mi impute poisson — Impute using Poisson regression

Description

`mi impute poisson` fills in missing values of a count variable using a Poisson 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 poisson` *ivar* [*indepvars*] [*if*] [*weight*] [, *impute_options options*]

impute_options	Description
Main	
* `add(`*#*`)`	specify number of imputations to add; required when no imputations exist
* `replace`	replace imputed values in existing imputations
`rseed(`*#*`)`	specify random-number seed
`double`	store imputed values in double precision; the default is to store them as `float`
`by(`*varlist* [, *byopts*]`)`	impute separately on each group formed by *varlist*
Reporting	
`dots`	display dots as imputations are performed
`noisily`	display intermediate output
`nolegend`	suppress all table legends
Advanced	
`force`	proceed with imputation, even when missing imputed values are encountered
`noupdate`	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.

options	Description
Main	
noconstant	suppress constant term
exposure(*varname*$_e$)	include ln(*varname*$_e$) in model with coefficient constrained to 1
offset(*varname*$_o$)	include *varname*$_o$ in model with coefficient constrained to 1
conditional(*if*)	perform conditional imputation
bootstrap	estimate model parameters using sampling with replacement
Maximization	
maximize_options	control the maximization process; seldom used

You must mi set your data before using mi impute poisson; see [MI] **mi set**.

You must mi register *ivar* as imputed before using mi impute poisson; see [MI] **mi set**.

indepvars may contain factor variables; see [U] **11.4.3 Factor variables**.

fweights, iweights, 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**.

exposure(*varname*$_e$), offset(*varname*$_o$); see [R] **estimation options**.

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 Poisson 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] **poisson**. These options are seldom 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

Remarks are presented under the following headings:

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 Poisson regression

The Poisson regression imputation method can be used to fill in missing values of a count variable (for example, Raghunathan et al. [2001] and van Buuren [2007]). It is a parametric method that assumes an underlying Poisson model for the imputed variable (given other predictors). For imputation of overdispersed count variables, see [MI] **mi impute nbreg**. The Poisson method is based on the asymptotic approximation of the posterior predictive distribution of the missing data.

Using mi impute poisson

To illustrate the use of mi impute poisson, we continue with our heart attack data analysis example in [MI] **intro substantive** and consider an additional predictor, npreg, which records the number of pregnancies:

```
. use http://www.stata-press.com/data/r15/mheartpois
(Fictional heart attack data; npreg missing)
. misstable summarize
```

				Obs<.		
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
npreg	10		144	6	0	5

```
. tab female if npreg==.
```

Gender	Freq.	Percent	Cum.
Male	7	70.00	70.00
Female	3	30.00	100.00
Total	10	100.00	

According to misstable summarize, npreg is the only variable containing missing values, and it has 10 out of 154 observations missing. The tabulation of missing values of npreg by gender reveals that most missing values (7) correspond to males.

In this example, we could replace missing `npreg` for males with 0 and proceed with complete-data analysis, disregarding the remaining three missing observations. Instead, as an illustration, we use `mi impute poisson` to impute missing values of `npreg`. Our dataset is not declared yet, so we use `mi set` to declare it. We also use `mi register` to register `npreg` as the imputed variable before using `mi impute poisson`:

```
. mi set mlong

. mi register imputed npreg
(10 m=0 obs. now marked as incomplete)

. mi impute poisson npreg attack smokes age bmi hsgrad, add(20)
> conditional(if female==1)

Univariate imputation                          Imputations =       20
Poisson regression                                   added =       20
Imputed: m=1 through m=20                          updated =        0

Conditional imputation:
  npreg: incomplete out-of-sample obs. replaced with value 0
```

	Observations per m			
Variable	Complete	Incomplete	Imputed	Total
npreg	144	10	10	154

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

The `npreg` variable is relevant to females only, so we used the `conditional()` option to restrict imputation to observations with `female==1`; see *Conditional imputation* in [MI] **mi impute**.

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

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

Stored results

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

Scalars

`r(M)`	total number of imputations
`r(M_add)`	number of added imputations
`r(M_update)`	number of updated imputations
`r(k_ivars)`	number of imputed variables (always 1)
`r(N_g)`	number of imputed groups (1 if `by()` is not specified)

Macros

`r(method)`	name of imputation method (`poisson`)
`r(ivars)`	names of imputation variables
`r(rngstate)`	random-number state used
`r(by)`	names of variables specified within `by()`

Matrices

`r(N)`	number of observations in imputation sample in each group
`r(N_complete)`	number of complete observations in imputation sample in each group
`r(N_incomplete)`	number of incomplete observations in imputation sample in each group
`r(N_imputed)`	number of imputed observations in imputation sample in each group

Methods and formulas

Consider a univariate variable $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ that follows a Poisson model

$$\Pr(x_i = x|\mathbf{z}_i) = \frac{e^{-\lambda_i}\lambda_i^x}{x!}, \; x = 0, 1, 2, \ldots \tag{1}$$

where $\lambda_i = \exp(\mathbf{z}_i'\boldsymbol{\beta} + \text{offset}_i)$ (see [R] **poisson**), $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iq})'$ records values of predictors of $\mathbf{x}$ for observation i and $\boldsymbol{\beta}$ is the $q \times 1$ vector of unknown regression coefficients. (When a constant is included in the model—the default—$z_{i1} = 1, \; i = 1, \ldots, n$.)

$\mathbf{x}$ contains missing values that are to be filled in. Consider the partition of $\mathbf{x} = (\mathbf{x}_o', \mathbf{x}_m')$ into $n_0 \times 1$ and $n_1 \times 1$ vectors containing the complete and the incomplete observations. Consider a similar partition of $\mathbf{Z} = (\mathbf{Z}_o, \mathbf{Z}_m)$ into $n_0 \times q$ and $n_1 \times q$ submatrices.

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

1. Fit a Poisson regression model (1) to the observed data $(\mathbf{x}_o, \mathbf{Z}_o)$ to obtain the maximum likelihood estimates, $\widehat{\boldsymbol{\beta}}$, and their asymptotic sampling variance, $\widehat{\mathbf{U}}$.
2. Simulate new parameters, $\boldsymbol{\beta}_\star$, from the large-sample normal approximation, $N(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{U}})$, to its posterior distribution assuming the noninformative prior $\Pr(\boldsymbol{\beta}) \propto \text{const}$.
3. Obtain one set of imputed values, $\mathbf{x}_m^1$, by simulating from a Poisson distribution (1) with $\lambda_i = \lambda_{i_m} = \exp(\mathbf{z}_{i_m}'\boldsymbol{\beta}_\star + \text{offset}_{i_m})$.
4. Repeat steps 2 and 3 to obtain M sets of imputed values $\mathbf{x}_m^1, \mathbf{x}_m^2, \ldots, \mathbf{x}_m^M$.

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 $\boldsymbol{\beta}_\star$ is drawn from the asymptotic approximation to its posterior distribution.

If weights are specified, a weighted Poisson regression model is fit to the observed data in step 1 (see [R] **poisson** 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.

Also see

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

[MI] **mi impute nbreg** — Impute using negative binomial regression

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

[MI] **intro** — Introduction to mi

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