Description

To demonstrate a finite mixture model (FMM), we use the following data:

```
. use https://www.stata-press.com/data/r16/gsem_mixture
(U.S. Medical Expenditure Panel Survey (2003))
```

Contains data from https://www.stata-press.com/data/r16/gsem_mixture.dta

| vars: 12 | 26 Jan 2019 08:46 |
| `_dta has notes` |

<table>
<thead>
<tr>
<th>variable name</th>
<th>storage</th>
<th>display</th>
<th>value</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>drvisits</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td>number of doctor visits</td>
</tr>
<tr>
<td>private</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>has private supplementary insurance</td>
</tr>
<tr>
<td>medicaid</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>has Medicaid public insurance</td>
</tr>
<tr>
<td>age</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>age in years</td>
</tr>
<tr>
<td>educ</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>years of education</td>
</tr>
<tr>
<td>actlim</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>has activity limitations</td>
</tr>
<tr>
<td>chronic</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>number of chronic conditions</td>
</tr>
<tr>
<td>income</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>income in $1,000s</td>
</tr>
<tr>
<td>offer</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>employer offers insurance</td>
</tr>
<tr>
<td>hpvisits</td>
<td>int</td>
<td>%8.0g</td>
<td></td>
<td>number of visits to health professionals other than doctors</td>
</tr>
<tr>
<td>female</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>female</td>
</tr>
<tr>
<td>phylim</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>has physical limitation</td>
</tr>
</tbody>
</table>

Sorted by:

```
. notes
```

_dta:

1. Data on annual number of doctor visits for individuals age 65 and older from the U.S. Medical Expenditure Panel Survey for 2003.
2. Data are analyzed in Cameron, A. C. and P. K. Trivedi, 2010, _Microeconometrics Using Stata_, Rev. Ed., College Station, TX: Stata Press.
3. Additional information on finite mixture models for count data and a similar example are found in Deb, P. and P. K. Trivedi, 1997, Demand for medical care by the elderly: A finite mixture approach, _Journal of Applied Econometrics_, vol. 12, 313--336.

See _Finite mixture models_ in [SEM] Intro 5 for background.
We are interested in fitting a Poisson regression to model the annual number of doctor visits as a function of whether an individual has private supplementary insurance, whether he or she has Medicaid, age, age squared, education level, whether he or she has activity limitations, and the number of chronic conditions. If we believed that the same model applied to the entire population, we could fit the model by typing

```
.poison drvisits private medicaid c.age##c.age educ actlim chronic
```

or, equivalently, by using gsem,

```
.gsem (drvisits <- private medicaid c.age##c.age educ actlim chronic), poisson
```

However, we believe that the model may differ across groups in the population. We do not have any information that identifies what these groups are or that tells us which individuals in our sample belong to each group. We can consider a categorical latent variable that identifies these groups and refer to the levels of this latent variable as latent classes. With an FMM, we can incorporate the categorical latent variable into our model to account for differences across the latent classes.

Following Cameron and Trivedi (2010), we will fit an FMM with a Poisson regression component for each latent class. We will estimate distinct coefficients for the Poisson model in each class, and we will estimate the probability of belonging to each of these classes using a multinomial logistic regression. We fit the model as follows:

```
.gsem (drvisits <- private medicaid c.age##c.age educ actlim chronic), > poisson lclass(C 2) startvalues(randomid, draws(5) seed(15))
```

Computing starting values using randomid:

```
(Iteration log omitted)
```

```
Generalized structural equation model Number of obs = 3,677
Log likelihood = -11502.686

| Coef. | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|-------|-----------|------|-----|---------------------|
| 1.C   | (base outcome) |
| 2.C   | _cons     | 0.877227 | 0.0494614 | 17.74 | 0.000 | 0.7802845 | 0.9741696 |
```

Class : 1
Response : drvisits
Family : Poisson
Link : log

| Coef. | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|-------|-----------|------|-----|---------------------|
| drvisits | private  | .138229 | .0247626 | 5.58 | 0.000 | .0896951 | .1867629 |
|        | medicaid | .1269723 | .0341525 | 3.72 | 0.000 | .0600345 | .19391 |
|        | age      | .2628874 | .0466774 | 5.63 | 0.000 | .1714014 | .3543735 |
|        | c.age#c.age | -.0017418 | .0003108 | -5.60 | 0.000 | -.002351 | -.0011326 |
|        | educ     | .0241679 | .0030705 | 7.87 | 0.000 | .0181499 | .030186 |
|        | actlim   | .1831598 | .0238817 | 7.67 | 0.000 | .1363525 | .2299671 |
|        | chronic  | .1970511 | .0088783 | 22.19 | 0.000 | .17965 | .2144523 |
|        | _cons    | -8.051256 | 1.741677 | -4.62 | 0.000 | -11.46488 | -4.637632 |
Example 53g — Finite mixture Poisson regression

Class : 2  
Response : drvisits  
Family : Poisson  
Link : log

|            | Coef.       | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|------------|-------------|-----------|------|------|----------------------|
| drvisits   |             |           |      |      |                      |
|            | private     | .2077415  | .0306353 | 6.78 | 0.000 | .1476974  | .2677856 |
|            | medicaid    | .1071618  | .0407211 | 2.63 | 0.008 | .02735    | .1869736 |
|            | age         | .3798087  | .0562035 | 6.76 | 0.000 | .269652   | .4899655 |
|            | c.age#c.age | -.0024869 | .0003736 | -6.66| 0.000 | -.0032191 | -.0017547 |
|            | educ        | .029099   | .003972  | 7.33 | 0.000 | .021314   | .0368841 |
|            | actlim      | .1244235  | .0310547 | 4.01 | 0.000 | .0635574  | .1852895 |
|            | chronic     | .3191166  | .0089757 | 35.55| 0.000 | .3015247  | .3367086 |
|            | _cons       | -14.25713 | 2.101964 | -6.78| 0.000 | -18.37691 | -10.13736 |

Notes:

1. We used the lclass(C 2) to specify that our categorical latent variable is named C and has two latent classes.

2. The first table in the output provides the estimated coefficients in the multinomial logit model for C.

3. The next two tables are the results for the Poisson regression models for the first and second classes. By default, the coefficients and intercepts vary across the classes. We can specify lcinvariant(cons) if we want intercepts to be constrained to be equal across classes, or we can specify lcinvariant(coef) if we want all coefficients constrained to be equal across classes. See [SEM] gsem lc options for details on the lcinvariant() option.

4. We added the startvalues(randomid), draws(5) seed(15)) option to request that starting values be computed using random class assignments. In this option, draws(5) specifies that five random draws be taken and that the one with the best log likelihood after the EM iterations be selected. If you fit FMMs and other models with categorical latent variables, taking multiple draws of random starting values can help to prevent convergence at a local maximum rather than the global maximum. gsem provides a variety of options for obtaining starting values. See [SEM] Intro 12 and [SEM] gsem estimation options for more information on starting values.

5. The fmm: prefix can be used to fit finite mixture regression models with a single response variable. We could have fit this same model with fmm: poisson by typing

```
fmm 2, startvalues(randomid, draws(5) seed(15)): ///
poisson drvisits private medicaid c.age##c.age educ actlim chronic
```
We can use \texttt{estat lcprob} to estimate the proportion of individuals in each class.

\begin{verbatim}
. estat lcprob

Latent class marginal probabilities
Number of obs = 3,677

+----------------+------------------+------------------+
<table>
<thead>
<tr>
<th>C</th>
<th>Margin</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.2937527</td>
<td>.0102614</td>
</tr>
<tr>
<td>2</td>
<td>.7062473</td>
<td>.0102614</td>
</tr>
</tbody>
</table>
+----------------+------------------+------------------+

We find that about 29% of the population is in class 1 and about 71% is in class 2.

To better understand these classes, we use \texttt{estat lcmean} to estimate the marginal predicted counts (means) for each class.

\begin{verbatim}
. estat lcmean

Latent class marginal means
Number of obs = 3,677

+----------------+----------------+------------------+------------------+
<table>
<thead>
<tr>
<th>C</th>
<th>drvisits</th>
<th>Margin</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.95943</td>
<td>.1767506</td>
<td>78.98</td>
</tr>
<tr>
<td>2</td>
<td>3.801692</td>
<td>.0587685</td>
<td>64.69</td>
</tr>
</tbody>
</table>
+----------------+----------------+------------------+------------------+

Class 1 appears to represent those who visit the doctor frequently and class 2, those who visit less frequently.

We can also predict the posterior probabilities of class membership and then use those to determine the predicted class for each individual.

\begin{verbatim}
. predict postpr_dr*, classposteriorpr
. generate pclass_dr = 1 + (postpr_dr2>0.5)
. tabulate pclass_dr

pclass_dr |   Freq. |  Percent |    Cum. |
-----------|--------|----------|--------|
     1      | 1,061  |   28.86  |  28.86 |
     2      | 2,616  |   71.14  | 100.00 |
Total      | 3,677  | 100.00   |        |
\end{verbatim}

We see that 1,061 individuals in our sample are predicted to be in class 1, the class that frequently visits the doctor.

Our dataset also includes the variable \texttt{hpvisits}, which records the number of visits individuals make to health professionals other than doctors. We fit a similar model to the one above but with \texttt{hpvisits} as our response variable.
Example 53g — Finite mixture Poisson regression

```
gsem (hpvisits <- private medicaid c.age##c.age educ actlim chronic),
> poisson lclass(C 2) startvalues(classid pclass_dr)

(iteration log omitted)
```

Generalized structural equation model

| Coef. | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|-------|-----------|------|------|----------------------|
| 1.C (base outcome) | | | | |
| 2.C _cons | 2.241837 | .059523 | 37.66 | 0.000 | 2.125174 2.3585 |

Class : 1
Response : hpvisits
Family : Poisson
Link : log

```
Class : 2
Response : hpvisits
Family : Poisson
Link : log
```

This time, we used the `startvalues(classid pclass_dr)` option to specify how starting values are calculated. This means that we are using the variable `pclass_dr` as an initial guess of class membership to be used when computing starting values.
We again use `estat lcprob` to estimate the predicted proportion of the population in each class.

```
estat lcprob
Latent class marginal probabilities

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Margin</td>
<td>Std. Err.</td>
<td>[95% Conf. Interval]</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>.0960559</td>
<td>.0051683</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.9039441</td>
<td>.0051683</td>
</tr>
</tbody>
</table>
```

This time about 10% is in class 1, and 90% is in class 2.

We can predict the class for each individual based on this model and compare the classifications from the two models.

```
.predict postpr_hp*, classposteriorpr
.generate pclass_hp = 1 + (postpr_hp2>0.5)
.tabulate pclass_hp pclass_dr
```

<table>
<thead>
<tr>
<th>pclass_dr</th>
<th>pclass_hp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>169</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>892</td>
<td>2,436</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1,061</td>
<td>2,616</td>
</tr>
</tbody>
</table>

Many individuals are predicted to be in class 2 based on both models, meaning that they are in the group that visits the doctor infrequently and in the group that visits other health professionals infrequently. However, there are also 892 that are classified differently by the two models. These individuals are in the class that visits the doctor frequently based on the first model but in the class that visits other healthcare professionals infrequently based on the second model.

In [SEM] Example 54g, we consider simultaneously modeling `drvisits` and `hpvisits` and using a single categorical latent variable that identifies groups in the population.

References


Also see

- [SEM] Example 54g — Finite mixture Poisson regression, multiple responses
- [SEM] Intro 5 — Tour of models
- [SEM] gsem — Generalized structural equation model estimation command
- [SEM] estat lcmean — Latent class marginal means
- [SEM] estat lcprob — Latent class marginal probabilities
- [FMM] fmm intro — Introduction to finite mixture models
- [FMM] fmm: poisson — Finite mixtures of Poisson regression models