

mi impute mlogit — Impute using multinomial logistic regression

| | | | |
|----------------------|----------------|----------------------|------------|
| Description | Menu | Syntax | Options |
| Remarks and examples | Stored results | Methods and formulas | References |
| Also see | | | |

Description

`mi impute mlogit` fills in missing values of a nominal variable by using the multinomial (polytomous) logistic regression imputation method. You can perform separate imputations on different subsets of the data by specifying the `by()` option. You can also account for frequency, importance, and sampling weights.

Menu

Statistics > Multiple imputation

Syntax

```
mi impute mlogit ivar [indepvars] [if] [weight] [, impute_options options]
```

| <i>impute_options</i> | Description |
|---|--|
| Main | |
| * <code>add(#)</code> | specify number of imputations to add; required when no imputations exist |
| * <code>replace</code> | replace imputed values in existing imputations |
| <code>rseed(#)</code> | specify random-number seed |
| <code>double</code> | store imputed values in double precision; the default is to store them as <code>float</code> |
| <code>by(<i>varlist</i> [, <i>byopts</i>])</code> | impute separately on each group formed by <i>varlist</i> |
| Reporting | |
| <code>dots</code> | display dots as imputations are performed |
| <code>noisily</code> | display intermediate output |
| <code>nolegend</code> | suppress all table legends |
| Advanced | |
| <code>force</code> | proceed with imputation, even when missing imputed values are encountered |
| <code>noupdate</code> | do not perform mi update; see [MI] noupdate option |

*`add(#)` is required when no imputations exist; `add(#)` or `replace` is required if imputations exist. `noupdate` does not appear in the dialog box.

| <i>options</i> | Description |
|------------------------------|--|
| <code>noconstant</code> | suppress constant term |
| <code>baseoutcome(#)</code> | specify value of <i>ivar</i> that will be the base outcome |
| <code>augment</code> | perform augmented regression in the presence of perfect prediction |
| <code>conditional(if)</code> | perform conditional imputation |
| <code>bootstrap</code> | estimate model parameters using sampling with replacement |

Main

Maximization

`maximize_options` control the maximization process; seldom used

You must `mi set` your data before using `mi impute mlogit`; see [MI] [mi set](#).

You must `mi register` *ivar* as imputed before using `mi impute mlogit`; see [MI] [mi set](#).

indepvars may contain factor variables; see [U] [11.4.3 Factor variables](#).

`collect` is allowed; see [U] [11.1.10 Prefix commands](#).

`fweights`, `iwweights`, and `pweights` are allowed; see [U] [11.1.6 weight](#).

Options

Main

`noconstant`; see [R] [Estimation options](#).

`add()`, `replace`, `rseed()`, `double`, `by()`; see [MI] [mi impute](#).

`baseoutcome(#)` specifies the value of *ivar* to be treated as the base outcome. The default is to choose the most frequent outcome.

`augment` specifies that augmented regression be performed if perfect prediction is detected. By default, an error is issued when perfect prediction is detected. The idea behind the augmented-regression approach is to add a few observations with small weights to the data during estimation to avoid perfect prediction. See *The issue of perfect prediction during imputation of categorical data* under *Remarks and examples* in [MI] [mi impute](#) for more information. `augment` is not allowed with importance weights.

`conditional(if)` specifies that the imputation variable be imputed conditionally on observations satisfying *exp*; see [U] [11.1.3 if exp](#). That is, missing values in a conditional sample, the sample identified by the *exp* expression, are imputed based only on data in that conditional sample. Missing values outside the conditional sample are replaced with a conditional constant, the value of the imputation variable in observations outside the conditional sample. As such, the imputation variable is required to be constant outside the conditional sample. Also, if any conditioning variables (variables involved in the conditional specification `if exp`) contain soft missing values (`.`), their missing values must be nested within missing values of the imputation variables. See *Conditional imputation* under *Remarks and examples* in [MI] [mi impute](#).

`bootstrap` specifies that posterior estimates of model parameters be obtained using sampling with replacement; that is, posterior estimates are estimated from a bootstrap sample. The default is to sample the estimates from the posterior distribution of model parameters or from the large-sample normal approximation of the posterior distribution. This option is useful when asymptotic normality of parameter estimates is suspect.

Reporting

`dots`, `noisily`, `nolegend`; see [MI] **mi impute**. `noisily` specifies that the output from the multinomial logistic regression fit to the observed data be displayed. `nolegend` suppresses all legends that appear before the imputation table. Such legends include a legend about conditional imputation that appears when the `conditional()` option is specified and group legends that may appear when the `by()` option is specified.

Maximization

`maximize_options`; see [R] **mlogit**. These options are seldom used. `difficult`, `technique()`, `gradient`, `showstep`, `hessian`, and `showtolerance` are not allowed when the `augment` option is used.

Advanced

`force`; see [MI] **mi impute**.

The following option is available with `mi impute` but is not shown in the dialog box:

`noupdate`; see [MI] **noupdate option**.

Remarks and examples

[stata.com](http://www.stata.com)

Remarks are presented under the following headings:

Univariate imputation using multinomial logistic regression

Using mi impute mlogit

See [MI] **mi impute** for a general description and details about options common to all imputation methods, *impute_options*. Also see [MI] **Workflow** for general advice on working with `mi`.

Univariate imputation using multinomial logistic regression

The multinomial logistic regression imputation method can be used to fill in missing values of a nominal variable (for example, [Raghunathan et al. \[2001\]](#) and [van Buuren \[2007\]](#)). It is a parametric method that assumes an underlying multinomial logistic model for the imputed variable (given other predictors). Similarly to the logistic imputation method, this method is based on the asymptotic approximation of the posterior predictive distribution of the missing data.

Using mi impute mlogit

Consider the heart attack data introduced in [MI] **Intro substantive** and discussed in [MI] **mi impute**. Suppose that we want our logistic model of interest to also include information about marital status (categorical variable `marstatus`)—`logit attack smokes age bmi female hsgrad i.marstatus`.

We first tabulate values of `marstatus`:

```
. use https://www.stata-press.com/data/r17/mheart3
(Fictional heart attack data; marstatus missing)
. tabulate marstatus, missing
```

| Marital status: single, married, divorced | Freq. | Percent | Cum. |
|---|-------|---------|--------|
| Single | 53 | 34.42 | 34.42 |
| Married | 48 | 31.17 | 65.58 |
| Divorced | 46 | 29.87 | 95.45 |
| . | 7 | 4.55 | 100.00 |
| Total | 154 | 100.00 | |

From the output, the `marstatus` variable has three unique categories and seven missing observations. Because `marstatus` is a categorical variable, we use the multinomial logistic imputation method to fill in its missing values.

We `mi set` the data, register `marstatus` as an imputed variable, and then create 10 imputations by specifying the `add(10)` option with `mi impute mlogit`:

```
. mi set mlong
. mi register imputed marstatus
(7 m=0 obs now marked as incomplete)
. mi impute mlogit marstatus attack smokes age bmi female hsgrad, add(10)
Univariate imputation                Imputations =    10
Multinomial logistic regression      added =    10
Imputed: m=1 through m=10           updated =     0
```

| Variable | Observations per <i>m</i> | | | Total |
|------------------------|---------------------------|------------|---------|-------|
| | Complete | Incomplete | Imputed | |
| <code>marstatus</code> | 147 | 7 | 7 | 154 |

(Complete + Incomplete = Total; Imputed is the minimum across *m* of the number of filled-in observations.)

We can now analyze these multiply imputed data using logistic regression via `mi estimate`:

```
. mi estimate: logit attack smokes age bmi female hsgrad i.marstatus
(output omitted)
```

Stored results

mi impute mlogit stores the following in `r()`:

Scalars

| | |
|--------------------------|--|
| <code>r(M)</code> | total number of imputations |
| <code>r(M_add)</code> | number of added imputations |
| <code>r(M_update)</code> | number of updated imputations |
| <code>r(k_ivars)</code> | number of imputed variables (always 1) |
| <code>r(pp)</code> | 1 if perfect prediction detected, 0 otherwise |
| <code>r(N_g)</code> | number of imputed groups (1 if <code>by()</code> is not specified) |

Macros

| | |
|--------------------------|---|
| <code>r(method)</code> | name of imputation method (<code>mlogit</code>) |
| <code>r(ivars)</code> | names of imputation variables |
| <code>r(rngstate)</code> | random-number state used |
| <code>r(by)</code> | names of variables specified within <code>by()</code> |

Matrices

| | |
|------------------------------|--|
| <code>r(N)</code> | number of observations in imputation sample in each group |
| <code>r(N_complete)</code> | number of complete observations in imputation sample in each group |
| <code>r(N_incomplete)</code> | number of incomplete observations in imputation sample in each group |
| <code>r(N_imputed)</code> | number of imputed observations in imputation sample in each group |

Methods and formulas

Consider a univariate variable $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ that contains K categories (without loss of generality, let $k = 1$ be the base outcome) and follows a multinomial logistic model

$$\Pr(x_i = k | \mathbf{z}_i) = \begin{cases} \frac{1}{1 + \sum_{l=2}^K \exp(\mathbf{z}'_i \beta_l)}, & \text{if } k = 1 \\ \frac{\exp(\mathbf{z}'_i \beta_k)}{1 + \sum_{l=2}^K \exp(\mathbf{z}'_i \beta_l)}, & \text{if } k > 1 \end{cases} \quad (1)$$

where $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{iq})'$ records values of predictors of \mathbf{x} for observation i and β_l is the $q \times 1$ vector of unknown regression coefficients for outcome $l = 2, \dots, K$. (When a constant is included in the model—the default— $z_{i1} = 1$, $i = 1, \dots, n$.)

\mathbf{x} contains missing values that are to be filled in. Consider the partition of $\mathbf{x} = (\mathbf{x}'_o, \mathbf{x}'_m)$ into $n_o \times 1$ and $n_m \times 1$ vectors containing the complete and the incomplete observations. Consider a similar partition of $\mathbf{Z} = (\mathbf{Z}_o, \mathbf{Z}_m)$ into $n_o \times q$ and $n_m \times q$ submatrices.

mi impute mlogit follows the steps below to fill in \mathbf{x}_m :

1. Fit a multinomial logistic model (1) to the observed data $(\mathbf{x}_o, \mathbf{Z}_o)$ to obtain the maximum likelihood estimates, $\hat{\beta} = (\hat{\beta}'_2, \dots, \hat{\beta}'_K)'$, and their asymptotic sampling variance, $\hat{\mathbf{U}}$.
2. Simulate new parameters, β_* , from the large-sample normal approximation, $N(\hat{\beta}, \hat{\mathbf{U}})$, to its posterior distribution assuming the noninformative prior $\Pr(\beta) \propto \text{const}$.
3. Obtain one set of imputed values, \mathbf{x}_m^1 , by simulating from the multinomial logistic distribution: one of K categories is randomly assigned to a missing category, i_m , using the cumulative probabilities computed from (1) with $\beta_l = \beta_{*l}$ and $\mathbf{z}_i = \mathbf{z}_{i_m}$.
4. Repeat steps 2 and 3 to obtain M sets of imputed values, $\mathbf{x}_m^1, \mathbf{x}_m^2, \dots, \mathbf{x}_m^M$.

Steps 2 and 3 above correspond to only approximate draws from the posterior predictive distribution of the missing data $\Pr(\mathbf{x}_m | \mathbf{x}_o, \mathbf{Z}_o)$ because β_* is drawn from the asymptotic approximation to its posterior distribution.

If weights are specified, a weighted multinomial logistic regression model is fit to the observed data in step 1 (see [R] **mlogit** for details).

References

- 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.
- van Buuren, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 16: 219–242. <https://doi.org/10.1177/0962280206074463>.

Also see

- [MI] **mi impute** — Impute missing values
- [MI] **mi impute ologit** — Impute using ordered logistic regression
- [MI] **mi estimate** — Estimation using multiple imputations
- [MI] **Intro** — Introduction to mi
- [MI] **Intro substantive** — Introduction to multiple-imputation analysis