Title

intro substantive — Introduction to multiple-imputation analysis

Description

Missing data arise frequently. Various procedures have been suggested in the literature over the last several decades to deal with missing data (for example, Anderson [1957]; Hartley and Hocking [1971]; Rubin [1972, 1987]; and Dempster, Laird, and Rubin [1977]). The technique of multiple imputation, which originated in early 1970 in application to survey nonresponse (Rubin 1976), has gained popularity increasingly over the years as indicated by literature (for example, Rubin [1976, 1987, 1996]; Little [1992]; Meng [1994]; Schafer [1997]; van Buuren, Boshuizen, and Knook [1999]; Little and Rubin [2002]; Carlin et al. [2003]; Royston [2004, 2005a, 2005b, 2007, 2009]; Reiter and Raghunathan [2007]; Carlin, Galati, and Royston [2008]; Royston, Carlin, and White [2009]; and White, Royston, and Wood [2011]).

This entry presents a general introduction to multiple imputation and describes relevant statistical terminology used throughout the manual. The discussion here, as well as other statistical entries in this manual, is based on the concepts developed in Rubin (1987) and Schafer (1997).

Remarks

Remarks are presented under the following headings:

We will use the following definitions and notation.

An imputation represents one set of plausible values for missing data, and so multiple imputations represent multiple sets of plausible values. With a slight abuse of the terminology, we will use the term *imputation* to mean the data where missing values are replaced with one set of plausible values.

We use M to refer to the number of imputations and m to refer to each individual imputation; that is, $m = 1$ means the first imputation, $m = 2$ means the second imputation, and so on.

Motivating example

Consider a fictional case–control study examining a relationship between smoking and heart attacks.

```
. use http://www.stata-press.com/data/r12/mheart0
(Fictional heart attack data; bmi missing)
. describe
Contains data from http://www.stata-press.com/data/r12/mheart0.dta
  obs:           154                          Fictional heart attack data;
                                                bmi missing
 vars:             9                          19 Jun 2011 10:50
 size:         2,310
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
attack          byte   %9.0g                  Outcome (heart attack)
smokes          byte   %9.0g                  Current smoker
age             float  %9.0g                  Age, in years
bmi             float  %9.0g                  Body Mass Index, kg/m^2
female          byte   %9.0g                  Gender
hsgrad          byte   %9.0g                  High school graduate
marstatus       byte   %9.0g       mar        Marital status: single, married,
                                                divorced
alcohol         byte   %24.0g      alc        Alcohol consumption: none, <2
                                                drinks/day, >=2 drinks/day
hightar         byte   %9.0g                  Smokes high tar cigarettes
-------------------------------------------------------------------------------
Sorted by:
```

In addition to the primary variables `attack` and `smokes`, the dataset contains information about subjects' ages, body mass indexes (BMIs), genders, educational statuses, marital statuses, alcohol consumptions, and the types of cigarettes smoked (low/high tar).

We will use logistic regression to study the relationship between `attack`, recording heart attacks, and `smokes`:

```
. logit attack smokes age bmi hsgrad female
Iteration 0:   log likelihood = -91.359017
Iteration 1:   log likelihood = -79.374749
Iteration 2:   log likelihood = -79.342218
Iteration 3:   log likelihood =  -79.34221
Logistic regression                               Number of obs   =        132
                                                  LR chi2(5)      =      24.03
                                                  Prob > chi2     =     0.0002
Log likelihood =  -79.34221                       Pseudo R2       =     0.1315
------------------------------------------------------------------------------
      attack |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smokes |    1.544053   .3998329     3.86   0.000     .7603945    2.327711
         age |     .026112    .017042     1.53   0.125    -.0072898    .0595137
         bmi |    .1129938   .0500061     2.26   0.024     .0149837     .211004
      hsgrad |    .4048251   .4446019     0.91   0.363    -.4665786    1.276229
      female |    .2255301   .4527558     0.50   0.618    -.6618549    1.112915
       _cons |   -5.408398   1.810603    -2.99   0.003    -8.957115   -1.85968
------------------------------------------------------------------------------
```

The above analysis used 132 observations out of the available 154 because some of the covariates contain missing values. Let's examine the data for missing values, something we could have done first:

```
. misstable summarize
```

Variable	Obs=.	Obs>.	Obs<.	Obs<. Unique values	Obs<. Min	Obs<. Max
bmi	22		132	132	17.22643	38.24214

We discover that `bmi` is missing in 22 observations. Our analysis ignored the information about the other covariates in these 22 observations. Can we somehow preserve this information in the analysis? The answer is yes, and one solution is to use multiple imputation.

What is multiple imputation?

Multiple imputation (MI) is a flexible, simulation-based statistical technique for handling missing data. Multiple imputation consists of three steps:

1. *Imputation step.* M imputations (completed datasets) are generated under some chosen imputation model.
2. *Completed-data analysis (estimation) step.* The desired analysis is performed separately on each imputation $m = 1, \ldots, M$. This is called completed-data analysis and is the primary analysis to be performed once missing data have been imputed.
3. *Pooling step.* The results obtained from M completed-data analyses are combined into a single multiple-imputation result.

The completed-data analysis step and the pooling step can be combined and thought of generally as the analysis step.

MI as a missing-data technique has two appealing main features: 1) the ability to perform a wide variety of completed-data analyses using existing statistical methods; and 2) separation of the imputation step from the analysis step. We discuss these two features in more detail in what follows.

Among other commonly used missing-data techniques that allow a variety of completed-data analyses are complete-case analysis or listwise (casewise) deletion, available-case analysis, and single-imputation methods. Although these procedures share one of MI's appealing properties, they lack some of MI's statistical properties.

For example, listwise deletion discards all observations with missing values and thus all information contained in the nonmissing values of these observations. With a large number of missing observations, this may lead to results that will be less efficient (larger standard errors, wider confidence intervals, less power) than MI results. In situations when the remaining complete cases are not representative of the population of interest, listwise deletion may also lead to biased parameter estimates.

In our opening logistic analysis of heart attacks, we used listwise deletion. The effect of age was not statistically significant based on the reduced sample. The MI analysis of these data (see *A brief introduction to MI using Stata* below) will reveal the statistical significance of age by using all available observations after imputing missing values for BMI.

Unlike listwise deletion, single-imputation methods do not discard missing values. They treat the imputed values as known in the analysis. This underestimates the variance of the estimates and so overstates precision and results in confidence intervals and significance tests that are too optimistic. MI rectifies this problem by creating multiple imputations and taking into account the sampling variability due to the missing data (between-imputation variability). See Little and Rubin (2002) and Allison (2001), among others, for a more detailed comparison of the methods.

The independence of the imputation step from the analysis step is the property MI shares with other imputation methods. The imputation step fills in missing values. The analysis step provides inference about multiply imputed results and does not require any information about the missing-data aspect of the problem.

The separation of the two steps allows different individuals, a data collector/imputer and a data analyst, to perform these steps independently of one another. The advantage is that the data collector/imputer usually has access to more information about the data than may be disclosed to the data analyst and thus can create more accurate imputations. The data analyst can use the imputed data released by the data collector in a number of different analyses. Of course, it is crucial that the imputer make the imputation model as general as possible to accommodate a wide variety of analyses that the data analyst might choose to perform; see *Proper imputation methods* below for details.

In our heart attack example, the imputer would create multiple imputations of missing values of BMI using, for example, a linear regression method, and then release the resulting data to the analyst. The analyst could then analyze these multiply imputed data using an ordinary logistic regression. That is, no adjustment is needed to the analysis model itself to account for missing BMI—the pooling portion of the analysis will account for the increased variability because of imputed missing data.

Theory underlying multiple imputation

MI was derived using the Bayesian paradigm yet was proved to be statistically valid from the frequentist (randomization-based) perspective. We use the definition from Rubin (1996) of statistical validity that implies approximately unbiased point estimates and implies confidence intervals achieving their nominal coverages when averaged over the randomization distributions induced by the known sampling and the posited missing-data mechanisms.

To explain the role the Bayesian and frequentist concepts play in MI, we need to consider the MI procedure in more detail. MI requires specification of two models—the imputation model and the analysis model. The imputation model is the model used to create imputations in the imputation step. The analysis model is the completed-data model used during the analysis step to obtain completed-data estimates, $\widehat{Q}$, of parameters of interest, Q, and the estimate, U, of sampling variability associated with $\widehat{Q}$. During the pooling step, the individual completed-data estimates $(\widehat{Q}, U)$ are combined into $(\widehat{Q}_{\rm MI}, T)$ to form one repeated-imputation inference. The statistical validity of the repeated-imputation inference is of interest.

Consider the case when both the imputation model and the analysis model are the same Bayesian models. Then the repeated imputations (multiple imputations) are repeated draws from the posterior predictive distribution of the missing data under a posited Bayesian model. The combined parameter estimates, $\widehat{Q}_{\rm MI}$, and their associated sampling variance estimate, $T = W + B$, are the approximations to the posterior mean and variance of Q. Here W represents the within-imputation variability (average of the completed-data variance estimates, U), and B represents the between-imputation variability (variance estimate of $\widehat{Q}_{\rm MI}$ over repeated imputations). Provided that the posterior mean and variance are adequate summaries of the posterior distribution, the repeated-imputation inference based on these combined estimates can be justified either from a purely Bayesian standpoint or from a purely frequentist standpoint. Thus a Bayesian apparatus is used to create imputations and also underlies the rules for combining parameter estimates.

In reality, the analysis model is rarely the same as the imputation model, and neither of them is an explicit Bayesian model. Repeated-imputation inference is still statistically valid in those cases. The rigorous justification is given in chapters 3 and 4 of Rubin (1987) from the frequentist perspective. Below we briefly summarize the conditions under which the repeated-imputation inference from the pooling step is statistically valid; also see Rubin (1987, 117–119) for more detail.

The repeated-imputation inference is statistically valid if 1) the multiple imputations from the imputation step are proper (see *Proper imputation methods* below) and 2) the completed-data inference based on $(\widehat{Q}, U)$ from the analysis step is randomization valid. Completed-data inference based on $(\widehat{Q}, U)$ is randomization valid if $\widehat{Q} \sim N\{Q, \mathrm{Var}(\widehat{Q})\}$ and U is a consistent estimate of $\mathrm{Var}(\widehat{Q})$ over the distribution of the sampling mechanism.

The randomization validity of MI was derived under the assumption of an infinite number of imputations. In practice, however, the number of imputations tends to be small and so the finite-M properties of the MI estimators must be explored. Rubin (1987) derives the fundamental result underlying the MI inference based on a finite M. We restate it below for a scalar Q:

$$T_M^{-1/2}(Q - \widehat{Q}_M) \sim t_{\nu_M}$$

where $\widehat{Q}_M$ is the average of M completed-data estimates of Q, $T_M = W + (1 + 1/M)B$, and t_{ν_M} is a Student's t distribution with degrees of freedom ν_M that depend on the number of imputations and rates of missing information (or the fraction of information missing because of nonresponse that measures the influence of the missing data on parameter estimates). Later, Li, Raghunathan, and Rubin (1991b) derived an improved procedure for multiple testing, and Barnard and Rubin (1999) and Reiter (2007) extended the MI inference to account for small samples. For computation details, see *Methods and formulas* in [MI] **mi estimate**.

How large should M be?

The theory underlying the validity of MI relies on an infinite number of imputations, M. The procedure is also known to have good statistical properties with finite M, but what values of M should we use in practice? Rubin (1987, 114) answers this question: the asymptotic relative efficiency (RE) of the MI procedure with finite M compared with infinite M is roughly 90% with only two imputations for a missing-information rate as high as 50%.

Most literature (for example, Rubin [1987] and van Buuren, Boshuizen, and Knook [1999]) suggests that $M = 5$ (corresponding to RE of 95% for 50% of information missing) should be sufficient to obtain valid inference. In general, however, the actual number of imputations necessary for MI to perform satisfactorily depends not only on the amount of information missing due to nonresponse but also on the analysis model and the data. Some analyses may require M to be 50 or more to obtain stable results (Kenward and Carpenter 2007; Horton and Lipsitz 2001).

Literature with formal recommendations on how to choose M is very sparse. Royston (2004), Royston, Carlin, and White (2009), and White, Royston, and Wood (2011) discuss the impact of the number of imputations on the precision of estimates and suggest ways of determining the required number of imputations by evaluating the sampling error of the MI estimates.

Because it is computationally feasible to obtain more imputations, we recommend using at least 20 imputations to reduce the sampling error due to imputations.

Assumptions about missing data

The theory underlying MI methodology makes no assumption about the missing-data mechanism. However, many imputation methods (including those provided by Stata) require that the missing-data mechanism be ignorable. Before we discuss the ignorability conditions, consider the following definitions.

Missing data are said to be missing completely at random (MCAR) if the probability that data are missing does not depend on observed or unobserved data. Under MCAR, the missing-data values are a simple random sample of all data values, and so any analysis that discards the missing values remains consistent, albeit perhaps inefficient.

Consider a hypothetical longitudinal study comparing different blood-pressure treatments. Suppose that the follow-up blood-pressure measurements were not collected from some subjects because they moved to a different area. These missing blood-pressure measurements can be viewed as MCAR as long as subjects' decisions to move were unrelated to any item in the study.

Missing data are said to be missing at random (MAR) if the probability that data are missing does not depend on unobserved data but may depend on observed data. Under MAR, the missing-data values do not contain any additional information given observed data about the missing-data mechanism. Note that MCAR can be viewed as a particular case of MAR. When missing data are MAR, listwise deletion may lead to biased results.

Suppose that some subjects decided to leave the study because of severe side effects from the assigned treatment of a high dosage of a medicine. Here it is unlikely that missing blood-pressure measurements are MCAR because the subjects who received a higher dosage of the medicine are more likely to suffer severe side effects than those who received a lower dosage and thus are more likely to drop out of the study. Missing blood-pressure measurements depend on the dosage of the received treatment and therefore are MAR.

On the other hand, if the subjects are withdrawn from the study for ethical reasons because of extremely high blood pressures, missing blood-pressure measurements would not be MAR. The measurements for the subjects with very high blood pressures will be missing and thus the reason for drop out will depend on the missing blood pressures. This type of missing-data mechanism is called missing not at random (MNAR). For such missing data, the reasons for its missingness must be accounted for in the model to obtain valid results.

Model parameters are said to be *distinct* from a Bayesian standpoint if their joint prior distribution can be factorized into independent marginal prior distributions.

The missing-data mechanism is said to be *ignorable* if missing data are MAR and the parameters of the data model and the parameters of the missing-data mechanism are distinct (Rubin 1976).

The ignorability assumption makes it possible to ignore the process that causes missing data in the imputation model—something not possible with MNAR—which simplifies the imputation step while still ensuring correct inference. The provided imputation methods assume that missing data are MAR.

In practice, it is difficult to test the ignorability assumption formally because the MAR mechanism can be distinguished from the MNAR mechanism only through the missing data that are not observed. Thus careful consideration is necessary before accepting this assumption. If in doubt, sensitivity analysis—analysis repeated under various missing-data models—needs to be performed to verify the stability of inference. In the context of MI, sensitivity analysis can be performed by modifying the imputation step to accommodate the nonignorable missing-data mechanism (for example, Kenward and Carpenter [2007] and van Buuren, Boshuizen, and Knook [1999]).

Patterns of missing data

Another issue we need to consider related to missing data is a pattern of missingness (or missing-data pattern).

Consider an $N \times p$ data matrix $Y = (Y_1, Y_2, \ldots, Y_p)'$ with p variables and N observations. Consider a permutation of column indices $(i_1, i_2, \ldots, i_p)$ such that Y_{i_1} is at least as observed as Y_{i_2}, which is at least as observed as Y_{i_3}, and so on. In other words, Y_{i_2} has missing values in the same

observations (and possibly more) as Y_{i_1}, Y_{i_3} has missing values (and possibly more) in the same observations as Y_{i_2}, and so on. If such a permutation exists, then the pattern of missingness in Y is said to be monotone. If the pattern of missingness is not monotone, it is assumed to be arbitrary.

For example, consider the following indicator matrix recording the missing pattern in Y:

$$R_1 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

where R_1^{ij} is 1 if variable Y_j is observed (complete) in observation i and 0 otherwise. We can see that Y has a monotone-missing pattern if we interchange the first and the third columns of R_1. In fact, if we also rearrange the rows such that

$$R_1 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

then the monotonicity of missing values becomes even more evident. An example of a nonmonotone missing-value pattern is

$$R_2 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

There is no ordering of the first two columns of R_2 such that the missing values in one column imply missing values in the other column.

Why is it important to consider the monotone missing-value pattern? A monotone-missing pattern greatly simplifies the imputation task. Under a monotone-missing pattern, a multivariate imputation task can be formulated as a sequence of independent univariate (conditional) imputation tasks, which allows the creation of a flexible imputation model; see [MI] **mi impute monotone** for details, and see Rubin (1987, 174) for more technical conditions under which such a formulation is applicable.

Proper imputation methods

As we mentioned earlier, a key concept underlying the randomization-based evaluations of the repeated-imputation inference is proper multiple imputation.

A multiple-imputation method is said to be proper if it produces proper multiple imputations, which we are about to define. Rubin (1987, 118–119) gives a full technical definition for proper multiple imputations. Ignoring the more technical definition, Rubin (1996) states the following main conditions. The multiple imputations are said to be proper if

1. MI estimates $\widehat{Q}_{\mathrm{MI}}$ are asymptotically normal with mean $\widehat{Q}$ and a consistent variance–covariance estimate B.
2. The within-imputation variance estimate W is a consistent estimate of the variance–covariance estimate U with variability of a lower order than $\mathrm{Var}(\widehat{Q}_{\mathrm{MI}})$.

The above statements assume a large number of imputations and the randomization distribution induced by the missing-data mechanism.

In general, it is difficult to determine if an imputation method is proper using the above definition. Rubin (1987, sec. 4.3) and Binder and Sun (1996) describe several examples of proper and improper imputation methods. Rubin (1987, 125–127) recommends drawing imputations from a Bayesian posterior predictive distribution (or an appropriate approximation to it) of missing values under the chosen model for the data and the missing-data mechanism. The chosen imputation model must also be appropriate for the completed-data statistics likely to be used at the analysis stage. Schafer (1997, 145) points out that from a practical standpoint, it is more important that the chosen imputation model performs well over the repeated samples than that it be technically proper. This can be checked via simulation.

With the exception of predictive mean matching and chained equations, the imputation methods available in Stata obtain imputations by simulating from a Bayesian posterior predictive distribution of the missing data (or its approximation) under the conventional (or chosen) prior distribution; see *Imputation methods* in [MI] **mi impute** for details. To ensure that the multiple imputations are proper, you must choose an appropriate imputation model, which we briefly discuss next.

The imputation model must include all predictors relevant to the missing-data mechanism, and it must preserve all data characteristics likely to be explored at the analysis stage. For example, if the analysis model explores a correlation between two variables, then omitting either of those variables from the imputation model will lead to estimates of the correlation biased toward zero. Another common mistake that may lead to biased estimates is when an outcome variable of the analysis model is not used in the imputation model. In the survey context, all structural variables such as sampling weights, strata, and cluster identifiers (or at least main strata and main clusters) need to be included in the imputation model.

In general, any predictors involved in the definition of the completed-data estimators and the sampling design should be included in the imputation model. If you intend to use the multiply imputed data in an analysis involving a wide range of completed-data estimators, you should include as many variables as possible.

Using our heart attack data, if we were to release the multiply imputed version of it for general analyses, we would have included all available covariates as predictors in the regression model used to impute BMI and not only the subset of covariates (heart attacks, smoking status, age, gender, and educational status) used in our specific data analysis.

The severity of the effect of a misspecified imputation model will typically depend on the amount of imputed data relative to the observed data—a small number of observations with improperly imputed values may not affect the inference greatly if there is a large number of observations with complete data.

For more details about imputation modeling, see Rubin (1996), Schafer (1997, 139–144), Schafer and Olsen (1998), Allison (2001), Schafer and Graham (2002), Kenward and Carpenter (2007), Graham (2009), and White, Royston, and Wood (2011), among others. For imputation modeling of large surveys, see, for example, Schafer, Khare, and Ezzati-Rice (1993) and Ezzati-Rice et al. (1995).

Analysis of multiply imputed data

Once we have multiply imputed data, we perform our primary analysis on each completed dataset and then use Rubin's combination rules to form one set of results. Assuming that the underlying imputation model is properly specified (see, for example, Abayomi, Gelman, and Levy [2008] and Gelman et al. [2005] for multiple-imputation diagnostics), we can choose from a variety of statistical methods. For example, the methods can include maximum likelihood methods, survey methods, nonparametric methods, and any other method appropriate for the type of data we have.

Each of the methods have certain concepts associated with them. For example, maximum likelihood methods use a likelihood function, whereas a deviance is associated with generalized linear models. While these concepts are well defined within each individual completed-data analysis, they may not have a clear interpretation when the individual analyses are combined in the pooling step. (Only in the special case when the imputation and analysis models are compatible Bayesian models can the estimated parameters be viewed as approximations to the mode of the posterior distribution.)

As a result, various statistical (postestimation) procedures based on these concepts, such as likelihood-ratio tests, goodness-of-fit tests, etc., are not directly applicable to MI results. Instead, their "MI" versions are being studied in the literature (Li et al. 1991a; Meng and Rubin 1992). Another concept that is not uniquely defined within MI is that of prediction; see Carlin, Galati, and Royston (2008) and White, Royston, and Wood (2011) for one definition.

Donald Bruce Rubin (1943–) was born in Washington, DC. He entered Princeton intending to become a physicist but ended up majoring in psychology. He entered Harvard intending to continue as a psychologist, but in the event, gained further degrees in computer science and statistics. After periods at the Educational Testing Service and the University of Chicago, Rubin returned to Harvard in 1984. He has had many visiting appointments and has carried out extensive consultancy work. Rubin has long been a leader in research on causal inference in experiments and observational studies, and problems of nonresponse and missing data. Among many major contributions is his formalization of the expectation-maximization algorithm with Arthur Dempster and Nan Laird. Rubin's work ranges over a wide variety of sciences and is often Bayesian in style. Rubin was elected a member of the National Academy of Sciences in 2010.

A brief introduction to MI using Stata

Stata offers full support for MI analysis from the imputation step to the pooling step.

The imputation step can be performed for one variable or multiple variables. A number of imputation methods, including flexible methods accommodating variables of different types and an iterative Markov chain Monte Carlo method based on multivariate normal, are available; see [MI] **mi impute** for details.

The analysis and pooling steps are combined into one step and performed by `mi estimate`; see [MI] **mi estimate**. You can fit many commonly used models and obtain combined estimates of coefficients (or transformed coefficients) (see [MI] **estimation** for a list of supported estimation commands), or you can create your own estimation command and use it with the `mi estimate` prefix.

In addition to the conventional estimation steps, Stata facilitates many data-manipulation routines for managing your multiply imputed data and verifying its integrity over the imputations; see [MI] **intro** for a full list of commands.

As a short demonstration of `mi`, let's analyze the heart attack data introduced earlier using MI; see [MI] **workflow** for more thorough guidelines.

The goals are 1) to fill in missing values of `bmi` using, for example, a linear regression imputation method (`mi impute regress`) to obtain multiply imputed data and 2) to analyze the multiply imputed data using logistic regression, which we will do using `mi estimate`. Before we can accomplish these two steps, we need to prepare the data so they can be used with `mi`. First, we declare the data to be `mi` data:

```
. use http://www.stata-press.com/data/r12/mheart0
(Fictional heart attack data; bmi missing)
. mi set mlong
```

We choose to use the data in the marginal long style (`mlong`) because it is a memory-efficient style; see [MI] **styles** for details.

To use `mi impute`, we must first register imputation variables. In general, we recommend that you register all variables relevant to the analysis as `imputed`, `passive`, or `regular` with `mi register` (see [MI] **mi set**), especially if you plan on doing any data management of your multiply imputed data.

```
. mi register imputed bmi
(22 m=0 obs. now marked as incomplete)
. mi register regular attack smokes age hsgrad female
```

We are now ready to use `mi impute`. To lessen the simulation (Monte Carlo) error, we arbitrarily choose to create 20 imputations (`add(20)` option). We also specify the `rseed()` option for reproducibility:

```
. mi impute regress bmi attack smokes age hsgrad female, add(20) rseed(2232)
Univariate imputation                       Imputations =       20
Linear regression                                 added =       20
Imputed: m=1 through m=20                       updated =        0
```

	Observations per *m*			
Variable	Complete	Incomplete	Imputed	Total
bmi	132	22	22	154

```
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

From the output, we see that all 22 incomplete values of `bmi` were successfully imputed. You may want to examine your imputations to verify that nothing abnormal occurred during imputation. For example, as a quick check, we can compare main descriptive statistics from some imputations (say, the first and the last one) to those from the observed data. We use `mi xeq` (see [MI] **mi xeq**) to execute Stata's `summarize` command on the original data ($m = 0$), the first imputation ($m = 1$), and the last imputation ($m = 20$):

```
. mi xeq 0 1 20: summarize bmi
m=0 data:
-> summarize bmi
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         bmi |       132    25.24136    4.027137   17.22643   38.24214

m=1 data:
-> summarize bmi
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         bmi |       154    25.11855    3.990918   15.47331   38.24214

m=20 data:
-> summarize bmi
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         bmi |       154    25.37117    4.051929    15.4505   38.24214
```

The summary statistics of the imputed datasets look reasonable.

We now fit the logistic regression using the `mi estimate` prefix command:

```
. mi estimate, dots: logit attack smokes age bmi hsgrad female
Imputations (20):
  .........10.........20 done
Multiple-imputation estimates                   Imputations     =         20
Logistic regression                             Number of obs   =        154
                                                Average RVI     =     0.0404
                                                Largest FMI     =     0.1678
DF adjustment:   Large sample                   DF:     min     =     694.17
                                                        avg     =  115477.35
                                                        max     =  287682.23
Model F test:       Equal FMI                   F(   5,43531.9) =       3.74
Within VCE type:          OIM                   Prob > F        =     0.0022
```

attack	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
smokes	1.239172	.3630877	3.41	0.001	.5275236	1.950821
age	.0354929	.0154972	2.29	0.022	.0051187	.065867
bmi	.1184188	.0495676	2.39	0.017	.0210985	.2157391
hsgrad	.185709	.4075301	0.46	0.649	-.6130435	.9844615
female	-.0996102	.4193583	-0.24	0.812	-.9215408	.7223204
_cons	-5.845855	1.72309	-3.39	0.001	-9.225542	-2.466168

Compared with the earlier `logit` analysis (using listwise deletion), we detect the significance of `age`, whose effect was apparently disguised by the missing data. See [MI] **mi estimate** for details.

We will be using variations of these data throughout the `mi` documentation.

Summary

- MI is a simulation-based procedure. Its purpose is not to re-create the individual missing values as close as possible to the true ones but to handle missing data in a way resulting in valid statistical inference (Rubin 1987, 1996).
- MI yields valid inference if 1) the imputation method is proper with respect to the posited missing-data mechanism (see *Proper imputation methods* above) and 2) completed-data analysis is valid in the absence of missing data.
- A small number of imputations (5 to 20) may be sufficient when fractions of missing data are low. High fractions of missing data as well as particular data structures may require up to 100 (or more) imputations. Whenever feasible to do so, we recommend that you vary the number of imputations to see if this affects your results.
- With a small number of imputations, the reference distribution for the MI inference is Student's t (or F in multiple-hypothesis testing). The residual degrees of freedom depend on M and the rates of missing information and thus are different for each parameter of interest.
- With a large number of imputations, the reference distribution for MI inference is approximately normal (or χ^2 in multiple-hypothesis testing).
- When the imputer's model is more restrictive than the analyst's model, the MI inference can be invalid if the imputer's assumptions are not true. On the other hand, when the analyst's model is more restrictive than the imputer's model, the MI results will be valid but somewhat conservative if the analyst's assumptions are true. If the analyst's assumptions are false, the results can be biased; see, for example, Schafer (1997) for details.

- MI is relatively robust to departures from the correct specification of the imputation model, provided the rates of missing information are low and the correct completed-data model is used in the analysis.
- Certain concepts, for example, likelihood and deviance, do not have clear interpretation within the MI framework. As such, various statistical (postestimation) procedures based on these concepts (for example, likelihood-ratio tests, goodness-of-fit tests) are not directly applicable to MI results.

References

Abayomi, K., A. Gelman, and M. Levy. 2008. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society, Series C* 57: 273–291.

Allison, P. D. 2001. *Missing Data*. Thousand Oaks, CA: Sage.

Anderson, T. W. 1957. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association* 52: 200–203.

Arnold, B. C., E. Castillo, and J. M. Sarabia. 1999. *Conditional Specification of Statistical Models*. New York: Springer.

——. 2001. Conditionally specified distributions: An introduction. *Statistical Science* 16: 249–274.

Barnard, J., and D. B. Rubin. 1999. Small-sample degrees of freedom with multiple imputation. *Biometrika* 86: 948–955.

Binder, D. A., and W. Sun. 1996. Frequency valid multiple imputation for surveys with a complex design. *Proceedings of the Survey Research Methods Section, American Statistical Association* 281–286.

Carlin, J. B., J. C. Galati, and P. Royston. 2008. A new framework for managing and analyzing multiply imputed data in Stata. *Stata Journal* 8: 49–67.

Carlin, J. B., N. Li, P. Greenwood, and C. Coffey. 2003. Tools for analyzing multiple imputed datasets. *Stata Journal* 3: 226–244.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39: 1–38.

Ezzati-Rice, T. M., W. Johnson, M. Khare, R. J. A. Little, D. B. Rubin, and J. L. Schafer. 1995. A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. *Proceedings of the Annual Research Conference*, 257–266. U.S. Bureau of the Census: Washington, DC.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. *Bayesian Data Analysis*. 2nd ed. London: Chapman & Hall/CRC.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* 7: 457–472.

Gelman, A., I. Van Mechelen, G. Verbeke, D. F. Heitjan, and M. Meulders. 2005. Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics* 61: 74–85.

Graham, J. W. 2009. Missing data analysis: Making it work in the real world. *Annual Review of Psychology* 60: 549–576.

Hartley, H. O., and R. R. Hocking. 1971. The analysis of incomplete data (with discussion). *Biometrics* 27: 783–823.

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.

Horton, N. J., and S. R. Lipsitz. 2001. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician* 55: 244–254.

Jenkins, S. P., R. V. Burkhauser, S. Feng, and J. Larrimore. 2011. Measuring inequality using censored data: a multiple-imputation approach to estimation and inference. *Journal of the Royal Statistical Society, Series A* 174: 63–81.

Kenward, M. G., and J. R. Carpenter. 2007. Multiple imputation: Current perspectives. *Statistical Methods in Medical Research* 16: 199–218.

Lee, K. J., and J. B. Carlin. 2010. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology* 171: 624–632.

Li, K.-H. 1988. Imputation using Markov chains. *Journal of Statistical Computation and Simulation* 30: 57–79.

Li, K.-H., X.-L. Meng, T. E. Raghunathan, and D. B. Rubin. 1991a. Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica* 1: 65–92.

Li, K.-H., T. E. Raghunathan, and D. B. Rubin. 1991b. Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association* 86: 1065–1073.

Little, R. J. A. 1988. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics* 6: 287–296.

——. 1992. Regression with missing X's: A review. *Journal of the American Statistical Association* 87: 1227–1237.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: Wiley.

Marchenko, Y. V., and J. P. Reiter. 2009. Improved degrees of freedom for multivariate significance tests obtained from multiply imputed, small-sample data. *Stata Journal* 9: 388–397.

Meng, X.-L. 1994. Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science* 9: 538–573.

Meng, X.-L., and D. B. Rubin. 1992. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* 79: 103–111.

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 27: 85–95.

Reiter, J. P. 2007. Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika* 94: 502–508.

——. 2008. Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika* 95: 933–946.

Reiter, J. P., and T. E. Raghunathan. 2007. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* 102: 1462–1471.

Royston, P. 2004. Multiple imputation of missing values. *Stata Journal* 4: 227–241.

——. 2005a. Multiple imputation of missing values: Update. *Stata Journal* 5: 188–201.

——. 2005b. Multiple imputation of missing values: Update of ice. *Stata Journal* 5: 527–536.

——. 2007. Multiple imputation of missing values: Further update of ice, with an emphasis on interval censoring. *Stata Journal* 7: 445–464.

——. 2009. Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables. *Stata Journal* 9: 466–477.

Royston, P., J. B. Carlin, and I. R. White. 2009. Multiple imputation of missing values: New features for mim. *Stata Journal* 9: 252–264.

Rubin, D. B. 1972. A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. *Journal of the Royal Statistical Society, Series C* 21: 136–141.

——. 1976. Inference and missing data. *Biometrika* 63: 581–592.

——. 1986. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics* 4: 87–94.

——. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

——. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 91: 473–489.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC.

Schafer, J. L., and J. W. Graham. 2002. Missing data: Our view of the state of the art. *Psychological Methods* 7: 147–177.

Schafer, J. L., M. Khare, and T. M. Ezzati-Rice. 1993. Multiple imputation of missing data in NHANES III. *Proceedings of the Annual Research Conference*, 459–487. U.S. Bureau of the Census: Washington, DC.

Schafer, J. L., and M. K. Olsen. 1998. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research* 33: 545–571.

Schenker, N., and J. M. G. Taylor. 1996. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis* 22: 425–446.

Tanner, M. A., and W. H. Wong. 1987. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* 82: 528–550.

van Buuren, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 16: 219–242.

van Buuren, S., H. C. Boshuizen, and D. L. Knook. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 18: 681–694.

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.

White, I. R., R. Daniel, and P. Royston. 2010. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical data. *Computational Statistics & Data Analysis* 54: 2267–2275.

White, I. R., P. Royston, and A. M. Wood. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 30: 377–399.

Also see

[MI] **intro** — Introduction to mi

[MI] **workflow** — Suggested workflow

[MI] **mi impute** — Impute missing values

[MI] **estimation** — Estimation commands for use with mi estimate

[MI] **mi estimate** — Estimation using multiple imputations

[MI] **Glossary**