Title

stata.com

Example 3 — Zero-inflated models

Description Remarks and examples References Also see

Description

In this example, we demonstrate how to fit a zero-inflated Poisson model as a two-component mixture model. We use `estat lcprob` to estimate marginal class probabilities and `estat lcmean` to estimate marginal predicted counts. A likelihood-ratio test is performed to compare models with and without predictors of class membership.

Remarks and examples

stata.com

Two-component mixture models are often used to model counts that include book sales through direct mail (Wedel et al. 1993), healthcare utilization (Deb and Trivedi 1997), and modeling of risk behavior (Lanza, Kugler, and Mathur 2011). In the FMM framework, a zero-inflated count model is represented by a mixture of a component that models both zero and nonzero counts and a degenerate point mass distribution that models the zeros; see [FMM] **fmm: pointmass** for details.

The most popular zero-inflated count model is the zero-inflated Poisson (ZIP) model. Here we fit this model to the data on the number of fish caught by park visitors. Almost 57% of visitors reported zero catch, but we do not know whether they fished in the first place. In other words, zero counts can either be from a Poisson distribution or are hard zeros from a point mass distribution. Using a zero-inflated FMM, we can make probabilistic statements about which distribution a given zero came from.

Using `fish2.dta`, we fit a two-component mixture model where the nonfishing group (class 1) is modeled using a degenerate point mass distribution with the default value zero and the fishing group (class 2) is modeled using a Poisson distribution. For the latter group, we model the number of fish caught as a function of whether the visitor brought a boat (`boat`) and the number of persons in the party (`persons`).

By default, the reference probability is the class 1 probability. We specify `lcbase(2)` to make the reference probability be the probability for class 2. This will allow us to more easily compare the mixing proportions when we add covariates to model the probability of being in the nonfishing group.

```
. use https://www.stata-press.com/data/r17/fish2
(Fictional fishing data)
. fmm, lcbase(2): (pointmass count) (poisson count persons boat)
  (iteration log omitted)
Finite mixture model                                    Number of obs = 250
Log likelihood = -882.31198
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.Class						
_cons	.0867958	.1390251	0.62	0.532	-.1856884	.35928
2.Class	(base outcome)					

```
Class:    2
Response: count
Model:    poisson
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
count						
persons	.750919	.0422907	17.76	0.000	.6680307	.8338072
boat	1.813785	.2648584	6.85	0.000	1.294672	2.332898
_cons	-2.024982	.2974941	-6.81	0.000	-2.608059	-1.441904

The first table in the output provides the estimated coefficients on the logit scale for the class probabilities. The coefficient on `1.Class` represents the probability of being in the nonfishing group which is about 52% [invlogit(0.087) $\approx$ 0.52]. Because we have only two groups, the fishing fraction is 48%. Recall that the fraction of zeros in the data is 0.57, thus the model suggests that some zero counts are due to the Poisson component.

The second output table presents the results for the Poisson model component. The coefficients here are interpreted just as those from a standard Poisson regression; see [R] **poisson**. For example, we see that having a boat increases the expected number of fish caught by around six [exp(1.814) $\approx$ 6.14] for those who did fish, holding other covariates constant.

We store our estimates for later use.

```
. estimates store model1
```

In the model above, we did not model class probabilities. By modeling class probabilities with covariates, we can further differentiate between visitors who did not fish and those who fished without success. Here we make the mixing probability for the point mass component depend on covariates by using the `lcprob()` option with covariates `child` and `camper`. The default reference probability now switches to the Poisson component; therefore, we no longer need to specify `lcbase(2)`.

```
. fmm: (pointmass count, lcprob(child camper)) (poisson count persons boat)
```

(*iteration log omitted*)

```
Finite mixture model                                    Number of obs = 250
Log likelihood = -850.70142
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.Class						
child	1.602571	.2797719	5.73	0.000	1.054228	2.150913
camper	-1.015698	.365259	-2.78	0.005	-1.731593	-.2998039
_cons	-.4922872	.3114562	-1.58	0.114	-1.10273	.1181558
2.Class	(base outcome)					

```
Class:    2
Response: count
Model:    poisson
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
count						
persons	.8068853	.0453288	17.80	0.000	.7180424	.8957281
boat	1.757289	.2446082	7.18	0.000	1.277866	2.236713
_cons	-2.178472	.2860289	-7.62	0.000	-2.739078	-1.617865

The coefficients for the Poisson component are close to those from the previous model.

The coefficients of interest for the class 1 probability are both significant. A positive coefficient on the `child` variable means people with children in their party tended do to something other than fish. A negative coefficient on the `camper` variable means people camping at the park were more likely to go fishing.

Because we modeled the probability of being in the point mass component with covariates, calculating the marginal probabilities of belonging to a given component is more involved than before. We use `estat lcprob` to display marginal class probabilities on a probability scale.

```
. estat lcprob
Latent class marginal probabilities                     Number of obs = 250
```

	Margin	Delta-method std. err.	[95% conf.	interval]
Class				
1	.4786335	.0341083	.4125554	.5454678
2	.5213665	.0341083	.4545322	.5874446

We find that about 48% of the park visitors are in the nonfishing group, which is slightly lower than the 52% we found previously.

We can use `lrtest` to compare the current model with the previous one.

```
. lrtest model1 .
Likelihood-ratio test
Assumption: model1 nested within .
 LR chi2(2) =  63.22
Prob > chi2 = 0.0000
```

The likelihood-ratio test favors the model that includes covariates in the modeling of the probability of being in the nonfishing group.

We can also estimate the marginal predicted counts (means) for the fishing group using `estat lcmean`.

```
. estat lcmean
Latent class marginal means                             Number of obs = 250
Expression: Predicted mean (number of fish caught in class 2.Class),
            predict(outcome(count) class(2))
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
2						
count	6.490014	.2361623	27.48	0.000	6.027144	6.952884

The marginal predicted count for the fishing group is 6.49. This is much higher than the sample mean of 3.30 that is based on the fishing and nonfishing populations combined. If we were advertising fishing opportunities in the park, we know which number we would use!

References

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.

Lanza, S. T., K. C. Kugler, and C. Mathur. 2011. Differential effects for sexual risk behavior: An application of finite mixture regression. *Open Family Studies Journal* 4 (Suppl. 1-M9): 81–88. https://doi.org/10.2174/1874922401104010081.

Wedel, M., W. S. DeSarbo, J. R. Bult, and V. Ramaswamy. 1993. A latent class poisson regression model for heterogeneous count data. *Journal of Applied Econometrics* 8: 397–411. https://doi.org/10.1002/jae.3950080407.

Also see

[FMM] **fmm** — Finite mixture models using the fmm prefix

[R] **zip** — Zero-inflated Poisson regression