example 33g — Logistic regression

Description

In this example, we demonstrate with `gsem` how to fit a standard logistic regression, which is often referred to as the logit model in generalized linear model (GLM) framework.

```
use http://www.stata-press.com/data/r13/gsem_lbw
(Hosmer & Lemeshow data)
describe
Contains data from http://www.stata-press.com/data/r13/gsem_lbw.dta
obs: 189  Hosmer & Lemeshow data
vars: 11   21 Mar 2013 12:28
size: 2,646 (_dta has notes)
```

<table>
<thead>
<tr>
<th>variable name</th>
<th>storage</th>
<th>display format</th>
<th>value label</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>int</td>
<td>%8.0g</td>
<td>subject id</td>
</tr>
<tr>
<td>low</td>
<td>byte</td>
<td>%8.0g</td>
<td>birth weight &lt; 2500g</td>
</tr>
<tr>
<td>age</td>
<td>byte</td>
<td>%8.0g</td>
<td>age of mother</td>
</tr>
<tr>
<td>lwt</td>
<td>int</td>
<td>%8.0g</td>
<td>weight, last menstrual period</td>
</tr>
<tr>
<td>race</td>
<td>byte</td>
<td>%8.0g</td>
<td>race</td>
</tr>
<tr>
<td>smoke</td>
<td>byte</td>
<td>%9.0g</td>
<td>smoked during pregnancy</td>
</tr>
<tr>
<td>ptl</td>
<td>byte</td>
<td>%8.0g</td>
<td>premature labor history (count)</td>
</tr>
<tr>
<td>ht</td>
<td>byte</td>
<td>%8.0g</td>
<td>has history of hypertension</td>
</tr>
<tr>
<td>ui</td>
<td>byte</td>
<td>%8.0g</td>
<td>presence, uterine irritability</td>
</tr>
<tr>
<td>ftv</td>
<td>byte</td>
<td>%8.0g</td>
<td># physician visits, 1st trimester</td>
</tr>
<tr>
<td>bwt</td>
<td>int</td>
<td>%8.0g</td>
<td>birth weight (g)</td>
</tr>
</tbody>
</table>

Sorted by:

`. notes
_dta:
2. Data from a study of risk factors associated with low birth weights.


Remarks and examples

Remarks are presented under the following headings:

- Fitting the logit model
- Obtaining odds ratios
- Fitting the model with the Builder
Fitting the logit model

The model we wish to fit is

That is, we wish to fit a model in which low birthweight is determined by a history of hypertension \((ht)\), mother’s age \((age)\), mother’s weight at last menstrual period \((lwt)\), mother’s race (white, black, or other; \textit{race}), whether the mother smoked during pregnancy \((\text{smoke})\), the number of premature babies previously born to the mother \((ptl)\), and whether the mother has suffered from the presence of uterine irritability \((ui)\).

The path diagram matches the variable names listed in parentheses above except for \textit{race}, where the path diagram contains not one but three boxes filled in with \textit{1b.race}, \textit{2.race}, and \textit{3.race}. This is because in our dataset, race is coded 1, 2, or 3, meaning white, black, or other. We want to include indicator variables for \textit{race} so that we have a separate coefficient for each race. Thus we need three boxes.

In Stata, \textit{1.race} means “an indicator for race equaling 1”. Thus it should not surprise you if you filled in the boxes with \textit{1.race}, \textit{2.race}, and \textit{3.race}, and that is almost what we did. The difference is that we filled in the first box with \textit{1b.race} rather than \textit{1.race}. We use the \textit{b} to specify the base category, which we specified as white. If we wanted the base category to be black, we would have specified \textit{2b.race} and left \textit{1.race} alone.
The above is called factor-variable notation. See [SEM] intro 3 for details on using factor-variable notation with *gsem*.

In the command language, we could type

```
. gsem (low <- age lwt 1b.race 2.race 3.race smoke ptl ht ui), logit
```

to fit the model. Written that way, there is a one-to-one correspondence to what we would type and what we would draw in the Builder. The command language, however, has a feature that will allow us to type `i.race` instead of `1b.race 2.race 3.race`. To fit the model, we could type

```
. gsem (low <- age lwt i.race smoke ptl ht ui), logit
```

`i.varname` is a command-language shorthand for specifying indicators for all the levels of a variable and using the first level as the base category. You can use `i.varname` in the command language but not in path diagrams because boxes can contain only one variable. In the Builder, however, you will discover a neat feature so that you can type `i.race` and the Builder will create however many boxes are needed for you, filled in, and with the first category marked as the base. We will explain below how you do that.

The result of typing our estimation command is

```
. gsem (low <- age lwt i.race smoke ptl ht ui), logit
Iteration 0:  log likelihood = -101.0213
Iteration 1:  log likelihood = -100.72519
Iteration 2:  log likelihood = -100.724
Iteration 3:  log likelihood = -100.724
Generalized structural equation model  Number of obs = 189
Log likelihood = -100.724

+-----------------+-----------------+-----------------+-----------------+-----------------+
|     low <-      |     Coef.      |     Std. Err.   |       z        |     P>|z|     |     [95% Conf. Interval] |
|-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+
|      age        | -.0271003       | .0364504        | -0.74          | .457          | -.0985418 -.0443412 |
|      lwt        | -.0151508       | .0069259        | -2.19          | .029          | -.0287253 -.0015763 |
|      race       |                |                |                |              |                |
|  black          | 1.262647        | .5264101        | 2.40           | .016          | .2309024 2.294392 |
|  other          | .8620792        | .4391532        | 1.96           | .050          | .0013548 1.722804 |
|      smoke      | .9233448        | .4008266        | 2.30           | .021          | .137739 1.708951 |
|      ptl        | .5418366        | .346249         | 1.56           | .118          | -.136799 1.220472 |
|      ht         | 1.832518        | .6916292        | 2.65           | .008          | .4769494 3.188086 |
|      ui         | .7585135        | .4593768        | 1.65           | .099          | -.1418484 1.658875 |
|     _cons       | .4612239        | 1.20459         | 0.38           | .702          | -.1899729 2.822176 |
+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+
```

**Obtaining odds ratios**

Some of you are looking at the output above, nodding your heads, and thinking to yourselves, “Yes, that’s right.” Others are shaking your heads sadly and thinking, “Where are the exponentiated coefficients, the odds ratios?” Researchers from different backgrounds are used to seeing logit results presented in two different ways.
If you want to see the odds ratios, type `estat eform` after fitting the model:

```
. estat eform
```

|    | exp(b)  | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|----|---------|-----------|-----|------|----------------------|
| age| 0.9732636 | 0.0354759 | -0.74 | 0.457 | 0.9061578 1.045339 |
| lwt| 0.9849634 | 0.0068217 | -2.19 | 0.029 | 0.9716834 0.9984249 |
| race|        |           |      |      |                      |
| white | 3.534767 | 1.860737 | 2.40 | 0.016 | 1.259736 9.918406 |
| black | 2.368079 | 1.039949 | 1.96 | 0.050 | 1.001356 5.600207 |
| other| 2.517698 | 1.009160 | 2.30 | 0.021 | 1.147676 5.253162 |
| smoke| 1.719161 | 0.595257 | 2.85 | 0.004 | 1.11152 24.24199 |
| ptl | 6.249602 | 4.322408 | 1.45 | 0.149 | 1.611152 24.24199 |
| ht  | 2.1351 | 0.9808153 | 1.65 | 0.099 | 0.8677528 5.2534 |
| ui  | 1.586014 | 1.910496 | 0.38 | 0.702 | 0.1496092 16.8134 |

Whichever way you look at the results above, they are identical to the results that would be produced by typing

```
. logit low age lwt i.race smoke ptl ht ui
```

or

```
. logistic low age lwt i.race smoke ptl ht ui
```

which are two other ways that Stata can fit logit models. `logit`, like `gsem`, reports coefficients by default. `logistic` reports odds ratios by default.

### Fitting the model with the Builder

Use the diagram in *Fitting the logit model* above for reference.

1. Open the dataset.

   In the Command window, type
   
   ```
   . use http://www.stata-press.com/data/r13/gsem_lbw
   ```

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 `G` button.

4. Enlarge the size of the canvas to accommodate the height of the diagram.

   Click on the Adjust Canvas Size button, in the Standard Toolbar, change the second size to 5 (inches), and then click on OK.

5. Create the logistic regression component for `low`.

   Select the Add Regression Component tool, and then click in the diagram about one-third of the way in from the left and halfway down.

   In the resulting dialog box,

   a. select `low` in the *Dependent variable* control;
b. check Make measurements generalized;
c. select Bernoulli, Logit in the Family/Link control;
d. select the Select variables radio button (it may already be selected);
e. use the Independent variables control to select the variables age and lwt;
f. include the levels of the factor variable race by clicking on the .... button next to the Independent variables control. In the resulting dialog box, select the Factor variable radio button, select Main effect in the Specification control, and select race in the Variables control for Variable 1. Click on Add to varlist, and then click on OK;
g. continue with the Independent variables control to select the variables smoke, pt1, ht, and ui;
h. select Left in the Independent variables’ direction control;
i. click on OK.
If you wish, move the component by clicking on any variable and dragging it.
6. Clean up.
The box for low is created closer to the independent variables than it is in the example diagram. Use the Select tool, , and click on the box for low. Drag it to the right to allow more space for results along the paths.
7. 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_logit

Reference


Also see

- [SEM] example 34g — Combined models (generalized responses)
- [SEM] example 35g — Ordered probit and ordered logit
- [SEM] example 37g — Multinomial logistic regression
- [SEM] gsem — Generalized structural equation model estimation command
- [SEM] estat eform — Display exponentiated coefficients
- [SEM] intro 3 — Learning the language: Factor-variable notation (gsem only)
- [SEM] intro 5 — Tour of models