Contents lists available at ScienceDirect

Journal of Econometrics

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

# Regression with imputed covariates: A generalized missing-indicator approach

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

[a] University of Palermo, Italy

[b] Tor Vergata University and EIEF, Italy

ARTICLE INFO

*Article history:*
Received 2 October 2009
Received in revised form
11 December 2010
Accepted 8 February 2011
Available online 25 February 2011

*JEL classification:*
C12
C13
C19

*Keywords:*
Missing covariates
Imputations
Bias-precision trade-off
Model reduction
Model averaging
BMI and income

ABSTRACT

A common problem in applied regression analysis is that covariate values may be missing for some observations but imputed values may be available. This situation generates a trade-off between bias and precision: the complete cases are often disarmingly few, but replacing the missing observations with the imputed values to gain precision may lead to bias. In this paper, we formalize this trade-off by showing that one can augment the regression model with a set of auxiliary variables so as to obtain, under weak assumptions about the imputations, the same unbiased estimator of the parameters of interest as complete-case analysis. Given this augmented model, the bias-precision trade-off may then be tackled by either model reduction procedures or model averaging methods. We illustrate our approach by considering the problem of estimating the relation between income and the body mass index (BMI) using survey data affected by item non-response, where the missing values on the main covariates are filled in by imputations.

## Introduction

A common problem in applied regression analysis is that covariate values may be missing for some observations but imputed values may be available, either values provided by the data-producing agency or directly constructed by the researcher. This problem has received little attention compared to the more general problem of missing covariate values, but is of considerable practical relevance as all empirical researchers know well. In many cases, it is safe to assume that the mechanism leading to missing covariate values does not depend on the outcome of interest. In these cases, one can ignore the missing data mechanism and focus on the problem of what use to make of the available imputations.

There are two main approaches to this problem. One is to simply ignore the imputations and only use the observations with complete data on all covariates—the so-called complete-case analysis. Although this may entail a loss of precision, it has the strong appeal of yielding an unbiased estimator of the parameters of interest when the missing data mechanism is ignorable. The other approach is more concerned with precision and replaces the missing covariate values with the imputations. A refined version of this approach corrects for incorporating the imputed values by some variant of the so-called missing-indicator method (Little, 1992; Horton and Kleinman, 2007; Little and Rubin, 2002), which consists of augmenting the regression model with a set of binary indicators for each covariate with missing values. Although frequently used in practice, this approach is known to produce biased estimates (Jones, 1996; Horton and Kleinman, 2007). It also raises the problem of how to assess precision of the estimators, a problem that we ignore in this paper because it can easily be handled by multiple imputation methods (Rubin, 1987).

Thus, when covariate values are missing we face a trade-off between bias and precision: the complete cases are often disarmingly few, but replacing the missing observations with the imputed values to gain precision may lead to bias. In this paper, we formalize the bias-precision trade-off by showing that one can augment the regression model with a set of auxiliary variables so as to obtain, under weak assumptions about the imputations, the same unbiased estimator of the parameters of interest as complete-case analysis. Given this augmented model, the bias-precision trade-off may then be tackled either by standard model reduction procedures or, more aptly in our view, by model averaging methods.

We illustrate our approach by considering the problem of estimating the relationship between income and the body mass

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

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.02.005

index (BMI) using survey data affected by item non-response, where the missing values on the main covariates are filled in by imputation.

The sequel of the paper is organized as follows. Section 1 presents the basic notation. Section 2 discusses complete-case analysis. Sections 3 and 4 present the augmented model with auxiliary variables and discuss its missing-indicator interpretation. Section 5 contains our main result. Section 6 discusses the trade-off between bias and precision. Section 7 presents our application to modeling the relation between BMI and income. Finally, Section 8 offers some concluding remarks.

## 1. Notation

Observations are indexed by $n = 1, \dots, N$, and covariates by $k = 0, 1, \dots, K - 1$, with $k = 0$ corresponding to the constant term and $K > 1$. We consider the classical linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{1}$$

where $\boldsymbol{y}$ is the $N \times 1$ vector of observations on the outcome of interest, $\boldsymbol{X}$ is an $N \times K$ matrix of observations on the covariates, $\boldsymbol{\beta}$ is the $K \times 1$ vector of coefficients and $\boldsymbol{u}$ is an $N \times 1$ vector of homoskedastic and serially uncorrelated regression errors with zero mean conditional on $\boldsymbol{X}$.

A subsample with incomplete data is a group of observations where one or more covariates are missing. Because the constant term is always observed, the number of possible subsamples with incomplete data is equal to $2^{K-1} - 1$. Not all such subsamples need be present in a data set. In addition to the subsample with complete data (indexed by $j = 0$), we assume to have $J \leq 2^{K-1} - 1$ subsamples with incomplete data, indexed by $j = 1, \dots, J$. This formulation covers both the case when some patterns of missing covariates are not present in the data and the case when the investigator decides to drop from the analysis some groups with incomplete data.

Let $N_j, K_j$ and $K_j^* = K - K_j$, respectively, denote the sample size, the number of available covariates (the covariates with no missing values, including the constant erm), and the number of missing covariates in the $j$th subsample. By construction $\sum_{j=0}^{J} N_j = N$, $K_0 = K$, $K_0^* = 0$ and $1 \leq K_j, K_j^* < K$ for $j = 1, \dots, J$. Let $\boldsymbol{y}^j$, $\boldsymbol{X}_a^j$ and $\boldsymbol{X}_m^j$, respectively, denote the $N_j \times 1$ outcome vector, the $N_j \times K_j$ submatrix containing the values of the available covariates, and the $N_j \times K_j^*$ submatrix containing the values of the missing covariates for the $j$th subsample. Also, let $\boldsymbol{X}^j = [\boldsymbol{X}_a^j, \boldsymbol{X}_m^j]$, an $N_j \times K$ matrix. We assume that $\boldsymbol{X}^0 = \boldsymbol{X}_a^0$ is of full column rank, which implies that $N_0 \geq K$.

## 2. Complete-case analysis

Our benchmark in dealing with missing values is the so-called complete-case method, which uses only the observations with complete data on all covariates.

Let $\boldsymbol{M}$ denote the $N \times K$ missing-data indicator matrix, whose $(n, k)$th element $m_{nk}$ takes value 1 if the $n$th observation contains a missing value on the $k$th covariate and value 0 otherwise. The following assumption is common to most approaches to the problem of missing covariate values and is maintained throughout this paper.

**Assumption 1** (*Ignorability*)**.** $\boldsymbol{M}$ and $\boldsymbol{y}$ are conditionally independent given $\boldsymbol{X}$.

By symmetry of conditional independence, it is easily seen that Assumption 1 is equivalent to the following two assumptions:

$$P(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{M}) = P(\boldsymbol{y} \mid \boldsymbol{X}) \tag{2}$$

and

$$P(\boldsymbol{M} \mid \boldsymbol{y}, \boldsymbol{X}) = P(\boldsymbol{M} \mid \boldsymbol{X}). \tag{3}$$

Assumption (2) basically says that if we knew the true values of the missing covariates, knowing the pattern of missing data would not help in predicting $\boldsymbol{y}$. Assumption (3) implies that the missing data mechanism, seen as a function of $\boldsymbol{y}$ and $\boldsymbol{X}$, depends on $\boldsymbol{X}$ only. Assumption 1 may fail if, for example, observations with missing covariate values have a different regression function than observations with no missing values. On the other hand, it does not place restrictions on how $\boldsymbol{M}$ is generated from $\boldsymbol{X}$. For example, $\boldsymbol{M}$ may exhibit patterns such that cases with low or high levels of some covariates systematically have a greater percentage of missing values.

Theorem 1 provides a formal proof of the fact that, under Assumption 1, the OLS estimator for the complete case is unbiased. This result has been known for long time, but may be considered a "folk theorem". Little (1992) and Little and Rubin (2002) attribute it to an unpublished 1986 technical report by William Glynn and Nan Laird. Private communication with Nan Laird however informs us that the report has never been published and is no longer available. Jones (1996) offers a proof for the case of two covariates, one of which has missing values, whereas Wooldridge (2002, p. 553) shows that the two-stage least-squares estimator for the complete case is consistent.

**Theorem 1** (*Complete-Case Estimation*)**.** *If Assumption 1 holds, then the OLS estimator of $\boldsymbol{\beta}$ obtained by using only the observations with complete data on all covariates is unbiased for $\beta$.*

**Proof.** The OLS estimator for the complete data may be written as follows:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{y},$$

where $\boldsymbol{D}$ is an $N \times N$ diagonal matrix whose $n$th diagonal element $d_n$ takes value 1 if no covariate is missing for the $n$th observation and value 0 otherwise. The elements of $\boldsymbol{D}$ are related to the elements of the missing-indicator matrix $\boldsymbol{M}$ through $d_n = \prod_{k=1}^{K}(1 - m_{nk})$. The Ignorability assumption implies that any function of $\boldsymbol{M}$, in particular $\boldsymbol{D}$, is independent of $\boldsymbol{y}$ conditional on $\boldsymbol{X}$. From (2),

$$\mathbb{E}(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}, \boldsymbol{D}) = (\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

and therefore $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. □

An implication of Theorem 1 is that the subsample with complete data satisfies

$$\boldsymbol{y}^0 = \boldsymbol{X}^0\boldsymbol{\beta} + \boldsymbol{u}^0, \tag{4}$$

where $\boldsymbol{u}^0$ is an $N_0 \times 1$ vector of homoskedastic and serially uncorrelated regression errors. This result supports the common practice of complete-case analysis, namely estimating $\boldsymbol{\beta}$ by regressing $\boldsymbol{y}^0$ on $\boldsymbol{X}^0$. However, severe loss of information, and hence of precision, may result unless the fraction of deleted cases is small.

## 3. The augmented model with auxiliary variables

Suppose that, for each subsample $j = 1, \dots, J$ with incomplete data, the values of the $K_j^*$ missing covariates are filled-in using some imputation procedure. A covariate with imputed values is called an imputed covariate. The $N_j \times K_j^*$ matrix corresponding to the set of imputed covariates is called the imputation matrix for the $j$th subsample and is denoted by $\boldsymbol{L}^j$. The $N_j \times K$ matrix $\boldsymbol{W}^j = [\boldsymbol{X}_a^j, \boldsymbol{L}^j]$, whose columns correspond to the $K_j$ available covariates and the $K_j^*$ imputed covariates, is called the completed design matrix for the $j$th subsample. Our treatment of imputation is very general and covers a variety of imputation procedures, including regression and donor-based methods such as nearest-neighbor and hot-deck imputations. It also allows for the possibility that different imputation procedures are used for different covariates, or for different subsamples with incomplete data.

Consider modeling the $N_j \times 1$ outcome vector $\boldsymbol{y}^j$ for the $j$th subsample as a linear function of the observed covariates in $\boldsymbol{W}^j$. The best (minimum mean-square error) linear predictor of $\boldsymbol{y}^j$ given $\boldsymbol{W}^j = [\boldsymbol{X}_a^j, \boldsymbol{L}^j]$ is

$$\begin{aligned}\mathbb{E}^*(\boldsymbol{y}^j \mid \boldsymbol{X}_a^j, \boldsymbol{L}^j) &= \mathbb{E}^*(\boldsymbol{X}^j\boldsymbol{\beta} \mid \boldsymbol{X}_a^j, \boldsymbol{L}^j) \\ &= \boldsymbol{X}_a^j\boldsymbol{\beta}_a^j + \mathbb{E}^*(\boldsymbol{X}_m^j \mid \boldsymbol{X}_a^j, \boldsymbol{L}^j)\boldsymbol{\beta}_m^j \\ &= \boldsymbol{X}_a^j\boldsymbol{\beta}_a^j + (\boldsymbol{X}_a^j\Delta^j + \boldsymbol{L}^j\Gamma^j)\boldsymbol{\beta}_m^j \\ &= \boldsymbol{X}_a^j\boldsymbol{\gamma}_a^j + \boldsymbol{L}^j\boldsymbol{\gamma}_m^j,\end{aligned}$$

where $\boldsymbol{\beta}_a^j$ and $\boldsymbol{\beta}_m^j$ are the subvectors of $\boldsymbol{\beta}$ associated with $\boldsymbol{X}_a^j$ and $\boldsymbol{X}_m^j$, respectively, $\mathbb{E}^*(\boldsymbol{X}_m^j \mid \boldsymbol{X}_a^j, \boldsymbol{L}^j) = \boldsymbol{X}_a^j\Delta^j + \boldsymbol{L}^j\Gamma^j$ is the best linear predictor of $\boldsymbol{X}_m^j$ given $\boldsymbol{X}_a^j$ and $\boldsymbol{L}^j$, and

$$\boldsymbol{\gamma}_a^j = \boldsymbol{\beta}_a^j + \Delta^j\boldsymbol{\beta}_m^j, \qquad \boldsymbol{\gamma}_m^j = \Gamma^j\boldsymbol{\beta}_m^j.$$

The resulting linear model for the $j$th subsample may be written, more compactly,

$$\boldsymbol{y}^j = \boldsymbol{W}^j\boldsymbol{\gamma}^j + \boldsymbol{u}^j, \quad j = 1, \dots, J, \tag{5}$$

where $\boldsymbol{\gamma}^j$ is the $K \times 1$ vector consisting of the coefficients associated with the observed and the imputed covariates in $\boldsymbol{W}^j = [\boldsymbol{X}_a^j, \boldsymbol{L}^j]$, and $\boldsymbol{u}^j$ is an $N_j \times 1$ vector of projection errors that, by construction, have mean zero and are uncorrelated with $\boldsymbol{W}^j$.

Two important features distinguish model (5) from the original model (1). First, the vector of population coefficients $\boldsymbol{\gamma}^j$ is generally different from $\boldsymbol{\beta}$ unless $\Delta^j = 0$ and $\Gamma^j$ is equal to the identity matrix or, equivalently, $\mathbb{E}^*(\boldsymbol{X}_m^j \mid \boldsymbol{X}_a^j, \boldsymbol{L}^j) = \boldsymbol{L}^j$, that is, given the imputations, the available covariates contain no further information about the missing covariates. Second, the elements of the error vector $\boldsymbol{u}^j$ are not necessarily homoskedastic, even when homoskedasticity holds for the elements of $\boldsymbol{u}$.

Letting $\delta^j = \boldsymbol{\gamma}^j - \boldsymbol{\beta}, j = 1, \dots, J$, and stacking on top of each other the complete-case model and the $J$ linear models for the subsamples with incomplete data give

$$\begin{bmatrix}\boldsymbol{y}^0 \\ \boldsymbol{y}^*\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}^0 \\ \boldsymbol{W}^*\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{0} \\ \boldsymbol{Z}^*\end{bmatrix}\delta + \begin{bmatrix}\boldsymbol{u}^0 \\ \boldsymbol{u}^*\end{bmatrix},$$

where

$$\boldsymbol{y}^* = \begin{bmatrix}\boldsymbol{y}^1 \\ \vdots \\ \boldsymbol{y}^J\end{bmatrix}, \qquad \boldsymbol{W}^* = \begin{bmatrix}\boldsymbol{W}^1 \\ \vdots \\ \boldsymbol{W}^J\end{bmatrix}, \qquad \boldsymbol{Z}^* = \begin{bmatrix}\boldsymbol{W}^1 & & \\ & \ddots & \\ & & \boldsymbol{W}^J\end{bmatrix},$$

$$\boldsymbol{u}^* = \begin{bmatrix}\boldsymbol{u}^1 \\ \vdots \\ \boldsymbol{u}^J\end{bmatrix},$$

and $\boldsymbol{\delta}$ is the $JK \times 1$ vector consisting of $\delta^1, \dots, \delta^J$. We can now write the model for the available and the imputed data as the grand model:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{Z}\delta + \boldsymbol{u}, \tag{6}$$

where $\boldsymbol{\beta}$ is the parameter of primary interest, $\boldsymbol{\delta}$ is a vector of nuisance parameters, and

$$\boldsymbol{y} = \begin{bmatrix}\boldsymbol{y}^0 \\ \boldsymbol{y}^*\end{bmatrix}, \qquad \boldsymbol{W} = \begin{bmatrix}\boldsymbol{X}^0 \\ \boldsymbol{W}^*\end{bmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix}\boldsymbol{0} \\ \boldsymbol{Z}^*\end{bmatrix}, \qquad \boldsymbol{u} = \begin{bmatrix}\boldsymbol{u}^0 \\ \boldsymbol{u}^*\end{bmatrix},$$

Respectively, an $N$-vector, an $N \times K$ matrix, an $N \times JK$ matrix, and an $N$-vector. Note that the matrix $\boldsymbol{W}$ is obtained by filling-in the missing covariate values with the available imputations. Model (6) includes all observations re-ordered groupwise: first, the complete cases, and then the first group with incomplete data, etc. Ordering of the groups is arbitrary and plays no role in the analysis. In the terminology of Danilov and Magnus (2004), the $K$ columns of $\boldsymbol{W}$ are the "focus" regressors, while the $JK$ columns of $\boldsymbol{Z}$ are the "auxiliary" regressors.

## 4. A missing-indicator interpretation

Before presenting our main result it is instructive to give a missing-indicator interpretation of model (6). Indeed, the $JK$ auxiliary variables in the matrix $\boldsymbol{Z}$ are obtained by multiplying the covariates in each group by the various indicators of group membership. To see this write $\boldsymbol{Z} = [\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_j, \dots, \boldsymbol{Z}_J]$, where $\boldsymbol{Z}_j$ is the $N \times K$ matrix that contains the auxiliary variables for the $j$th group. Let $\boldsymbol{1}_K$ denote the $1 \times K$ vector whose elements are all equal to one and let $\boldsymbol{d}_j$ denote the $N \times 1$ vector of group-membership indicators for the $j$th group (the elements of $\boldsymbol{d}_j$ are equal to one for observations in group $j$ and zero otherwise). Then

$$\boldsymbol{Z}_j = [\boldsymbol{1}_K \otimes \boldsymbol{d}_j] \cdot \boldsymbol{W}, \qquad j = 1, \dots, J,$$

where $\otimes$ denotes the Kronecker product and $\cdot$ the Hadamard (elementwise) product.

As an illustration, consider the linear model

$$\mathbb{E}(y_n \mid x_{n1}, x_{n2}) = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2},$$

with a constant term and two covariates, $x_1$ and $x_2$. Suppose that, in addition to the group with complete data, one has two groups with incomplete data: in group 1 [resp. 2] only the first [resp. second] covariate is missing. If $d_{n0}$, $d_{n1}$ and $d_{n2}$ denote the group-membership indicators, and $L_{n1}^1$ and $L_{n2}^2$ denote the imputed values in each group with incomplete data, then we may write

$$\begin{aligned}y_n = {} & d_{n0}(\beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2}) + d_{n1}(\gamma_0^1 + \gamma_1^1 L_{n1}^1 + \gamma_2^1 x_{n2}) \\ & + d_{n2}(\gamma_0^2 + \gamma_1^2 x_{n1} + \gamma_2^2 L_{n2}^2) + u_n.\end{aligned}$$

Let $w_{nk}$ be equal to $x_{nk}$ if the $k$th covariate is observed for the $n$th observation and to its imputed value otherwise. Then, the last relation may be written as

$$\begin{aligned}y_n = {} & \beta_0 + \beta_1 w_{n1} + \beta_2 w_{n2} + \delta_0^1 d_{n1} + \delta_0^2 d_{n2} \\ & + \delta_1^1 d_{n1} w_{n1} + \delta_1^2 d_{n2} w_{n1} + \delta_2^1 d_{n1} w_{n2} + \delta_2^2 d_{n2} w_{n2} + u_n,\end{aligned}$$

where $\delta_k^j = \gamma_k^j - \beta_k$. This is exactly model (6) for this special case, where the auxiliary variables added to the $w_k$'s are the group-membership indicators and their interactions with the constant term and the observed or imputed covariates.

## 5. Main result

The following result shows that, no matter which imputation procedure is chosen, the OLS estimate of $\boldsymbol{\beta}$ in the grand model (6) and that in the complete-case model (4) are numerically the same. Thus, the statistical properties of the two estimators are also the same. In particular, if the latter is unbiased (for example, the conditions of Theorem 1 hold), so is the former.

**Theorem 2.** *Suppose that the matrix $\boldsymbol{W}$ is of full column rank $K$ and that $N \geq K(J + 1)$. Then, for any choice of imputation matrices $\boldsymbol{L}^1, \dots, \boldsymbol{L}^J$, the OLS estimate of $\boldsymbol{\beta}$ in the "grand' model" (6) coincides with the OLS estimate of $\boldsymbol{\beta}$ in the complete-case model (4).*

**Proof.** Given any matrix $\boldsymbol{A}$, let $\boldsymbol{R_A} = \boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-}\boldsymbol{A}'$, where $(\boldsymbol{A}'\boldsymbol{A})^{-}$ denotes a $g$-inverse of $\boldsymbol{A}'\boldsymbol{A}$. Since $\boldsymbol{Z}'\boldsymbol{Z}$ and $\boldsymbol{Z}$ have the same rank, the rank of $\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-}\boldsymbol{Z}'$ is equal to the rank of $\boldsymbol{Z}$ (Rao and Mitra, 1971). The fact that the rank of $\boldsymbol{Z}$ may be less than $JK$ implies that the rank of $\boldsymbol{R_Z}$ must be at least $N - JK$. Because $N \geq K(J+1)$ implies that $K \leq N - JK$, it follows that the rank of $\boldsymbol{W}$ cannot exceed the rank of $\boldsymbol{R_Z}$, so the matrix $\boldsymbol{W}'\boldsymbol{R_Z}\boldsymbol{W}$ must be nonsingular. Thus, by the Frisch-Waugh-Lovell (Partitioned Regression) Theorem, the OLS estimate of $\boldsymbol{\beta}$ in model (6) is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{W}'\boldsymbol{R_Z}\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{R_Z}\boldsymbol{y} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{y},$$

where $\tilde{\boldsymbol{X}} = \boldsymbol{R_Z}\boldsymbol{W}$. Next notice that

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{I}_{N_0} & & & \\ & \boldsymbol{R}_{\boldsymbol{W}^1} & & \\ & & \ddots & \\ & & & \boldsymbol{R}_{\boldsymbol{W}^J} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^0 \\ \boldsymbol{W}^1 \\ \vdots \\ \boldsymbol{W}^J \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^0 \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix},$$

where we used the fact that $\boldsymbol{W}^j(\boldsymbol{W}^{j\prime}\boldsymbol{W}^j)^{-}\boldsymbol{W}^{j\prime}\boldsymbol{W}^j = \boldsymbol{W}^j$ for all $j$ and any choice of $g$-inverse (Rao and Mitra, 1971). Therefore, $\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} = \boldsymbol{X}^{0\prime}\boldsymbol{X}^0$ and $\tilde{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}^{0\prime}\boldsymbol{y}^0$. Hence

$$\widehat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{y} = (\boldsymbol{X}^{0\prime}\boldsymbol{X}^0)^{-1}\boldsymbol{X}^{0\prime}\boldsymbol{y}^0,$$

which is the complete-case estimate. □

The matrix $\boldsymbol{W}$ is of full column rank if, as we already assumed, the $K$ columns of $\boldsymbol{X}^0$ are linearly independent. The use of a $g$-inverse in the proof of the theorem is necessary because some of the completed design matrices $\boldsymbol{W}^j$ may be singular, which happens if $N_j < K$ or if $N_j \geq K$ but the columns of $\boldsymbol{W}^j$ are linearly dependent. The latter is for example the case when a missing covariate value is replaced by its average value for the available cases (mean imputation) or by its predicted values based on the observed covariates $\boldsymbol{X}_a^j$ and the coefficients from an OLS regression using the subsample with complete data (deterministic regression imputation). One can replace a $g$-inverse with the regular inverse when the $J$ subsamples with incomplete data are such that all $\boldsymbol{W}^j$'s have full column rank. In practice, this may be achieved by dropping groups that contain too few observations and avoiding mean imputation or deterministic regression imputation. After all, these two imputation methods are known to produce completed data sets with undesirable properties, for example they have less variability than a set of truly observed values (see e.g. Lundström and Särndal, 2002).

The complete-case model (4) and the grand model (6) may at first appear as two polar approaches to the problem of handling missing data in model (1). At one extreme is complete-case analysis. Under the assumption of Theorem 1, this gives an unbiased estimate of $\boldsymbol{\beta}$ but may throw away too much information by retaining only the observations in the subsample with complete data. At the other extreme, all observations are retained but some imputation procedure is adopted to fill-in the missing data. In fact, Theorem 2 shows that if $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ in (6) are left unconstrained then this second approach is equivalent to complete-case analysis as far as estimation of $\boldsymbol{\beta}$ is concerned.

A referee offered the following heuristic. Our model places no restrictions (equivalently, uses no information) on the imputation method. In the decomposition $\boldsymbol{\gamma}^j = \boldsymbol{\beta} + \boldsymbol{\delta}^j$, it is only the observations from the complete case that sort out the part that should be $\boldsymbol{\beta}$. Since the remaining cases provide absolutely no information, the estimates are the same. "No information added, no change".

The standard practice of regressing $\boldsymbol{y}$ only on the completed design matrix $\boldsymbol{W}$ omitting the variable in $\boldsymbol{Z}$ corresponds to using a restricted version of the grand model (6) where all elements of the vector $\boldsymbol{\delta}$ are set equal to zero. This is the same as assuming that the missing data mechanism satisfies Assumption 1 and the imputation procedure is such that $\boldsymbol{\beta} = \boldsymbol{\gamma}^j$ for each $j = 1, \ldots, J$. The less frequent practice of regressing $\boldsymbol{y}$ on $\boldsymbol{W}$ and the set of group-membership indicators (which we shall refer to as the simple missing-indicator method) corresponds to using another restricted version of the grand model, where all interactions between the group membership indicators and the observed or derived covariates are set equal to zero.

Both sets of restrictions are testable. Testing the first set of restrictions corresponds to testing the hypothesis that all regression coefficients do not change across the $J$ groups containing missing covariates, while testing the second set of restrictions corresponds to testing the hypothesis that, except for the intercepts, all other regression coefficients do not change across the $J$ groups containing missing covariates.

The precise nature of these tests, in particular the form of the test statistics, depends on the properties of the error vector $\boldsymbol{u}$ in model (6). Given OLS estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$ of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ in the grand model, classical $F$-tests would be appropriate if it can safely be assumed that $\boldsymbol{u}$ is a vector of homoskedastic and serially uncorrelated regression errors. If this assumption cannot be justified, then one could use a "robustified" version of these tests based on an estimator of the sampling variance of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$ that is consistent under heteroskedasticity or autocorrelation of unknown form in the elements of $\boldsymbol{u}$.

In our view, however, the key issue is not what statistic to use for testing, but whether it makes sense to ask questions such as: Is it true that $\boldsymbol{\delta} = 0$? Following Leamer (1978) and Magnus and Durbin (1999), we think that asking such questions in this context is wrong. The right question is: What is the best available estimator of $\boldsymbol{\beta}$?

## 6. Bias versus precision

Theorem 2 says that unbiased estimates of $\boldsymbol{\beta}$ may be obtained in two equivalent ways, either by using the $N_0$ observations in the subsample with complete data, or by using all $N$ observations and the grand model (6) which includes the imputed values of the missing covariates in the matrix $\boldsymbol{W}$ and the auxiliary variables in the matrix $\boldsymbol{Z}$. We also know from standard results that placing restrictions on the elements of $\boldsymbol{\delta}$ may lead to biased but more precise estimates of $\boldsymbol{\beta}$.

Two approaches may be followed to handle this trade-off between bias and precision in the estimation of $\boldsymbol{\beta}$: model reduction and model averaging. Either approach can be applied to model (6).

Model reduction involves first selecting an intermediate model between the grand model and the fully restricted model where $\boldsymbol{\delta} = 0$, and then estimating the parameter of interest $\boldsymbol{\beta}$ conditional on the selected model. Model reduction may be carried out through variable selection methods, such as stepwise regression (see e.g. Kennedy and Bancroft, 1971), or more complex general-to-specific procedures (see Campos et al., 2005, for a survey). The details of the model reduction procedure may also depend on whether one allows dropping arbitrary subsets of auxiliary variables in $\boldsymbol{Z}$, or only subsets of auxiliary variables corresponding to specific subsamples with missing covariates. Dropping one of the columns of $\boldsymbol{Z}$ amounts to selecting a group $j$ and, in the corresponding equation (5), restricting one element of $\boldsymbol{\delta}^j$ to zero. This in turn corresponds to forcing the coefficient of that particular covariate in the completed design matrix $\boldsymbol{W}^j$ to be the same as in the subsample with complete data. Dropping the columns of $\boldsymbol{Z}$ corresponding to the $j$th subsample amounts instead to restricting all element of $\boldsymbol{\delta}^j$ to be zero, which in turn corresponds to forcing the relationship between $\boldsymbol{y}^j$ and the completed design matrix $\boldsymbol{W}^j$ to be the same as that between $\boldsymbol{y}^0$ and $\boldsymbol{X}^0$ in the subsample with complete data.

One well-known problem with this approach is pre-testing. A second problem is that model selection and estimation are completely separated. As a result, the reported conditional estimates tend to be interpreted as if they were unconditional. A third problem is that, since there are $J$ subsamples with incomplete data and $K$ covariates (including the constant term), the model space may contain up to $2^{JK}$ models. Thus, the model space is huge, unless both $J$ and $K$ are small.

Model averaging is different. Instead of selecting a model out of the available set of models, one first estimates the parameter of interest $\boldsymbol{\beta}$ conditional on each model in the model space, and then computes the estimator of $\boldsymbol{\beta}$ as a weighted average of these conditional estimators. When the model space contains $I$ models, a model averaging estimator of $\boldsymbol{\beta}$ is of the form

$$\bar{\boldsymbol{\beta}} = \sum_{i=1}^{I} \lambda_i \widehat{\boldsymbol{\beta}}_i, \tag{7}$$

where the $\lambda_i$ are non-negative weights that add up to one, and $\widehat{\boldsymbol{\beta}}_i$ is the estimator of $\boldsymbol{\beta}$ obtained by conditioning on the $i$th model.

In Bayesian model averaging (BMA), each $\widehat{\boldsymbol{\beta}}_i$ is weighted by the posterior probability of the corresponding model. If equal prior probabilities are assigned to each model under consideration, then the $\lambda_i$ are just proportional to the marginal likelihood of each model. Bayesian averaging of both classical (least-squares) and Bayesian estimators have been considered, with the posterior mean of $\boldsymbol{\beta}$ for the model under consideration as the typical Bayesian estimator. Bayesian averaging of Bayesian estimators has been popularized by Raftery et al. (1997), while Bayesian averaging of classical estimators has been popularized by Sala-i-Martin et al. (2004). The choice between the different approaches involves considering the computational burden and the statistical properties of the resulting estimators and, in the case of BMA, the nature of the assumed priors. The role of priors would also arise if a Bayesian model reduction approach is taken.

Magnus et al. (forthcoming) study the properties of model averaging estimators of the same form as (7) with $\lambda_i = \lambda_i(\widetilde{\boldsymbol{u}})$, where $\widetilde{\boldsymbol{u}}$ is the vector of OLS residuals from the regression of $\boldsymbol{y}$ on $\boldsymbol{W}$ only. Their class of weighted-average least-squares (WALS) estimators generalizes to the case when $I \geq 2$ the class of estimators introduced by Magnus and Durbin (1999), which contains the classical pre-test estimator as a special case. Although WALS estimators are in fact BMA estimators, they differ from standard BMA in three important respects: their computational burden, the choice of prior for $\boldsymbol{\delta}$, and their statistical properties.

The main advantage of WALS is that, although we may have up to $I = 2^{JK}$ models, the computational burden is only proportional to $JK$. With medium or large values of $J$ or $K$, the computation burden is minimal compared to standard BMA.

Like standard BMA, WALS assume a classical Gaussian linear model for (6) and noninformative priors for $\boldsymbol{\beta}$ and the error variance $\sigma^2$. The assumption that the regression errors are homoskedastic and serially uncorrelated is not crucial for WALS, and the method can be generalized to non-spherical errors (Magnus et al., forthcoming). The key step in WALS is to reparameterize the model replacing $\boldsymbol{Z\delta}$ by $\boldsymbol{Z}^*\boldsymbol{\delta}^*$, with $\boldsymbol{Z}^* = \boldsymbol{ZP\Lambda}^{-1/2}$ and $\boldsymbol{\delta}^* = \boldsymbol{\Lambda}^{1/2}\boldsymbol{P}'\boldsymbol{\delta}$, where $\boldsymbol{P}$ is an orthonormal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix such that $\boldsymbol{P}'\boldsymbol{Z}'\boldsymbol{R_W ZP} = \boldsymbol{\Lambda}$. The main difference with respect to standard BMA is that, instead of a multivariate Gaussian prior for $\boldsymbol{\delta}$, WALS use a Laplace distribution with zero mean for the independently and identically distributed elements of the transformed parameter vector $\boldsymbol{\eta} = \boldsymbol{\delta}^*/\sigma$, whose $i$th element, $\eta_i$ is the population $t$-ratio on $\delta_i$, the $i$th element of $\boldsymbol{\delta}$. In this formulation, ignorance is a situation where it is equally likely for these population $t$-ratios to be larger or smaller than one in absolute value.

Finally, unlike standard BMA, WALS have bounded risk and are near-optimal in terms of a well-defined regret criterion (Magnus et al., 2010).

## 7. An application

In this section, we apply our approach in the context of a concrete example with missing data. The problem at hand is that of estimating the relation between body-mass and income using survey data affected by item non-response. We first present the estimates one obtains by the complete-case approach, by using raw data (no dummies), and by the simple indicator method. We then compare them with the estimates one obtains using different model-selection or model-averaging techniques on the basis of the grand model (6).

The BMI, namely the ratio of weight (in kg) to squared height (in meters), is one way of combining weight and height into a single measure. Due to its ease of measurement and calculation, the BMI is the most common diagnostic tool to identify obesity problems within a population. As such, it has received lots of attention in the recent literature on the obesity epidemic and its economic and public health consequences (Cutler et al., 2003; Philipson and Posner, 2008).

The obesity epidemic is essentially an imbalance between food intake and energy expenditure. It has been argued that this imbalance may be linked to income (see e.g. Drewnowski and Specter, 2004). The available empirical evidence – Cawley et al. (2008) for elderly people in the USA and Sanz-de-Galdeano (2005), and García Villar and Quintana-Domeque (2009) for Europe – is inconclusive for men, whereas for women there appears to be a more clear indication of a negative correlation between BMI and income.

Our data are from Release 2 of the first wave of the Survey of Health, Ageing, and Retirement in Europe (SHARE), a multidisciplinary and cross-national household panel survey designed to investigate several aspects of the elderly population in Europe. The target population of SHARE consists of people aged above 50 living in residential households, plus their co-resident partners irrespective of age. The first wave, conducted in 2004, covered 15,544 households and 22,431 individuals in 11 European countries. All national samples are selected through probability sampling.

The key to ensure comparability is the adoption of a common survey instrument. The physical health module of the questionnaire collects self-reported height and weight, the income module collects information on 25 income components, which are then aggregated into a measure of household income, and the consumption module collects household expenditure on four consumption categories (food at home, food outside the home, telephone, and all goods and services) in the last month.

Non-response to household income and food expenditure is substantial, and in this case we use the imputations provided by SHARE. Complete or partial non-response to household income occurs for as much as 60 percent of the observations, such a high fraction being due to the fact that this variable is obtained by aggregating a large number of income components across household members. Non-response to food expenditure occurs for about 15 percent of the observations.

To impute missing values, SHARE uses a complex two-stage multivariate procedure (Kalwij and van Soest, 2005). Imputations are first obtained recursively for a few core variables. In the second stage, the imputed values from the first stage are used to impute the other variables. This procedure essentially employs only univariate regression imputation methods. It is important to note that height and weight are never used to impute missing variables. To allow multiple imputation methods, SHARE provides five imputations for each missing value. SHARE imputes total household income by separately imputing each income component and then aggregating them. Imputations are provided for individual incomes of all eligible partners who did not agree to participate to the survey.

We focus on the income-BMI relationship for males. We model the mean of log BMI as a function of age and age squared, log

**Table 1**
Estimated coefficients.

| | Complete case | Fully restricted | Simple indicator | Stata's stepwise | WALS | BMA |
|---|---|---|---|---|---|---|
| age | 0.0008 | 0.0023$^{***}$ | 0.0023$^{***}$ | 0.0020$^{***}$ | 0.0012$^{*}$ | 0.0023$^{***}$ |
| | (0.0008) | (0.0004) | (0.0004) | (0.0005) | (0.0007) | (0.0004) |
| agesq | −0.0082$^{***}$ | −0.0107$^{***}$ | −0.0107$^{***}$ | −0.0108$^{***}$ | −0.0087$^{***}$ | −0.0107$^{***}$ |
| | (0.0020) | (0.0012) | (0.0012) | (0.0012) | (0.0018) | (0.0012) |
| lowed | 0.0144$^{***}$ | 0.0201$^{***}$ | 0.0201 $^{***}$ | 0.0197$^{***}$ | 0.0160$^{***}$ | 0.0201$^{***}$ |
| | (0.0047) | (0.0027) | (0.0027) | (0.0027) | (0.0041) | (0.0027) |
| lypc | −0.0196$^{***}$ | −0.0097$^{***}$ | −0.0098$^{***}$ | −0.0174$^{***}$ | −0.0165$^{***}$ | −0.0101$^{***}$ |
| | (0.0030) | (0.0016) | (0.0016) | (0.0026) | (0.0027) | (0.0020) |
| lfpc | 0.0054 | 0.0042 | 0.0044$^{*}$ | 0.0049$^{*}$ | 0.0046 | 0.0042 |
| | (0.0045) | (0.0026) | (0.0026) | (0.0026) | (0.0039) | (0.0026) |
| *N* | 4067 | 11,475 | 11,475 | 11,475 | 11,475 | 11,475 |

Note: Observed $p$-values: $^{*}$ $p < 0.10$; $^{**}$ $p < 0.05$; $^{***}$ $p < 0.01$.

household income per capita, log household food expenditure per capita, and a dummy indicator for low educational attainment. In addition to the subsample with complete data (4067 obs., 35.5%), we have three subsamples with incomplete data: one where only food expenditure is missing (287 obs., 2.5%), one where only household income is missing (5891 obs., 51.3%), and one where both household income and food expenditure are missing (1230 obs., 10.7%). For each variable, we use the first of the five available imputations.

Table 1 shows the estimated coefficients for age and its square (`agesq`), log household income per capita (`lypc`), log food expenditure per capita (`lfpc`), and a dummy for not having a high-school degree (`lowed`). The first three columns contain estimates for the complete-case/grand model, the fully restricted estimator corresponding to $\boldsymbol{\delta} = 0$, and the simple missing-indicator method. The other three are obtained by model-selection or averaging on the basis of the grand model: Stata's `stepwise` procedure with $p$-value equal to.05, WALS and BMA. Estimates of the coefficients for the 18 auxiliary regressors are not presented but are available upon request. For simplicity, all estimates are based on the assumption of spherically distributed errors in the grand model (6).

The BMA and WALS estimates and their standard errors have been computed using the Matlab code downloaded from Jan Magnus's web page at http://center.uvt.nl/staff/magnus/wals/. This BMA implementation estimates all possible models, so it becomes very time consuming when $J$ or $K$ are large. In our case, with $J = 3$ subsamples with incomplete data and $K = 6$ focus regressors (including the constant term), examining all possible $2^{18}$ models required about one day on our desktop computer. Faster implementations are available, but they estimate only a randomly chosen subset of all possible models and have the important disadvantage of not using the distinction between focus and auxiliary regressors, which is key to our analysis.

As for WALS, it is worth discussing briefly the concept and treatment of uncertainty implicit in the choice of a Laplace prior for the elements of the transformed parameter vector $\boldsymbol{\eta}$. Assuming this particular prior means that we think that it is equally likely that the observed value of the $t$-statistic on any element of $\boldsymbol{\delta}$ is greater or smaller than one. This is equivalent to say that we are agnostic about the quality of the imputation: it could be either good or bad. Since we are simply users, not producers, of the imputations, this may not be a bad assumption.

There is agreement between the different methods on the qualitative effect of the various variables: concave for age, negative for education and income, and positive for food expenditure. The magnitude of the estimated coefficients, however, differs considerably across methods. At one extreme are the fully restricted estimator, the simple missing-indicator method and BMA that produce nearly identical results: they assign more importance to age and less importance to income. At the other extreme are the complete-case estimator and WALS: they assign less importance to age and more importance to income. It is noteworthy that, in this example, WALS is close to complete-case (all dummies in the model), while BMA is close to fully restricted (no dummy). Thus, starting with the grand model, WALS seems to give more weight to the auxiliary dummies than BMA. The stepwise procedure gives estimates of the relative effects of age and income that are somewhat in between these two extremes.

## 8. Concluding remarks

In this paper, we formalized the trade-off between bias and efficiency that arises when there are missing covariate values in a regression relationship of interest and showed how to tackle this trade-off by model reduction procedures or model averaging methods. In future work, we plan to extend our approach to generalized linear models (GLM), such as logit, probit and Poisson regression, for which we conjecture that versions of Theorems 1 and 2 also hold. Our conjecture is motivated by the fact that maximum-likelihood estimators of exponential family models may be obtained by iteratively reweighted least squares.

## Acknowledgement

We thank Jan Magnus for comments and insightful discussions, and an associate editor and a referee for helpful suggestions.

## References

Campos, J., Ericsson, N.R., Hendry, D.F., 2005. General-to-specific modeling: an overview and selected bibliography. FRB International Finance Discussion Paper No. 838. Available at SSRN: http://ssrn.com/abstract=791684.

Cawley, J., Moran, J.R., Simon, K.I., 2008. The impact of income on the weight of elderly Americans. NBER Working Paper No. 14104.

Cutler, D.A., Glaeser, E.L., Shapiro, J.M., 2003. Why have Americans become more obese? Journal of Economic Perspectives 17, 93–118.

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

Drewnowski, A., Specter, S., 2004. Poverty and obesity: the role of energy density and costs. American Journal of Clinical Nutrition 79, 6–16.

García Villar, J., Quintana-Domeque, C., 2009. Income and body mass index in Europe. Economics & Human Biology 7, 73–83.

Horton, N.J., Kleinman, K.P., 2007. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. The American Statistician 61, 79–90.

Jones, M.P., 1996. Indicator and stratification methods for missing explanatory variables in multiple linear regression. Journal of the American Statistical Association 91, 222–230.

Kalwij, A., van Soest, A., 2005. Item non-response and alternative imputation procedures. In: Börsch-Supan, Axel, Jürges, Hendrik (Eds.), The Survey of Health, Aging, and Retirement in Europe-Methodology. MEA, Mannheim, pp. 128–150.

Kennedy Jr., W.J., Bancroft, T.A., 1971. Model-building for prediction in regression based on repeated significance tests. Annals of Mathematical Statistics 42, 1273–1284.

Leamer, E.E., 1978. Specification Searches. Ad Hoc Inference with Nonexperimental Data. Wiley.

Little, R.J.A., 1992. Regression with missing X's: a review. Journal of the American Statistical Association 87, 1227–1237.

Little, R.J.A., Rubin, D., 2002. Statistical Analysis with Missing Data, 2nd ed. Wiley.

Lundström, S., Särndal, C.-E., 2002. Estimation in the Presence of Nonresponse and Frame Imperfections. Statistics Sweden.

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. Journal of Econometrics 154, 139–153.

Magnus, J.R., Wan, A.T.K., Zhang, X., 2011. WALS estimation with nonspherical disturbances and an application to the Hong Kong housing market. Computational Statistics & Data Analysis (forthcoming).

Philipson, T., Posner, R., 2008. Is the obesity epidemic a public health problem? A decade of research on the economics of obesity. NBER Working Paper No. 14010.

Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92, 179–191.

Rao, C.R., Mitra, S.K., 1971. Generalized Inverse of Matrices and its Applications. Wiley.

Rubin, D., 1987. Multiple Imputations. Wiley.

Sala-i-Martin, X., Doppelhofer, G., Miller, R.I., 2004. Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) aproach. American Economic Review 94, 813–835.

Sanz-de-Galdeano, A., 2005. The obesity epidemic in Europe. IZA Discussion Paper No. 1814.

Wooldridge, J.M., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press.