Example 39g — Three-level model (multilevel, generalized response)

Description

To demonstrate three-level models, we use the following data:

```
. use https://www.stata-press.com/data/r16/gsem_melanoma
(Skin cancer (melanoma) data)
. describe
Contains data from https://www.stata-press.com/data/r16/gsem_melanoma.dta
    obs: 354  Skin cancer (melanoma) data
    vars: 6    25 Mar 2018 15:28
           (_dta has notes)
storage  display  value
variable name  type    format  label variable label
nation     byte  %12.0g  nation  Nation ID
region     byte  %9.0g  Region ID: EEC level-I areas
county     int    %9.0g  County ID: EEC level-II/level-III areas
deads      int    %9.0g  No. deaths during 1971-1980
expected   float  %9.0g  No. expected deaths
uv          float  %9.0g  UV dose, mean-centered
```

Sorted by: nation  region  county

```
_dta:
2. Data on 7 nations, 3-95 regions w/i nation, 1-13 counties w/i region.
4. Variable expected contains # of expected male deaths based on crude rates for the combined counties.


Data are stored in the long form. Observations are counties within regions within nation. These data and some of the models fit below are also demonstrated in [ME] _membreg_.

See _Structural models 4: Count models_ and _Multilevel mixed-effects models_ in [SEM] _Intro 5_ for background.
Remarks and examples

Remarks are presented under the following headings:

- Three-level negative binomial model
- Three-level Poisson model
- Testing for overdispersion
- Fitting the models with the Builder

Three-level negative binomial model

The model we wish to fit is

Deaths due to malignant melanoma at the county level are modeled as being affected by ultraviolet exposure with random region and nation effects.
To fit this model, we type

```
.gsem (deaths <- uv M1[nation] M2[nation>region]), nbreg exposure(expected)
```

Fitting fixed-effects model:

<table>
<thead>
<tr>
<th>Iteration</th>
<th>log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1361.855</td>
</tr>
<tr>
<td>1</td>
<td>-1230.021</td>
</tr>
<tr>
<td>2</td>
<td>-1211.049</td>
</tr>
<tr>
<td>3</td>
<td>-1202.5641</td>
</tr>
<tr>
<td>4</td>
<td>-1202.5329</td>
</tr>
<tr>
<td>5</td>
<td>-1202.5329</td>
</tr>
</tbody>
</table>

Refining starting values:

<table>
<thead>
<tr>
<th>Grid node</th>
<th>log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1209.6951</td>
</tr>
</tbody>
</table>

Fitting full model:

<table>
<thead>
<tr>
<th>Iteration</th>
<th>log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1209.6951 (not concave)</td>
</tr>
<tr>
<td>1</td>
<td>-1195.0761 (not concave)</td>
</tr>
<tr>
<td>2</td>
<td>-1189.7235 (not concave)</td>
</tr>
<tr>
<td>3</td>
<td>-1167.58 (not concave)</td>
</tr>
<tr>
<td>4</td>
<td>-1145.4325 (not concave)</td>
</tr>
<tr>
<td>5</td>
<td>-1138.4471</td>
</tr>
<tr>
<td>6</td>
<td>-1088.3882</td>
</tr>
<tr>
<td>7</td>
<td>-1086.7992</td>
</tr>
<tr>
<td>8</td>
<td>-1086.4085</td>
</tr>
<tr>
<td>9</td>
<td>-1086.3903</td>
</tr>
<tr>
<td>10</td>
<td>-1086.3902</td>
</tr>
<tr>
<td>11</td>
<td>-1086.3902</td>
</tr>
</tbody>
</table>

Generalized structural equation model

<table>
<thead>
<tr>
<th>Response</th>
<th>Number of obs = 354</th>
</tr>
</thead>
<tbody>
<tr>
<td>deaths</td>
<td></td>
</tr>
<tr>
<td>Family</td>
<td>nbinomial</td>
</tr>
<tr>
<td>Dispersion</td>
<td>mean</td>
</tr>
<tr>
<td>Link</td>
<td>log</td>
</tr>
</tbody>
</table>

Log likelihood = -1086.3902

( 1) [deaths]M1[nation] = 1
( 2) [deaths]M2[nation>region] = 1

|                | Coef.  | Std. Err. | z      | P>|z|  | [95% Conf. Interval] |
|----------------|--------|-----------|--------|------|----------------------|
| deaths         |        |           |        |      |                      |
| uv             | -.0335933 | .0113725 | -2.95  | 0.003| -.055883 -.0113035   |
| M1[nation]     | 1 (constrained) |         |        |      |                      |
| M2[nation>region] | 1 (constrained) |         |        |      |                      |
| _cons          | -.0790606 | .1295931 | -0.61  | 0.542| -.3330538 .1749372   |
| ln(expected)   | 1 (exposure) |         |        |      |                      |
| /deaths       |        |           |        |      |                      |
| lnsalpha      | -4.182603 | .3415036 |        |      | -4.851937 -3.513268 |

|                |           |           |        |      |                      |
| var(M1[nation])| .1283614 | .0678971 |        |      | .0455187 .3619758    |
| var(M2[nation>region]) | .0401818 | .0104855 |        |      | .0240938 .067012    |
Notes:

1. This is a three-level model of counties nested within region nested within nation, so we specified the latent variables as $M_1[nation] M_2[nation>region]$. Actually, we did the same thing in the diagram when we used the SEM Builder to define the latent variables, but the nesting information does not show in the double rings.

2. We fit this model by using negative binomial regression, also known as a mean-dispersion model. In the command, we typed `nbreg`, which is shorthand for `family(nbinomial mean) link(log)`.

3. A negative binomial distribution can be regarded as a gamma mixture of Poisson random variables, where said gamma distribution has mean 1 and variance $\alpha$. The estimated ln($\alpha$) is $-4.183$, which is small; $\alpha$ is estimated as 0.0153.

4. Zero does not mean lack of overdispersion, because we are including random effects that also allow for extra dispersion. For a discussion on these issues, see [ME] `menbreg`.

5. Notice that we specified `exposure(expected)`, where variable `expected` contains the expected number of deaths based on crude rates.

   The `exposure()` option is allowed with Poisson and negative binomial models. If we specify `exposure(varname)`, we are usually saying that each observation’s time at risk is recorded in variable `varname`. When we omit the option, we are saying that each observation has the same time at risk. Obviously, if one observation had twice the time at risk of another observation, but was otherwise identical, we would expect twice the number of events in the first observation.

   In this case, however, we are using `exposure()` differently. We have a variable called expected containing the expected number of deaths from crude rates, and we are claiming `exposure(expected)`. What this is doing is saying that in two otherwise identical observations, if the number of expected deaths differed, we would expect the number of deaths due to melanoma to differ, too, and by the same proportion. See [SEM] `gsem family-and-link options`.

**Three-level Poisson model**

The same model, fit with Poisson, is

$$
\text{Poisson} \hbox{log} \text{uv} \text{region2} \text{nation1}
$$
To fit the model in the command language, we type

```
.gsem (deaths <- uv M1[nation] M2[nation>region]), poisson exposure(expected)
```

Fitting fixed-effects model:
Iteration 0: log likelihood = -2136.5847
Iteration 1: log likelihood = -1723.8955
Iteration 2: log likelihood = -1723.7727
Iteration 3: log likelihood = -1723.7727
Refining starting values:
Grid node 0: log likelihood = -1166.6536
Refining starting values (unscaled likelihoods):
Grid node 0: log likelihood = -1166.6536
Fitting full model:
Iteration 0: log likelihood = -1166.6536 (not concave)
Iteration 1: log likelihood = -1152.2741 (not concave)
Iteration 2: log likelihood = -1146.3094 (not concave)
Iteration 3: log likelihood = -1119.8479 (not concave)
Iteration 4: log likelihood = -1108.0129 (not concave)
Iteration 5: log likelihood = -1098.8067
Iteration 6: log likelihood = -1095.7563
Iteration 7: log likelihood = -1095.3164
Iteration 8: log likelihood = -1095.31
Iteration 9: log likelihood = -1095.31

Response : deaths
Family : Poisson
Link : log
Log likelihood = -1095.31
( 1) [deaths]M1[nation] = 1
( 2) [deaths]M2[nation>region] = 1

|                      | Coef.  | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|----------------------|--------|-----------|-------|-------|----------------------|
| deaths               |        |           |       |       |                      |
| uv                   | -0.0282041 | 0.0113998 | -2.47 | 0.013 | -0.0505473 -0.0058608 |
| M1[nation]           | 1 (constrained) |       |       |       |                      |
| M2[nation>region]    | 1 (constrained) |       |       |       |                      |
| _cons                | -0.0639672 | 0.1335515 | -0.48 | 0.632 | -0.3257234 0.197789  |
| ln(expected)         | 1 (exposure) |       |       |       |                      |
| var(                 |        |           |       |       |                      |
| M1[nation])          | 0.1371732 | 0.0723303 |       |       | 0.048802 0.3855676  |
| var(                 |        |           |       |       |                      |
| M2[nation>region])   | 0.0483483 | 0.0109079 |       |       | 0.0310699 0.0752353  |

Testing for overdispersion

The negative binomial model allows for overdispersion, or in a multilevel framework, allows for conditional overdispersion. The Poisson model has no overdispersion, or in a multilevel model, no overdispersion beyond that predicted by the latent variables. We can test whether there is dispersion beyond what Poisson would predict:
Example 39g — Three-level model (multilevel, generalized response)

```
gsem (deaths <- uv M1[nation] M2[nation>region]), nbreg exposure(expected)
(output omitted)
egress
.gsem (deaths <- uv M1[nation] M2[nation>region]), poisson exposure(expected)
(output omitted)
egoing
. estimates store nbreg
. estimates store poisson
. lrtest nbreg poisson
```

Likelihood-ratio test
LR chi2(1) = 17.84
(Assumption: poisson nested in nbreg)
Prob > chi2 = 0.0000

We can reject at any reasonable level that the Poisson model adequately accounts for the dispersion in these data. Be aware that this test is conservative, because we are testing whether a variance goes to 0. `lrtest` usually issues a warning in such cases, but `lrtest` does not know that the relationship between negative binomial regression and Poisson regression involves a variance going to 0.

Fitting the models with the Builder

Use the diagram in *Three-level negative binomial model* above for reference.

1. Open the dataset.
   In the Command window, type
   ```
   . use https://www.stata-press.com/data/r16/gsem_melanoma
   ```
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. Create the generalized response variable.
   a. Select the Add generalized response variable tool, `.`
   b. Click in the diagram about one-third of the way in from the right and one-fourth of the way up from the bottom.
   c. In the Contextual Toolbar, select `Nbinomial mean`, Log in the Family/Link control.
   d. In the Contextual Toolbar, select `deaths` in the Variable control.
5. Create the observed exogenous variable.
   a. Select the Add observed variable 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.
   b. In the Contextual Toolbar, select `uv` with the Variable control.
6. Create the level-three latent variable.
   a. Select the Add multilevel latent variable tool, `°`, and click above the rectangle for `uv` about one-fourth of the way down from the top.
   b. In the Contextual Toolbar, click on the `°` button.
   c. Select the nesting level and nesting variable by selecting 2 from the Nesting depth control and selecting nation > Observations in the next line.
d. Specify M1 as the Base name.
e. Click on OK.

7. Create the level-two latent variable.
a. Select the Add multilevel latent variable tool, 

b. In the Contextual Toolbar, click on the button.
c. Select the nesting level and nesting variable by selecting 3 from the Nesting depth control and selecting nation > region > Observations in the next control.
d. Specify M2 as the Base name.
e. Click on OK.

8. Create the paths from the exogenous variables to deaths.
a. Select the Add path tool, 

b. Click in the right side of the uv rectangle (it will highlight when you hover over it), and drag a path to the left side of the deaths rectangle (it will highlight when you can release to connect the path).
c. Continuing with the tool, draw paths from the right side of the double ovals for nation1 and region2 to the left side of the deaths rectangle.

9. Specify the level of exposure.

Use the Select tool, , and double-click in the deaths rectangle. In the resulting dialog box, select expected in the Exposure control, and click on OK.

10. Clean up the location of the paths.

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

11. 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_3lev`

References


Also see

[SEM] Example 38g — Random-intercept and random-slope models (multilevel)
[SEM] Example 34g — Combined models (generalized responses)
[SEM] Intro 5 — Tour of models
[SEM] gsem — Generalized structural equation model estimation command