Example 2 — Mixture of Poisson regression models

Description Remarks and examples References Also see

Description

In this example, we demonstrate how to fit a two-component mixture of Poisson regressions models. We also use `estat lcmean` to estimate marginal predicted counts and `estat lcprob` to estimate the proportion of individuals in each class.

Remarks and examples

We are interested in fitting a Poisson regression to model the annual number of doctor visits. We hypothesize that there are two distinct groups or classes in the population that differ in their healthcare utilization—frequent users and infrequent users—and we believe that the model may differ across these two groups.

We do not have any information that tells us which individuals in our sample belong to which group. With FMM, we can specify two latent classes in our model to identify these groups. To account for differences between the latent classes, we include predictor variables in our model to fit potentially different Poisson distributions for each class.

Here we replicate the finite mixture Poisson regression example from [SEM] **Example 53g**. We use the following data:

```
. use https://www.stata-press.com/data/r17/gsem_mixture
(U.S. Medical Expenditure Panel Survey (2003))

. describe

Contains data from https://www.stata-press.com/data/r17/gsem_mixture.dta
 Observations:         3,677                 U.S. Medical Expenditure Panel
                                               Survey (2003)
    Variables:            12                 26 Jan 2021 08:46
                                             (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    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
income          float   %9.0g                 Income in $1,000s
offer           byte    %8.0g                 Employer offers insurance
hpvisits        int     %8.0g                 Number of visits to health
                                                professionals other than
                                                doctors
female          byte    %8.0g                 Female
phylim          byte    %8.0g                 Has physical limitation
-------------------------------------------------------------------------------
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 12: 313-336.
      https://doi.org/10.1002/(SICI)1099-1255(199705)12:3<313::AID-JAE440>3.0.C
> O;2-G.
```

Following Cameron and Trivedi (2010), we fit an FMM with a Poisson regression component for each latent class. We model the 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.

We add the `startvalues(randomid, draws(5) seed(15))` option to specify that five random draws are taken when computing starting values. The class assignment is selected from the draw that has the best log likelihood after the EM iterations. When fitting FMMs, taking multiple draws of random starting values can help to prevent convergence at a local maximum rather than the global maximum. `fmm` provides a variety of options for obtaining starting values; see [FMM] **fmm** for more information on starting values.

```
. fmm 2, startvalues(randomid, draws(5) seed(15)):
> poisson drvisits private medicaid c.age##c.age educ actlim chronic
```

(iteration log omitted)

```
Finite mixture model                               Number of obs = 3,677
Log likelihood = -11502.686
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.Class	(base outcome)					
2.Class						
_cons	.877227	.0494614	17.74	0.000	.7802845	.9741696

```
Class:     1
Response:  drvisits
Model:     poisson
```

	Coefficient	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

```
Class:    2
Response: drvisits
Model:    poisson
```

	Coefficient	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

The first table in the output provides the estimated coefficients in the multinomial logit model for the latent class probabilities. The next two tables are the results for the Poisson regression models for the first and second classes. The estimated coefficients from these tables are interpreted just as you would coefficients from `poisson`; see [R] **poisson**.

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

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

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1						
drvisits	13.95943	.1767506	78.98	0.000	13.613	14.30585
2						
drvisits	3.801692	.0587685	64.69	0.000	3.686508	3.916876

We see that class 1 represents those who visit the doctor frequently and class 2 represents those who visit the doctor less frequently. We can use `estat lcprob` to estimate the proportion of individuals in each class.

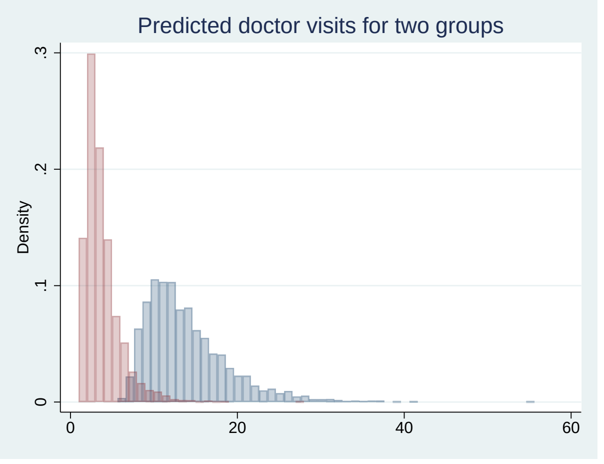
```
. estat lcprob
Latent class marginal probabilities             Number of obs = 3,677
```

	Margin	Delta-method std. err.	[95% conf.	interval]
Class				
1	.2937527	.0102614	.2740502	.3142586
2	.7062473	.0102614	.6857414	.7259498

We find that about 29% of the population is in the group that visits the doctor frequently (class 1) and about 71% is in the group that visits the doctor less frequently (class 2).

We can visually compare the resulting distributions of the means by plotting the predicted number of doctor visits.

```
. predict mu*
(option mu assumed)
. twoway histogram mu1, width(1) color(navy)   fcolor(%25) lcolor(%25) ||
>        histogram mu2, width(1) color(maroon) fcolor(%25) lcolor(%25)
>        legend(off) title("Predicted doctor visits for two groups")
```



We can clearly see the two groups. The frequent user group exhibits more variability, which is expected in a Poisson process where the variance is equal to the mean.

References

Cameron, A. C., and P. K. Trivedi. 2010. *Microeconometrics Using Stata*. Rev. ed. College Station, TX: Stata Press.

Deb, P., and P. K. Trivedi. 1997. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* 12: 313–336. https://doi.org/10.1002/(SICI)1099-1255(199705)12:3⟨313::AID-JAE440⟩3.0.CO;2-G.

Also see

[FMM] **fmm intro** — Introduction to finite mixture models

[FMM] **fmm: poisson** — Finite mixtures of Poisson regression models

[FMM] **estat lcmean** — Latent class marginal means

[FMM] **estat lcprob** — Latent class marginal probabilities

[SEM] **Example 53g** — Finite mixture Poisson regression

[SEM] **Example 54g** — Finite mixture Poisson regression, multiple responses

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