EIEF WORKING PAPER SERIES

EINAUDI INSTITUTE FOR ECONOMICS AND FINANCE

EIEF Working Paper 11/11
May 2011

A Generalized Missing-Indicator Approach to Regression with Imputed Covariates

by
Valentino Dardanoni
(University of Palermo)
Giuseppe De Luca
(ISFOL)
Salvatore Modica
(University of Palermo)
Franco Peracchi
(University of Rome "Tor Vergata" and EIEF)

A Generalized Missing-Indicator Approach to Regression with Imputed Covariates*

Valentino Dardanoni
University of Palermo

Giuseppe De Luca
ISFOL

Salvatore Modica
University of Palermo

Franco Peracchi
Tor Vergata University and EIEF

This version: May 27, 2011

Abstract

This paper considers estimation of a linear regression model using data where some covariate values are missing but imputations are available to fill-in the missing values. The availability of imputations generates a trade-off between bias and precision in the estimators of the regression parameters: the complete cases are often too few, so precision is lost, but filling-in the missing values with imputations may lead to bias. We provide the new Stata command `gmi` which allows handling such bias-precision trade-off using either model reduction or model averaging techniques in the context of the generalized missing-indicator approach recently proposed by Dardanoni et al.(2011). If multiple imputations are available, our `gmi` command can be also combined with the built-in Stata prefix `mi estimate` to account for the extra variability due to the imputation process. The `gmi` command is illustrated with an empirical application which investigates the relationship between an objective health indicator and a set of socio-demographic and economic covariates affected by substantial item nonresponse.

Keywords: Missing covariates; Imputation; Bias-precision trade-off; Model reduction; Model averaging.

* We thank Jan Magnus for extensive and insightful discussions. The SHARE data collection has been primarily funded by the European Commission through the 5th, 6th and 7th framework programmes. Additional funding from the U.S. National Institute on Aging as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions).

1 Introduction

When trying to run a regression of interest, researchers often face the problem of missing values on some of the variables. We focus on the case when only the covariates contain missing values and the data are Missing-At-Random (MAR). As argued by Little(1992), the related problem of missing values in the dependent variable is less interesting because if the covariates are complete and the missing values on the dependent variable are MAR, then the incomplete cases contribute no information about the regression parameters of interest.

One approach to this problem–complete-case analysis–is to drop all cases with missing values and run the regression using only the complete cases. Another approach, when imputations are available, is to fill-in the missing values with the imputations and run the regression using all the data, whether observed or imputed. This second approach–which we call the 'naive' approach–is actually becoming quite common, as public-use data files increasingly include imputations of key variables affected by missing data problems. Specialized software for carrying out imputations is also becoming increasingly available. One example is the `mi` suite of commands developed in Stata 11.

From the view point of inference about the regression parameter of interest, the availability of imputations generates a trade-off between bias and precision: the complete cases are often too few, so precision is lost, but filling-in the missing values with the imputations may lead to bias. Dardanoni et al. (2011), henceforth DMP, show that this trade-off is in fact equivalent to that arising in an extended or 'grand' regression model that includes two subsets of regressors: the focus regressors corresponding to the observed or imputed covariates, and a set of auxiliary regressors representing all possible interactions between the focus regressors and the missing-data indicators. In the 'grand' model, the trade-off is between bias and precision in estimating the coefficients on the focus regressors when we drop subsets of the auxiliary regressors. As discussed in DMP, this second trade-off is easier to deal with than the first, because a variety of methods are available. This paper presents the command `gmi` that implements several methods corresponding to two alternative strategies for handling such trade-off: model reduction and model averaging.

The remainder of this paper is organized as follows. Section 2 reviews the theoretical background in DMP. Section 3 describes the two alternative strategies for estimating the regression parameters of interest. Section 4 provides a detailed description of our `gmi` command. Section 5 illustrates the `gmi` command using data available on the Stata website. Finally, Section 6 use data from the first wave of SHARE (Survey of Health, Ageing and Retirement in Europe) to provide an empirical

application on the relationship between an objective health indicator and a set of socio-demographic and economic covariates affected by substantial item nonresponse.

2 Background

Consider modeling the relationship between an outcome Y and a set of covariates X using data where some covariate values are missing. We assume that, in the absence of missing values, the data would satisfy 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 regression parameters, and $\boldsymbol{u}$ is an $N \times 1$ vector of regression errors that are homoskedastic, serially uncorrelated and have zero mean conditional on $\boldsymbol{X}$. This means that the full-information OLS estimator from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ would be unbiased for $\boldsymbol{\beta}$ and efficient in the Gauss-Markov sense.

We also assume that all missing covariate values can be replaced by imputations. These imputations may be provided by the data-producing agency or may be constructed by the researcher, for example by using the Stata command `mi impute`.

Because the first element of X is taken to be the constant term, which is always observed, the number of possible missing-data patterns is equal to 2^{K-1} (no missing data, only the first covariate missing, only the first and the second missing, etc.). A particular data set need not contain all the possible patterns, so we simply index by $j = 0, \ldots, J$ the patterns that are present in the data, with $j = 0$ corresponding to the subsample with complete data, which is assumed to be always available, and $J \leq 2^{K-1} - 1$. To keep track of exactly which covariate values are missing, we introduce the $N \times K$ missing-data indicator matrix $\boldsymbol{M}$, whose (n, k) element is equal to one if the nth case has a missing value on the kth covariate and is equal to zero otherwise.

We are concerned with the problem of how to combine the observed and the imputed values in order to estimate the regression parameter $\boldsymbol{\beta}$.

2.1 Complete-case analysis

This approach amounts to ignore the imputed values and use only the subsample with complete data. Complete-case analysis is our benchmark because, under two key assumptions, it delivers an unbiased estimator of the regression parameter $\boldsymbol{\beta}$.

Denoting the subsample with complete data by $[\boldsymbol{X}^0, \boldsymbol{y}^0]$, where $\boldsymbol{X}^0$ is an $N_0 \times K$ matrix and $\boldsymbol{y}^0$ is an $N_0 \times 1$ vector, the two key assumptions are full rank of $\boldsymbol{X}^0$ and ignorability of the missing-data process.

Assumption 1 (Complete-case full rank) *$\boldsymbol{X}^0$ has full column rank.*

Assumption 2 (Ignorability) *$\boldsymbol{M}$ and $\boldsymbol{y}$ are conditionally independent given $\boldsymbol{X}$.*

For Assumption 1 to hold, there must be enough cases (at least K) with non-missing covariate values. Assumption 2 is weaker than the standard MAR assumption because it only requires mean independence and not independence. Thus, it admits patterns where cases with low or high levels of some covariates systematically have a greater percentage of missing values. This assumption fails if, for example, observations with missing covariate values have a different regression function than observations with no missing values. In this case, an alternative is some type of sample selection model (see for example the Stata command `heckman`). Under these two assumptions, we have the following result, which represents the main justification for complete-case analysis:

Result 1 *If Assumptions 1 and 2 hold, then the complete-case OLS estimator from a regression of $\boldsymbol{y}^0$ on $\boldsymbol{X}^0$ is unbiased for $\boldsymbol{\beta}$.*

Although unbiased, the complete-case OLS estimator has the drawback of being much less precise than the full-information OLS estimator, except when the fraction of complete cases is large.

2.2 The 'naive' and the simple missing-indicator approaches

A common alternative to complete-case analysis is to use *all* cases and regress $\boldsymbol{y}$ on the completed design matrix $\boldsymbol{W}$, whose (n, k) element is equal to the corresponding element of $\boldsymbol{X}$ if a covariate value is not missing and is equal to the imputed value otherwise. This *'naive' approach* ignores the fact that the imputations are not the same as the missing covariate values, so it gives an estimator of $\boldsymbol{\beta}$ that is more precise than the complete-case OLS estimator but is also biased.

Another alternative, the so-called *simple missing-indicator approach*, consists of regressing $\boldsymbol{y}$ on the completed design matrix $\boldsymbol{W}$ and a set of J dummies $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_J$, where the elements of $\boldsymbol{d}_j$ are equal to one for cases that belong to the jth missing-data pattern and are equal to zero otherwise (the subsample with complete data represents the baseline). Adding dummies for the missing-data patterns increases the flexibility of the model by allowing the intercepts to differ across patterns but, again, unbiasedness is lost (Horton and Kleinman 2007, Jones 1996, Little 1992).

2.3 The generalized missing-indicator approach

The bias arising from the use of imputations may be eliminated by fully interacting the columns of the completed design matrix $\boldsymbol{W}$ with the dummies for the missing-data patterns. DMP call this a *generalized missing-indicator approach*.

They show that, if $\boldsymbol{y}^j$ and $\boldsymbol{W}^j$ respectively denote the $N_j \times 1$ subvector of $\boldsymbol{y}$ and the $N_j \times K$ submatrix of $\boldsymbol{W}$ corresponding to the jth missing-data pattern, then the generalized missing-indicator approach corresponds to using the following *'grand' model*

$$\begin{bmatrix} \boldsymbol{y}^0 \\ \boldsymbol{y}^1 \\ \vdots \\ \boldsymbol{y}^J \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^0 \\ \boldsymbol{W}^1 \\ \vdots \\ \boldsymbol{W}^J \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{W}^1 & & \\ & \ddots & \\ & & \boldsymbol{W}^J \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}^1 \\ \vdots \\ \boldsymbol{\delta}^J \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}^0 \\ \boldsymbol{v}^1 \\ \vdots \\ \boldsymbol{v}^J \end{bmatrix},$$

where $\boldsymbol{\beta}$ is the regression parameter of interest, the $\boldsymbol{\delta}^j$ are $K \times 1$ vectors of nuisance parameters that may be interpreted as the asymptotic bias in the regression of $\boldsymbol{y}^j$ on $\boldsymbol{W}^j$, and the $\boldsymbol{v}^j$ are $N_j \times 1$ vectors of projection errors that have mean zero and are orthogonal to the columns of $\boldsymbol{W}^j$. A more compact representation of the 'grand' model is

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{v}, \tag{2}$$

where

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{X}^0 \\ \boldsymbol{W}^1 \\ \vdots \\ \boldsymbol{W}^J \end{bmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{W}^1 & & \\ & \ddots & \\ & & \boldsymbol{W}^J \end{bmatrix}, \qquad \boldsymbol{\delta} = \begin{bmatrix} \boldsymbol{\delta}^1 \\ \vdots \\ \boldsymbol{\delta}^J \end{bmatrix}, \qquad \boldsymbol{v} = \begin{bmatrix} \boldsymbol{u}^0 \\ \boldsymbol{v}^1 \\ \vdots \\ \boldsymbol{v}^J \end{bmatrix},$$

respectively an $N \times K$ matrix of observed or imputed covariates, an $N \times JK$ matrix of auxiliary variables, a $JK \times 1$ vector of nuisance parameters, and an $N \times 1$ error vector. The matrix of auxiliary variables $\boldsymbol{Z}$ consists of JK interactions between the set of J dummies $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_J$ for the missing-data patterns and the K columns of the completed design matrix $\boldsymbol{W}$. Notice that this matrix is not required to have full column rank. This occurs when some of the $\boldsymbol{W}^j$ does not have full column rank, either because $N_j < K$ or because $N_j \geq K$ but the columns of $\boldsymbol{W}^j$ are linearly dependent, as when mean imputation or deterministic regression imputation is used. Incidentally, such imputation methods are known to produce data sets with undesirable properties, see e.g. Lundstrom and Sarnndal (2002). When some of the $\boldsymbol{W}^j$ does not have full column rank, only a subset of the coefficients in $\boldsymbol{\delta}^j$ is identifiable but this does not affect the estimates of $\boldsymbol{\beta}$. Also notice that the regression errors in model (2) need not have constant variance because the projections errors $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^J$ may be heteroskedastic.

The main result in DMP is the following:

Result 2 *If Assumption 1 holds then, for any choice of imputations, the OLS estimate of* $\boldsymbol{\beta}$ *in model* (2) *is numerically the same as the complete-case OLS estimate of* $\boldsymbol{\beta}$*.*

Thus, if the ignorability assumption holds, regressing $\boldsymbol{y}$ on $\boldsymbol{W}$ and $\boldsymbol{Z}$ allows one to fully exploit the available information *and* to obtain an unbiased estimator of the regression parameter $\boldsymbol{\beta}$.

3 Alternative strategies for estimating $\boldsymbol{\beta}$

Both the 'naive' and the simple missing-indicator approach correspond to using restricted versions of model (2) obtained by placing restrictions on the vector $\boldsymbol{\delta}$. The 'naive' approach restricts $\boldsymbol{\delta}$ to be equal to zero, while the simple missing-indicator approach restricts all the $\boldsymbol{\delta}^j$ to be equal to zero except for their first element. When these restrictions are at odds with the data, imposing them leads to an estimator of $\boldsymbol{\beta}$ that is biased but more precise (less variable) than the OLS estimator of $\boldsymbol{\beta}$ in model (2) which, in turn, is numerically the same as the complete-case estimator of $\boldsymbol{\beta}$. This suggest that, by placing restrictions on $\boldsymbol{\delta}$, or equivalently by excluding some of the auxiliary variables in $\boldsymbol{Z}$, one may obtain an estimator of $\boldsymbol{\beta}$ that is better in the mean squared error (MSE) sense than the complete-case estimator. The Stata command in this paper implements two alternative strategies for obtaining such an estimator of $\boldsymbol{\beta}$: model reduction and model averaging.

3.1 Model reduction

Model reduction involves selecting first an intermediate model between the 'grand' model (2) and the 'naive' model corresponding to $\boldsymbol{\delta} = 0$, and then estimating the parameter of interest $\boldsymbol{\beta}$ conditional on the selected model. Because the variables in the completed design matrix $\boldsymbol{W}$ are treated as focus regressors and are always included, an intermediate model corresponds to one of the 2^{JK} possible subsets of auxiliary regressors in $\boldsymbol{Z}$.

Model reduction may be carried out through a number of variable selection methods, such as those implemented by the built-in Stata command `stepwise` or by the `vselect` command discussed in Lindsey and Sheather (2010). Dropping one of the variables in $\boldsymbol{Z}$ amounts to restricting one element of $\boldsymbol{\delta}^j$ to zero. This in turn corresponds to selecting one of the J missing-data patterns and forcing the coefficient on a particular covariate for that data pattern to be the same as for the subsample with complete data. The various methods differ depending on how one explores the set of all the possible models (e.g. via a general-to-specific or via a specific-to-general approach) and

the decision rule used to judge validity of each model considered (e.g. a fixed significance level or an information criterion such as AIC or BIC).

One well known problem with this strategy is pretesting.[1] Another is the fact that model reduction 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. Simple model reduction techniques, such as backward and forward selection, analyze at most $JK(JK+1)/2$ models. More complicated model reduction techniques, such as the leaps and bounds technique implemented in `vselect`, usually analyze a larger number of models.

3.2 Model averaging

Model averaging takes a different route. 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 estimate of $\boldsymbol{\beta}$ as a weighted average of these conditional estimates. When the model space contains I models, a model averaging estimate of $\boldsymbol{\beta}$ is of the form

$$\bar{\boldsymbol{\beta}} = \sum_{i=1}^{I} \lambda_i \widehat{\boldsymbol{\beta}}_i, \tag{3}$$

where the λ_i are non-negative random weights that add up to one and $\widehat{\boldsymbol{\beta}}_i$ is the estimate of $\boldsymbol{\beta}$ obtained by conditioning on the ith 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, then λ_i is proportional to the marginal likelihood of $\boldsymbol{y}$ under model i. The BMA literature is vast and we refer the reader to Raftery et al. (1997) for a starting point.

Our Stata implementation of standard BMA is based on the `bma` command provided by De Luca and Magnus (2011). This approach assumes a classical Gaussian linear model for (2), non-informative priors for $\boldsymbol{\beta}$ and the error variance, and a multivariate Gaussian prior for $\boldsymbol{\delta}$. Notice that the computational burden required to obtain a standard BMA estimate is proportional to the dimension I of the model space. In our case $I = 2^{JK}$, so this computational burden is substantial unless both J and K are small.

Another type of BMA is Weighted-Average Least Squares (WALS), introduced by Magnus et al. (2010). WALS also assumes a classical Gaussian linear model for (2) and noninformative priors for $\boldsymbol{\beta}$

[1] See Magnus (1999) and the FAQ http://www.stata.com/support/faqs/stat/stepwise.html.

and the error variance but, instead of a multivariate Gaussian prior for $\boldsymbol{\delta}$, it uses a distribution with zero mean for the independently and identically distributed elements of the transformed parameter vector $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\delta})$, whose hth element η_h is the population t-statistic for testing the significance of the hth element of $\boldsymbol{\delta}$. Magnus et al. (2010) use the Laplace distribution, while Einmahl et al. (2011) use the Subbotin family which leads to estimators with better asymptotic properties. The assumption that the regression errors in (2) are homoskedastic and serially uncorrelated is not crucial for WALS, and the method can be generalized to non-spherical errors (Magnus et al. 2011).

WALS has three main advantages over standard BMA. First, its computational burden is only proportional to JK. Second, its choice of priors corresponds to a more intuitive concept of uncertainty about the role of the auxiliary variables. Third, WALS estimates have bounded risk and are near-optimal in terms of a well-defined regret criterion (Magnus et al. 2010). Our Stata implementation of WALS, for both Laplace and Subbotin priors, is based on the `wals` command provided by De Luca and Magnus (2011).

3.3 Standard errors of the estimators

Like standard Stata estimation commands, we provide estimated coefficients, standard errors and t-ratios. We do not provide p-values and confidence intervals because our estimators are generally biased and their distribution need not be Gaussian, not even asymptotically. On the other hand, the hth regressor may be considered to be robustly correlated with the outcome if the t-ratio η_h on its coefficient is greater than one in absolute value, in which case the MSE of the unrestricted OLS estimator of the coefficient is lower than that of the restricted OLS estimator (see e.g. Magnus 2002). On the basis of this criterion, we also provide two-standard error bands for the estimated coefficients.

Computation and interpretation of the standard errors differ depending on the estimation strategy (model reduction vs. model averaging) and the general approach to estimation (frequentist vs. Bayesian).

For model reduction, the default is 'classical' standard errors of the OLS estimator of the selected model. These standard errors do not take into account heteroskedasticity or serial correlation in the data and, most importantly, ignore the additional sampling variability induced by the model selection step. The option `bootstrap` gives standard errors based on the wild bootstrap which are valid under conditional heteroskedasticity and also take into account the additional variability due to model selection.

For BMA, the default standard errors have the usual Bayesian interpretation of measuring the spread of the posterior distribution of the parameters of interest given the data and take model uncertainty explicitly into account. In this case, the option `bootstrap` provides a frequentist measure of the variability due to sampling, including the variability due to model selection.

Notice that neither model reduction nor model averaging take into account the additional sampling variability due to imputation. This problem could be addressed by multiple imputation methods (Rubin 1987). As illustrated in Sections 5 and 6, our `gmi` command can be combined with the build-in Stata prefix `mi estimate` (see `mi estimate`).

4 Stata command

The new Stata command `gmi` allows handling the trade-off between bias and precision when estimating a classical linear regression model with imputed covariates. The earliest version of Stata required to run this command is `version 11.1`. The syntax is as follows:

`gmi` *depvar* [*varlist*] [*if*] [*in*] `,` `imputed(`*varlist*`)` `missing(`*varlist*`)` [`summarize cc naive smi sw vs bma wals` *stepwise_options vselect_options bma_options wals_options* `full vce(bootstrap [,`*bootstrap_options*`]) auxiliary(`*string*`) keep nowarn`]

where *depvar* is the dependent variable, *varlist* is an optional list of observed covariates (i.e. covariates whose values are fully observed), `imputed` is the list of imputed covariates (i.e. covariates whose missing values are replaced by imputed values) and `missing` is the relevant list of missing-data indicators (i.e. the non-zero columns of the matrix $\boldsymbol{M}$ corresponding to the set of imputed covariates). Missing-data indicators take value 0 for observed cases and value 1 for imputed cases. The number of imputed covariates must coincide with the number of missing-data indicators. The first variable in `missing` is paired with the first indicator in `imputed`, the second variable in `missing` is paired with the second indicator in `imputed`, and so on. The constant term (which is always included) plus the set of observed and imputed covariates correspond to the K columns of the completed design matrix $\boldsymbol{W}$. The auxiliary regressors in $\boldsymbol{Z}$ (i.e. the JK interactions between the J dummies for the missing-data patterns and the K columns of $\boldsymbol{W}$) are instead automatically generated by the command using the information from `missing`. The `gmi` command shares the same features of all Stata estimation commands, including access to the estimation results. Factor variables, time-series operators and weights are not allowed. A description of the options that are specific to this command is provided in the following sections.

4.1 Options of the gmi command

`summarize`, the default, provides a description of the 'grand' model (number of observations, number of observed and imputed covariates, number of focus and auxiliary regressors, number of missing-data patterns, and dimension of the model space), plus summaries of the distribution of *depvar* (number of observations, mean and standard deviation) for the complete-case and each missing-data pattern.

`cc` provides the complete-case estimate of $\boldsymbol{\beta}$, namely the OLS estimate from a regression of *depvar* on the K focus regressors in $\boldsymbol{W}$ using only the complete cases. This is numerically the same as the OLS estimate of $\boldsymbol{\beta}$ in the grand model (2).

`naive` provides the 'naive' estimate of $\boldsymbol{\beta}$, namely the OLS estimate from a regression of *depvar* on the K focus regressors in $\boldsymbol{W}$ using all cases.

`smi` provides the simple missing-indicator estimate of $\boldsymbol{\beta}$, namely the OLS estimate of $\boldsymbol{\beta}$ from a regression of *depvar* on the K focus regressors in $\boldsymbol{W}$ and the J dummies for the missing-data patterns using all cases.

`sw` provides the OLS estimate of $\boldsymbol{\beta}$ from a regression of *depvar* on the K focus regressors in $\boldsymbol{W}$ and the subset of auxiliary regressors in $\boldsymbol{Z}$ selected through the build-in Stata command `stepwise`. This estimate of $\boldsymbol{\beta}$ is conditional on the selected model. A brief description of the options for the `stepwise` command is given in Section 4.2.

`vs` provides the OLS estimate of $\boldsymbol{\beta}$ from a regression of *depvar* on the K focus regressors in $\boldsymbol{W}$ and the subset of auxiliary regressors in $\boldsymbol{Z}$ selected through the `vselect` command by Lindsey and Sheather (2010). As for the `sw` option, this estimate of $\boldsymbol{\beta}$ is conditional on the selected model. A brief description of the options for the `vselect` command is given in Section 4.3.

`bma` provides the BMA estimate of $\boldsymbol{\beta}$ in the 'grand' model (2) using the `bma` command implemented by De Luca and Magnus (2011). This option assumes a classical Gaussian linear model for (2), noninformative priors for the regression parameter $\boldsymbol{\beta}$ and the error variance, and a multivariate Gaussian prior for the auxiliary parameter $\boldsymbol{\delta}$. This estimate is obtained as a weighted average of the estimates of $\boldsymbol{\beta}$ from each of the 2^{JK} possible models in the model space with weights proportional to the marginal likelihood of *depvar* in each model. A brief description of the options for the `bma` command is given in Section 4.4.

`wals` provides the WALS estimate of $\boldsymbol{\beta}$ in the 'grand' model (2) using the `wals` command implemented by De Luca and Magnus (2011). Like BMA, this option assumes a classical Gaussian linear model for (2) and noninformative priors for the regression parameter $\boldsymbol{\beta}$ and the error

variance. Unlike BMA, WALS uses orthogonal transformations of the auxiliary regressors and their parameters, which reduces to JK the order of magnitude of the required calculations. Further, the transformed auxiliary parameters are assumed to be identically and independently distributed according to either a Laplace or a Subbotin prior. A brief description of the options for the `wals` command is given in Section 4.5.

`full` requires to display the estimation results for all model parameters (i.e. focus and auxiliary parameters) and to return the associated estimates and their variance-covariance matrix in the vector `e(b)` and the matrix `e(V)` respectively. By default, display of the estimation results is restricted to the focus parameters of interest, the associated estimates and their variance-covariance matrix are returned in the vector `e(b)` and the matrix `e(V)` respectively, while estimates of the auxiliary parameters and their variance-covariance matrix are returned in the vector `e(b_aux)` and the matrix `e(V_aux)` respectively.

`vce(bootstrap [,`*bootstrap_options*`])` uses wild bootstrap to estimate the variance-covariance matrix of the parameter estimates (see [R] **bootstrap**). By default, bootstrap estimates of the variance-covariance matrix are computed only for the focus parameters. To obtain bootstrap estimates of the variance-covariance matrix the focus and the auxiliary parameters, the option `vce(bootstrap)` must be combined with the option `full`. In any case, `vce(bootstrap)` and `full` cannot be jointly specified when applying model reduction techniques (i.e. the options `sw` and `vs`) because the subset of selected auxiliary regressors can vary across bootstrap replicates. Standard options for bootstrap estimation can be specified as sub-options within `vce(bootstrap)` (see [R] **vce_option**).

`auxiliary(`*string*`)` specifies the prefix for the name of the auxiliary regressors. The default is D. Thus, auxiliary regressors are named as Dj and $Dj\_varname$ where $j = 1, \ldots, J$ is an index for the sub-samples of missing data and *varname* is the name of each variable listed in *varlist* and `imputed`.

`keep` specifies whether auxiliary regressors have to be kept in the data after estimation. By default, they are dropped.

`nowarn` suppresses the display of a warning message on dropped collinear regressors.

4.2 Options for stepwise

By specifying the `sw` option, `gmi` carries out model reduction through the build-in Stata command `stepwise` (see [R] **stepwise** for details). The relevant options of the `stepwise` command are `pr(#)`

(significance level for backward selection), `pe(#)` (significance level for forward selection), `forward` (backward stepwise) and `lr` (likelihood-ratio test of term significance). Since the auxiliary regressors in $\boldsymbol{Z}$ have no hierarchical ordering, backward hierarchical selection and forward hierarchical selection are not allowed.

4.3 Options for vselect

By specifying the `vs` option, `gmi` carries out model reduction through the `vselect` command implemented by Lindsey and Sheather (2010). This command offers three model reduction techniques: backward selection (the default), forward selection (`forward`), and leaps-and-bounds selection (`best`). An information criterion is used to judge the validity of each model through the options `r2adj` (adjusted R^2), `aic` (AIC), `aicc` (corrected AIC), `bic` (BIC), `cp1` or `cp2` (Mallow's C_p). Mallows's C_p criterion can only be used with leaps-and-bounds selection and the decision rule can be either a value of C_p close to zero (`cp1`) or a value close to the number of covariates (`cp2`). For additional information see Lindsey and Sheather (2010).

4.4 Options for BMA

By specifying the `bma` option, `gmi` carries out BMA through the `bma` command implemented by De Luca and Magnus (2011). In this case, one can use two additional options. The option `scaling` provides an alternative way of scaling the model weights λ_i when the default scaling procedure suffers from numerical problems. Although scaling of the model weights does not affect BMA estimates, the computing time required by this alternative procedure is almost double. The option `nodots` suppresses the display of the dots to track the progress of `bma` estimation. By default, dots are displayed only if the model space consists of more than 128 models. One dot means that 1% of the models in the model space have been estimated.

4.5 Options for WALS

By specifying the `wals` option, `gmi` carries out model averaging through the `wals` command implemented by De Luca and Magnus (2011). As for the prior on the transformed auxiliary parameters, one can choose between Laplace or Subbotin priors through the option `q(#)` which defines the free parameter $0 < q \leq 1$ of a Subbotin density with the prior median of η_h equal to zero and the prior median of η_h^2 equal to one. The default is $q = 1$ corresponding to a Laplace prior. Values of q in the interval $(0, 1)$ give instead a class of Subbotin priors. Einmahl et al. (2011) argue that values of q close to zero are unappealing from the point of view of ignorance. For empirical applications,

they recommend $q = 0.5$. For a Subbotin prior with $q \neq 1$ and $q \neq .5$, one can also specify a set of additional options (i.e. intpoints(#), eps(#) and iterate(#)) to control the accuracy of the numerical process for approximating the constrained parameter of a Subbotin density. Additional information can be found in De Luca and Magnus (2011).

5 Example

This section illustrates the gmi command using data available on the Stata website.

```
. quietly use "http://www.stata-press.com/data/r11/mhouses1993s30", clear
. describe
Contains data from http://www.stata-press.com/data/r11/mhouses1993s30.dta
  obs:         1,647                          Albuquerque Home Prices Feb15-Apr 30, 1993
 vars:            13                          19 Jun 2009 10:50
 size:        54,351 (99.9% of memory free)   (_dta has notes)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
price           int    %8.0g                  Sale price (hundreds)
sqft            int    %8.0g                  Square footage of living space
age             float  %10.0g                 Home age (years)
nfeatures       byte   %8.0g                  Number of certain features
ne              byte   %8.0g                  Located in northeast (largest
                                                residential) sector of the city
custom          byte   %8.0g                  Custom build
corner          byte   %8.0g                  Corner location
tax             float  %10.0g                 Tax amount (dollars)
lnage           float  %9.0g
lntax           float  %9.0g
_mi_miss        byte   %8.0g
_mi_m           int    %8.0g
_mi_id          int    %12.0g
-------------------------------------------------------------------------------
Sorted by:  _mi_m  _mi_id
```

We want to estimate a classical linear regression model for the relationship between home sale price (price) and home characteristics (sqft, nfeatures, ne, custom, corner, lnage and lntax). Since there are cases with age and tax missing, lnage and lntax are affected by a missing-data problem and their missing values have been imputed using a multivariate normal regression model (see mi impute mvn).

```
. mi describe
  Style:  mlong
          last mi update 19jun2009 10:50:22, 243 days ago
  Obs.:   complete           66
          incomplete         51  (M = 30 imputations)
          ---------------------
          total             117
  Vars.:  imputed:  2; lnage(49) lntax(10)
          passive: 2; age(49) tax(10)
          regular: 6; price sqft nfeatures ne custom corner
```

```
         system:  3; _mi_m _mi_id _mi_miss
        (there are no unregistered variables)

. mi misstable summarize lnage lntax
```

Variable	Obs=.	Obs>.	Obs<.	Obs<. Unique values	Obs<. Min	Obs<. Max
lnage	49		68	30	0	3.970292
lntax	10		107	95	5.407172	7.475906

Thus, the data contain 117 observations plus 30 multiple imputations stored in the `mlong` style (see `styles`) for each of the 51 incomplete cases. Below, we generate the missing-data indicators for `lnage` and `lntax` and the local `first_imp` which is used to restrict the estimation sample to the first imputation. Continuous covariates are centered to their median values to obtain meaningful estimates of the constant term.

```
. generate mis_lnage=(lnage==.)
. generate mis_lntax=(lntax==.)
. bys _mi_id: egen M_lnage=max(mis_lnage)
. bys _mi_id: egen M_lntax=max(mis_lntax)
. local first_imp "(_mi_miss==0|_mi_m==1)"
. foreach x of varlist sqft nfeatures lnage lntax {
  2.         quietly summarize `x´ if `first_imp´, d
  3.         quietly replace `x´=`x´-r(p50)
  4. }
```

The `gmi` command with its default option `summarize` produces the following output:

```
. gmi price sqft nfeatures ne custom corner if `first_imp´,       ///
>         imp(lnage lntax) mis(M_lnage M_lntax)
note: D1_nfeatures D1_ne D1_custom D1_corner D1_lnage D1_lntax D3_corner omitte
> d because of collinearity
------------------------------------------------------------------------------
Grand model
------------------------------------------------------------------------------
Number of obs                    : 117
Number of observed covariates    : 6
Number of imputed covariates     : 2
Number of focus covariates       : 8
Number of missing data patterns : 3
Number of auxiliary covariates   : 17
Dimension of model space         : 131072
```

Summary of price by missing data pattern: Group	Freq.	Percent	Cum.	Mean	Std.Dev.	Missing data patterns
0	66	56.41	56.41	1168.61	404.38	1 1
1	2	1.71	58.12	1010.00	452.55	1 .
2	41	35.04	93.16	930.44	298.59	. 1
3	8	6.84	100.00	880.50	307.17	. .

Our model includes 8 focus regressors, of which 6 (including the constant term) are observed and 2 are imputed. Excluding the subset of complete cases (66 observations), there are $2^2 - 1 = 3$

missing-data patterns: (i) `lnage` observed and `lntax` missing (2 observations), (ii) `lnage` missing and `lntax` observed (41 observations), and (iii) `lnage` and `lntax` both missing (8 observations). The grand model therefore includes $3 \cdot 8 = 24$ auxiliary variables, but 7 of them are dropped because of perfect collinearity. In particular, because the variable `corner` is constant for the third missing-data pattern, the auxiliary variables `D3` and `D3_corner` are perfectly collinear, so the latter is dropped.

```
. tab corner if `first_imp´ & M_lnage==1 & M_lntax==1
```

Corner location	Freq.	Percent	Cum.
0	8	100.00	100.00
Total	8	100.00	

Other 6 auxiliary variables are dropped because the first missing-data pattern includes only 2 observations, so we can identify at most 2 of the 8 associated auxiliary parameters. After dropping from $\boldsymbol{Z}$ all collinear variables, the dimension of the model space reduces to $2^{17} = 131072$. The summary statistics for the dependent variable across missing-data patterns reveal that both the mean and the variance of `price` are considerably higher for the subsample with complete cases.

We obtain the complete-case OLS estimator of the focus parameters $\boldsymbol{\beta}$ by specifying the `cc` option.

```
. gmi price sqft nfeatures ne custom corner if `first_imp´,       ///
>         imp(lnage lntax) mis(M_lnage M_lntax) cc nowarn
Complete-case OLS estimates                   Number of obs =      66
                                              df_m          =       7
```

price	Coef.	Std. Err.	t	[2 Std. Err. Bands]	
_cons	1000.288	39.59419	25.26	960.6942	1039.883
sqft	.4357152	.0983648	4.43	.3373504	.5340799
nfeatures	.3227029	18.34047	0.02	-18.01776	18.66317
ne	7.398968	46.91899	0.16	-39.52002	54.31796
custom	181.0344	54.37951	3.33	126.6549	235.4139
corner	-78.70756	49.85979	-1.58	-128.5673	-28.84777
lnage	-39.2261	27.55061	-1.42	-66.77671	-11.67549
lntax	302.2674	145.0322	2.08	157.2353	447.2996

These estimates could also be obtained through the built-in Stata command `regress` after restricting the estimation sample to the subset of complete data. They are also numerically the same as the OLS estimate of $\boldsymbol{\beta}$ in the grand model (2). Result 1 implies that, under the ignorability assumption, the complete-case OLS estimator is unbiased for $\boldsymbol{\beta}$. Our findings suggest that home sale price is positively related to square footage of living space, log of taxes paid and whether the home is located in a custom building. On the other side, there is negative association with log of home

age and whether the home has a corner location. The effects of the other covariates are not robust because the corresponding t-ratios are smaller than one in absolute value. It is also worth noticing that the complete-case estimator is likely to be highly inefficient as it discards about 44 percent of the sample observations.

To explore the trade-off between bias and precision, consider now the 'naive' and the simple missing-indicator approaches. The former ignores the fact that missing values have been imputed by restricting all auxiliary parameters to zero, while the latter restrict all auxiliary parameters to zero except the coefficients on the dummies for the missing-data patterns.

```
. gmi price sqft nfeatures ne custom corner if `first_imp´,       ///
>         imp(lnage lntax) mis(M_lnage M_lntax) naive nowarn
Naive OLS estimates                             Number of obs =     117
                                                df_m          =       7
```

price	Coef.	Std. Err.	t	[2 Std. Err. Bands]	
_cons	984.3707	35.50699	27.72	948.8638	1019.878
sqft	.382786	.0729738	5.25	.3098122	.4557598
nfeatures	3.622533	13.89274	0.26	-10.27021	17.51527
ne	28.93578	37.16146	0.78	-8.225679	66.09725
custom	145.1389	46.45179	3.12	98.68716	191.5907
corner	-85.8675	42.73586	-2.01	-128.6034	-43.13164
lnage	-26.48807	21.62821	-1.22	-48.11628	-4.859864
lntax	262.9705	106.5927	2.47	156.3778	369.5632

```
. gmi price sqft nfeatures ne custom corner if `first_imp´,       ///
>         imp(lnage lntax) mis(M_lnage M_lntax) smi nowarn
SMI OLS estimates                               Number of obs =     117
                                                df_m          =      10
```

price	Coef.	Std. Err.	t	[2 Std. Err. Bands]	
_cons	1007.357	35.88174	28.07	971.4752	1043.239
sqft	.3993985	.0718978	5.56	.3275006	.4712963
nfeatures	-5.977141	14.29397	-0.42	-20.27111	8.316833
ne	49.92553	37.20047	1.34	12.72506	87.12601
custom	157.4772	47.33692	3.33	110.1403	204.8141
corner	-103.4662	42.61305	-2.43	-146.0793	-60.85319
lnage	-30.55087	21.47985	-1.42	-52.03073	-9.071018
lntax	204.6133	108.1598	1.89	96.45353	312.7731

```
. matrix list e(b_aux)
e(b_aux)[1,3]
            D1          D2          D3
y1  -119.71306  -82.584248   -164.8674
. matrix list e(V_aux)
symmetric e(V_aux)[3,3]
           D1         D2         D3
D1  18343.655
D2  826.58661  1662.3468
D3   527.1418  680.08892  4520.6579
```

Both approaches impose arbitrary restrictions on the auxiliary parameter $\boldsymbol{\delta}$, so they are likely to result in biased estimates of the focus parameter $\boldsymbol{\beta}$. However, as suggested by their considerably

lower standard errors, these estimators are more precise than the complete-case OLS estimator. The most striking differences are in the estimated coefficients of `corner` and `lntax`. Notice that, to force users to treat the auxiliary parameters as nuisance parameters, their estimates and the associated variance-covariance matrix are returned in the vector `e(b_aux)` and the matrix `e(V_aux)` respectively.

The `gmi` command provides two alternative strategies for finding a better estimator of $\boldsymbol{\beta}$ in the MSE sense: model reduction and model averaging. Although the choice between these two strategies is left to the users, we strongly encourage choosing model averaging in order to avoid the problems caused by pretesting.

Model reduction can be carried out through the built-in Stata command `stepwise` or the `vselect` command by Lindsey and Sheather (2010). There are reasons to prefer the latter, as model reduction is based on an information criterion instead of an arbitrary significance level, and the leaps-and-bounds algorithm is expected to select the best model. To save space, we only present the OLS estimates of the model selected by `vselect` with the `best` and the `bic` options.

```
. gmi price sqft nfeatures ne custom corner if `first_imp´,       ///
>         imp(lnage lntax) mis(M_lnage M_lntax) vs best bic full nowarn
Model reduction: leaps and bounds with bic   Number of obs =     117
                                             df_m          =       9
```

price	Coef.	Std. Err.	t	[2 Std. Err. Bands]	
_cons	983.4677	32.52224	30.24	950.9454	1015.99
sqft	.4911947	.0722097	6.80	.418985	.5634044
nfeatures	1.022459	12.73723	0.08	-11.71477	13.75968
ne	6.726864	34.54129	0.19	-27.81442	41.26815
custom	163.2298	43.00966	3.80	120.2202	206.2395
corner	-80.96139	39.22133	-2.06	-120.1827	-41.74006
lnage	-25.25726	19.84414	-1.27	-45.1014	-5.413129
lntax	257.7811	98.19124	2.63	159.5898	355.9723
D2_sqft	-.2688726	.0622148	-4.32	-.3310874	-.2066578
D3_custom	-400.7815	168.7942	-2.37	-569.5757	-231.9873

In this case, we specified the `full` option to display estimates of the focus and the auxiliary parameters. The selected model includes two auxiliary variables: the interaction between `sqft` and the dummy `D2` for the second missing-data pattern, and the interaction between `custom` and the dummy `D3` for the third missing-data pattern. Notice that, the standard errors are conditional on the model selected by `vselect` and therefore should be treated with caution.

Next, we focus on model averaging using BMA and WALS respectively.

```
. gmi price sqft nfeatures ne custom corner if `first_imp´,       ///
>         imp(lnage lntax) mis(M_lnage M_lntax) bma nowarn
Model space: 131072 models
```

```
Estimation
----+--- 10% ----+--- 20% ----+--- 30% ----+--- 40% ----+--- 50%
..................................................     50%
..................................................    100%
Model averaging: BMA                              Number of obs =     117
                                                  df_m          =      24
```

price	Coef.	Std. Err.	t	[2 Std. Err. Bands]	
_cons	981.8827	35.19818	27.90	946.6845	1017.081
sqft	.4379617	.0994773	4.40	.3384844	.5374391
nfeatures	2.712441	13.75563	0.20	-11.04319	16.46807
ne	14.97688	38.36088	0.39	-23.384	53.33776
custom	157.6969	44.93556	3.51	112.7614	202.6325
corner	-77.18778	41.95601	-1.84	-119.1438	-35.23177
lnage	-31.51599	21.10457	-1.49	-52.62056	-10.41142
lntax	318.0173	138.4967	2.30	179.5206	456.514

```
. gmi price sqft nfeatures ne custom corner if `first_imp´,       ///
>         imp(lnage lntax) mis(M_lnage M_lntax) wals nowarn
WALS estimates - Laplace priors                   Number of obs =     117
                                                  df_m          =      24
```

price	Coef.	Std. Err.	t	[2 Std. Err. Bands]	
_cons	994.0145	37.68866	26.37	956.3258	1031.703
sqft	.420371	.0885567	4.75	.3318143	.5089278
nfeatures	.5016072	16.63116	0.03	-16.12955	17.13277
ne	18.17247	43.4971	0.42	-25.32463	61.66958
custom	175.4686	51.53303	3.40	123.9356	227.0016
corner	-80.34054	46.61626	-1.72	-126.9568	-33.72429
lnage	-35.90108	25.46287	-1.41	-61.36395	-10.4382
lntax	298.6159	130.3276	2.29	168.2883	428.9434

Magnus et al. (2010) argue that WALS is theoretically superior to BMA in the choice of priors for the auxiliary parameters, and is practically superior because of the substantially lower computational burden. Although the Stata command `bma` is much faster than the Magnus' original Matlab routine, we recognize that BMA can be very time consuming when the covariates or missing-data patterns is moderate or large. In such circumstances, users are encouraged to rely on WALS, at least when performing preliminary model specification search. In this example, we find that most of the estimated coefficients from BMA and WALS are similar, suggesting that differences in the priors for the auxiliary parameters play a minor role. Similar findings are also supported by the estimates from WALS with a Subbotin prior for the auxiliary parameters.

```
. gmi price sqft nfeatures ne custom corner if `first_imp´,       ///
>         imp(lnage lntax) mis(M_lnage M_lntax) wals q(.5)        ///
>         vce(bootstrap, rep(100)) nowarn
(running gmi on estimation sample)
Bootstrap replications (100)
----+--- 1 ----+--- 2 ----+--- 3 ----+--- 4 ----+--- 5
..................................................     50
..................................................    100
WALS estimates - Subbotin(q=.5) priors            Number of obs =     117
                                                  Replications  =     100
```

```
                                        df_m           =        24
```

price	Observed Coef.	Bootstrap Std. Err.	t	Bootstrap [2 Std. Err. Bands]	
_cons	992.4352	45.40523	21.86	947.03	1037.84
sqft	.4183898	.0968779	4.32	.3215119	.5152676
nfeatures	.1203615	17.2886	0.01	-17.16823	17.40896
ne	21.76857	56.17089	0.39	-34.40233	77.93946
custom	177.8062	71.95092	2.47	105.8553	249.7571
corner	-80.24304	48.86178	-1.64	-129.1048	-31.38126
lnage	-35.72275	38.17886	-0.94	-73.90161	2.456116
lntax	302.3153	166.3357	1.82	135.9796	468.651

In the above example, standard errors are estimated by the wild bootstrap with 100 replications. Bootstrapped standard errors are usually larger than traditional ones because they account for heteroskedasticity of unknown form. As argued in Section 3.3, the wild bootstrap also provides an easy way to ensure comparability of the standard errors across the different estimation methods.

Finally, we can use the 30 multiple imputations on `lnage` and `lntax` to account for the sampling variability induced by imputation of missing values. This can be done by combining our `gmi` command with the built-in Stata prefix `mi estimate`.

```
. mi estimate: gmi price sqft nfeatures ne custom corner,         ///
>         imp(lnage lntax) mis(M_lnage M_lntax) wals q(.5) nowarn full

Multiple-imputation estimates                     Imputations     =         30
WALS estimates - Subbotin(q=.5) priors            Number of obs   =        117
                                                  Average RVI     =     0.1202
                                                  Complete DF     =         92
DF adjustment:   Small sample                     DF:     min     =      53.04
                                                          avg     =      81.87
                                                          max     =      89.46
Model F test:       Equal FMI                     F(  24,   89.4) =      18.93
                                                  Prob > F        =     0.0000
```

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_cons	991.3835	37.04141	26.76	0.000	917.7875	1064.979
sqft	.4317465	.0867584	4.98	0.000	.2592937	.6041993
nfeatures	-.2094938	16.38439	-0.01	0.990	-32.76653	32.34754
ne	21.21545	42.39104	0.50	0.618	-63.00865	105.4396
custom	169.9147	50.06324	3.39	0.001	70.44587	269.3836
corner	-78.7322	46.13668	-1.71	0.091	-170.4003	12.93594
lnage	-42.43074	26.08771	-1.63	0.108	-94.33088	9.469404
lntax	280.3926	127.9489	2.19	0.031	26.02861	534.7565
D1	-113.878	110.0119	-1.04	0.304	-332.663	104.9071
D1_sqft	.3657415	.4004362	0.91	0.364	-.430267	1.16175
D2	-54.71213	69.21678	-0.79	0.431	-192.3361	82.91185
D2_sqft	-.0689371	.1178286	-0.59	0.560	-.3035881	.1657139
D2_nfeatures	-3.48209	25.28178	-0.14	0.891	-53.76192	46.79774
D2_ne	17.66156	65.09508	0.27	0.787	-111.7321	147.0552
D2_custom	-33.78444	80.12972	-0.42	0.674	-193.1165	125.5476
D2_corner	-42.7017	74.37813	-0.57	0.567	-190.6397	105.2363
D2_lnage	-3.959991	41.30198	-0.10	0.924	-86.7998	78.87982
D2_lntax	-151.7439	171.2648	-0.89	0.378	-492.8674	189.3797
D3	-167.1674	275.5164	-0.61	0.546	-716.1743	381.8394
D3_sqft	.1767931	.8775505	0.20	0.841	-1.56922	1.922806
D3_nfeatures	-25.63685	77.73594	-0.33	0.742	-180.3954	129.1217
D3_ne	145.6077	276.0254	0.53	0.599	-404.0849	695.3002

```
D3_custom |  -310.4791   394.2583    -0.79   0.433    -1095.022    474.0637
 D3_lnage |   35.51389   252.4899     0.14   0.889    -467.0959    538.1236
 D3_lntax |  -205.2547   1406.123    -0.15   0.884    -3003.184    2592.675
```

The prefix `mi estimate` runs the specified `gmi` command on each imputed dataset to obtain a set of alternative estimates of the model parameters and their variance-covariance matrix. Multiple imputation estimates are then obtained by applying the combination rules of Rubin (1987) on the resulting set of alternative estimates (see `mi estimate`). Here, it is worth noticing that the prefix `mi estimate` has its own reporting output and does not respect the reporting output of the `gmi` command. As discussed in Section 3.3, p-values and confidence intervals must then be treated with caution especially when the number multiple imputations is small. Also notice that the option `full` of the `gmi` command is always needed to obtain valid information on the average relative variance increase (RVI) due to nonresponse and the summaries about parameter-specific degrees of freedom (DF). In any case, the prefix `mi estimate` and the option `full` of the `gmi` command cannot be jointly combined when using model reduction techniques (i.e. options `sw` and `vs`) because the subset of selected auxiliary regressors can vary across imputations.

6 Empirical application

This application investigates the relationship between hand grip strength (GS) and a set of socio-demographic and economic covariates using data on the elderly European population. As argued by Andersen et al. (2009), GS is an important measure of health because it is objectively measured, it directly affects every day activity functions, it is known to decline linearly with age, and it is a strong predictor of disability, morbidity, frailty and mortality. Furthermore, measuring GS is cheap and can be carried out by trained survey interviewers in non-clinical studies.

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 coordinated by the Mannheim Research Institute for the Economics of Aging (MEA).[2] SHARE collects data on self-reported and objective measures of health, socio-economic status, and social and family networks for nationally representative samples of elderly people in the participating countries. The first wave, conducted in 2004, covers about 28,500 individuals in 11 European countries (Austria, Belgium, Denmark, France, Germany, Greece, Italy, the Netherlands, Spain, Sweden and Switzerland). In each country, the target population consists of people aged 50 and older, plus their possibly younger

[2] Data can be downloaded free of charge from the SHARE Research Data Center: http://www.share-project.org. To get access to the dataset, researchers have to complete a statement concerning the use of the microdata.

partners. Although all national samples are selected through probability sampling, the sampling procedures are not completely standardized across countries because of the lack of a common sampling frame. The unweighted household nonresponse rate in the first wave ranges between a maximum of 62 percent in Switzerland and a minimum of 26 percent in France, and is equal to 45 percent on average. The data collection mode is computer-assisted personal interview (CAPI), supplemented by show-cards and a self-administered paper-and-pencil questionnaire.[3]

The GS test in SHARE involves two measurements on each hand (alternating between the two hands) using a hand-grip dynamometer. Before the fieldwork period, interviewers participated in centrally designed training sessions to learn a common protocol for measuring GS. They are instructed to demonstrate the use of the hand-grip dynamometer, help the respondent to assume a suitable position before performing the test, and verbally encourage the respondent to squeeze the handles of the hand-grip dynamometer as hard as possible. Respondents are excluded from the GS test only for swelling, inflammation, severe pain, recent injury or surgery to both hands in the last 6 months. For respondents with problems in one hand, the GS test is performed on the other hand only. The measurement of GS on each hand is considered valid if the two assessments on the same hand were greater than 0 Kg, lower than 100 Kg and they do not differ from each other by more than 20 kg. The overall GS test is considered valid if there is at least one valid measurement on one hand. Following Andersen et al. (2009), our dependent variable is the maximum GS ('maxgrip') measurement resulting from valid test.

Our set of socio-demographic and economic covariates includes age, gender, macro-region of residence (Northern, Central or Southern countries), self-reported weight and height, an indicator for educational attainment, per-capita household income and household net worth. To ensure cross-country comparability of the information on educational attainment, the original values have been recoded using the 1997 International Standard Classification of Education (ISCED-97). For similar reasons, per-capita household income and household net worth have been adjusted for the differences in purchasing power across countries. Thus, nominal amounts have been divided by the national purchasing power parity to obtain real amounts denominated in German prices for the year 2005.[4]

Unlike Andersen et al. (2009), who use imputed values of household income and household net worth by relying on the estimates from the 'naive' approach, we are interested in investigating

[3] For additional information on survey design and response rates see Börsch-Supan et al. (2005).

[4] Data for these calculations are obtained from the purchasing power parity survey carried out by OECD in 2005. Further information can be found in the SHARE documentation of release 2.4.0.

the trade-off between bias and precision when replacing the missing values on these two variables with imputations. This is an important issue to consider because these covariates are affected by substantial item nonresponse. The item nonresponse rates for household income and household net worth range, respectively, between a maximum of 76 and 77 percent in Belgium and a minimum of 49 and 52 percent in Greece, and are equal to 62 and 64 percent on average. The substantial amount of item nonresponse reflects three problems. First, these variables are not asked directly to respondents but are obtained by aggregating a large number of income and wealth components (27 and 13 respectively), collected at both the individual and the household level. Second, information about incomes, real and financial assets, mortgage 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.

To deal with the potential selectivity effects generated by item nonresponse, the public-use SHARE data include 5 multiple imputations of the key survey variables. As discussed at length in Christelis (2011), these imputations are constructed by the multivariate iterative procedure of Buuren et al (2006) which attempts to preserve the correlation structure of the imputed data. In what follows, we account for the additional sampling variability induced by the imputation process using the combination rules proposed by Rubin (1987) on the 5 multiple imputations of household income and household net worth.

Another important difference with respect to Andersen et al. (2009) is that we focus on respondents aged between 50 and 80 years who do not report serious health problems. This choice is primarily motivated by the need of compensating for cross country differences in coverage of the institutionalized target population. While the sampling frames of Denmark, Netherlands and Sweden cover people living in institutions for the elderly, this segment of the 50+ population is excluded by the national sampling frames of the other SHARE countries. Moreover, Southern European countries are known to have fewer nursing home than Northern and Continental European countries and a cultural tradition of old parents living with a child. To limit the impact of these cross country differences, we select respondents who have at most one limitation with activities of daily living, at most one chronic disease, and whose self-reported health status is at least fair. After applying this sample selection criterion, dropping the invalid measurements of maxgrip (about 5 percent of the cases) and the few missing data on weight, height and education (about 1 percent of

the cases), our working sample consists of 13,724 observations. Summary statistics for the outcome and the covariates are presented in Table 1, separately by gender and macro-region.

Given the high level of comparability of the SHARE data, we pool data from countries in the same macro-region, and estimate our linear regression model of interest separately by gender and macro-region. For simplicity, we assume that the errors in the grand model are independent and spherically distributed. The model specification in each subgroup includes 7 focus regressors, of which 5 (age, weight, height, education and the constant term) are observed and 2 (per-capita household income and household net worth) are imputed, 3 subsamples with incomplete data, and 21 non-collinear auxiliary variables. The resulting dimension of the model space is 2,097,152. After centering the focus covariates on the corresponding medians of each subgroup, we compare the estimates from five alternative approaches: complete case, 'naive', model reduction, BMA and WALS. Model reduction estimation is carried out using the `vs` estimation option of the `gmi` command with leaps-and-bounds selection and AIC as model information criteria, while WALS estimation is carried out using a Subbotin prior with parameter $q = 0.5$.[5]

The estimated coefficients and their standard errors are presented in Tables 2 and 3, separately by gender and macro-region. Qualitatively, our results are consistent with the empirical findings in Andersen et al. (2009). In all specifications, maxgrip is negatively related to age and positively related to self-reported weight and height. Women have a lower level of maxgrip than men, but they also present a considerably flatter decline with advancing age. The positive gradient between Northern-Continental and Southern countries persists even after focusing on the healthier segment of the elderly European population. For men, the age-related decline in maxgrip is steeper for those living in Southern countries. For women, it is instead steeper for those living in Northern and Continental countries. Education, per-capita household income and household net worth do not seem to be robustly correlated with maxgrip. The only exceptions are the positive correlations between maxgrip and education for men and women living in Continental countries, between maxgrip and per-capita household income for women living in Southern countries, and between maxgrip and household net worth for men and women living in Southern countries.

Notice that, although there is broad agreement with previous studies on the sign of the estimated associations, their magnitude and the size of the standard errors are subject to non-negligible differences. For example, the point estimate of the coefficient on weight in the specification Male-

[5] Estimates from the simple missing-data indicator approach are omitted because similar to those obtained from the 'naive' approach. Estimates from WALS with a Laplace prior are instead omitted because very similar to those obtained with a Subbotin prior.

North ranges between a minimum 0.106 with a standard error of 0.022 using the 'naive' approach to a maximum of 0.214 with a standard error of 0.057 using complete-case analysis. Similar differences are observed for the estimated coefficients on education in the specifications Male-North and Male-Center, household net worth in the specification Male-South, weight in the specification Female-Center, and per-capita household income in the specification Female-South. The estimates from model reduction and model averaging are somewhat in-between the estimates from the complete-case and the 'naive' approach. In particular, the conditional estimates from model reduction are quite close to the unconditional estimates from BMA. This suggests that, in this example, the effects of pretesting are not very important. The differences in the unconditional estimates from BMA and WALS suggest instead that alternative assumptions on the prior distributions for the auxiliary parameters may matter. From this view point, WALS has the advantage of using priors that ensure bounded risk and a coherent treatment of ignorance about the auxiliary parameters.

References

Andersen-Ranberg K., I. Petersen, H. Frederiksen, J. P. Mackenbach, and K. Christensen (2009), "Cross-National Differences in Grip Strength among 50+ Year-Old Europeans: Results from the SHARE Study", *European Journal of Ageing*, 6: 227-236.

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

Buuren van S., J.P.L. Brand, C.G.M. Groothuis-Oudshoorn, D.B. Rubin (2006), "Fully Conditional Specification in Multivariate Imputation", *Journal of Statistical Computation and Simulation*, 76: 1049-1064.

Christelis D. (2011), "Imputation of Missing Data in Waves 1 and 2 of SHARE", *SHARE Working Paper Series*, 1–50.

Dardanoni V., S. Modica, and F. Peracchi (2011), "Regression with Imputed Covariates: A Generalized Missing Indicator Approach", *Journal of Econometrics*, doi: 10.1016/j.jeconom.2011.02.005.

De Luca G. and J.R. Magnus (2011), "Bayesian Model Averaging and Weighted Average Least Squares Estimators", *mimeo*.

Einmahl J.H.J., K. Kumar, and J.R. Magnus (2011), "On the Choice of Prior in Bayesian Model Averaging", Technical report, Tilburg University.

Horton N.J. and K.P. Kleinman (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.

Lindsey C. and S. Sheather (2010), "Variable Selection in Linear Regression", *The Stata Journal*, 10: 650–669.

Little R.J.A. (1992), "Regression with Missing X's: A Review", *Journal of the American Statistical Association*, 87: 1227–1237.

Lundström S. and C.E. Särndal (2002), "Estimation in the Presence of Nonresponse and Frame Imperfections", *Statistics Sweden*.

Magnus J.R. (1999), "The Traditional Pretest Estimator", *Theory of Probability and Its Applications*, 44: 293–308.

Magnus J.R. (2002), "Estimation of the Mean of a Univariate Normal Distribution with Known Variance", *Econometrics Journal*, 5: 225–236.

Magnus J.R., O. Powell, and P. Prüfer (2010), "A Comparison of Two Averaging Techniques With an Application to Growth Empirics", *Journal of Econometrics*, 154: 139–153.

Magnus J.R., A.T.K. Wan, and X. Zhang (2011), "Weighted Average Least Squares Estimation with Nonspherical Disturbances and an Application to the Hong Kong Housing Market", *Computational Statistics & Data Analysis*, 55: 1331–1341.

Raftery A.E., D. Madigan, and J.A. Hoeting (1997), "Bayesian Model Averaging for Linear Regression Models", *Journal of the American Statistical Association*, 92: 179–191.

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

Table 1: Descriptive statistics for the dependent and the independent variables (weight is in kilograms, height is in centimeters, PPP-adjusted per-capita household income is in 10,000 Euro and household net worth is 100,000 Euro).

		Male			Female		
Region	Variable	P25	Med	P75	P25	Med	P75
North	maxgrip	42.00	49.00	55.00	25.00	29.00	33.00
	age	55.00	60.00	67.00	55.00	59.00	66.00
	weight	74.00	81.00	90.00	60.00	66.00	74.00
	height	174.00	178.00	183.00	161.00	165.00	169.00
	education	0.00	1.00	1.00	0.00	1.00	1.00
	income	1.56	2.33	3.32	1.52	2.25	3.15
	net worth	0.54	1.42	2.77	0.42	1.25	2.56
	complete obs.			204			238
	imputed obs.			1123			1203
Center	maxgrip	41.00	47.00	53.00	25.00	30.00	34.00
	age	55.00	60.00	67.00	54.00	59.00	66.00
	weight	73.00	80.00	88.00	60.00	67.00	75.00
	height	171.00	176.00	180.00	160.00	164.00	168.00
	education	0.00	1.00	1.00	0.00	1.00	1.00
	income	1.12	1.84	3.08	1.11	1.80	3.07
	net worth	0.84	2.20	4.08	0.61	2.01	3.88
	complete obs.			730			799
	imputed obs.			3798			4057
South	maxgrip	35.00	43.00	50.00	22.00	26.00	30.00
	age	55.00	60.00	68.00	53.00	58.00	65.00
	weight	72.00	79.00	85.00	60.00	66.00	75.00
	height	167.00	170.00	175.00	157.00	161.00	165.00
	education	0.00	0.00	1.00	0.00	0.00	1.00
	income	0.56	0.94	1.60	0.54	0.93	1.64
	net worth	0.88	1.72	3.29	0.83	1.62	3.14
	complete obs.			485			470
	imputed obs.			1785			1758

Table 2: Estimated coefficients and standard errors (in parentheses) for males by macro-region. Estimation is based on $M = 5$ multiple imputations for income and net worth. Results for the auxiliary regressors are omitted to save space.

Region	Variable	CC	NAIVE	VS	BMA	WALS
North	constant	49.758	49.236	49.429	49.228	49.485
		(1.015)	(0.414)	(0.437)	(0.424)	(0.830)
	age	-0.410	-0.446	-0.442	-0.444	-0.423
		(0.067)	(0.031)	(0.031)	(0.032)	(0.059)
	weight	0.214	0.106	0.111	0.108	0.166
		(0.057)	(0.022)	(0.022)	(0.022)	(0.049)
	height	0.265	0.265	0.256	0.266	0.267
		(0.101)	(0.040)	(0.040)	(0.043)	(0.087)
	education	-2.595	-1.075	-1.158	-1.091	-1.893
		(1.161)	(0.496)	(0.496)	(0.503)	(0.956)
	income	0.290	0.021	0.201	0.072	0.179
		(0.304)	(0.126)	(0.145)	(0.158)	(0.265)
	net worth	0.138	0.033	-0.002	0.029	0.087
		(0.174)	(0.043)	(0.045)	(0.046)	(0.135)
Center	constant	46.670	47.013	47.005	47.019	46.841
		(0.584)	(0.247)	(0.247)	(0.252)	(0.468)
	age	-0.382	-0.436	-0.437	-0.436	-0.407
		(0.041)	(0.017)	(0.017)	(0.018)	(0.031)
	weight	0.082	0.119	0.119	0.119	0.096
		(0.028)	(0.012)	(0.012)	(0.013)	(0.025)
	height	0.252	0.209	0.208	0.209	0.237
		(0.053)	(0.022)	(0.022)	(0.022)	(0.048)
	education	1.694	0.779	1.112	0.813	1.277
		(0.686)	(0.291)	(0.334)	(0.313)	(0.550)
	income	0.014	0.045	0.048	0.046	0.030
		(0.132)	(0.059)	(0.059)	(0.061)	(0.100)
	net worth	0.063	0.012	0.017	0.013	0.038
		(0.061)	(0.015)	(0.016)	(0.016)	(0.045)
South	constant	42.006	42.670	42.391	42.553	42.295
		(0.583)	(0.286)	(0.352)	(0.329)	(0.474)
	age	-0.560	-0.536	-0.587	-0.539	-0.552
		(0.055)	(0.028)	(0.036)	(0.032)	(0.045)
	weight	0.105	0.113	0.114	0.113	0.105
		(0.039)	(0.021)	(0.021)	(0.021)	(0.031)
	height	0.245	0.226	0.226	0.225	0.236
		(0.068)	(0.034)	(0.034)	(0.035)	(0.054)
	education	0.646	0.193	0.395	0.184	0.409
		(0.966)	(0.466)	(0.486)	(0.489)	(0.781)
	income	-0.266	0.270	-0.098	0.207	-0.053
		(0.331)	(0.159)	(0.216)	(0.210)	(0.291)
	net worth	0.248	0.022	0.216	0.049	0.175
		(0.098)	(0.025)	(0.088)	(0.082)	(0.074)

Table 3: Estimated coefficients and standard errors (in parentheses) for females by macro-region. Estimation is based on $M = 5$ multiple imputations for income and net worth. Results for the auxiliary regressors are omitted to save space.

Region	Variable	CC	NAIVE	VS	BMA	WALS
North	constant	28.805	29.170	29.141	29.161	28.986
		(0.654)	(0.288)	(0.287)	(0.291)	(0.511)
	age	-0.284	-0.259	-0.255	-0.259	-0.271
		(0.051)	(0.022)	(0.022)	(0.023)	(0.040)
	weight	0.070	0.067	0.077	0.067	0.068
		(0.033)	(0.016)	(0.017)	(0.017)	(0.026)
	height	0.250	0.250	0.281	0.251	0.247
		(0.067)	(0.030)	(0.033)	(0.039)	(0.052)
	education	0.147	-0.028	-0.000	-0.023	0.055
		(0.781)	(0.353)	(0.352)	(0.358)	(0.611)
	income	-0.130	0.117	0.129	0.116	-0.006
		(0.371)	(0.108)	(0.108)	(0.111)	(0.284)
	net worth	0.062	-0.003	0.005	-0.001	0.031
		(0.109)	(0.036)	(0.043)	(0.040)	(0.083)
Center	constant	29.449	29.429	29.291	29.338	29.446
		(0.376)	(0.156)	(0.161)	(0.186)	(0.267)
	age	-0.303	-0.262	-0.259	-0.261	-0.284
		(0.030)	(0.012)	(0.012)	(0.013)	(0.024)
	weight	0.091	0.070	0.092	0.070	0.080
		(0.020)	(0.008)	(0.013)	(0.010)	(0.016)
	height	0.200	0.227	0.244	0.236	0.213
		(0.037)	(0.016)	(0.017)	(0.020)	(0.029)
	education	0.807	0.823	0.810	0.803	0.822
		(0.476)	(0.199)	(0.198)	(0.208)	(0.340)
	income	0.116	0.028	0.046	0.033	0.078
		(0.124)	(0.039)	(0.055)	(0.042)	(0.094)
	net worth	0.003	0.004	0.005	0.005	0.008
		(0.040)	(0.010)	(0.010)	(0.010)	(0.028)
South	constant	25.245	25.859	25.630	25.845	25.496
		(0.407)	(0.194)	(0.234)	(0.205)	(0.355)
	age	-0.219	-0.237	-0.236	-0.237	-0.227
		(0.039)	(0.020)	(0.020)	(0.020)	(0.031)
	weight	0.046	0.036	0.036	0.036	0.042
		(0.028)	(0.014)	(0.014)	(0.014)	(0.022)
	height	0.142	0.181	0.180	0.180	0.160
		(0.052)	(0.025)	(0.025)	(0.027)	(0.039)
	education	1.264	0.401	0.399	0.406	0.909
		(0.715)	(0.342)	(0.343)	(0.348)	(0.604)
	income	0.249	0.235	0.223	0.234	0.227
		(0.267)	(0.102)	(0.103)	(0.105)	(0.200)
	net worth	0.069	0.047	0.031	0.047	0.064
		(0.071)	(0.024)	(0.026)	(0.025)	(0.054)