The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
1-year subscription	$ 79	1-year subscription	$115
2-year subscription	$155	2-year subscription	$225
3-year subscription	$225	3-year subscription	$329
3-year subscription (electronic only)	$210	3-year subscription (electronic only)	$210
1-year student subscription	$ 48	1-year student subscription	$ 79
1-year university library subscription	$ 99	1-year university library subscription	$135
2-year university library subscription	$195	2-year university library subscription	$265
3-year university library subscription	$289	3-year university library subscription	$395
1-year institutional subscription	$225	1-year institutional subscription	$259
2-year institutional subscription	$445	2-year institutional subscription	$510
3-year institutional subscription	$650	3-year institutional subscription	$750

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The Stata Journal (2012)
12, Number 4, pp. 575–604

A generalized missing-indicator approach to regression with imputed covariates

Valentino Dardanoni
University of Palermo
Palermo, Italy

Giuseppe De Luca
University of Palermo
Palermo, Italy
delucgius@yahoo.it

Salvatore Modica
University of Palermo
Palermo, Italy

Franco Peracchi
Tor Vergata University and EIEF
Rome, Italy

Abstract. We consider estimation of a linear regression model using data where some covariate values are missing but imputations are available to fill in the missing values. This situation generates a tradeoff between bias and precision when estimating the regression parameters of interest. Using only the subsample of complete observations does not cause bias but may imply a substantial loss of precision because the complete cases may be too few. On the other hand, filling in the missing values with imputations may cause bias. We provide the new Stata command `gmi`, which handles such tradeoff by using either model reduction or Bayesian model averaging techniques in the context of the generalized missing-indicator approach recently proposed by Dardanoni, Modica, and Peracchi (2011, *Journal of Econometrics* 162: 362–368). If multiple imputations are available, `gmi` can also be combined with the built-in Stata prefix `mi estimate` to account for extra variability due to imputation. We illustrate the use of `gmi` with an empirical application in the health domain, where item nonresponse is substantial.

Keywords: st0273, gmi, missing covariates, imputation, bias–precision tradeoff, model reduction, model averaging

1 Introduction

In applied regression analysis, the values of some covariates are often missing for some observations. We focus on the case when the outcome of interest is always observed, and the missing-data mechanism satisfies a conditional independence assumption that we will make precise in section 2. This case has been studied extensively, starting with the seminal work of Little (1992). Our novel contribution is to consider the situation when imputations are available for the missing covariate values. This situation is becoming quite common because public data files increasingly include imputations of key variables affected by missing-data problems. Specialized software for carrying out imputations directly, such as the `mi` suite of commands in Stata, is also becoming increasingly available.

One approach to this problem—complete-case analysis—drops all observations with missing covariate values, thus ignoring the imputations altogether. An alternative approach uses all the observations without distinguishing between observed and imputed values. We call this the "filling-in approach" because the missing values are simply filled in with the imputations. A variant of this approach—the so-called simple missing-indicator approach—adds a set of indicators to the covariates for the different patterns of missing data.

From the viewpoint of inference about the regression parameter of interest, the availability of imputations generates a tradeoff between bias and precision: the complete cases are often too few, so precision is lost, but just filling in the missing values with the imputations may lead to bias (Jones 1996). In this article, we present a Stata command that handles the tradeoff by implementing the "generalized missing-indicator approach" proposed by Dardanoni, Modica, and Peracchi (2011), henceforth DMP. Their approach exploits the fact that complete-case analysis and the filling-in approach correspond to using two extreme versions of the same model. Complete-case analysis corresponds to using a "grand model" that includes two subsets of regressors: 1) the focus regressors consisting of the observed or imputed covariates; and 2) a set of auxiliary regressors consisting of the missing-data indicators (as in the simple missing-indicator approach) and their interactions with the covariates. The filling-in approach corresponds to using a restricted version of the grand model that includes only the focus regressors.

The key idea of the DMP approach is to also consider all the intermediate models between these two extremes, namely, all models obtained from the grand model by dropping alternative subsets of auxiliary regressors. Expanding the model space in this way has two advantages. First, the original tradeoff between bias and precision is transformed into a problem of model uncertainty, for which a variety of alternative strategies is available. Second, any intermediate model in the expanded model space may now play a role in dealing with the tradeoff between bias and precision.

The Stata command `gmi` (acronym for generalized missing indicator) presented in this article implements several methods corresponding to two alternative strategies: model reduction and Bayesian model averaging. In general, these methods may be regarded as providing a compromise that avoids dropping the incomplete cases while using the available imputations in a sensible way. The extreme choice of using either the complete-case or the filling-in approach is still available but is unlikely to emerge as the best one. Bayesian model averaging avoids the pretesting problem that plagues model reduction techniques. It also allows one to formally incorporate, through the choice of priors, the researcher's beliefs on the reliability of the imputations—on which the estimates must ultimately depend.

The remainder of this article is organized as follows. In section 2, we review theoretical background. In section 3, we describe the two alternative strategies for estimating the regression parameters of interest: model reduction and Bayesian model averaging. Section 4 provides a detailed description of the `gmi` command, and section 5 illustrates `gmi` using data available on the Stata website. In section 6, we use data from the first wave of the Survey of Health, Ageing, and Retirement in Europe (SHARE) to provide an empirical application on the relationship between an objective health indicator and a set of sociodemographic and economic covariates affected by substantial item nonresponse. We conclude the article in section 7.

2 Background

Consider modeling the relationship between an outcome Y and a set of covariates $\mathbf{X}$ using data where some covariate values are missing. We assume that in the absence of these values, the data would satisfy the classical linear model

$$Y = \mathbf{X}\beta + U$$

where Y is the $N \times 1$ vector of observations on the outcome of interest, $\mathbf{X}$ is an $N \times K$ matrix of observations on the covariates, β is the $K \times 1$ vector of regression parameters, and U is an $N \times 1$ vector of regression errors that are homoskedastic, serially uncorrelated, and have zero mean conditional on $\mathbf{X}$. This means that the full-information estimator—the unfeasible ordinary least-squares (OLS) estimator from the regression of Y on $\mathbf{X}$—is unbiased for β 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 constructed by the researcher by using, for example, the Stata command `mi impute`.

Because the first element of $\mathbf{X}$ is considered 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 dataset need not contain all the possible patterns, so we simply index the patterns present in the data by $j = 0, \ldots, J$, 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 $\mathbf{M}$, whose (n, k)th element is equal to 1 if the nth case has a missing value on the kth covariate and is equal to 0 otherwise.

We are concerned with the problem of how to combine the observed and the imputed values to estimate the regression parameter β. We shall introduce the generalized missing-indicator approach starting with two building blocks of the theory: complete-case analysis and the filling-in approach. The results in this section are taken from DMP (2011), where proofs can be found.

2.1 Complete-case analysis

This approach ignores the imputed values and uses only the subsample with complete data, denoted by $[\mathbf{X}^0, Y^0]$, where $\mathbf{X}^0$ is an $N_0 \times K$ matrix and Y^0 is an $N_0 \times 1$ vector. Complete-case analysis is a benchmark because, under two key conditions, it delivers an unbiased estimate of the regression parameter β.

The two key conditions are full rank of $\mathbf{X}^0$ and a conditional independence assumption on the missing-data process.

Assumption 2.1 $\mathbf{X}^0$ *has full column rank.*

Assumption 2.2 $\mathbf{M}$ *and* Y *are independent conditional on* $\mathbf{X}$*.*

Assumption 2.1 implies that the complete-case estimator, the OLS estimator from the regression of Y^0 on $\mathbf{X}^0$, exists and is unique. For this assumption to hold, there must be enough cases (at least K) without missing covariate values. Assumption 2.2 says that given the true values of the covariates, the pattern of missing data can be ignored when predicting Y. This assumption is different from the standard missing-at-random (MAR) assumption, which in our setting would require the missing-data process to be independent of the missing covariates given the observed outcome and the nonmissing covariates.

A simple example where assumption 2.2 is satisfied but MAR is not is when health is the outcome of interest, income is the only regressor, and missing income depends on true income but not on health. On the other hand, an example where MAR is satisfied but assumption 2.2 is not is when missing income depends on health but not on true income. Thus assumption 2.2 admits patterns where cases with low or high levels of income systematically have greater percentages of missing values, but the assumption fails if the health–income relationship is different for observations with and without missing income values.

Under assumptions 2.1 and 2.2, we have the following result, which represents the main justification for complete-case analysis:

Result 1 *If assumptions 2.1 and 2.2 hold, then the complete-case estimator from a regression of* Y^0 *on* $\mathbf{X}^0$ *is unbiased for* β*.*

Even if unbiased, the complete-case estimator has the drawback of being less precise than the full-information estimator except when the fraction of complete cases is close to 1. In the rest of this section, we review alternative uses of the observations with missing data.

2.2 The filling-in and the simple missing-indicator approaches

A common alternative to complete-case analysis is to use all cases and regress Y on the completed design matrix $\mathbf{W}$, whose (n, k)th element is equal to the corresponding element of $\mathbf{X}$ if a covariate value is not missing and is equal to the imputed value otherwise. This approach, which we call the filling-in approach, ignores that the imputations are not the same as the missing covariate values; thus it gives an estimator of β that is more precise than the complete-case estimator, but it may also be biased if MAR is not satisfied or the imputation model is not "congenial" in the sense of Meng (1994).

Another alternative, often called the simple missing-indicator approach, consists of regressing Y on the completed design matrix $\mathbf{W}$ and a set of J indicators, $D_1, \ldots, D_J$, where the elements of D_j are equal to 1 for cases that belong to the jth missing-data pattern and are equal to 0 otherwise (the subsample with complete cases represents the baseline). Adding the indicators for the missing-data patterns allows the intercepts to differ across patterns but not across the other coefficients. This increases the flexibility of the model but does not guarantee unbiasedness (Little 1992; Jones 1996; Horton and Kleinman 2007).

2.3 The generalized missing-indicator approach

The problem with complete-case analysis is that one may end up with too few observations. On the other hand, the filling-in approach ignores that the population regression of Y on the completed design matrix $\mathbf{W}$ may differ across missing-data patterns and that all of these regressions may be different from the full-information regression of Y on $\mathbf{X}$. The aim of the generalized missing-indicator approach is to account for these differences.

The intuition is the following. By introducing a set of indicators for the missing-data patterns, one only controls for differences in the intercepts of these regressions. But if one adds enough auxiliary regressors to also control for differences in the slope coefficients, then one may hope to obtain an unbiased estimate of β, the regression parameter of interest. This is precisely what our grand model does. In practice, a one-to-one relation exists between the auxiliary regressors included in the grand model and the subsets of imputed missing values. The grand model coincides with the model used in complete-case analysis; and excluding some auxiliary variables from the grand model is equivalent to assuming that for some subsamples of imputed missing values, there is no difference in the regression coefficients of interest. If we exclude all auxiliary variables from the grand model, then one obtains the same model used in the filling-in approach.

The formal result is as follows. Let Y^j and $\mathbf{W}^j$, respectively, denote the $N_j \times 1$ subvector of Y and the $N_j \times K$ submatrix of $\mathbf{W}$ corresponding to the jth missing-data pattern. The generalized missing-indicator approach is based on the grand model

$$\begin{bmatrix} Y^0 \\ Y^1 \\ \vdots \\ Y^J \end{bmatrix} = \begin{bmatrix} \mathbf{X}^0 \\ \mathbf{W}^1 \\ \vdots \\ \mathbf{W}^J \end{bmatrix} \beta + \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{W}^1 & & \\ & \ddots & \\ & & \mathbf{W}^J \end{bmatrix} \begin{bmatrix} \delta^1 \\ \vdots \\ \delta^J \end{bmatrix} + \begin{bmatrix} U^0 \\ V^1 \\ \vdots \\ V^J \end{bmatrix}$$

where β is the regression parameter of interest, the δ^j are $K \times 1$ vectors of nuisance parameters that may be interpreted as the asymptotic bias in the regression of Y^j on $\mathbf{W}^j$, and the V^j are $N_j \times 1$ vectors of projection errors that have mean zero and are orthogonal to the columns of $\mathbf{W}^j$. A compact representation of the grand model is

$$Y = \mathbf{W}\beta + \mathbf{Z}\delta + V \tag{1}$$

where

$$\mathbf{W} = \begin{bmatrix} \mathbf{X}^0 \\ \mathbf{W}^1 \\ \vdots \\ \mathbf{W}^J \end{bmatrix}, \qquad \mathbf{Z} = \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{W}^1 & & \\ & \ddots & \\ & & \mathbf{W}^J \end{bmatrix}, \qquad \delta = \begin{bmatrix} \delta^1 \\ \vdots \\ \delta^J \end{bmatrix}, \qquad V = \begin{bmatrix} U^0 \\ V^1 \\ \vdots \\ V^J \end{bmatrix}$$

respectively, an $N \times K$ matrix of observed or imputed covariates, an $N \times JK$ matrix of auxiliary regressors, a $JK \times 1$ vector of nuisance parameters, and an $N \times 1$ error vector. The variables in matrix $\mathbf{Z}$ consist of JK interactions between the set of J indicators, $D_1, \ldots, D_J$, for the missing-data patterns and the K columns of the completed design matrix $\mathbf{W}$. This matrix is not required to have full column rank. This occurs when some of the $\mathbf{W}^j$ do not have full column rank, either because $N_j < K$ or because $N_j \geq K$, but the columns of $\mathbf{W}^j$ are linearly dependent, as when mean imputation or deterministic regression imputation is used. Incidentally, such imputation methods are known to produce datasets with undesirable properties (see, for example, Lundström and Särndal [2001]). When some of the $\mathbf{W}^j$ do not have full column rank, only a subset of the coefficients in δ^j is identifiable, but this does not affect the estimates of β. Additionally, regression errors in (1) need not have constant variance, because the projection errors $V^1, \ldots, V^J$ may be heteroskedastic.

The main result in DMP (2011) is the following:

Result 2 *If assumption 2.1 holds, then, for any choice of imputations, the* OLS *estimate of β in* (1) *equals the complete-case estimate of β.*

If assumption 2.2 holds, regressing Y on $\mathbf{W}$ and $\mathbf{Z}$ allows one to fully exploit the available information and to obtain an unbiased estimator of the regression parameter β. In addition, in this case, the filling-in approach gives an unbiased estimator of β only if the elements of δ—the coefficients on the auxiliary regressors—all equal 0.

3 Alternative strategies for estimating β

Both the filling-in and the simple missing-indicator approaches correspond to using restricted versions of (1) obtained by placing restrictions on the vector δ. The former restricts δ to equal 0; the latter restricts all the elements of δ to equal 0 except the first one. When these restrictions are at odds with the data, imposing them leads to an estimator of β that is biased but more precise than the OLS estimator of β in

(1), which, from result 2, is numerically the same as the complete-case estimator of β. This suggests that by placing restrictions on δ or, equivalently, by excluding some of the auxiliary regressors in $\mathbf{Z}$, one may obtain an estimator of β that is better in the sense of mean squared error (MSE) than the complete-case estimator. Our `gmi` command implements two alternative strategies for obtaining an estimator of β in this way: model reduction and model averaging.

3.1 Model reduction

Model reduction involves selecting first an intermediate model between the grand model (1) and the model corresponding to $\delta = 0$ and then estimating the parameter of interest β conditional on the selected model. Because the variables in the completed design matrix $\mathbf{W}$ 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 $\mathbf{Z}$.

The conceptually simplest and most transparent model reduction procedure is stepwise selection, through either backward elimination or forward selection. Backward elimination (general to specific) starts from the model that includes all the auxiliary regressors (the grand model) and drops them one at a time if their p-value is above a threshold chosen by the modeler. Forward selection (specific to general) starts from the model without auxiliary regressors and adds them one at a time if their p-value is below a chosen threshold.

In either case, the threshold on the p-value may reflect prior beliefs about the quality of the imputations: the more one trusts the imputations—that is, the less weight one wants to place on the auxiliary regressors—the lower one may set the threshold. Further, comparing the results obtained with different values of the threshold may give some indication about the quality of the available imputations. For example, stepwise results that are close to the estimates obtained from the filling-in approach even for high values of the threshold may be interpreted as favorable evidence for the quality of the imputations. An improvement over the standard stepwise procedure is the variable selection method recently introduced by Lindsey and Sheather (2010), where instead of a fixed significance level, an information criterion such as Akaike's information criterion (AIC) or the Bayesian information criterion is used to gauge each model.

One well-known problem with this strategy is pretesting.[1] Another is that model reduction and estimation are completely separated. Thus the reported conditional estimates tend to be interpreted as if they were unconditional. A third problem is that with 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 Lindsey and Sheather (2010), usually analyze a larger number of models.

[1] See Magnus (2000) and the FAQ titled *What are some of the problems with stepwise regression?*, available at http://www.stata.com/support/faqs/stat/stepwise.html.

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 β conditional on each model in the model space and then computes the estimate of β as a weighted average of these conditional estimates. When the model space contains I models, a model averaging estimate of β is of the form

$$\overline{\beta} = \sum_{i=1}^{I} \lambda_i \widehat{\beta}_i$$

where the λ_i are nonnegative random weights that add up to 1 and $\widehat{\beta}_i$ is the estimate of β obtained by conditioning on the ith model. In Bayesian model averaging (BMA), each $\widehat{\beta}_i$ 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 Y under model i. The BMA literature is vast, and we refer the reader to Raftery, Madigan, and Hoeting (1997) for a starting point.

Our `gmi` command implements two BMA procedures in the options `bma` and `wals`: standard BMA and weighted-average least squares (WALS). The 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 (1), noninformative priors for β and the error variance, and a multivariate Gaussian prior for δ. Notice that the computational burden required to obtain a standard BMA estimate is proportional to the dimension of the model space. Because this is equal to 2^{JK} in our case, the computational burden is substantial unless both J and K are small.

WALS was introduced by Magnus, Powell, and Prüfer (2010). It also assumes a classical Gaussian linear model for (1) and noninformative priors for β and the error variance. However, instead of a multivariate Gaussian prior for δ, it uses a distribution with zero mean for the independent and identically distributed elements of the transformed parameter vector $\eta = \eta(\delta)$, whose hth element is the population t statistic for testing the significance of the hth element of δ. Magnus, Powell, and Prüfer (2010) use the Laplace distribution, while Kumar and Magnus (2011) use the Subbotin family, which leads to estimators with better asymptotic properties. Our Stata implementation for both Laplace and Subbotin priors is again based on the `wals` command provided by De Luca and Magnus (2011). The assumption that the regression errors in (1) are homoskedastic and serially uncorrelated is not crucial for WALS, and the method can be generalized to nonspherical errors (Magnus, Wan, and Zhang 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 regressors. Third, WALS estimates have bounded risk and are near optimal in terms of a well-defined regret criterion (Magnus, Powell, and Prüfer 2010).

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 robustly correlated with the outcome if the t ratio on its coefficient is greater than 1 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, for example, Magnus [2002]). On the basis of this criterion, we also provide one-standard-error bands for the estimated coefficients.

Computation and interpretation of the standard errors differ depending on the estimation strategy (model reduction versus model averaging) and the general approach to estimation (frequentist versus 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 `bootstrap` option gives standard errors based on the wild bootstrap that are valid under conditional heteroskedasticity and also consider the additional variability due to model selection.

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

Neither model reduction nor model averaging considers 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, the `gmi` command can be combined with the built-in Stata prefix `mi estimate` (see [MI] **mi estimate**).

4 The gmi command

The new Stata command `gmi` handles the tradeoff between bias and precision when fitting a classical linear regression model with imputed covariates. The earliest version of Stata required to run this command is version 11.1.

4.1 Syntax

The syntax for the `gmi` command is

```
gmi depvar [varlist] [if] [in], impvar(varlist) misind(varlist) [summarize cc
   fi smi sw vs bma wals full vce(bootstrap [, bootstrap_options])
   mimputations auxiliary(string) keep nowarn stepwise_options vselect_options
   bma_options wals_options mi_options]
```

where *depvar* is the dependent variable; *varlist* is an optional list of observed covariates (covariates whose values are fully observed); `impvar()` is the list of imputed covariates (covariates whose missing values are replaced by imputed values); and `misind()` is the relevant list of missing-data indicators (the nonzero columns of the matrix **M** corresponding to the set of imputed covariates). Missing-data indicators take on the value 0 for observed cases and the value 1 for imputed cases. The number of imputed covariates must coincide with the number of missing-data indicators. The first variable in `impvar()` is paired with the first indicator in `misind()`, the second variable in `impvar()` is paired with the second indicator in `misind()`, and so on.

The constant term (which is always included) and the set of observed and imputed covariates correspond to the K columns of the completed design matrix $\mathbf{W}$. The auxiliary regressors in $\mathbf{Z}$ (the JK interactions between the J indicators for the missing-data patterns and the K columns of $\mathbf{W}$) are instead automatically generated by the command with the information from `misind()`. `misind()` and `impvar()` are required. 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. Descriptions of the options specific to this command are provided in the next sections.

4.2 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) and summaries of the distribution of *depvar* (number of observations, mean, and standard deviation) for the complete-case estimate and each missing-data pattern.

`cc` provides the complete-case estimate of β, the OLS estimate from a regression of *depvar* on the K focus regressors in $\mathbf{W}$ using only the complete cases. From result 2, this is numerically the same as the OLS estimate of β in the grand model (1).

`fi` provides the filling-in estimate of β, the OLS estimate from a regression of *depvar* on the K focus regressors in $\mathbf{W}$ using all cases.

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

`sw` provides the OLS estimate of β from a regression of *depvar* on the K focus regressors in $\mathbf{W}$ and the subset of auxiliary regressors in $\mathbf{Z}$ selected through the built-in Stata command `stepwise`. This estimate of β is conditional on the selected model. A brief description of the options for the `stepwise` command is given in section 4.3.

`vs` provides the OLS estimate of β from a regression of *depvar* on the K focus regressors in $\mathbf{W}$ and the subset of auxiliary regressors in $\mathbf{Z}$ selected through the `vselect` command by Lindsey and Sheather (2010). Similarly to the `sw` option, this estimate of β is conditional on the selected model. A brief description of the options for the `vselect` command is given in section 4.4.

`bma` provides the BMA estimate of β in the grand model (1) using the `bma` command implemented by De Luca and Magnus (2011). This option assumes a classical Gaussian linear model for (1), noninformative priors for the regression parameter β and the error variance, and a multivariate Gaussian prior for the auxiliary parameter δ. This estimate is obtained as a weighted average of the estimates of β 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 4.5.

`wals` provides the WALS estimate of β in the grand model (1) using the `wals` command implemented by De Luca and Magnus (2011). Like `bma`, this option assumes a classical Gaussian linear model for (1) and noninformative priors for the regression parameter β 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 in η are assumed to be independent and identically distributed according to either a Laplace or a Subbotin prior. A brief description of the options for the `wals` command is given in section 4.6.

`full` displays the estimation results for all model parameters (focus and auxiliary parameters) and returns 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 for the focus and the auxiliary parameters, one must combine the option `vce(bootstrap)` with the option `full`. In any case, `vce(bootstrap)` and `full` cannot be jointly specified when applying model reduction techniques (the options `sw` and `vs`), because the subset of selected regressor variables can vary across bootstrap replicates. Furthermore, the option `vce(bootstrap)` cannot be combined with the option `mimputations`. Standard options for bootstrap estimation can be specified as suboptions within `vce(bootstrap)` (see [R] ***vce_option***).

`mimputations` runs the specified `gmi` command on multiply imputed data by using the built-in Stata prefix `mi estimate` (see [MI] **mi estimate**). By default, multiple-imputation estimates are computed only for the focus parameters. To obtain multiple-imputation estimates of the focus and the auxiliary parameters, one must combine the option `mimputations` with the option `full`. In any case, `mimputations` and `full` cannot be jointly specified when applying model reduction techniques (the options `sw` and `vs`), because the subset of selected auxiliary regressors may vary across imputations. Moreover, the option `mimputations` cannot be combined with the options `cc` and `vce(bootstrap)`. A brief description of the options for `mi estimate` is given in section 4.7.

`auxiliary(`*string*`)` specifies the prefix for the name of the auxiliary regressors. The default is `auxiliary(D)`. Thus auxiliary regressors are named `D`*j* and `D`*j_varname*, where $j = 1, \dots, J$ is an index for the subsamples of missing data and *varname* is the name of each variable listed in *varlist* and `impvar()`.

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

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

4.3 Options for stepwise

With the `sw` option, `gmi` carries out model reduction through the built-in Stata command `stepwise` (see [R] **stepwise** for details). The relevant options of the `stepwise` command are `pr(`#`)` (significance level for backward elimination), `pe(`#`)` (significance level for forward selection), `forward` (backward stepwise), and `lr` (likelihood-ratio test-of-term significance). Because the auxiliary regressors in **Z** have no hierarchical ordering, backward hierarchical selection and forward hierarchical selection are not allowed.

4.4 Options for vselect

With the `vs` option, `gmi` carries out model reduction through the `vselect` command provided by Lindsey and Sheather (2010). This command offers three model reduction techniques: backward elimination (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` (Bayesian information criterion), `cp1`, or `cp2` (Mallows'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 0 (cp1) or a value close to the number of covariates (cp2). For additional information, see Lindsey and Sheather (2010).

4.5 Options for BMA

With the bma option, gmi carries out standard BMA through the bma command provided by De Luca and Magnus (2011). In this case, one can use the option nodots to suppress the display of the dots that 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 fit.

4.6 Options for WALS

With the wals option, gmi carries out BMA through the wals command provided by De Luca and Magnus (2011). As for the prior on the transformed auxiliary parameters, one can choose between Laplace and Subbotin priors through the option q(#). This option defines the free parameter $0 < q \leq 1$ of a Subbotin density for the elements η_h of the transformed parameter vector η. This density is standardized to have a prior median of η_h equal to 0 and a prior median of η_h^2 equal to 1. The default is $q = 1$, corresponding to the Laplace prior. Values of q in the interval $(0, 1)$ give instead a class of Subbotin priors. Kumar and Magnus (2011) argue that values of q close to 0 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 $q \neq 0.5$, one can also specify a set of additional options (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).

4.7 Options for multiple imputations

With the mimputations option, gmi computes multiple-imputation estimates through the built-in Stata prefix mi estimate (see [MI] **mi estimate** for details). One can specify these options with mi estimate: nimputations(#), imputations(*numlist*), saving(*miestfile*[, replace]), vartable, noisily, trace, esample(*newvar*), and dots. The remaining options are suppressed because they can be inappropriate for most of the estimation methods implemented by the gmi command. Furthermore, we forced the built-in Stata prefix mi estimate to respect the reporting output of the gmi command to avoid misleading interpretations of the estimation results.

5 Examples

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

```
. quietly use "http://www.stata-press.com/data/r11/mhouses1993s30"
. describe
Contains data from http://www.stata-press.com/data/r11/mhouses1993s30.dta
  obs:         1,647                            Albuquerque Home Prices
                                                  Feb15-Apr30, 1993
 vars:            13                            19 Jun 2009 10:50
 size:        47,763                            (_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
. set seed 1234
```

We want to fit a classical linear regression model for the relationship between home sale price (price) and home characteristics (sqft, nfeatures, ne, custom, corner, lnage, and lntax). Because 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 by using a multivariate normal regression model (see [MI] **mi impute mvn**).

```
. mi describe
  Style:  mlong
          last mi update 19jun2009 10:50:22, 386 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
                                                         Obs<.
                                                +------------------------------
               |                                | Unique
      Variable |     Obs=.     Obs>.     Obs<.  | values        Min         Max
  -------------+--------------------------------+------------------------------
         lnage |        49                  68  |     30          0    3.970292
         lntax |        10                 107  |     95   5.407172    7.475906
  -----------------------------------------------------------------------------
```

Thus the data contain 117 observations and 30 multiple imputations stored in the `mlong` style (see [MI] **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==.)
. by _mi_id, sort: egen M_lnage=max(mis_lnage)
. by _mi_id, sort: egen M_lntax=max(mis_lntax)
. foreach x of varlist sqft nfeatures lnage lntax {
  2.         quietly summarize `x´ if _mi_miss==0|_mi_m==1, 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´,
> impvar(lnage lntax) misind(M_lnage M_lntax)
note: D1_nfeatures D1_ne D1_custom D1_corner D1_lnage D1_lntax D3_corner
> omitted 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
-------------------------------------------------------------------------
 Missing data patterns  |    Summary of price by missing data pattern
(1 complete, 0 imputed) |   Freq.  Percent     Cum.     Mean   Std.Dev.
------------------------+------------------------------------------------
1 1                     |      66    56.41    56.41  1168.61     404.38
1 0                     |       2     1.71    58.12  1010.00     452.55
0 1                     |      41    35.04    93.16   930.44     298.59
0 0                     |       8     6.84   100.00   880.50     307.17
-------------------------------------------------------------------------
```

Our model includes eight focus regressors, of which six (including the constant term) are observed and two are imputed. Excluding the subset of complete cases (66 observations), there are $2^2 - 1 = 3$ missing-data patterns: 1) lnage observed and lntax missing (2 observations); 2) lnage missing and lntax observed (41 observations); and 3) lnage and lntax both missing (8 observations). The grand model therefore includes $3 \times 8 = 24$ auxiliary regressors, 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 regressors 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
```

Six other auxiliary regressors are dropped because the first missing-data pattern includes only two observations, so we can identify at most two of the eight associated auxiliary parameters. After dropping from **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 estimator of the focus parameters β by specifying the cc option.

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

         price        Coef.   Std. Err.       t     [1 Std. Err. Bands]

          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
         _cons     1000.288    39.59419    25.26      960.6942   1039.883
```

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 β in the grand model (1). Result 1 implies that under our assumptions, the complete-case estimator is unbiased for β. Our findings suggest that home sale price is positively related to the square footage of living space, the log of taxes paid, and whether the home is located in a custom building. On the other side, there is negative association with the 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 1 in absolute value. Also notice that the complete-case estimator is likely to be highly inefficient because it discards about 44% of the sample observations.

To explore the tradeoff between bias and precision, consider now the filling-in and the simple missing-indicator approaches. The former ignores that missing values have been imputed by restricting all auxiliary parameters to 0, while the latter restricts all auxiliary parameters to 0 except the coefficients on the dummies for the missing-data patterns.

```
. gmi price sqft nfeatures ne custom corner if `first_imp´,
> impvar(lnage lntax) misind(M_lnage M_lntax) fi nowarn
Filling-in estimates                          Number of obs =     117
                                              df_m          =       7

------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t     [1 Std. Err. Bands]
-------------+----------------------------------------------------------
        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
       _cons |   984.3707   35.50699    27.72     948.8638   1019.878
------------------------------------------------------------------------

. gmi price sqft nfeatures ne custom corner if `first_imp´,
> impvar(lnage lntax) misind(M_lnage M_lntax) smi nowarn
Simple missing indicator estimates            Number of obs =     117
                                              df_m          =      10

------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t     [1 Std. Err. Bands]
-------------+----------------------------------------------------------
        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
       _cons |   1007.357   35.88174    28.07     971.4752   1043.239
------------------------------------------------------------------------

. 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 δ, so they are likely to result in biased estimates of the focus parameter β. However, as suggested by their considerably lower standard errors, these estimators are more precise than the complete-case estimator. The most striking differences are in the estimated coefficients of `corner` and `lntax`. 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 β in the MSE sense: model reduction and model averaging. Although the choice between these two strategies is left to the user, we strongly encourage choosing model averaging 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: 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 present only 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´,
> impvar(lnage lntax) misind(M_lnage M_lntax) vs best bic full nowarn

Model reduction: L&B with bic                Number of obs =     117
                                             df_m          =       9

---------------------------------------------------------------------
       price |      Coef.   Std. Err.      t     [1 Std. Err. Bands]
-------------+-------------------------------------------------------
        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
       _cons |   983.4677   32.52224    30.24     950.9454    1015.99
-------------+-------------------------------------------------------
     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 regressors: 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. 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´,
> impvar(lnage lntax) misind(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     [1 Std. Err. Bands]
-------------+----------------------------------------------------------
        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
       _cons |   981.8827   35.19818    27.90      946.6845   1017.081
------------------------------------------------------------------------
```

```
. gmi price sqft nfeatures ne custom corner if `first_imp´,
> impvar(lnage lntax) misind(M_lnage M_lntax) wals nowarn
Model averaging: WALS - Lap. prior           Number of obs =      117
                                             df_m          =       24

------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t     [1 Std. Err. Bands]
-------------+----------------------------------------------------------
        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
       _cons |   994.0145   37.68866    26.37      956.3258   1031.703
------------------------------------------------------------------------
```

Magnus, Powell, and Prüfer (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 Magnus' original Matlab command, we recognize that BMA can be very time consuming when the covariates or missing-data patterns are moderate or large. In such circumstances, users are encouraged to rely on WALS, at least when performing a preliminary model-specification search. In this example, BMA estimation requires about 45 seconds on a standard desktop computer. As for the estimated coefficients, we find that BMA and WALS estimates are similar, which suggests 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´, impvar(lnage lntax)
> misind(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

Model averaging: WALS - Sub.(q=.5) prior     Number of obs =      117
                                             Replications  =      100
                                             df_m          =       24

─────────────┬────────────────────────────────────────────────────────────
             │   Observed   Bootstrap                 Bootstrap
       price │      Coef.   Std. Err.       t     [1 Std. Err. Bands]
─────────────┼────────────────────────────────────────────────────────────
        sqft │   .4183898   .0968522     4.32      .3215375    .515242
   nfeatures │   .1203615    16.5871     0.01     -16.46674   16.70746
          ne │   21.76857   47.98438     0.45     -26.21581   69.75295
      custom │   177.8062   70.68029     2.52      107.1259   248.4865
      corner │  -80.24304   46.45896    -1.73      -126.702  -33.78408
       lnage │  -35.72275   32.14764    -1.11     -67.87038  -3.575108
       lntax │   302.3153   145.1049     2.08      157.2104   447.4202
       _cons │   992.4352   38.75039    25.61      953.6848   1031.186
─────────────┴────────────────────────────────────────────────────────────
```

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 we 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 the imputation of missing values. This can be done by specifying the `mimputations` option.

```
. gmi price sqft nfeatures ne custom corner, impvar(lnage lntax)
> misind(M_lnage M_lntax) wals q(.5) nowarn mi full

Multiple-imputation estimates                 Imputations    =      30
Model averaging: WALS - Sub.(q=.5) prior      Number of obs =      117
                                              Average RVI    =  0.1202

------------------------------------------------------------------------
       price |      Coef.   Std. Err.       t     [1 Std. Err. Bands]
-------------+----------------------------------------------------------
        sqft |   .4317465    .0867584    4.98      .3449881   .5185049
   nfeatures |  -.2094938    16.38439   -0.01     -16.59388   16.17489
          ne |   21.21545    42.39104    0.50     -21.17559   63.60649
      custom |   169.9147    50.06324    3.39      119.8515    219.978
      corner |   -78.7322    46.13668   -1.71     -124.8689  -32.59552
       lnage |  -42.43074    26.08771   -1.63     -68.51845  -16.34302
       lntax |   280.3926    127.9489    2.19      152.4437   408.3415
       _cons |   991.3835    37.04141   26.76      954.3421   1028.425
-------------+----------------------------------------------------------
          D1 |   -113.878    110.0119   -1.04     -223.8899  -3.866018
     D1_sqft |   .3657415    .4004362    0.91     -.0346947   .7661776
          D2 |  -54.71213    69.21678   -0.79     -123.9289   14.50465
     D2_sqft |  -.0689371    .1178286   -0.59     -.1867657   .0488915
D2_nfeatures |   -3.48209    25.28178   -0.14     -28.76387   21.79969
       D2_ne |   17.66156    65.09508    0.27     -47.43352   82.75664
   D2_custom |  -33.78444    80.12972   -0.42     -113.9142   46.34528
   D2_corner |   -42.7017    74.37813   -0.57     -117.0798   31.67643
    D2_lnage |  -3.959991    41.30198   -0.10     -45.26197   37.34199
    D2_lntax |  -151.7439    171.2648   -0.89     -323.0086   19.52088
          D3 |  -167.1674    275.5164   -0.61     -442.6838    108.349
     D3_sqft |   .1767931    .8775505    0.20     -.7007574   1.054344
D3_nfeatures |  -25.63685    77.73594   -0.33     -103.3728   52.09909
       D3_ne |   145.6077    276.0254    0.53     -130.4177   421.6331
   D3_custom |  -310.4791    394.2583   -0.79     -704.7373   83.77917
    D3_lnage |   35.51389    252.4899    0.14      -216.976   288.0038
    D3_lntax |  -205.2547    1406.123   -0.15     -1611.378   1200.868
------------------------------------------------------------------------
```

This option 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] **mi estimate**). Although `mi estimate` has its own reporting output, we forced this built-in Stata prefix to respect the reporting output of the `gmi` command to avoid misleading interpretations of the estimation results. As we discussed in section 3.3, this is important because p-values and confidence intervals must be treated with caution.

6 Empirical application

This application investigates the relationship between hand grip strength (GS) and a set of sociodemographic and economic characteristics by using data on the elderly European population. As argued by Andersen-Ranberg et al. (2009), GS is an important measure of health because it is objectively measured, it directly affects everyday 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 interviewers in nonclinical studies.

Our data are from release 2.4.0 of the first wave of SHARE, a multidisciplinary and cross-national household panel survey that provides information on self-reported and objective measures of health, socioeconomic status, and social and family networks for nationally representative samples of people aged 50 and over in the participating countries.[2] The first wave, conducted in 2004–2005, covers about 28,500 individuals in 11 European countries (Austria, Belgium, Denmark, France, Germany, Greece, Italy, the Netherlands, Spain, Sweden, and Switzerland).[3]

The data include two GS measurements on each hand obtained using a hand-grip dynamometer. Respondents are excluded in case of 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 did 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-Ranberg et al. (2009), our dependent variable is the maximum GS (`maxgrip`) measurement resulting from a valid test. Our set of sociodemographic and economic covariates includes age, gender, macroregion 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, we recoded the original values by using the 1997 International Standard Classification of Education. For similar reasons, per capita household income and household net worth have been adjusted for the differences in purchasing power across countries.

Unlike Andersen-Ranberg et al. (2009), who use imputed values of household income and household net worth by relying on the estimates from the filling-in approach, we are interested in investigating the tradeoff between bias and precision when replacing the missing values on these two variables with imputations. This is important 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% in Belgium and a minimum of 49% and 52% in Greece and are equal to 62% and 64% 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). 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 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

[2] Data can be freely downloaded from the SHARE website: http://www.share-project.org.

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

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 five multiple imputations of the key survey variables. As discussed at length in Christelis (2010), these imputations are constructed by the multivariate iterative procedure of van Buuren et al. (2006), which attempts to preserve the correlation structure of the imputed data. In what follows, we account for the additional sampling variability induced by imputation by using the combination rules proposed by Rubin (1987) on the five multiple imputations of household income and household net worth.

Also unlike Andersen-Ranberg et al. (2009), we focus on respondents between 50 and 80 years old 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. Accordingly, we select respondents who have at most one limitation with activities in daily living, who have at most one chronic disease, and whose self-reported health status is at least fair. After we apply this sample selection criterion, dropping the invalid measurements of `maxgrip` (about 5% of the cases) and the few missing data on `weight`, `height`, and `education` (about 1% 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 macroregion.

Table 1. Descriptive statistics for the outcome of interest and the covariate. Weight is in kilograms, height is in centimeters, purchasing power parity-adjusted per capita household income is in 10,000 Euros, and household net worth is 100,000 Euros.

		Male			Female		
Region	Variable	Median	Mean	Standard deviation	Median	Mean	Standard deviation
North	`maxgrip`	49.0	48.3	9.0	29.0	29.1	6.3
	`age`	60.0	61.2	7.9	59.0	60.9	7.8
	`weight`	81.0	82.7	11.8	66.0	67.7	10.9
	`height`	178.0	178.3	6.5	165.0	165.2	5.9
	`education`	1.0	0.6	0.5	1.0	0.6	0.5
	`income`	2.3	2.7	2.0	2.2	2.6	1.7
	`net worth`	1.4	2.9	5.6	1.2	2.4	4.7
	Complete obs.		204			238	
	Imputed obs.		1123			1203	
Center	`maxgrip`	47.0	47.1	9.3	30.0	29.8	6.7
	`age`	60.0	61.4	7.9	59.0	60.9	8.0
	`weight`	80.0	81.3	11.9	67.0	68.0	12.1
	`height`	176.0	175.6	7.0	164.0	164.0	6.3
	`education`	1.0	0.7	0.5	1.0	0.6	0.5
	`income`	1.8	2.5	2.5	1.8	2.6	2.7
	`net worth`	2.2	4.1	9.1	2.0	3.8	9.9
	Complete obs.		730			799	
	Imputed obs.		3798			4057	
South	`maxgrip`	43.0	42.3	10.3	26.0	26.3	6.6
	`age`	60.0	61.7	8.0	58.0	60.1	7.7
	`weight`	79.0	79.3	11.4	66.0	67.8	10.9
	`height`	170.0	171.3	7.2	161.0	161.5	6.2
	`education`	0.0	0.4	0.5	0.0	0.3	0.5
	`income`	0.9	1.3	1.4	0.9	1.4	1.5
	`net worth`	1.7	3.5	9.2	1.6	3.0	6.8
	Complete obs.		48			470	
	Imputed obs.		1785			1758	

Given the high level of comparability of the SHARE data, we pool data from countries in the same macroregion and estimate our linear regression model of interest separately by gender and macroregion. 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 (household income per capita and net worth) are imputed; 3 subsamples with incomplete data; and 21 noncollinear auxiliary regressors. The resulting dimension of the model space is 2,097,152. After centering the focus covariates on their median for each subgroup, we compare the estimates from five alternative approaches: complete-case (CC), filling-in (FI), model reduction (VS), 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; WALS estimation is carried out using a Subbotin prior with parameter $q = 0.5$.[4]

The estimated coefficients and their standard errors are presented in tables 2 and 3, separately by gender and macroregion.[5] Qualitatively, our results are consistent with the empirical findings in Andersen-Ranberg 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 population. For men, the age-related decline in `maxgrip` is steeper for those living in Southern countries. For women, it is 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.

[4] Estimates from the simple missing-indicator approach are omitted because they are similar to those obtained from the filling-in approach. Estimates from WALS with a Laplace prior are omitted because they are very similar to those obtained with a Subbotin prior.

[5] Using a desktop computer with 2 quad-core Intel Xeon E5504/2 GHz processors and Stata/MP4 version 11.2, the computer time required for BMA estimation varies between a minimum of 10 hours in the specification Male–North and a maximum of 1 day in the specification Female–Center.

Table 2. Estimated coefficients and standard errors (in parentheses) for males by macroregion. Estimation is based on $M = 5$ multiple imputations for income and net worth. Results for the auxiliary regressors are omitted to save space. * denotes a t ratio greater than 1 in absolute value.

Region	Variable	CC	FI	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 macroregion. Estimation is based on $M = 5$ multiple imputations for income and net worth. Results for the auxiliary regressors are omitted to save space. * denotes a t ratio greater than 1 in absolute value.

Region	Variable	CC	FI	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)

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 different. For example, the point estimate of the coefficient on weight in the specification Male–North ranges between a minimum of 0.106 with a standard error of 0.022 using the filling-in 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 filling-in approaches. 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 that alternative assumptions on the prior distributions for the auxiliary parameters may matter. From this viewpoint, WALS has the advantage of using priors that ensure bounded risk and a coherent treatment of ignorance about the auxiliary parameters.

7 Conclusions

In this article, we introduced a Stata command that implements the generalized missing-indicator approach of Dardanoni, Modica, and Peracchi (2011) for fitting a regression model with imputed covariates. The command enables one to go beyond the alternative of either dropping the observations with imputed values (the complete-case approach) or using all the observations without distinguishing between observed and imputed values (the filling-in approach). The command essentially expands the model space by including all the intermediate cases between the model that contains only the observed or imputed covariates and a “grand model” that adds to them a full set of auxiliary regressors.

In the expanded model space, the proposed command offers two alternative strategies for obtaining a best estimate of the regression parameters of interest: model reduction and BMA. The second strategy avoids the pretesting problem that plagues model reduction techniques and allows one to formally incorporate, through the choice of priors, the researcher’s uncertainty about the role of the auxiliary regressors.

The proposed command also offers two different BMA implementations: standard BMA and WALS. Relative to standard BMA, the advantages of WALS are its more intuitive concept of uncertainty about the role of the auxiliary regressors, the bounded risk and near optimality of its estimates, and most importantly for practitioners, its substantially lower computational burden.

8 Acknowledgments

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

9 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, ed. 2005. *Health, Ageing and Retirement in Europe: First Results from the Survey of Health, Ageing and Retirement in Europe*. Mannheim: Mannheim Research Institute for the Economics of Aging.

Christelis, D. 2010. Imputation of missing data in waves 1 and 2 of SHARE. Working Paper 01-2011, SHARE. http://share-dev.mpisoc.mpg.de/uploads/tx_sharepublications/WP_Series_01-2011_Christelis.pdf.

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

De Luca, G., and J. R. Magnus. 2011. Bayesian model averaging and weighted-average least squares: Equivariance, stability, and numerical issues. *Stata Journal* 11: 518–544.

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. *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.

Kumar, K., and J. R. Magnus. 2011. A characterization of Bayesian robustness for a normal location parameter. http://www.janmagnus.nl/wips/wip23.pdf.

Lindsey, C., and S. Sheather. 2010. Variable selection in linear regression. *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. 2001. *Estimation in the Presence of Nonresponse and Frame Imperfections*. Örebro, Sweden: Statistics Sweden.

Magnus, J. R. 2000. The traditional pretest estimator. *Theory of Probability and its Applications* 44: 293–308.

———. 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 model 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 and Data Analysis* 55: 1331–1341.

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

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.

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

About the authors

Valentino Dardanoni is a professor of economics at the University of Palermo.

Giuseppe De Luca is an assistant professor at the University of Palermo.

Salvatore Modica is a professor of economics at the University of Palermo.

Franco Peracchi is a professor of econometrics at Tor Vergata University and a fellow of EIEF.