Macros

e(mi)	mi
e(cmdline_mi)	command as typed
e(prefix_mi)	mi estimate
e(cmd_mi)	name of *estimation_command*
e(cmd)	mi estimate (equals e(cmd_mi) when post is used)
e(title_mi)	"Multiple-imputation estimates"
e(wvce_mi)	title used to label within-imputation variance in the table header
e(modeltest_mi)	title used to label the model test in the table header
e(dfadjust_mi)	title used to label the degrees-of-freedom adjustment in the table header
e(expnames_mi)	names of expressions specified in *spec*
e(exp#_mi)	expressions of the transformed coefficients specified in *spec*
e(rc_mi)	return codes for each imputation
e(m_mi)	specified imputation numbers
e(m_est_mi)	imputation numbers used in the computation
e(names_vvl_mi)	command-specific e() macro names that contents varied across imputations
e(names_vvm_mi)	command-specific e() matrix names that values varied across imputations (excluding b, V, and Cns)
e(names_vvs_mi)	command-specific e() scalar names that values varied across imputations

Matrices

e(b)	MI estimates of coefficients (equals e(b_mi), stored only if post is used)
e(V)	variance–covariance matrix (equals e(V_mi), stored only if post is used)
e(Cns)	constraint matrix (for constrained estimation only; equals e(Cns_mi), stored only if post is used)
e(N_g_mi)	group counts
e(g_min_mi)	group-size minimums
e(g_avg_mi)	group-size averages
e(g_max_mi)	group-size maximums
e(b[_Q]_mi)	MI estimates of coefficients (or transformed coefficients)
e(V[_Q]_mi)	variance–covariance matrix (total variance)
e(Cns_mi)	constraint matrix (for constrained estimation only)
e(W[_Q]_mi)	within-imputation variance matrix
e(B[_Q]_mi)	between-imputation variance matrix
e(re[_Q]_mi)	parameter-specific relative efficiencies
e(rvi[_Q]_mi)	parameter-specific RVIs
e(fmi[_Q]_mi)	parameter-specific FMIs
e(df[_Q]_mi)	parameter-specific degrees of freedom
e(pise[_Q]_mi)	parameter-specific percentages increase in standard errors
e(*vs_names*_vs_mi)	values of command-specific e() scalar *vs_names* that varied across imputations

vs_names include (but are not restricted to) df_r, N, N_strata, N_psu, N_pop, N_sub, N_postrata, N_stdize, N_subpop, N_over, and converged.

Results N_g_mi, g_min_mi, g_avg_mi, and g_max_mi are stored for panel-data models only. The results are stored as matrices for mixed-effects models and as scalars for other panel-data models.

If transformations are specified, the corresponding estimation results are stored with the _Q_mi suffix, as described above.

Command-specific e() results that remain constant across imputations are also stored. Command-specific results that vary from imputation to imputation are posted as missing, and their names are stored in the corresponding macros e(names_vvl_mi), e(names_vvm_mi), and e(names_vvs_mi). For some command-specific e() scalars (see *vs_names* above), their values from each imputation are stored in a corresponding matrix with the _vs_mi suffix.

Methods and formulas

Let $\mathbf{q}$ define a column vector of parameters of interest. For example, $\mathbf{q}$ may be a vector of coefficients (or functions of coefficients) from a regression model. Let $\{(\widehat{\mathbf{q}}_i, \widehat{\mathbf{U}}_i) : i = 1, 2, \ldots, M\}$

be the completed-data estimates of $\mathbf{q}$ and the respective variance–covariance estimates from M imputed datasets.

The MI estimate of $\mathbf{q}$ is

$$\overline{\mathbf{q}}_M = \frac{1}{M}\sum_{i=1}^{M}\widehat{\mathbf{q}}_i$$

The variance–covariance estimate (VCE) of $\overline{\mathbf{q}}_M$ (total variance) is

$$\mathbf{T} = \overline{\mathbf{U}} + (1+\frac{1}{M})\mathbf{B}$$

where $\overline{\mathbf{U}} = \sum_{i=1}^{M}\widehat{\mathbf{U}}_i/M$ is the within-imputation variance–covariance matrix and $\mathbf{B} = \sum_{i=1}^{M}(\mathbf{q}_i - \overline{\mathbf{q}}_M)(\mathbf{q}_i - \overline{\mathbf{q}}_M)'/(M-1)$ is the between-imputation variance–covariance matrix.

Methods and formulas are presented under the following headings:

Univariate case

Let Q, $\overline{Q}_M$, B, $\overline{U}$, and T correspond to the scalar analogues of the above formulas. Univariate inferences are based on the approximation

$$T^{-1/2}(Q - \overline{Q}_M) \sim t_\nu \tag{1}$$

where t_ν denotes a Student's t distribution with ν degrees of freedom, which depends on the number of imputations, M, and the increase in variance of estimates due to missing data. Under the large-sample assumption with respect to complete data, the degrees of freedom is

$$\nu_{\text{large}} = (M-1)\left(1+\frac{1}{r}\right)^2 \tag{2}$$

where

$$r = \frac{(1+M^{-1})B}{\overline{U}} \tag{3}$$

is an RVI due to missing data. Under the small-sample assumption, the degrees of freedom is

$$\nu_{\text{small}} = \left(\frac{1}{\nu_{\text{large}}} + \frac{1}{\widehat{\nu}_{\text{obs}}}\right)^{-1} \tag{4}$$

where $\widehat{\nu}_{\text{obs}} = \nu_c(\nu_c+1)(1-\gamma)/(\nu_c+3)$, $\gamma = (1+1/M)B/T$, and ν_c are the complete degrees of freedom, the degrees of freedom used for inference when data are complete (Barnard and Rubin 1999).

The small-sample adjustment (4) is applied to the degrees of freedom ν when the specified command stores the residual degrees of freedom in `e(df_r)`. This number of degrees of freedom is used as the complete degrees of freedom, ν_c, in the computation. (If `e(df_r)` varies across imputations, the smallest is used in the computation, resulting in conservative inference.) If `e(df_r)` is not set by the specified command or if the `nosmall` option is specified, then (2) is used to compute the degrees of freedom, ν.

Parameter-specific significance levels, confidence intervals, and degrees of freedom as reported by `mi estimate` are computed using the formulas above.

The percentage of standard-error increase due to missing data, as reported by `mi estimate, dftable`, is computed as $\{(T/\overline{U})^{1/2} - 1\} \times 100\%$.

The FMIs due to missing data and relative efficiencies reported by `mi estimate, vartable` are computed as follows.

In the large-sample case, the fraction of information about Q missing due to nonresponse (Rubin 1987, 77) is

$$\lambda = \frac{r + 2/(\nu_{\text{large}} + 3)}{r + 1}$$

where the RVI, r, is defined in (3). In the small-sample case, the fraction of information about Q missing due to nonresponse (Barnard and Rubin 1999, 953) is

$$\lambda = 1 - \frac{\lambda(\nu_{\text{small}})}{\lambda(\nu_c)} \frac{\overline{U}}{T}$$

where $\lambda(u) = (u + 1)/(u + 3)$.

The relative (variance) efficiency of using M imputations versus the infinite number of imputations is $\text{RE} = (1 + \lambda/M)^{-1}$ (Rubin 1987, 114).

Also see Rubin (1987, 76–77) and Schafer (1997, 109–111) for details.

Multivariate case

The approximation (1) can be generalized to the multivariate case:

$$(\mathbf{q} - \overline{\mathbf{q}}_M)\mathbf{T}^{-1}(\mathbf{q} - \overline{\mathbf{q}}_M)'/k \sim F_{k,\nu} \tag{5}$$

where $F_{k,\nu}$ denotes an F distribution with $k = \text{rank}(T)$ numerator degrees of freedom and ν denominator degrees of freedom defined as in (2), where the RVI, r, is replaced with the average RVI, r_{ave}:

$$r_{\text{ave}} = (1 + 1/M)\text{tr}(\mathbf{B}\overline{\mathbf{U}}^{-1})/k$$

The approximation (5) is inadequate with a small number of imputations because the between-imputation variance, $\mathbf{B}$, cannot be estimated reliably based on small M. Moreover, when M is smaller than the number of estimated parameters, $\mathbf{B}$ does not have a full rank. As such, the total variance, $\mathbf{T}$, may not be a valid variance–covariance matrix for $\overline{\mathbf{q}}_M$.

One solution is to assume that the between-imputation and within-imputation matrices are proportional, that is $B = \overline{\lambda} \times \overline{\mathbf{U}}$ (Rubin 1987, 78). This assumption implies that FMIs of all estimated parameters are equal. Under this assumption, approximation (5) becomes

$$(1 + r_{\text{ave}})^{-1}(\mathbf{q} - \overline{\mathbf{q}}_M)\overline{\mathbf{U}}^{-1}(\mathbf{q} - \overline{\mathbf{q}}_M)'/k \sim F_{k,\nu_\star} \tag{6}$$

where $k = \text{rank}(U)$ and $\nu_\star$ is computed as described in Li et al. (1991, 1067).

Also see Rubin (1987, 77–78) and Schafer (1997, 112–114) for details.

We refer to (6) as an equal FMI test and to (5) as the unrestricted FMI test. By default, `mi estimate` uses the approximation (6) for the model test. If the `ufmitest` option is specified, it uses the approximation (5) for the model test.

Similar to the univariate case, the degrees of freedom $\nu_\star$ and ν are adjusted for small samples when the command stores the completed-data residual degrees of freedom in `e(df_r)`.

In the small-sample case, the degrees of freedom $\nu_\star$ is computed as described in Reiter (2007) (in the rare case, when $k(M-1) \leq 4$, $\nu_\star = (k+1)\nu_1/2$, where ν_1 is the degrees of freedom from Barnard and Rubin [1999]). In the small-sample case, the degrees of freedom ν is computed as described in Barnard and Rubin (1999) and Marchenko and Reiter (2009).

Acknowledgments

The `mi estimate` command was inspired by the user-written command `mim` by John Carlin and John Galati, both of the Murdoch Children's Research Institute and University of Melbourne; Patrick Royston of the MRC Clinical Trials Unit, London, and coauthor of the Stata Press book *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*; and Ian White of the MRC Biostatistics Unit, London. We greatly appreciate the authors for their extensive body of work in Stata in the multiple-imputation area.

References

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

Cleves, M. A., W. W. Gould, R. G. Gutierrez, and Y. V. Marchenko. 2010. *An Introduction to Survival Analysis Using Stata*. 3rd ed. College Station, TX: Stata Press.

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

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.

Mortimore, P., P. Sammons, L. Stoll, D. Lewis, and R. Ecob. 1988. *School Matters*. Berkeley, CA: University of California Press.

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

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

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

Wagstaff, D. A., and O. Harel. 2011. A closer examination of three small-sample approximations to the multiple-imputation degrees of freedom. *Stata Journal* 11: 403–419.

White, I. R., and P. Royston. 2009. Imputing missing covariate values for the Cox model. *Statistics in Medicine* 28: 1982–1998.

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] **mi estimate postestimation** — Postestimation tools for mi estimate

[MI] **mi estimate using** — Estimation using previously saved estimation results

[MI] **intro** — Introduction to mi

[MI] **intro substantive** — Introduction to multiple-imputation analysis

[MI] **Glossary**