Latent Class Analysis (LCA) in Stata

Kristin MacDonald

Director of Statistical Services
StataCorp LLC

2018 London Stata Conference
What is latent class analysis (LCA)?

- We believe that there are groups in a population and that individuals in these groups behave differently.
- We often have variables in our dataset that record group membership.
- For instance, we might have variables indicating:
  - age group
  - male or female
  - employed or unemployed
  - has high blood pressure or not
- When groupings are known, we can test for differences in other variables across groups, allow regression models to differ across groups, and make other comparisons of the groups.
What is latent class analysis (LCA)?

- We believe that there are groups in a population and that individuals in these groups behave differently.
- We often have variables in our dataset that record group membership.
- For instance, we might have variables indicating
  - age group
  - male or female
  - employed or unemployed
  - has high blood pressure or not
- When groupings are known, we can test for differences in other variables across groups, allow regression models to differ across groups, and make other comparisons of the groups.
What is latent class analysis (LCA)?

- We believe that there are groups in a population and that individuals in these groups behave differently.
- We often have variables in our dataset that record group membership.
- For instance, we might have variables indicating
  - age group
  - male or female
  - employed or unemployed
  - has high blood pressure or not
- When groupings are known, we can test for differences in other variables across groups, allow regression models to differ across groups, and make other comparisons of the groups.
Sometimes we believe groups exist, but we do not have a variable that records group membership.

For instance, we might believe that there exist

- groups of consumers with different buying preferences
- groups of adolescents with different propensities for delinquent behaviors
- groups of individuals who respond differently to a treatment
- groups of ...
Sometimes we believe groups exist, but we do not have a variable that records group membership.

For instance, we might believe that there exist

- groups of consumers with different buying preferences
- groups of adolescents with different propensities for delinquent behaviors
- groups of individuals who respond differently to a treatment
- groups of ...
Sometimes we believe groups exist, but we do not have a variable that records group membership.

For instance, we might believe that there exist

- groups of consumers with different buying preferences
- groups of adolescents with different propensities for delinquent behaviors
- groups of individuals who respond differently to a treatment
- groups of ...
Sometimes we believe groups exist, but we do not have a variable that records group membership.

For instance, we might believe that there exist

- groups of consumers with different buying preferences
- groups of adolescents with different propensities for delinquent behaviors
- groups of individuals who respond differently to a treatment
- groups of ...
Sometimes we believe groups exist, but we do not have a variable that records group membership.

For instance, we might believe that there exist

- groups of consumers with different buying preferences
- groups of adolescents with different propensities for delinquent behaviors
- groups of individuals who respond differently to a treatment
- groups of ...
Using LCA we can fit a model and try to determine which individuals are likely to belong to each group based on information available in other variables.

One common use of LCA is as a model-based method of clustering.
We believe that there are different types of people who attend Stata conferences.

We hypothesize that there are three groups. Our intuition tells us the groups might be characterized as

1. Stata promoters—those who love Stata, encourage others to use Stata, and provide resources for others
2. Stata researchers—those who use Stata regularly for their own research
3. Stata novices—those who have used Stata for a short time and want to learn more
We have a sample of individuals who have attended conferences around the world.

We don’t have a variable that records the whether each individual is a Stata promoter, researcher, or novice. Instead, attendee classification can be considered a latent (unobserved) variable.
Each conference attendee in our sample answered the following questions:

1. Do you use Stata at least once per week?
2. Have you ever written and distributed a Stata command?
3. Have you used Stata for more than 5 years?
4. Have you presented at a previous Stata conference?
5. Do you teach a course using Stata?
6. Have you published a paper based on data analyzed using Stata?
7. Have you published an article in the Stata Journal?
8. Do you regularly participate in discussions on Statalist?
9. Do you live within 50 miles of the conference?
### What is latent class analysis (LCA)?

### Example of classic LCA

```stata
. summarize

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>weekly</td>
<td>576</td>
<td>.52083</td>
<td>.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>command</td>
<td>576</td>
<td>.29861</td>
<td>.45805</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>years5</td>
<td>576</td>
<td>.48264</td>
<td>.50013</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>presenter</td>
<td>576</td>
<td>.34028</td>
<td>.47421</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>teacher</td>
<td>576</td>
<td>.42014</td>
<td>.49401</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>published</td>
<td>576</td>
<td>.49306</td>
<td>.50039</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sjauthor</td>
<td>576</td>
<td>.31423</td>
<td>.46461</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>statalist</td>
<td>576</td>
<td>.36285</td>
<td>.48124</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>location</td>
<td>576</td>
<td>.51563</td>
<td>.50019</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
```

K. L. MacDonald (StataCorp)  6-7 September 2018  8 / 52
Do our data support our hypothesized grouping?

- Have we proposed the correct number of groups?
- Do our descriptions accurately characterize the types of people who attend Stata conferences?
- Can we predict who is likely to belong to each group?
- We use the `gsem` command to fit a latent class model.

  . gsem (weekly command years5 presenter teacher //
  published sjauthor statlist location <- ), //
  logit lclass(C 3)

- The `lclass(C 3)` option specifies that we want to allow for differences in these logistic regression models across the levels of a categorical latent variable named C with three classes.

- Our observed variables are all binary, and we use the `logit` option to model each one using a constant-only logistic regression.
Latent Class Analysis

What is latent class analysis (LCA)?

Example of classic LCA

- We will not look at the `gsem` output yet. It is easier to interpret results using `estat lcprob` and `estat lcmean`.

- Based on this model, what are the expected proportions of the population in each group?

```
. estat lcprob
Latent class marginal probabilities
Number of obs = 576

                  Delta-method
             Margin  Std. Err.  [95% Conf. Interval]

   C1     .1057509   .0582876   .0341272   .2835627
   C2     .4187809   .0704887   .2900013   .5596688
   C3     .4754682   .0397848   .3987046   .5534088
```

- We estimate that 10.6% of the population is in class 1, 41.9% is in class 2, and 47.5% is in class 3.

- But what do those classes represent?
For individuals in Class 1, what is the probability of responding positively to each question?

```
. estat lcmean
Latent class marginal means
Number of obs = 576
```

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Margen</td>
<td>Std. Err.</td>
<td>[95% Conf. Interval]</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>weekly</td>
<td>.5594732</td>
<td>.1144653</td>
<td>.338218 .759382</td>
</tr>
<tr>
<td>command</td>
<td>.703362</td>
<td>.1655266</td>
<td>.3336843 .9182112</td>
</tr>
<tr>
<td>years5</td>
<td>.9462668</td>
<td>.1099533</td>
<td>.2644505 .9988421</td>
</tr>
<tr>
<td>presenter</td>
<td>.5892076</td>
<td>.1128971</td>
<td>.3650511 .7815784</td>
</tr>
<tr>
<td>teacher</td>
<td>.596822</td>
<td>.0986313</td>
<td>.3986389 .7677449</td>
</tr>
<tr>
<td>published</td>
<td>.8785688</td>
<td>.0824458</td>
<td>.6140342 .9705049</td>
</tr>
<tr>
<td>sjauthor</td>
<td>.7467327</td>
<td>.1777284</td>
<td>.3185127 .9489785</td>
</tr>
<tr>
<td>statalist</td>
<td>.4410877</td>
<td>.1074878</td>
<td>.2513733 .6497189</td>
</tr>
<tr>
<td>location</td>
<td>.1202751</td>
<td>.0922665</td>
<td>.0241521 .4302775</td>
</tr>
</tbody>
</table>

The marginal probabilities of answering yes are high for all questions except the one about living nearby. This might be our hypothesized "Stata Promoters" group.
What about individuals in Class 2?

<table>
<thead>
<tr>
<th></th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>weekly</td>
<td>.7953942</td>
</tr>
<tr>
<td>command</td>
<td>.2682777</td>
</tr>
<tr>
<td>years5</td>
<td>.7053751</td>
</tr>
<tr>
<td>presenter</td>
<td>.5136087</td>
</tr>
<tr>
<td>teacher</td>
<td>.5796951</td>
</tr>
<tr>
<td>published</td>
<td>.6302565</td>
</tr>
<tr>
<td>sjauthor</td>
<td>.3026139</td>
</tr>
<tr>
<td>statalist</td>
<td>.5908731</td>
</tr>
<tr>
<td>location</td>
<td>.4509978</td>
</tr>
</tbody>
</table>

The marginal probabilities of using Stata weekly, having used Stata for more than five years, and publishing articles based on data analyzed in Stata are fairly large.

These individuals are less likely to have written a Stata command or to have published in the Stata Journal.

This class might be our hypothesized “Stata Researchers”.
What do we expect in Class 3?

<table>
<thead>
<tr>
<th></th>
<th>3weekly</th>
<th>command</th>
<th>years5</th>
<th>presenter</th>
<th>teacher</th>
<th>published</th>
<th>sjauthor</th>
<th>statalist</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.270413</td>
<td>.2353055</td>
<td>.1833394</td>
<td>.1322467</td>
<td>.2403093</td>
<td>.2864695</td>
<td>.2282789</td>
<td>.1446059</td>
<td>.6604777</td>
</tr>
<tr>
<td></td>
<td>.0382115</td>
<td>.0288825</td>
<td>.0370618</td>
<td>.0255786</td>
<td>.0312686</td>
<td>.0349021</td>
<td>.029189</td>
<td>.0295687</td>
<td>.0334121</td>
</tr>
<tr>
<td></td>
<td>.2022746</td>
<td>.1834426</td>
<td>.1214216</td>
<td>.089635</td>
<td>.1844201</td>
<td>.2231754</td>
<td>.1761288</td>
<td>.0956889</td>
<td>.592279</td>
</tr>
<tr>
<td></td>
<td>.3513939</td>
<td>.2965067</td>
<td>.2672279</td>
<td>.1908686</td>
<td>.3067651</td>
<td>.3594091</td>
<td>.290427</td>
<td>.2126493</td>
<td>.7226114</td>
</tr>
</tbody>
</table>

These individuals are likely to live close to the conference, but they have lower probabilities of answering yes to all other questions.

This class might be our hypothesized "Stata Novice" group.
Did this model fit well?

- **`estat lcgof`** reports goodness-of-fit statistics.

```
. estat lcgof

<table>
<thead>
<tr>
<th>Fit statistic</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood ratio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chi2_ms(482)</td>
<td>460.457</td>
<td>model vs. saturated</td>
</tr>
<tr>
<td>p &gt; chi2</td>
<td>0.753</td>
<td></td>
</tr>
<tr>
<td>Information criteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>6624.113</td>
<td>Akaike’s information criterion</td>
</tr>
<tr>
<td>BIC</td>
<td>6750.441</td>
<td>Bayesian information criterion</td>
</tr>
</tbody>
</table>
```

- We fail to reject the null hypothesis that our model fits as well as a saturated model.
- The AIC and BIC are useful when we want to compare models.
We can use `predict, classposteriorpr` to estimate probabilities of belonging to class 1, class 2, and class 3.

Let’s select the class with the highest predicted probability as being the predicted class.

```
predict cpost*, classposteriorpr
egen max = rowmax(cpost*)
generate predclass = 1 if cpost1==max
(528 missing values generated)
replace predclass = 2 if cpost2==max
(250 real changes made)
replace predclass = 3 if cpost3==max
(278 real changes made)
tabulate predclass
```

<table>
<thead>
<tr>
<th>predclass</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48</td>
<td>8.33</td>
<td>8.33</td>
</tr>
<tr>
<td>2</td>
<td>250</td>
<td>43.40</td>
<td>51.74</td>
</tr>
<tr>
<td>3</td>
<td>278</td>
<td>48.26</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total 576 100.00
Let's take a look at these predictions for some individuals in our sample.

```
. list in 1/2, abbrev(10)

 1. weekly 0  command 0  years5 1  presenter 0  teacher 0
     published 1  sjauthor 0  statalist 1  location 1  sjeditor 0
     cpost1 .0145142  cpost2 .6011773  cpost3 .3843085  predclass 2

 2. weekly 1  command 1  years5 1  presenter 1  teacher 1
     published 1  sjauthor 1  statalist 1  location 0  sjeditor 1
     cpost1 .7521391  cpost2 .2477402  cpost3 .0001208  predclass 1
```
Now that we have seen some of the ways we can interpret the results, let’s take a step back and look at the output of the `gsem` command. Four tables are reported.

The first table reports the result of a multinomial logistic regression for the latent categorical variable $C$.

```
gsem (weekly command years5 presenter teacher published ///
>     sjauthor statalist location<-), logit lclass(C 3)
```

Generalized structural equation model
Number of obs = 576
Log likelihood = -3283.0567

|       | Coef. | Std. Err. | z   | P>|z|  | [95% Conf. Interval] |
|-------|-------|-----------|-----|------|----------------------|
| 1.C   | (base outcome) |       |     |      |                      |
| 2.C   | _cons | 1.376261  | .696632 | 1.98 | 0.048          | .0108875    | 2.741635    |
| 3.C   | _cons | 1.503213  | .5577001 | 2.70 | 0.007          | .4101412    | 2.596285    |
We also have a table of results for each class. These tables report class-specific, constant-only logistic regression results for each of our observed variables.
## Latent Class Analysis

What is latent class analysis (LCA)?

### Example of classic LCA

| Class | Coef.   | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|-------|---------|-----------|-------|-----|----------------------|
| weekly  | _cons  | 0.2390244 | 0.464432 | 0.51 | 0.607 | -0.6712456 1.149294 |
| command  | _cons  | 0.8633593 | 0.7933449 | 1.09 | 0.276 | -0.6915682 2.418287 |
| years5  | _cons  | 2.868493  | 1.985474 | 1.44 | 0.149 | -1.022964 6.75995  |
| presenter | _cons | 0.3606906 | 0.4664361 | 0.77 | 0.439 | -0.5535073 1.274889 |
| teacher  | _cons  | 0.3922409 | 0.4098956 | 0.96 | 0.339 | -0.4111397 1.195621 |
| published | _cons | 1.978947  | 0.7727922 | 2.56 | 0.010 | 0.4643019 3.493592 |

(output omitted)
### Example of classic LCA

| Class     | Coef.    | Std. Err. | z    | P>|z|   | [95% Conf. Interval] |
|-----------|----------|-----------|------|-------|---------------------|
| weekly    |          |           |      |       |                     |
| _cons     | 1.357752 | .3013059  | 4.51 | 0.000 | .7672035 1.948301   |
| command   |          |           |      |       |                     |
| _cons     | -1.003379| .2652515  | -3.78| 0.000 | -1.523262 -.4834952|
| years5    |          |           |      |       |                     |
| _cons     | .8730265 | .2221644  | 3.93 | 0.000 | .4375923 1.308461  |
| presenter |          |           |      |       |                     |
| _cons     | .0544483 | .1997721  | 0.27 | 0.785 | -.3370978 .4459945 |
| teacher   |          |           |      |       |                     |
| _cons     | .3215218 | .1895961  | 1.70 | 0.090 | -.0500796 .6931232 |
| published |          |           |      |       |                     |
| _cons     | .5333175 | .2177424  | 2.45 | 0.014 | .1065502 .9600848  |

(output omitted)
A classic example of latent class analysis (LCA) is as follows:

| Class   | Coef.    | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|---------|----------|-----------|-------|------|---------------------|
| weekly  | -.9925281 | .1936823  | -5.12 | 0.000 | -1.372138 - .6129178 |
| command | -1.178592 | .1605149  | -7.34 | 0.000 | -1.493195 - .8639885 |
| years5  | -1.493884 | .2475309  | -6.04 | 0.000 | -1.979036 - 1.008733 |
| presenter| -1.881238 | .2228929  | -8.44 | 0.000 | -2.3181 - 1.444376  |
| teacher | -1.150985 | .1712778  | -6.72 | 0.000 | -1.486683 - .8152864 |
| published| -.9125932 | .1707498  | -5.34 | 0.000 | -1.247257 - .5779298 |

(output omitted)

K. L. MacDonald (StataCorp) 6-7 September 2018
The model we fit is

\[
\Pr( C = 1 ) = \frac{e^{\gamma_1}}{e^{\gamma_1} + e^{\gamma_2} + e^{\gamma_3}} \\
\Pr( C = 2 ) = \frac{e^{\gamma_2}}{e^{\gamma_1} + e^{\gamma_2} + e^{\gamma_3}} \\
\Pr( C = 3 ) = \frac{e^{\gamma_3}}{e^{\gamma_1} + e^{\gamma_2} + e^{\gamma_3}}
\]

where \( \gamma_1, \gamma_2, \) and \( \gamma_3 \) are intercepts in the multinomial logit model for \( C \). By default, the first class will be treated as the base, so \( \gamma_1 = 0 \).
In addition, we have logistic regression models for each of the nine observed variables, conditional on being in class 1:

\[
\Pr(\text{weekly} = 1 \mid C = 1) = \frac{e^{\alpha_{11}}}{1 + e^{\alpha_{11}}}
\]

\[
\ldots
\]

\[
Pr(\text{location} = 1 \mid C = 1) = \frac{e^{\alpha_{91}}}{1 + e^{\alpha_{91}}}
\]

where \(\alpha_{11}, \ldots, \alpha_{91}\) are the intercepts in the logistic regression models.
We also have logistic regression models, conditional on being in class 2:

$$\Pr(\text{weekly} = 1 \mid C = 2) = \frac{e^{\alpha_{12}}}{1 + e^{\alpha_{12}}}$$

... 

$$\Pr(\text{location} = 1 \mid C = 2) = \frac{e^{\alpha_{92}}}{1 + e^{\alpha_{92}}}$$

And conditional on being in class 3:

$$\Pr(\text{weekly} = 1 \mid C = 3) = \frac{e^{\alpha_{13}}}{1 + e^{\alpha_{13}}}$$

... 

$$\Pr(\text{location} = 1 \mid C = 3) = \frac{e^{\alpha_{93}}}{1 + e^{\alpha_{93}}}$$
This is the classic latent class model.
Extensions

Because LCA is implemented through *gsem*, we can extend this basic model in many ways.

- We can include continuous, binary, ordinal, categorical, count, fractional, and even survival-time observed variables.
- We can include predictors of the latent classes.
- We can allow regression models to vary across classes.
- We can allow multiple-equation path models to vary across classes.
Continuous outcomes

- When all of the observed variables are continuous, latent class analysis is sometimes referred to as latent profile analysis.

- To fit a latent profile model using `gsem`, we simply need to model the observed outcomes using linear regression instead of logistic. This is `gsem`'s default.
To demonstrate, let's look at an example from Masyn (2013) where we are interested in identifying unobserved groupings of diabetes patients based on three continuous variables glucose, insulin, sspg.

Let's describe the patient glucose insulin sspg:

```
. describe patient glucose insulin sspg
```

<table>
<thead>
<tr>
<th>variable name</th>
<th>type</th>
<th>format</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>patient</td>
<td>int</td>
<td>%9.0g</td>
<td>Patient ID</td>
</tr>
<tr>
<td>glucose</td>
<td>float</td>
<td>%9.0g</td>
<td>Glucose area (mg/10mL/hr)</td>
</tr>
<tr>
<td>insulin</td>
<td>float</td>
<td>%9.0g</td>
<td>Insulin area (mIU/10mL/hr)</td>
</tr>
<tr>
<td>sspg</td>
<td>float</td>
<td>%9.0g</td>
<td>Steady-state plasma glucose</td>
</tr>
</tbody>
</table>
We fit a model with two classes and a model with three classes. We store the results of each model.

```
. gsem (glucose insulin sspg <- ), lclass(C 2)
. estimates store twoclass
. gsem (glucose insulin sspg <- ), lclass(C 3)
. estimates store threeclass
```
We can use AIC and BIC to determine which of these models fits best.

<table>
<thead>
<tr>
<th>Model</th>
<th>Obs</th>
<th>ll(null)</th>
<th>ll(model)</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>twoclass</td>
<td>145</td>
<td></td>
<td>-1702.554</td>
<td>10</td>
<td>3425.108</td>
<td>3454.876</td>
</tr>
<tr>
<td>threeclass</td>
<td>145</td>
<td></td>
<td>-1653.238</td>
<td>14</td>
<td>3334.476</td>
<td>3376.15</td>
</tr>
</tbody>
</table>

Note: N=Obs used in calculating BIC; see [R] BIC note.

The three-class model has smaller values of AIC and BIC.
We can again use **estat lcmean** to obtain marginal means of the observed variables, conditional on being in class 1, class 2, and class 3.

```
estat lcmean
Latent class marginal means
Number of obs = 145

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Margin</td>
<td>Std. Err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>glucose</td>
<td>39.51632</td>
<td>1.576263</td>
<td>25.07</td>
<td>0.000</td>
<td>36.4269</td>
</tr>
<tr>
<td></td>
<td>insulin</td>
<td>16.95918</td>
<td>0.9219973</td>
<td>18.39</td>
<td>0.000</td>
<td>15.15209</td>
</tr>
<tr>
<td></td>
<td>sspg</td>
<td>13.03127</td>
<td>0.9119668</td>
<td>14.29</td>
<td>0.000</td>
<td>11.24385</td>
</tr>
<tr>
<td>2</td>
<td>glucose</td>
<td>49.87783</td>
<td>3.38311</td>
<td>14.74</td>
<td>0.000</td>
<td>43.24706</td>
</tr>
<tr>
<td></td>
<td>insulin</td>
<td>42.28255</td>
<td>4.489995</td>
<td>9.42</td>
<td>0.000</td>
<td>33.48232</td>
</tr>
<tr>
<td></td>
<td>sspg</td>
<td>25.04299</td>
<td>1.468301</td>
<td>17.06</td>
<td>0.000</td>
<td>22.16517</td>
</tr>
<tr>
<td>3</td>
<td>glucose</td>
<td>115.5237</td>
<td>2.698185</td>
<td>42.82</td>
<td>0.000</td>
<td>110.2354</td>
</tr>
<tr>
<td></td>
<td>insulin</td>
<td>7.574585</td>
<td>1.373028</td>
<td>5.52</td>
<td>0.000</td>
<td>4.883499</td>
</tr>
<tr>
<td></td>
<td>sspg</td>
<td>34.53398</td>
<td>1.308423</td>
<td>26.39</td>
<td>0.000</td>
<td>31.96952</td>
</tr>
</tbody>
</table>
**estat lcmean** is really just a wrapper for **margins**. If we want to graph these means, we can use **margins** and **marginsplot**.

```
. margins, predict(outcome(glucose) class(1)) ///
   predict(outcome(insulin) class(1)) ///
   predict(outcome(sspg) class(1))
```

```
. marginsplot, recast(bar) title("Class 1") xtitle(NULL) ///
   xlabel(1 "glucose" 2 "insulin" 3 "sspg", angle(45)) ///
   ytitle("Predicted mean") ylabel(0(25)125) name(class1)
```
. margins, predict(outcome(glucose) class(2)) ///
    predict(outcome(insulin) class(2)) ///
    predict(outcome(sspg) class(2))

. marginsplot, recast(bar) title("Class 2") xtitle(""") ///
    xlabel(1 "glucose" 2 "insulin" 3 "sspg", angle(45)) ///
    ytitle("") ylabel(0(25)125) name(class2) ///

. margins, predict(outcome(glucose) class(3)) ///
    predict(outcome(insulin) class(3)) ///
    predict(outcome(sspg) class(3))

. marginsplot, recast(bar) title("Class 2") xtitle(""") ///
    xlabel(1 "glucose" 2 "insulin" 3 "sspg", angle(45)) ///
    ytitle("") ylabel(0(25)125) name(class3) ///

. graph combine class1 class2 class3, cols(3)
We have seen latent class models for binary and continuous outcomes.

What if the observed variables are counts?

- `gsem (y1 y2 y3 y4 <- ), poisson lclass(C 3)`
- `gsem (y1 y2 y3 y4 <- ), nbreg lclass(C 3)`

What if they are ordinal?

- `gsem (y1 y2 y3 y4 <- ), ologit lclass(C 3)`
- `gsem (y1 y2 y3 y4 <- ), oprobit lclass(C 3)`
What if the items are ...?

*gsem* supports many family and link combinations to allow for outcomes that are continuous, binary, ordinal, categorical, count, fractional, and survival times.

Observed variables in latent class models can be of one of these types or a combination of them.

For instance, for a combination of binary, ordinal, and count variables, we could type

```
. gsem (y1 y2 <- , logit) ///
   (y3 <- , ologit) ///
   (y4 <- , poisson), lclass(C 3)
```
We can also have variables that are predictors of class membership.

\[
. \texttt{gsem (y1 y2 y3 y4 <- , logit) ///}
. (C <- x1), lclass(C 3)
\]

Now \texttt{x1} is included as a regressor in the multinomial logit model for \texttt{C}.
Regression models

- We might want to go further than classifying individuals into unobserved groupings.
- Maybe the parameters of a regression models have differ across unknown groups.
We have data on annual number of doctor visits for individuals age 65 and older from the U.S. Medical Expenditure Panel Survey for 2003.

```
. describe drvisits private medicaid age educ actlim chronic

<table>
<thead>
<tr>
<th>variable name</th>
<th>type</th>
<th>format</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>drvisits</td>
<td>int</td>
<td>%9.0g</td>
<td>number of doctor visits</td>
</tr>
<tr>
<td>private</td>
<td>byte</td>
<td>%8.0g</td>
<td>has private supplementary insurance</td>
</tr>
<tr>
<td>medicaid</td>
<td>byte</td>
<td>%8.0g</td>
<td>has Medicaid public insurance</td>
</tr>
<tr>
<td>age</td>
<td>byte</td>
<td>%8.0g</td>
<td>age in years</td>
</tr>
<tr>
<td>educ</td>
<td>byte</td>
<td>%8.0g</td>
<td>years of education</td>
</tr>
<tr>
<td>actlim</td>
<td>byte</td>
<td>%8.0g</td>
<td>has activity limitations</td>
</tr>
<tr>
<td>chronic</td>
<td>byte</td>
<td>%8.0g</td>
<td>number of chronic conditions</td>
</tr>
</tbody>
</table>
```
We could use the **poisson** command to fit a Poisson model for the number of doctor visits.

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

We could fit the same model using **gsem**.

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

If we want to allow the parameters to differ across two unobserved groups of individuals, we simply add the **lclass(C 2)** option.

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

(output omitted)

Generalized structural equation model

Number of obs = 3,677
Log likelihood = -12100.185

|            | Coef.  | Std. Err. |   z    | P>|z|  | [95% Conf. Interval] |
|------------|--------|-----------|--------|------|----------------------|
| 1.C        | (base outcome) |                      |        |      |                     |
| 2.C        | _cons  | -.5980831 | .050677| -11.80| -.6974083 -.4987579 |
Class : 1
Response : drvisits
Family : Poisson
Link : log

|                      | Coef.  | Std. Err. | z      | P>|z| | [95% Conf. Interval] |
|----------------------|--------|-----------|--------|------|---------------------|
| drvisits             |        |           |        |      |                     |
| private              | .2393558 | .0312351  | 7.66   | 0.000 | .1781361 | .3005756          |
| medicaid             | .0463821 | .040343  | 1.15   | 0.250 | -.0326888 | .125453           |
| age                  | -.6233526 | .0583698 | -10.68 | 0.000 | -.7377554 | -.5089499         |
| c.age#c.age          | .0045366 | .0003904 | 11.62  | 0.000 | .0037714 | .0053019          |
| educ                 | .0284599 | .0039608 | 7.19   | 0.000 | .0206969 | .0362229          |
| actlim               | .1723268 | .0318187 | 5.42   | 0.000 | .1099633 | .2346903          |
| chronic              | .3286694 | .0097798 | 33.61  | 0.000 | .3095014 | .3478374          |
| _cons                | 21.35464 | 2.164152 | 9.87   | 0.000 | 17.11298 | 25.5963           |
Latent Class Analysis

Extensions

Regression models

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

|                        | Coef.   | Std. Err. |    z  | P>|z| | [95% Conf. Interval] |
|------------------------|---------|-----------|-------|-----|----------------------|
| drvisits               |         |           |       |     |                      |
| private                | .1566873| .0252956  | 6.19  | 0.000| .1071088 .2062658    |
| medicaid               | .1924436| .0337855  | 5.70  | 0.000| .1262252 .258662     |
| age                    | 1.232368| .0485717  | 25.37 | 0.000| 1.137169 1.327567    |
| c.age#c.age            | -.0085471| .0003268 | -26.15| 0.000| -.0091876 -.0079065  |
| educ                   | .0219929| .0032055  | 6.86  | 0.000| .0157102 .0282756    |
| actlim                 | .1486859| .0260608  | 5.71  | 0.000| .0976077 .1997641    |
| chronic                | .1898829| .009189   | 20.66 | 0.000| .1718728 .207893     |
| _cons                  | -42.46506| 1.795471 | -23.65| 0.000| -45.98412 -38.946    |
We use `estat lcmean` to obtain marginal counts for each class.

```
estat lcmean
Latent class marginal means Number of obs = 3,677

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Margin Std. Err. z P&gt;</td>
<td>z</td>
<td>[95% Conf. Interval]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>drvisits 5.050474 .0828385 60.97 0.000 4.888113 5.212834</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>drvisits 11.65096 .1689544 68.96 0.000 11.31982 11.98211</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

We see that individuals in class 1 visit the doctor less frequently, and individuals in class 2 visit the doctor more frequently.
We again use **estat lcprob** to estimate the proportion of individuals in each class.

```
. estat lcprob
Latent class marginal probabilities
Number of obs = 3,677

<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>.6452176</td>
<td>.0116006</td>
<td>.6221674 .6676129</td>
</tr>
<tr>
<td>1</td>
<td>.3547824</td>
<td>.0116006</td>
<td>.3323871 .3778326</td>
</tr>
</tbody>
</table>
```

We estimate that 65% of the population is in class 1.
For each individual, we can predict the number of doctor visits, conditional on being in class 1 and conditional on being in class 2. We can plot the distributions of the two to compare them.

. predict mu*
(option mu assumed)

. histogram mu1, width(1) xtitle("Expected number of visits") ///
>   name(class1) title(Class 1)
>   (bin=50, start=.95077324, width=1)

. histogram mu2, width(1) xtitle("Expected number of visits") ///
>   name(class2) title(Class 2)
>   (bin=58, start=.66974765, width=1)

. graph combine class1 class2, ycommon xcommon ///
>   title("Predicted number of doctor visits for two classes") ///
>
Predicted number of doctor visits for two classes

Class 1

Class 2
For the model we just fit, we didn’t really need to use **gsem**. We could have instead used the new **fmm** prefix.

```
. fmm 2: poisson drvisits private medicaid ///
    c.age##c.age educ actlim chronic
```

The **estat** and **predict** commands work after **fmm** just like they do after **gsem**.
**fmm** is very convenient if you are fitting single-equation models. It works with many of Stata’s estimation commands.

- Continuous outcomes: `regress` and `ivregress`,
- Truncated and censored outcomes: `truncreg`, `intreg`, and `tobit`
- Binary outcomes: `logit`, `probit`, and `cloglog`
- Ordered outcomes: `ologit` and `oprobit`
- Categorical outcomes: `mlogit`
- Count outcomes: `poisson`, `nbreg`, and `tpoisson`
- Fractional outcomes: `betareg`
- Survival-time outcomes: `streg`
- Generalized linear models: `glm`
Why learn about `gsem, lclass()` if we are interested in regression models?

The usual answer: Extensions!

In this case, you can fit multiple-equation (path) models using `gsem`. Each parameter can be estimated separately across classes or constrained to be equal.

```plaintext
. gsem ( y1 y2 <- x1 x2 x3), lclass(C 3)
. gsem (y1 <- x1 y2) (y2 <- x1 x2), lclass(C 3)
. gsem (y1 <- x1@cns1 y2) (y2 <- x1 x2), lclass(C 3)
...
Conclusion

- LCA is a powerful and flexible method for identifying and understanding unobserved groups in a population.
- `gsem`’s `lclass()` option allows for fitting a wide variety of latent class models.
- In the special case of regression models that vary across groups, try the convenient `fmm` prefix.