formulation of the CIC prior is attractive since it provides a general Bayesian justification for the entire family of model selection criteria.

### 6.3. Approximating the posteriors

As for problem (iii), an additional difficulty of BMA in the context of GLMs is that the integrals implicit in the marginal likelihood (6) are hard to compute analytically, except for linear regression models where closed form solutions are sometimes available (see, for instance, Magnus et al. 2010). Thus, some approximation to the marginal likelihood of each model is usually needed. For regular statistical models these approximations can be obtained through 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). On the basis of this approximation, Clyde (2000) shows that the posterior probability of model $M_r$ is

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

where $D_r$ is the deviance of model $M_r$ (-2 times the log-likelihood ratio between the fully restricted model with $\delta = 0$ and model $M_r$). Hence, under CIC priors, the logarithm of the posterior probability of each model is approximately proportional to its deviance minus a complexity penalty term related to 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 viewpoint, CIC priors represent an attractive family of prior distributions for sensitivity analysis in BMA estimation.

### 6.4. Block BMA

Our last issue is how to handle the large number of candidate models that may be included in the model space $\mathcal{M}$. In empirical applications with $J$ distinct subsamples of incomplete data and $K$ available or imputed covariates (including the constant term), the number of possible models to be considered is $R = 2^{JK}$. Hence, unless both $J$ and $K$ are small,
the dimension of the model space is huge and complete enumeration of all possible models is computationally infeasible.

In linear regression models, this issue can be addressed in various ways. One possibility is the weighted-average least squares (WALS) estimator introduced by Magnus et al. (2010), whose computational burden is greatly reduced to the order of $JK$ through preliminary orthogonal transformations of the auxiliary covariates and their parameters. A generalization of WALS for GLMs has been proposed by Heumann and Grenke (2010) using a strategy similar to that originally suggested by Magnus et al. (2011) for linear regression models with nonspherical errors. Unfortunately, this generalization ignores the fact that, in GLMs, the conditional mean and variance of the outcome variable both depend on the linear predictor. Thus, one cannot use a two-step procedure to find the necessary orthogonal transformation of the auxiliary variables in the first step and then obtain the WALS estimates of the parameters of the linear predictor in the second step.

Another possibility is an approximate BMA estimator that considers only a suitable subset of models supported by the data, typically chosen either by stochastic search methods based on Markov Chain Monte Carlo (MCMC) techniques, or by deterministic search methods such as the Occam’s window (Madigan and Raftery 1994) or the leaps and bounds algorithm (Furnival and Wilson 1974).

In the context of our generalized missing-data approach, a third possibility is to confine model uncertainty to the $J$ blocks of $K$ auxiliary variables in $Z$ corresponding to the different missing-data patterns. The great advantage of this BMA procedure is that the number of competing models is reduced to $R = 2^J$. Further, as shown in Theorem 2, the procedure is justified by the fact that the $K$ auxiliary variables in a given block capture the asymptotic bias of the fill-in estimator of $\beta$ due to the imputation of the missing values. Thus, instead of being interested to the separate contribution of the $JK$ auxiliary variables in $Z$, we want to assert which block of auxiliary variables should be jointly included or excluded from the grand model in order to obtain an estimator of $\beta$ with lower asymptotic mean squared error.

### 7. The Multivariate Case

The results of Section 5 extend to multivariate settings where the outcome $Y$ is $Q$-dimensional with a conditional distribution that belongs to the multivariate exponential family. This includes
models for unordered and ordered multinomial outcomes where the outcome of interest can take
$Q+1$ possible values corresponding to $Q+1$ mutually exclusive categories, leading to multinomial,
conditional and ordered logit and probit regressions.

The expression for the density of $Y$ is now

$$f(Y; \gamma) = \exp \left[ \gamma^\top Y - b(\gamma) + c(Y) \right],$$

where $\gamma \in \Theta \subset \mathbb{R}^Q$ is a vector of canonical parameters that may depend on a vector of covariates,
and $b(\cdot) : \Theta \to \mathbb{R}$ and $c(\cdot) : \mathbb{R}^Q \to \mathbb{R}$ are known functions which satisfy the standard regularity
conditions in Fahrmeir and Kaufmann (1985). The mean and variance of $Y$ are $b'(\gamma)$ and $b''(\gamma)$,
the $Q \times 1$ gradient vector and the $Q \times Q$ Hessian matrix of $b(\gamma)$ respectively.

We next define the linear predictor for a given $K$-vector of covariates $X$. For the $q$th component
$Y_q$ of $Y$ the linear predictor is $X^\top \beta_q$, with $\beta_q \in \mathbb{R}^K$. Stacking all the $\beta_q$’s on top of each other
as $\beta = [\beta_1^\top, \ldots, \beta_Q^\top]^\top$ and defining the $QK \times Q$ matrix $X = I_Q \otimes X$, where $I_Q$ is the $Q \times Q$
unit matrix and $\otimes$ is Kronecker’s product, the linear predictor of $Y$ is $X^\top \beta$.

The missing-free data consist of the $Q \times 1$ vectors $Y_n$ and the $K \times 1$ vector of covariates $X_n$,
with $n = 1, \ldots, N$. The 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 $b'(\gamma) = h(X_n^\top \beta)$, with $h : \mathbb{R}^Q \to \mathbb{R}^Q$ and $X_n = I_Q \otimes X_n$. The
full-information ML estimator of $\beta$ is obtained by solving the $QK$ likelihood equations

$$l'(\beta) = \sum_{n=1}^N X_n v(\beta) [Y_n - h(X_n^\top \beta)] = 0,$$

where $v(\beta)$ is the transpose of the $Q \times Q$ matrix $[b''(\gamma_n(\beta))]^{-1} h(X_n^\top \beta)$ and depends on $\beta$
only through the linear predictor $X_n^\top \beta$. Conditions for uniqueness and asymptotic properties of
the ML estimator of $\beta$ are as before (Fahrmeir and Kaufmann 1985).

The property that $\beta$ enters the likelihood equations only through the linear predictor $X_n^\top \beta$, as
in the univariate case, is all we need to adapt the proofs of the equivalence theorem in Section 5
to this case. To show this result, is enough to write the grand model as an augmented GLM
with linear predictor of the form $\eta = W \beta + Z \delta$, with

$$W = \begin{bmatrix} W_1^\top \\ \vdots \\ W_N^\top \end{bmatrix}, \quad Z = \begin{bmatrix} Z_1^\top \\ \vdots \\ Z_N^\top \end{bmatrix},$$
where \( W_n = I_Q \otimes W_n, Z_n = I_Q \otimes Z_n \), and \( \delta = (\delta_1^\top, \ldots, \delta_Q^\top)^\top \) is a \( QJK \times 1 \) vector of nuisance parameters to be estimated jointly with the vector of parameters \( \beta \). Notice that, unlike the univariate case, the grand model now contains \( Q \) different \( JK \times 1 \) vectors of auxiliary parameters. As before, the aim of the generalized missing-indicator approach is to handle the trade-off between asymptotic bias and precision in the estimation of \( \beta \) by considering all intermediate models obtained from the grand model by simultaneously dropping arbitrary subsets of the \( JK \) auxiliary covariates in \( Z \) from the \( Q \) equations for the various components of \( Y \). This corresponds to restrict arbitrary subsets of elements in \( \delta_q \) to be equal to zero for all \( q \). As before, the dimension of the model space to consider in BMA estimation is \( R = 2^{JK} \) with standard BMA and \( R = 2^J \) with block-BMA.

8. **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+ in the participating countries. 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

---

2 Data can be freely downloaded from the SHARE web site: http://www.share-project.org. For information on survey design, target population, country coverage, response rates and other survey methodology issues see Börsch-Supan et al. (2005).
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 the number of limitations with activities of daily living and the number of chronic diseases as self-reported measures of physical health, hand grip strength as an objective measure of physical health, and age, gender, an indicator for educational attainment, per-capita household income and household net worth as socio-economic variables. To ensure cross-country comparability, the information on educational attainment has been recoded using the 1997 International Standard Classification of Education (ISCED-97), and per-capita household income and household net worth have been adjusted for the differences in purchasing power across countries. Summary statistics for the outcomes and the covariates are presented in Table 1, separately for three macro-region: 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 equal to 62 and 64 percent, respectively. The substantial amount of item nonresponse on these variables reflects three problems. First, these variables 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 about 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 by 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 since 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 the MAR assumption holds, the model of interest and the imputation model are likely to be uncongenial as they are based on different sets of explanatory variables. Finally, we produce our own imputations for the missing values on this variable using a simple hot-deck procedure.

Given the high level of cross-country comparability of SHARE, we pool data within each macro-region and estimate a Poisson regression model for verbal fluency and an ordered probit model for numeracy, separately by macro-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. Our aim is to investigate the trade-off between bias and precision when replacing the missing values on hand grip strength, household income and household net word with their imputed values. The Poisson model for verbal fluency includes \( K = 9 \) covariates (a constant term, five available covariates, and three imputed covariates), while the ordered probit model for numeracy includes \( K = 12 \) covariates (four constant terms capturing the thresholds for the five categories of the outcome, five available covariates, and three imputed covariates). The number of missing-data patterns is \( J = 7 \) in all models. Thus, the dimension of the model space is \( R = 2^{63} \) for verbal fluency and \( R = 2^{84} \) for numeracy. In both cases, exact BMA estimation over all possible models is infeasible. To deal with this issue, we rely on our block-BMA procedure which requires to consider only \( R = 2^7 = 128 \) models for each outcome and macro-region.

The estimates for the two models are presented in Tables 2 and 3. For each outcome and macro-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, RIC and BIC 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 of the ML estimators of the selected models which ignore the additional sampling variability induced by the model selection step. For the generalized missing-data approach, standard errors of the BMA estimators have the usual Bayesian interpretation of measuring the spread of the posterior distribution of the parameters of interest given the data.
discussed in Section 6, 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, macro-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. Verbal fluency is higher for women than for men, while numeracy is higher for men than for women. The size of the coefficients and the standard errors are instead subject to non-negligible differences depending on the estimation method. Complete-case and fill-in ML estimates are considerably different in the specifications of all outcomes and macro-regions, and one can easily notice the substantial loss of precision resulting from complete-case analysis. As expected, block-BMA estimates based on BIC and RIC priors are closer to the more parsimonious fill-in model, while block-BMA estimates with AIC are closer to the less parsimonious complete-case model. The differences are particularly striking in the Poisson models for fluency, especially for the Central and the Southern macro-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 raise the issue of validity of the SHARE imputations when studying cognitive functioning. Our analysis suggests that this issue is particularly important for verbal fluency and for countries belonging to the central and the south macro-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.

9. 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. We approach this problem using the generalized missing-indicator method originally applied to linear regression by Dardanoni et al. (2011, 2012). This method is attractive because it suggests handling model uncertainty through a simple BMA approach. In addition to applying the method to the general class of GLMs, we make two extensions. First,
we allow the observed outcome to be multivariate, thus covering the case of seemingly unrelated regression equations (SURE) models and ordered or multinomial logit and probit models. Second, we propose a computationally simple block-BMA strategy that incorporates the information on the available patterns of missing data. An empirical application using the first wave of the Survey on Health, Aging and Retirement in Europe illustrates the practical use of our approach.
References


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

<table>
<thead>
<tr>
<th>Region</th>
<th>Variable</th>
<th>Median</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>fluency</td>
<td>23.0</td>
<td>22.9</td>
<td>7.0</td>
<td>.0</td>
<td>53.0</td>
</tr>
<tr>
<td></td>
<td>numeracy</td>
<td>4.0</td>
<td>3.6</td>
<td>1.0</td>
<td>1.0</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>adl</td>
<td>.0</td>
<td>.1</td>
<td>.5</td>
<td>.0</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>chronic</td>
<td>1.0</td>
<td>1.5</td>
<td>1.4</td>
<td>.0</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>grip strength</td>
<td>35.0</td>
<td>36.6</td>
<td>12.5</td>
<td>4.0</td>
<td>75.0</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>62.0</td>
<td>62.8</td>
<td>8.4</td>
<td>50.0</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>.0</td>
<td>.5</td>
<td>.5</td>
<td>.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>1.0</td>
<td>.6</td>
<td>.5</td>
<td>.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>2.1</td>
<td>2.5</td>
<td>1.8</td>
<td>.0</td>
<td>47.5</td>
</tr>
<tr>
<td></td>
<td>net worth</td>
<td>1.2</td>
<td>2.3</td>
<td>4.9</td>
<td>-20.3</td>
<td>83.6</td>
</tr>
<tr>
<td></td>
<td>Complete obs.</td>
<td>813</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Imputed obs.</td>
<td>4,088</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Center</td>
<td>fluency</td>
<td>20.0</td>
<td>20.3</td>
<td>7.1</td>
<td>.0</td>
<td>90.0</td>
</tr>
<tr>
<td></td>
<td>numeracy</td>
<td>4.0</td>
<td>3.5</td>
<td>1.1</td>
<td>1.0</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>adl</td>
<td>.0</td>
<td>.1</td>
<td>.6</td>
<td>.0</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>chronic</td>
<td>1.0</td>
<td>1.4</td>
<td>1.3</td>
<td>.0</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td>grip strength</td>
<td>35.0</td>
<td>36.1</td>
<td>12.2</td>
<td>1.0</td>
<td>92.0</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>62.0</td>
<td>63.1</td>
<td>8.4</td>
<td>50.0</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>.0</td>
<td>.5</td>
<td>.5</td>
<td>.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>1.0</td>
<td>.6</td>
<td>.5</td>
<td>.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>1.7</td>
<td>2.4</td>
<td>2.6</td>
<td>.0</td>
<td>44.7</td>
</tr>
<tr>
<td></td>
<td>net worth</td>
<td>1.9</td>
<td>3.5</td>
<td>8.9</td>
<td>-20.7</td>
<td>291.1</td>
</tr>
<tr>
<td></td>
<td>Complete obs.</td>
<td>2,575</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Imputed obs.</td>
<td>13,635</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South</td>
<td>fluency</td>
<td>14.0</td>
<td>14.6</td>
<td>5.5</td>
<td>.0</td>
<td>88.0</td>
</tr>
<tr>
<td></td>
<td>numeracy</td>
<td>3.0</td>
<td>3.0</td>
<td>1.1</td>
<td>1.0</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>adl</td>
<td>.0</td>
<td>.2</td>
<td>.6</td>
<td>.0</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>chronic</td>
<td>1.0</td>
<td>1.6</td>
<td>1.5</td>
<td>.0</td>
<td>10.0</td>
</tr>
<tr>
<td></td>
<td>grip strength</td>
<td>30.0</td>
<td>31.7</td>
<td>11.7</td>
<td>3.0</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>63.0</td>
<td>63.3</td>
<td>8.4</td>
<td>50.0</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>.0</td>
<td>.5</td>
<td>.5</td>
<td>.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>.0</td>
<td>.3</td>
<td>.4</td>
<td>.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>.9</td>
<td>1.3</td>
<td>1.4</td>
<td>.0</td>
<td>27.5</td>
</tr>
<tr>
<td></td>
<td>net worth</td>
<td>1.5</td>
<td>3.1</td>
<td>9.5</td>
<td>-22.5</td>
<td>260.9</td>
</tr>
<tr>
<td></td>
<td>Complete obs.</td>
<td>1,754</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Imputed obs.</td>
<td>6,745</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Estimated coefficients and standard errors (in parentheses) of Poisson regression models for fluency by macro-region. Results for the constant term and the auxiliary regressors are omitted to save space.

<table>
<thead>
<tr>
<th>Region</th>
<th>Variable</th>
<th>CC</th>
<th>FI</th>
<th>Block BMA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AIC</td>
<td>RIC</td>
<td>BIC</td>
</tr>
<tr>
<td>North</td>
<td>adl</td>
<td>-0.0625</td>
<td>-0.525</td>
<td>-0.0564</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0147)</td>
<td>(.0072)</td>
<td>(.0134)</td>
</tr>
<tr>
<td></td>
<td>chronic</td>
<td>-0.0117</td>
<td>-0.0162</td>
<td>-0.0115</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0058)</td>
<td>(.0026)</td>
<td>(.0058)</td>
</tr>
<tr>
<td></td>
<td>grip strength</td>
<td>0.0051</td>
<td>0.0039</td>
<td>0.0045</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0010)</td>
<td>(.0004)</td>
<td>(.0011)</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>-0.0055</td>
<td>-0.0049</td>
<td>-0.0056</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0011)</td>
<td>(.0005)</td>
<td>(.0011)</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>-0.0873</td>
<td>-0.0817</td>
<td>-0.0750</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0238)</td>
<td>(.0128)</td>
<td>(.0235)</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>0.0406</td>
<td>0.0794</td>
<td>0.0389</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0159)</td>
<td>(.0070)</td>
<td>(.0160)</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>0.0287</td>
<td>0.0124</td>
<td>0.0286</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0054)</td>
<td>(.0017)</td>
<td>(.0062)</td>
</tr>
<tr>
<td></td>
<td>net worth</td>
<td>0.0061</td>
<td>0.0032</td>
<td>0.0060</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0023)</td>
<td>(.0006)</td>
<td>(.0024)</td>
</tr>
<tr>
<td>Center</td>
<td>adl</td>
<td>-0.0325</td>
<td>-0.561</td>
<td>-0.0400</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0097)</td>
<td>(.0038)</td>
<td>(.0094)</td>
</tr>
<tr>
<td></td>
<td>chronic</td>
<td>-0.0029</td>
<td>-0.0031</td>
<td>-0.0024</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0036)</td>
<td>(.0016)</td>
<td>(.0030)</td>
</tr>
<tr>
<td></td>
<td>grip strength</td>
<td>0.0043</td>
<td>0.0043</td>
<td>0.0044</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0006)</td>
<td>(.0003)</td>
<td>(.0005)</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>-0.0054</td>
<td>-0.0058</td>
<td>-0.0056</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0006)</td>
<td>(.0003)</td>
<td>(.0005)</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>-0.0718</td>
<td>-0.0770</td>
<td>-0.0770</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0134)</td>
<td>(.0058)</td>
<td>(.0116)</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>0.1388</td>
<td>0.1574</td>
<td>0.1438</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0095)</td>
<td>(.0041)</td>
<td>(.0086)</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>0.0092</td>
<td>0.0073</td>
<td>0.0101</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0021)</td>
<td>(.0007)</td>
<td>(.0018)</td>
</tr>
<tr>
<td></td>
<td>net worth</td>
<td>-0.0006</td>
<td>-0.0005</td>
<td>-0.0005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0009)</td>
<td>(.0002)</td>
<td>(.0010)</td>
</tr>
<tr>
<td>South</td>
<td>adl</td>
<td>-0.0567</td>
<td>-0.567</td>
<td>-0.0604</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0131)</td>
<td>(.0060)</td>
<td>(.0122)</td>
</tr>
<tr>
<td></td>
<td>chronic</td>
<td>0.0073</td>
<td>0.128</td>
<td>0.0688</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0049)</td>
<td>(.0024)</td>
<td>(.0048)</td>
</tr>
<tr>
<td></td>
<td>grip strength</td>
<td>0.0063</td>
<td>0.0067</td>
<td>0.0060</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0008)</td>
<td>(.0004)</td>
<td>(.0008)</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>-0.0069</td>
<td>-0.0071</td>
<td>-0.0072</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0009)</td>
<td>(.0005)</td>
<td>(.0009)</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>-0.0488</td>
<td>-0.0559</td>
<td>-0.0441</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0185)</td>
<td>(.0091)</td>
<td>(.0183)</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>0.1651</td>
<td>0.1574</td>
<td>0.1678</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0158)</td>
<td>(.0075)</td>
<td>(.0157)</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>0.0105</td>
<td>0.0059</td>
<td>0.0095</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0058)</td>
<td>(.0021)</td>
<td>(.0058)</td>
</tr>
<tr>
<td></td>
<td>net worth</td>
<td>0.0048</td>
<td>0.0008</td>
<td>0.0049</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0017)</td>
<td>(.0003)</td>
<td>(.0017)</td>
</tr>
</tbody>
</table>
Table 3. Estimated coefficients and standard errors (in parentheses) of ordered probit models for numeracy by macro-region. Results for the thresholds of the outcome and the auxiliary regressors are omitted to save space.

<table>
<thead>
<tr>
<th>Region</th>
<th>Variable</th>
<th>Block BMA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CC</td>
</tr>
<tr>
<td>North</td>
<td>adl</td>
<td>-.1373</td>
</tr>
<tr>
<td></td>
<td>chronic</td>
<td>-.0472</td>
</tr>
<tr>
<td></td>
<td>grip strength</td>
<td>.0143</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>-.0068</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>.1458</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>.1782</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>.1055</td>
</tr>
<tr>
<td></td>
<td>net worth</td>
<td>.0518</td>
</tr>
<tr>
<td></td>
<td>chronic</td>
<td>-.0386</td>
</tr>
<tr>
<td></td>
<td>grip strength</td>
<td>.0138</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>-.0084</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>.1293</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>.5447</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>.0446</td>
</tr>
<tr>
<td></td>
<td>net worth</td>
<td>.0072</td>
</tr>
<tr>
<td>Center</td>
<td>adl</td>
<td>-.0386</td>
</tr>
<tr>
<td></td>
<td>chronic</td>
<td>-.0281</td>
</tr>
<tr>
<td></td>
<td>grip strength</td>
<td>.0138</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>-.0084</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>.1293</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>.5447</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>.0446</td>
</tr>
<tr>
<td></td>
<td>net worth</td>
<td>.0072</td>
</tr>
<tr>
<td>South</td>
<td>adl</td>
<td>-.0179</td>
</tr>
<tr>
<td></td>
<td>chronic</td>
<td>-.0358</td>
</tr>
<tr>
<td></td>
<td>grip strength</td>
<td>.0177</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>-.0152</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>.1730</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>.9448</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>.0248</td>
</tr>
<tr>
<td></td>
<td>net worth</td>
<td>.0007</td>
</tr>
</tbody>
</table>
Appendix A. Asymptotic properties of complete-case and fill-in ML estimators

From Section 5.1, the complete-case ML estimator coincides with the ML estimator of $\beta$ in the grand model with linear predictor $\eta = W\beta + Z\delta$. The ML estimator $\hat{\theta} = (\hat{\beta}, \hat{\delta})$ in this model converges in probability to the true population value $\theta^0 = (\beta^0, \delta^0)$, which solves the equation system

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

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

where $s_\beta$ and $s_\delta$ denote the elements of the score vector corresponding to $\beta$ and $\delta$ respectively.

Further, $\sqrt{N}(\hat{\beta} - \beta^0) \Rightarrow N(0, I_0^{-1})$, where

$$I_0 = \begin{bmatrix} I_{\beta\beta} & I_{\beta\delta} \\ I_{\beta\delta} & I_{\delta\delta} \end{bmatrix}$$

is the Fisher information matrix evaluated at $\theta^0$. Because the asymptotic variance of $\hat{\beta}$ is the top-left block of the inverse of $I_0$, it follows that

$$\sqrt{N}(\hat{\beta} - \beta^0) \Rightarrow N(0, (I_{\beta\beta} - I_{\beta\delta} I_{\delta\delta}^{-1} I_{\beta\delta})^{-1}).$$

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

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

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

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

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

$$\sqrt{N}(\tilde{\beta} - \beta^*) = \left[ -\frac{1}{N} \sum_{n=1}^{N} s_{\beta\beta}(\beta^*, 0; W_n, Z_n) \right]^{-1} \frac{1}{\sqrt{N}} \sum_{n=1}^{N} s_\beta(\beta^*, 0; W_n, Z_n) + o_p(1),$$

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

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

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

$$\text{plim} \frac{1}{N} \sum_{n=1}^{N} s_{\beta\beta}(\beta^*, 0; W_n, Z_n) = H_{\beta\beta}^*,$$

where $H_{\beta\beta}^*$ is a positive definite matrix. Therefore,

$$\sqrt{N}(\tilde{\beta} - \beta^0) \Rightarrow N(\beta^* - \beta^0, (H_{\beta\beta}^*)^{-1} V_{\beta\beta}^* (H_{\beta\beta}^*)^{-1}).$$
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 $\tilde{\beta}$ is consistent and asymptotically more precise than the complete-case ML estimator $\hat{\beta}$, that is

$$AV(\tilde{\beta})^{-1} - AV(\hat{\beta})^{-1} \geq 0.$$  

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

$$AV(\tilde{\beta})^{-1} - AV(\hat{\beta})^{-1} = T_{\beta\delta}^0 T_{\delta\delta}^{-1} T_{\delta\beta},$$

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 $\beta$ in terms of the imputations. To keep the notation simple, we 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 that where $X_1$ is missing, $X_2$ is fully observed, and the missing values of $X_1$ are replaced by the imputations $L_1$.

In this case, the fill-in ML estimator of $\beta$ solves the following system of first-order conditions

$$E[Y - h(\beta^*_0 + \beta^*_1 L_1 + \beta^*_2 X_2)] = 0,$$

$$E[L_1(Y - h(\beta^*_0 + \beta^*_1 L_1 + \beta^*_2 X_2))] = 0,$$

$$E[X_2(Y - h(\beta^*_0 + \beta^*_1 L_1 + \beta^*_2 X_2))] = 0,$$

where the $\beta^*_k = \beta_k + \delta_k$, $k = 0, 1, 2$, are the pseudo-true parameter values, the $\beta_k$ are the true parameter values, and the $\delta_k$ are the asymptotic biases. 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 $E[Y - h(\eta)] = E[X_2(Y - h(\eta))] = 0$, our system can be re-written as

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

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

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

where $m_0 = E[h'(\eta)]$, $m_k = E[h'(\eta)X_k]$ ($k = 1, 2$), and $\tilde{m}_1 = E[h'(\eta)L_1]$ are weighted first order moments, and $\tilde{m}_{1u} = E[L_1(Y - h(\eta))]$, $\tilde{m}_{11} = E[h'(\eta)L^2_1]$, $m_{12} = E[h'(\eta)X_1 X_2]$, $\tilde{m}_{12}$ =
E[h′(η)L′1X′2], m_{22} = E[h′(η)X′2], and \( \tilde{m}_{11} \) = E[h′(η)L′1L1] are weighted uncentered second order moments. Solving this system with respect to the components of the asymptotic bias gives

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

\[
\delta_1 \simeq - [c_{22} \tilde{c}_{1u} - \beta_1 (c_{22} \Delta_{11} - \tilde{c}_{12} \Delta_{12})] / \tilde{\rho}_{12}^2,
\]

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

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

The asymptotic bias of the fill-in ML estimator of \( \beta \) then depends on the asymptotic correlation between the imputations and the generalized residual, and on the differences between weighted first and second order moments for 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 \( \tilde{m}_{1u}, \Delta_{11} \) and \( \Delta_{12} \) converge to zero and \( \tilde{\beta} \) is a consistent estimator of \( \beta \). In general, the fill-in ML estimator of \( \beta \) is consistent if the moments of the joint distribution of \( (Y^j, W^j) \) converge to the moments of the joint distribution of \( (Y^j, X^j), j = 1, \ldots, J \).

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 \( v_u(\eta) \). 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 \( \tilde{c}_{1u} = 0 \). Then the asymptotic biases of the OLS estimators of \( \beta_1 \) and \( \beta_2 \) are given by

\[
\delta_1 = (\Gamma^1 - 1) \beta_1, \quad \delta_2 = \Delta^1 \beta_1,
\]
where
\[
\Gamma^1 = \frac{\tilde{c}_{11} - c_{12}\tilde{c}_{12}}{1 - \tilde{c}_{12}^2}, \quad \Delta^1 = \frac{c_{12} - \tilde{c}_{11}\tilde{c}_{12}}{1 - \tilde{c}_{12}^2},
\]
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. As expected, after imposing the restrictions \( \tilde{c}_{1u} = 0, \tilde{c}_{11} = 1 \) and \( c_{22} = 1 \), the asymptotic bias of the ML estimators of \( \beta_1 \) and \( \beta_2 \) reduces to
\[
\delta_1 = \left( \frac{\tilde{c}_{11} - c_{12}\tilde{c}_{12}}{1 - \tilde{c}_{12}^2} - 1 \right) \beta_1, \quad \delta_2 = \frac{c_{12} - \tilde{c}_{11}\tilde{c}_{12}}{1 - \tilde{c}_{12}^2} \beta_1.
\]