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 variale 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 variale 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 variale 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 variale 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 variale 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.

Example of classic LCA

- 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?

```
. summarize
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      weekly |        576    .5208333          .5          0          1
     command |        576    .2986111    .4580467          0          1
      years5 |        576    .4826389    .5001328          0          1
   presenter |        576    .3402778    .4742143          0          1
     teacher |        576    .4201389      .49401          0          1
-------------+---------------------------------------------------------
   published |        576    .4930556    .5003863          0          1
    sjauthor |        576    .3142361    .4646144          0          1
   statalist |        576    .3628472    .4812392          0          1
    location |        576     .515625    .5001902          0          1
```

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.

- 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
```

	Margin	Delta-method Std. Err.	[95% Conf.	Interval]
C				
1	.1057509	.0582876	.0341272	.2835627
2	.4187809	.0704887	.2900013	.5596688
3	.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

--------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
1            |
      weekly |   .5594732   .1144653       .338218     .759382
     command |    .703362   .1655266      .3336843    .9182112
      years5 |   .9462668   .1009533      .2644505    .9988421
   presenter |   .5892076   .1128971      .3650511    .7815784
     teacher |    .596822   .0986313      .3986389    .7677449
   published |   .8785688   .0824458      .6140342    .9705049
    sjauthor |   .7467327   .1777284      .3185127    .9489785
   statalist |   .4410877   .1074878      .2513733    .6497189
    location |   .1202751   .0922665      .0241521    .4302775
--------------------------------------------------------------
```

- 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?

2				
weekly	.7953942	.0490352	.6829157	.8752613
command	.2682777	.0520701	.1789817	.3814271
years5	.7053751	.0461704	.6076852	.7872555
presenter	.5136087	.049906	.4165146	.6096865
teacher	.5796951	.0461948	.4874827	.6666613
published	.6302565	.0507412	.5266124	.7231388
sjauthor	.3026139	.051335	.2122123	.4114143
statalist	.5908731	.0555132	.479385	.6937391
location	.4509978	.0559189	.3454076	.5611936

- 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?

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

- 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

----------------------------------------------------------------------------
Fit statistic        |     Value   Description
---------------------+------------------------------------------------------
Likelihood ratio     |
        chi2_ms(482) |   460.457   model vs. saturated
            p > chi2 |     0.753
---------------------+------------------------------------------------------
Information criteria |
                 AIC |  6624.113   Akaike's information criterion
                 BIC |  6750.441   Bayesian information criterion
----------------------------------------------------------------------------
```

- We fail to reject the null hypothesis that our model fits as well as a satruated 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
```

predclass	Freq.	Percent	Cum.
1	48	8.33	8.33
2	250	43.40	51.74
3	278	48.26	100.00
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 | command | years5 | presenter | teacher |
     |      0 |       0 |      1 |         0 |       0 |
     |-------------------------------------------------------------|
     | published | sjauthor | statalist | location | sjeditor |
     |         1 |        0 |         1 |        1 |        0 |
     |-------------------------------------------------------------|
     |   cpost1 |   cpost2 |   cpost3 | predclass |
     | .0145142 | .6011773 | .3843085 |         2 |
     +-------------------------------------------------------------+

     +-------------------------------------------------------------+
  2. | weekly | command | years5 | presenter | teacher |
     |      1 |       1 |      1 |         1 |       1 |
     |-------------------------------------------------------------|
     | published | sjauthor | statalist | location | sjeditor |
     |         1 |        1 |         1 |        0 |        1 |
     |-------------------------------------------------------------|
     |   cpost1 |   cpost2 |   cpost3 | predclass |
     | .7521391 | .2477402 | .0001208 |         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-sepcific, constant-only logistic regression results for each of our observed variables.

```
Class          : 1
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weekly						
_cons	.2390244	.464432	0.51	0.607	-.6712456	1.149294
command						
_cons	.8633593	.7933449	1.09	0.276	-.6915682	2.418287
years5						
_cons	2.868493	1.985474	1.44	0.149	-1.022964	6.75995
presenter						
_cons	.3606906	.4664361	0.77	0.439	-.5535073	1.274889
teacher						
_cons	.3922409	.4098956	0.96	0.339	-.4111397	1.195621
published						
_cons	1.978947	.7727922	2.56	0.010	.4643019	3.493592

(*output omitted*)

Class : 2

	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*)						

```
Class          : 3
```

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

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 γ_1, γ_2, and γ_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:

$$\Pr(\textit{weekly} = 1 \mid C = 1) = \frac{e^{\alpha_{11}}}{1 + e^{\alpha_{11}}}$$

$$\dots$$

$$Pr(\textit{location} = 1 \mid C = 1) = \frac{e^{\alpha_{91}}}{1 + e^{\alpha_{91}}}$$

where $\alpha_{11}, \dots, \alpha_{91}$ are the intercepts in the logistic regression models.

We also have logistic regression models, conditional on being in class 2:

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

$$\dots$$

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

And conditional on being in class 3:

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

$$\dots$$

$$Pr(location = 1 \,|\, C = 3) = \frac{e^{\alpha_{93}}}{1 + e^{\alpha_{93}}}$$

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 refered 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 example from Masyn (2013) where we are interested in identifying unobserved groupings of diabetes patients based on three continuous variables **glucose**, **insulin**, **sspg**.

```
. describe patient glucose insulin sspg

              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
patient         int     %9.0g                 Patient ID
glucose         float   %9.0g                 Glucose area (mg/10mL/hr)
insulin         float   %9.0g                 Insulin area (mIU/10mL/hr)
sspg            float   %9.0g                 Steady-state plasma glucose
```

- 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.

```
. estimates stats twoclass threeclass
Akaike´s information criterion and Bayesian information criterion
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
twoclass	145	.	-1702.554	10	3425.108	3454.876
threeclass	145	.	-1653.238	14	3334.476	3376.15

```
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, contidional on being in class 1, class 2, and class 3.

```
. estat lcmean

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

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
1						
glucose	39.51632	1.576263	25.07	0.000	36.4269	42.60574
insulin	16.95918	.9219973	18.39	0.000	15.15209	18.76626
sspg	13.03127	.9119668	14.29	0.000	11.24385	14.8187
2						
glucose	49.87783	3.38311	14.74	0.000	43.24706	56.50861
insulin	42.28255	4.489995	9.42	0.000	33.48232	51.08277
sspg	25.04299	1.468301	17.06	0.000	22.16517	27.92081
3						
glucose	115.5237	2.698185	42.82	0.000	110.2354	120.8121
insulin	7.574585	1.373028	5.52	0.000	4.883499	10.26567
sspg	34.53398	1.308423	26.39	0.000	31.96952	37.09845

- **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("")          ///
        xlabel(1 "glucose" 2 "insulin" 3 "sspg", angle(45))    ///
        ytitle("Predicted mean") ylabel(0(25)125) name(class1)
```

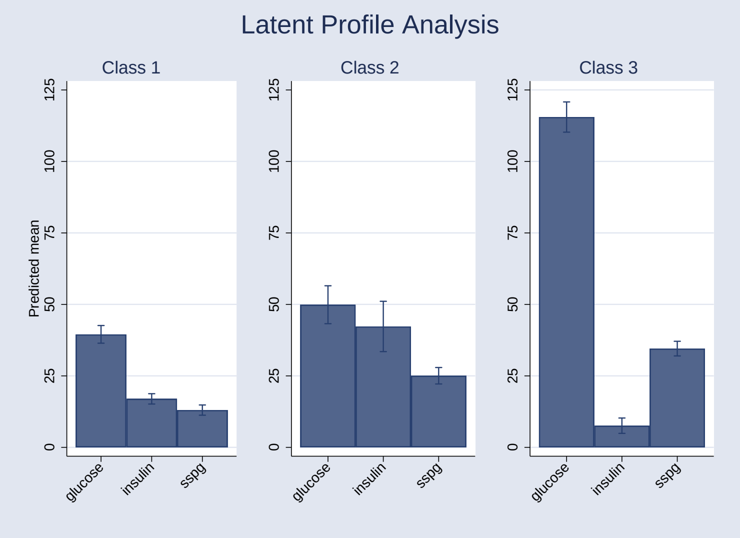
```
. margins, predict(outcome(glucose) class(2)) ///
          predict(outcome(insulin) class(2)) ///
          predict(outcome(sspg) class(2))

. 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)
```


Latent Profile Analysis
Class 1
Class 2
Class 3
Predicted mean
0
25
50
75
100
125
glucose
insulin
sspg

- 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 insance, 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.

```
. gsem (y1 y2 y3 y4 <- , logit)    ///
       (C <- x1), lclass(C 3)
```

- Now **x1** is included as a regressor in the multinomial logit model for **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

              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
drvisits        int     %9.0g                 number of doctor visits
private         byte    %8.0g                 has private supplementary
                                                insurance
medicaid        byte    %8.0g                 has Medicaid public insurance
age             byte    %8.0g                 age in years
educ            byte    %8.0g                 years of education
actlim          byte    %8.0g                 has activity limitations
chronic         byte    %8.0g                 number of chronic conditions
```

- 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),  ///
    poisson lclass(C 2)
```

```
. 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	0.000	-.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

```
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

--------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+------------------------------------------------------------------
1            |
    drvisits |   5.050474   .0828385    60.97   0.000     4.888113    5.212834
-------------+------------------------------------------------------------------
2            |
    drvisits |   11.65096   .1689544    68.96   0.000     11.31982    11.98211
--------------------------------------------------------------------------------
```

- 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

--------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
           C |
           1 |   .6452176   .0116006      .6221674    .6676129
           2 |   .3547824   .0116006      .3323871    .3778326
--------------------------------------------------------------
```

- We estimate that 65% of the population is in class 1.

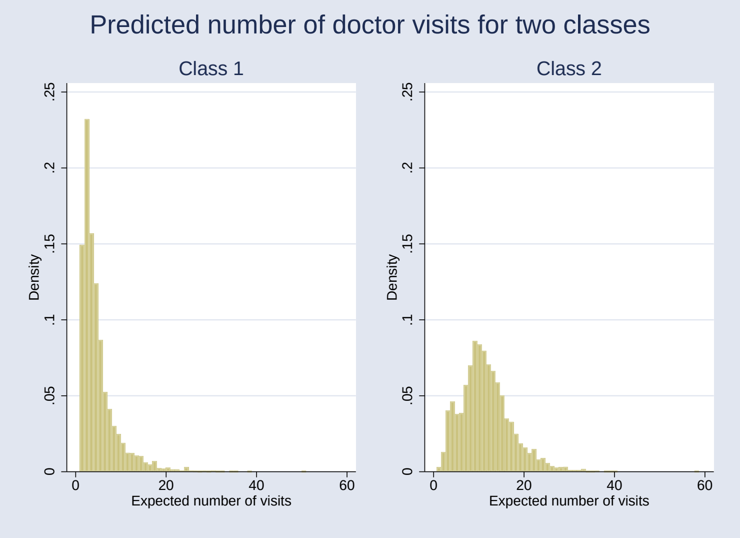
- For each individual, we can predict the number of doctor visits, conditional on being in class 1 and conditional on being in class 2. We can plot the distributions of the two to compare them.

```
. 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
Density
0
.05
.1
.15
.2
.25
0
20
40
60
Expected number of visits

- 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.

```
. 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.