Example 33g — Logistic regression

Description Remarks and examples Reference Also see

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 https://www.stata-press.com/data/r19/gsem_lbw
(Hosmer & Lemeshow data)
. describe
Contains data from https://www.stata-press.com/data/r19/gsem_lbw.dta
 Observations:           189                  Hosmer & Lemeshow data
    Variables:            11                  21 Mar 2024 12:28
                                              (_dta has notes)
```

Variable name	Storage type	Display format	Value label	Variable label
id	int	%8.0g		Subject ID
low	byte	%8.0g		Birthweight < 2500g
age	byte	%8.0g		Age of mother
lwt	int	%8.0g		Weight, last menstrual period
race	byte	%8.0g	race	Race
smoke	byte	%9.0g	smoke	Smoked during pregnancy
ptl	byte	%8.0g		Premature labor history (count)
ht	byte	%8.0g		Has history of hypertension
ui	byte	%8.0g		Presence, uterine irritability
ftv	byte	%8.0g		# physician visits, 1st trimester
bwt	int	%8.0g		Birthweight (g)

```
Sorted by:
. notes
_dta:
  1.  Source: Data from Hosmer, Jr., D. W., S. A. Lemeshow, and R. X.
      Sturdivant. 2013. Applied Logistic Regression. 3rd ed. Hoboken, NJ:
      Wiley.
  2.  Data from a study of risk factors associated with low birthweights.
```

See *Structural models 3: Binary-outcome models* in [SEM] **Intro 5** for background.

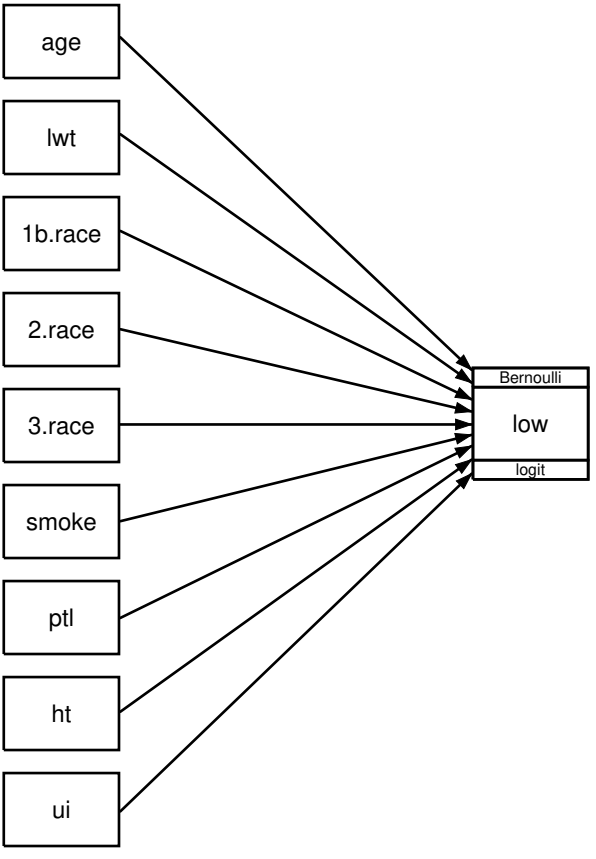
Remarks and examples

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; `race`), whether the mother smoked during pregnancy (`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 `race`, where the path diagram contains not one but three boxes filled in with `1b.race`, `2.race`, and `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 `race` so that we have a separate coefficient for each race. Thus we need three boxes.

In Stata, `1.race` means "an indicator for race equaling 1". Thus it should not surprise you if you filled in the boxes with `1.race`, `2.race`, and `3.race`, and that is almost what we did. The difference is that we filled in the first box with `1b.race` rather than `1.race`. We use the `b` to specify the base category, which we specified as white. If we wanted the base category to be black, we would have specified `2b.race` and left `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
Response: low
Family:   Bernoulli
Link:     Logit
Log likelihood = -100.724
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
low						
age	-.0271003	.0364504	-0.74	0.457	-.0985418	.0443412
lwt	-.0151508	.0069259	-2.19	0.029	-.0287253	-.0015763
race						
Black	1.262647	.5264101	2.40	0.016	.2309024	2.294392
Other	.8620792	.4391532	1.96	0.050	.0013548	1.722804
smoke	.9233448	.4008266	2.30	0.021	.137739	1.708951
ptl	.5418366	.346249	1.56	0.118	-.136799	1.220472
ht	1.832518	.6916292	2.65	0.008	.4769494	3.188086
ui	.7585135	.4593768	1.65	0.099	-.1418484	1.658875
_cons	.4612239	1.20459	0.38	0.702	-1.899729	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]
low						
age	.9732636	.0354759	-0.74	0.457	.9061578	1.045339
lwt	.9849634	.0068217	-2.19	0.029	.9716834	.9984249
race						
Black	3.534767	1.860737	2.40	0.016	1.259736	9.918406
Other	2.368079	1.039949	1.96	0.050	1.001356	5.600207
smoke	2.517698	1.00916	2.30	0.021	1.147676	5.523162
ptl	1.719161	.5952579	1.56	0.118	.8721455	3.388787
ht	6.249602	4.322408	2.65	0.008	1.611152	24.24199
ui	2.1351	.9808153	1.65	0.099	.8677528	5.2534
_cons	1.586014	1.910496	0.38	0.702	.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 https://www.stata-press.com/data/r19/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 GSEM 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`, `ptl`, `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

Hosmer, D. W., Jr., S. A. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.

Also see

[SEM] **Example 34g** — Combined models (generalized responses)

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

[SEM] **Example 37g** — Multinomial logistic regression

[SEM] **Intro 3** — Learning the language: Factor-variable notation (gsem only)

[SEM] **Intro 5** — Tour of models

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

[SEM] **estat eform** — Display exponentiated coefficients

Stata, Stata Press, Mata, NetCourse, and NetCourseNow are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. StataNow is a trademark of StataCorp LLC. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985–2025 StataCorp LLC, College Station, TX, USA. All rights reserved.

For suggested citations, see the FAQ on citing Stata documentation.