Title

stata.com

example 37g — Multinomial logistic regression

Description Remarks and examples Reference Also see

Description

With the data below, we demonstrate multinomial logistic regression, also known as multinomial logit, mlogit, and family multinomial, link logit:

```
. use http://www.stata-press.com/data/r15/gsem_sysdsn1
(Health insurance data)
. describe
Contains data from http://www.stata-press.com/data/r15/gsem_sysdsn1.dta
  obs:           644                          Health insurance data
 vars:            12                          28 Mar 2016 13:46
 size:        11,592                          (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
site            byte    %9.0g                 study site (1-3)
patid           float   %9.0g                 patient id
insure          byte    %9.0g      insure     insurance type
age             float   %10.0g                NEMC (ISCNRD-IBIRTHD)/365.25
male            byte    %8.0g                 NEMC PATIENT MALE
nonwhite        byte    %9.0g                 race
noinsur0        byte    %8.0g                 no insurance at baseline
noinsur1        byte    %8.0g                 no insurance at year 1
noinsur2        byte    %8.0g                 no insurance at year 2
ppd0            byte    %8.0g                 prepaid at baseline
ppd1            byte    %8.0g                 prepaid at year 1
ppd2            byte    %8.0g                 prepaid at year 2
-------------------------------------------------------------------------------
Sorted by: patid
. notes
_dta:
  1.  Data on health insurance available to 644 psychologically depressed
      subjects.
  2.  Data from Tarlov, A.R., et al., 1989, "The Medical Outcomes Study. An
      application of methods for monitoring the results of medical care." _J.
      of the American Medical Association_ 262, pp. 925-930.
  3.  insure: 1=indemnity, 2=prepaid, 3=uninsured.
```

See *Structural models 6: Multinomial logistic regression* in [SEM] **intro 5** for background.

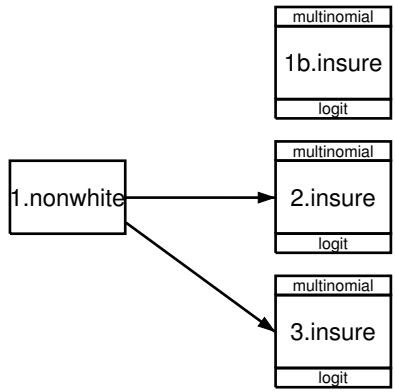
Remarks and examples

stata.com

Remarks are presented under the following headings:

Simple multinomial logistic regression model

In a multinomial logistic regression model, there are multiple unordered outcomes. In our case, these outcomes are recorded in variable `insure`. This variable records three different outcomes—indemnity, prepaid, and uninsured—recorded as 1, 2, and 3. The model we wish to fit is



The response variables are `1.insure`, `2.insure`, and `3.insure`, meaning `insure` $= 1$ (code for indemnity), `insure` $= 2$ (code for prepaid), and `insure` $= 3$ (code for uninsured). We specified that `insure` $= 1$ be treated as the mlogit base category by placing a `b` on `1.insure` to produce `1b.insure` in the variable box.

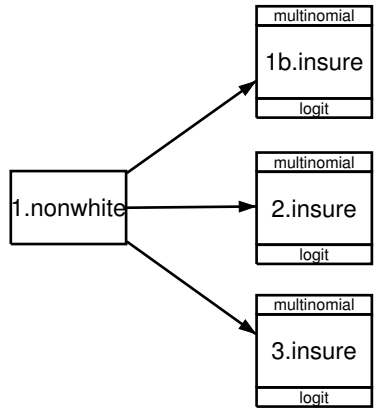
Notice that there are no paths into `1b.insure`. We could just as well have diagrammed the model with a path arrow from the explanatory variable into `1b.insure`. It would have made no difference.

In one sense, omitting the path is more mathematically appropriate, because multinomial logistic base levels are defined by having all coefficients constrained to be 0.

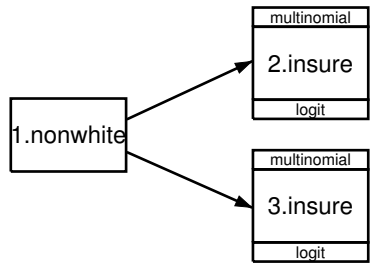
In another sense, drawing the path would be more appropriate because, even with `insure` $= 1$ as the base level, we are not assuming that outcome `insure` $= 1$ is unaffected by the explanatory variables. The probabilities of the three possible outcomes must sum to 1, and so any predictor that increases one probability of necessity causes the sum of the remaining probabilities to decrease. If a predictor x has positive effects (coefficients) for both `2.insure` and `3.insure`, then increases in x must cause the probability of `1.insure` to fall.

The choice of base outcome specifies that the coefficients associated with the other outcomes are to be measured relative to that base. In multinomial logistic regression, the coefficients are logs of the probability of the category divided by the probability of the base category, a mouthful also known as the log of the relative-risk ratio.

We drew the diagram one way, but we could just as well have drawn it like this:



In fact, we could just as well have chosen to indicate the base category by omitting it entirely from our diagram, like this:



Going along with that, we could type three different commands, each exactly corresponding to one of the three diagrams:

```
. gsem (1b.insure) (2.insure 3.insure <- i.nonwhite), mlogit
. gsem (1b.insure 2.insure 3.insure <- i.nonwhite), mlogit
. gsem (2.insure 3.insure <- i.nonwhite), mlogit
```

In the command language, however, we would probably just type

```
. gsem (i.insure <- i.nonwhite), mlogit
```

See [SEM] **intro 3** for a complete description of factor-variable notation. It makes no difference which diagram we draw or which command we type.

This model can be fit using the command syntax by typing

```
. gsem (i.insure <- i.nonwhite), mlogit
Iteration 0:   log likelihood = -556.59502
Iteration 1:   log likelihood = -551.78935
Iteration 2:   log likelihood = -551.78348
Iteration 3:   log likelihood = -551.78348
Generalized structural equation model           Number of obs     =        616
Response       : insure
Base outcome   : 1
Family         : multinomial
Link           : logit
Log likelihood = -551.78348
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.insure	(base outcome)					
2.insure						
1.nonwhite	.6608212	.2157321	3.06	0.002	.2379942	1.083648
_cons	-.1879149	.0937644	-2.00	0.045	-.3716896	-.0041401
3.insure						
1.nonwhite	.3779586	.407589	0.93	0.354	-.4209011	1.176818
_cons	-1.941934	.1782185	-10.90	0.000	-2.291236	-1.592632

Notes:

1. The above results say that nonwhites are more likely to have `insure` $= 2$ relative to 1 than whites, and that nonwhites are more likely to have `insure` $= 3$ relative to 1 than whites, which obviously implies that whites are more likely to have `insure` $= 1$.

2. For a three-outcome multinomial logistic regression model with the first outcome set to be the base level, the probability of each outcome is

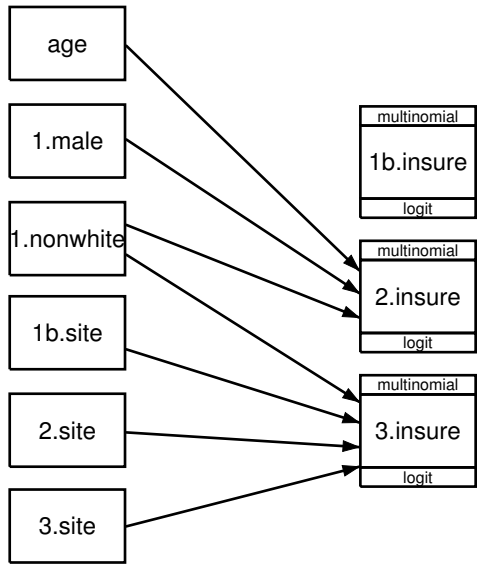
$$\Pr(y = 1) = 1/D$$
$$\Pr(y = 2) = \exp(\mathbf{X}_2\boldsymbol{\beta}_2)/D$$
$$\Pr(y = 2) = \exp(\mathbf{X}_3\boldsymbol{\beta}_3)/D$$

   where $D = 1 + \exp(\mathbf{X}_2\boldsymbol{\beta}_2) + \exp(\mathbf{X}_3\boldsymbol{\beta}_3)$.

3. For whites—that is, for `1.nonwhite` $= 0$—we have $\mathbf{X}_2\boldsymbol{\beta}_2 = -0.1879$ and $\mathbf{X}_3\boldsymbol{\beta}_3 = -1.9419$. Thus $D = 1.9721$, and the probabilities for each outcome are 0.5071, 0.4202, and 0.0727. Those probabilities sum to 1. You can make the similar calculations for nonwhites—that is, for `1.nonwhite` $= 1$—for yourself.

Multinomial logistic regression model with constraints

Using the same data, we wish to fit the following model:



In the above, `insure` = 2 and `insure` = 3 have paths pointing to them from different sets of predictors. They share predictor `1.nonwhite`, but `insure` = 2 also has paths from `age` and `1.male`, whereas `insure` = 3 also has paths from the `site` variables. When we fit this model, we will not obtain estimates of the coefficients on `age` and `1.male` in the equation for `insure` = 3. This is equivalent to constraining the coefficients for `age` and `1.male` to 0 in this equation. In other words, we are placing a constraint that the relative risk of choosing `insure` = 3 rather than `insure` = 1 is the same for males and females and is the same for all ages.

This model can be fit using command syntax by typing

```
. gsem (2.insure <- i.nonwhite age i.male)
>      (3.insure <- i.nonwhite i.site), mlogit
Iteration 0:   log likelihood = -555.85446
Iteration 1:   log likelihood = -541.20487
Iteration 2:   log likelihood = -540.85219
Iteration 3:   log likelihood = -540.85164
Iteration 4:   log likelihood = -540.85164
Generalized structural equation model           Number of obs     =        615
Response       : insure
Base outcome   : 1
Family         : multinomial
Link           : logit
Log likelihood = -540.85164
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.insure	(base outcome)					
2.insure						
1.nonwhite	.7219663	.2184994	3.30	0.001	.2937153	1.150217
age	-.0101291	.0058972	-1.72	0.086	-.0216874	.0014292
1.male	.5037961	.1912717	2.63	0.008	.1289104	.8786818
_cons	.1249932	.2743262	0.46	0.649	-.4126763	.6626627
3.insure						
1.nonwhite	.0569646	.4200407	0.14	0.892	-.7663001	.8802293
site						
2	-1.273576	.4562854	-2.79	0.005	-2.167879	-.3792728
3	.0434253	.3470773	0.13	0.900	-.6368337	.7236843
_cons	-1.558258	.2540157	-6.13	0.000	-2.056119	-1.060396

We could have gotten identical results from Stata's `mlogit` command for both this example and the previous one. To fit the first example, we would have typed

```
. mlogit insure i.nonwhite
```

To obtain the results for this second example, we would have been required to type a bit more:

```
. constraint 1 [Uninsure]age = 0
. constraint 2 [Uninsure]1.male = 0
. constraint 3 [Prepaid]2.site = 0
. constraint 4 [Prepaid]3.site = 0
. mlogit insure i.nonwhite age i.male i.site, constraints(1/4)
```

Having `mlogit` embedded in `gsem`, of course, also provides the advantage that we can combine the mlogit model with measurement models, multilevel models, and more. See [SEM] **example 41g** for a two-level multinomial logistic regression with random effects.

Fitting the simple multinomial logistic model with the Builder

Use the first diagram in *Simple multinomial logistic regression model* above for reference.

1. Open the dataset.

   In the Command window, type

   ```
   . use http://www.stata-press.com/data/r15/gsem_sysdsn1
   ```

2. Open a new builder diagram.

   Select menu item **Statistics > SEM (structural equation modeling) > Model building and estimation**.

3. Put the Builder in `gsem` mode by clicking on the GSEM button.

4. Create the rectangles for each possible outcome of the multinomial endogenous variable.

   Select the Add observed variables set tool, , and then click in the diagram about one-third of the way in from the right and one-fourth of the way up from the bottom.

   In the resulting dialog box,

   a. select the *Select variables* radio button (it may already be selected);

   b. check *Make variables generalized responses*;

   c. select `Multinomial, Logit` in the *Family/Link* control;

   d. select `insure` in the *Variable* control;

   e. select `Vertical` in the *Orientation* control;

   f. click on **OK**.

   If you wish, move the set of variables by clicking on any variable and dragging it.

5. Create the independent variable.

   a. Select the Add observed variable tool, , and then click in the diagram to the left of `2.insure`.

   b. In the Contextual Toolbar, type `1.nonwhite` in the *Variable* control and press *Enter*.

6. Create the paths from the independent variable to the rectangles for outcomes `insure` = 2 and `insure` = 3.

   a. Select the Add path tool, .

   b. Click in the right side of the `1.nonwhite` rectangle (it will highlight when you hover over it), and drag a path to the left side of the `2.insure` rectangle (it will highlight when you can release to connect the path).

   c. Continuing with the tool, click in the right side of the `1.nonwhite` rectangle and drag a path to the left side of the `3.insure` rectangle.

7. Clean up the location of the paths.

   If you do not like where the paths have been connected to the rectangles, use the Select tool, , to click on the path, and then simply click on where it connects to a rectangle and drag the endpoint.

8. Estimate.

   Click on the **Estimate** button, , in the Standard Toolbar, and then click on **OK** in the resulting *GSEM estimation options* dialog box.

You can open a completed diagram in the Builder by typing

```
. webgetsem gsem_mlogit1
```

Fitting the multinomial logistic model with constraints with the Builder

Use the diagram in *Multinomial logistic regression model with constraints* above for reference.

1. Open the dataset.

   In the Command window, type

   ```
   . use http://www.stata-press.com/data/r15/gsem_sysdsn1
   ```

2. Open a new Builder diagram.

   Select menu item **Statistics > SEM (structural equation modeling) > Model building and estimation**.

3. Put the Builder in `gsem` mode by clicking on the GSEM button.

4. Create the rectangles for each possible outcome of the multinomial endogenous variable.

   Select the Add observed variables set tool, , and then click in the diagram about one-third of the way in from the right and one-fourth of the way up from the bottom.

   In the resulting dialog box,

   a. select the *Select variables* radio button (it may already be selected);

   b. check *Make variables generalized responses*;

   c. select `Multinomial, Logit` in the *Family/Link* control;

   d. select `insure` in the *Variable* control;

   e. select `Vertical` in the *Orientation* control;

   f. click on **OK**.

   If you wish, move the set of variables by clicking on any variable and dragging it.

5. Create the independent variables.

   Select the Add observed variables set tool, , and then click in the diagram about one-third from the left and one-fourth from the bottom.

   In the resulting dialog box,

   a. select the *Select variables* radio button (it may already be selected);

   b. uncheck *Make variables generalized responses*;

   c. use the *Variables* control and select `age`;

   d. type `1.male 1.nonwhite` in the *Variables* control after `age` (typing `1.varname` rather than using the ... button to create them as `i.varname` factor variables prevents rectangles corresponding to the base categories for these binary variables from being created);

   e. include the levels of the factor variable `site` by clicking on the ... button next to the *Variables* control. In the resulting dialog box, select the *Factor variable* radio button, select `Main effect` in the *Specification* control, and select `site` in the *Variables* control for *Variable 1*. Click on **Add to varlist**, and then click on **OK**;

f. select `Vertical` in the *Orientation* control;

g. click on **OK**.

6. Create the paths from the independent variables to the rectangles for outcomes `insure` = 2 and `insure` = 3.

   a. Select the Add path tool, .

   b. Click in the right side of the `age` rectangle (it will highlight when you hover over it), and drag a path to the left side of the `2.insure` rectangle (it will highlight when you can release to connect the path).

   c. Continuing with the tool, create the following paths by clicking first in the right side of the rectangle for the independent variable and dragging it to the left side of the rectangle for the given outcome of the dependent variable:

```
1.male -> 2.insure
1.nonwhite -> 2.insure
1.nonwhite -> 3.insure
1b.site -> 3.insure
2.site -> 3.insure
3.site -> 3.insure
```

7. Clean up the location of the paths.

   If you do not like where the paths have been connected to the rectangles, use the Select tool,  , to click on the path, and then simply click on where it connects to a rectangle and drag the endpoint.

8. Estimate.

   Click on the **Estimate** button,  , in the Standard Toolbar, and then click on **OK** in the resulting *GSEM estimation options* dialog box.

You can open a completed diagram in the Builder by typing

```
. webgetsem sem_mlogit2
```

Reference

Tarlov, A. R., J. E. Ware, Jr., S. Greenfield, E. C. Nelson, E. Perrin, and M. Zubkoff. 1989. The medical outcomes study. An application of methods for monitoring the results of medical care. *Journal of the American Medical Association* 262: 925–930.

Also see

[SEM] **example 35g** — Ordered probit and ordered logit

[SEM] **example 41g** — Two-level multinomial logistic regression (multilevel)

[SEM] **gsem** — Generalized structural equation model estimation command

[SEM] **intro 5** — Tour of models