Generalized Linear Models

Lecture 10. Count data models IV.
Zero-truncated models. Generalized Poisson model

Count data without zeros

Situation: data has less zeros than model (or no zeros at all)

Missing zeros. Motivating examples

- Problems in healthcare economics: use of medical services (days in hospital, use of x-ray, etc.)
- Medical problems: alcohol use, drug use, sudden deaths
- Problems in ecology: dead trees, animals hit on highways

Possible solution: Zero Truncated (ZT) model

Zero Truncated Poisson (ZTP) model

To model count data without zeros, we need to estimate pmf, taking into account that there are no zeros

Corresponding conditional probability:

$$\mathbf{P}\{Y_i = y_i | Y_i > 0\} = \frac{\mathbf{P}\{Y_i = y_i\}}{1 - \mathbf{P}\{Y_i = 0\}}$$

For Poisson distribution

$$p(y_i; \mu_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}$$

Probability of having zero: $\mathbf{P}(Y_i = 0) = \exp(-\mu_i)$
Probability of not having zero: $1 - \mathbf{P}(Y_i = 0) = 1 - \exp(-\mu_i)$

ZTP as conditional model

$$p(y_i; \mu_i | Y_i > 0) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{[1 - \exp(-\mu_i)]y_i!},$$

where $\mu_i = \exp(\boldsymbol{x}_i^T \beta_i)$

Log-likelihood for ZTP model

Let us start with the pmf:

$$p(y_i; \mu_i | Y_i > 0) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{[1 - \exp(-\mu_i)]y_i!}, \quad \mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$$

Log-likelihood for i-th observation (with $\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$):

$$l(y_i; \mu_i | Y_i > 0) = y_i[\mathbf{x}_i^T \boldsymbol{\beta}] - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - \ln y_i! - \ln[1 - \exp\{-\exp(\mathbf{x}_i^T \boldsymbol{\beta})\}]$$

Log-likelihood for ZTP model

$$l(y_i; \mu_i | Y_i > 0) = \sum \left\{ y_i[\mathbf{x}_i^T \boldsymbol{\beta}] - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - \ln y_i! - \ln[1 - \exp\{-\exp(\mathbf{x}_i^T \boldsymbol{\beta})\}] \right\}$$

Zero truncated NB (ZTNB) model

Let us start with the pmf of NB distribution

$$p(y_i; \mu_i, k) = \frac{\Gamma(k + y_i)}{y_i!\, \Gamma(k)} \, (\frac{k}{k + \mu_i})^k (1 - \frac{k}{k + \mu_i})^{y_i}$$

and apply the condition of that there are no zeros

The probability that NB-distributed r.v. Y_i takes value zero is

$$\mathbf{P}\{Y_i = 0\} = \left(\frac{k}{k + \mu_i} \right)^k$$

The conditional probability is thus

$$p(y_i; \mu_i, k | Y_i > 0) = \frac{p(y_i; \mu_i, k)}{1 - P\{Y_i = 0\}}$$

The log-likelihood for ZTNB model can be derived analogously to ZTP model

Example. Length of stay in a hospital

Example. Length of hospital stay

A study of length of hospital stay, in days, as a function of age, kind of health insurance and whether or not the patient died while in the hospital. Length of hospital stay is recorded as a minimum of at least one day. Corresponding hypothetical data file has 1,493 observations and the variables are

- `stay` – length of stay
- `age` – group from 1 to 9 (will be treated as interval in this example)
- `hmo` – indicator (1/0, has HMO insurance?)
- `died` – indicator (1/0, died in hospital?)

Question of interest is whether and how the length of hospital stay depends on the mentioned variables.

Source: `https://stats.idre.ucla.edu/r/dae/zero-truncated-poisson/`

Example. ZTP for length of stay

```
> library(VGAM)
> m1 <- vglm(stay ~ age + hmo + died,
             family = pospoisson(), data = data)
> summary(m1)
...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.435808   0.027332  89.119  < 2e-16 ***
age         -0.014442   0.005035  -2.869  0.00412 **
hmo1        -0.135903   0.023741  -5.724 1.04e-08 ***
died1       -0.203771   0.018372 -11.091  < 2e-16 ***
...
Number of linear predictors:  1
Name of linear predictor: loge(lambda)
Log-likelihood: -6908.799 on 1489 degrees of freedom
Number of iterations: 3
No Hauck-Donner effect found in any of the estimates
```

Example. ZTNB for length of stay

```
> m2 <- vglm(stay ~ age + hmo + died,
             family = posnegbinomial(), data = data)
> summary(m2)
...
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  2.40833    0.07158  33.645  < 2e-16 ***
(Intercept):2  0.56864    0.05489  10.359  < 2e-16 ***
age           -0.01569    0.01304  -1.204    0.229
hmo1          -0.14706    0.05922  -2.483    0.013 *
died1         -0.21777    0.04615  -4.719 2.38e-06 ***
...
Number of linear predictors:  2
Names of linear predictors: loge(munb), loge(size)
...

> AIC(m1)
[1] 13825.6
> AIC(m2)
[1] 9520.559
```

Example. Counting skunks

Example. Skunks

Special traps were used to count the number of skunks in an area. Skunks that did not fall to the trap were not counted (thus the dataset does not contain zeros). The number of times each individual skunk was captured over several trappings was recorded. Individual skunks never trapped could not be recorded, so only counts of one or more appear in the data. One goal is to estimate the number of skunks in the area for each sex in each of two years. The dataset contains the following variables:

- `y` – the number of times a skunk was captured (response)
- `year` – 1977 or 1978
- `sex`

Source: Zelterman, D. (2002). Advanced Log-Linear Models using SAS

Skunk

Skunks are nocturnal animals. They hunt at night and sleep in the daytime. Skunks are omnivorous, eating both plant and animal material and changing their diets as the seasons change. Skunks are North and South American mammals. Not related to polecats which are in the weasel family, the closest Old World relative to the skunk is the stink badger. The animals are known for their ability to spray a liquid with a strong unpleasant smell.

Example. Skunks, ZTP model (1)

```
> modelZTP=vglm(y~year*sex,freq=freq,family="pospoisson",data=skunk)
> summary(modelZTP)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0372     0.2014   5.149 2.62e-07 ***
year78       -0.4187     0.2742  -1.527    0.127
sexM          0.1933     0.2553   0.757    0.449
year78:sexM  -0.7492     0.5242  -1.429    0.153
...
Number of linear predictors:  1
Name of linear predictor: loge(lambda)
Log-likelihood: -78.9688 on 47 degrees of freedom
Number of iterations: 5
No Hauck-Donner effect found in any of the estimates
```

Example. Skunks, ZTP model (2)

```
> modelZTP=vglm(y~year,freq=freq,family="pospoisson",data=skunk)
> summary(modelZTP)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1521     0.1237   9.317   <2e-16 ***
year78       -0.6566     0.2104  -3.121   0.0018 **
...
Number of linear predictors:  1
Name of linear predictor: loge(lambda)
Log-likelihood: -80.115 on 49 degrees of freedom
Number of iterations: 4
No Hauck-Donner effect found in any of the estimates

> exp(coef(skunkM)[2])
   year78
0.5186258
```

Interpretation?

Example. Skunks, ZTNB model

```
> modelZTNB=vglm(y~year,family="posnegbinomial",data=skunk)
There were 50 or more warnings (use warnings() to see the first 50)
> summary(modelZTNB)
...
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1   1.1653     0.1250   9.324  < 2e-16 ***
(Intercept):2   7.4287   157.0156   0.047  0.96226
year78         -0.6444     0.2093  -3.080  0.00207 **
...
Number of linear predictors:  2
Names of linear predictors: loge(munb), loge(size)
Log-likelihood: -80.136 on 99 degrees of freedom
Number of iterations: 10
Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):1'
```

Interpretation?

Example. Skunks, conclusion

```
> exp(coef(modelZTNB)[2])
(Intercept):2
     1683.699

> AIC(modelZTP)
[1] 164.2299

> AIC(modelZTNB)
[1] 166.2719
```

Conclusion?

Summary. Models with "zero-problems"

- ZI – zero inflated models, too many zeros, count process zeros are of interest, resulting model is a mixture
- ZA – zero altered (hurdle models) – too many zeros, zeros are not of interest, zeros are analyzed separately (two-step model)
- ZT – zero truncated models – no zeros, conditional model

Models with Poisson distribution: ZIP, ZAP, ZTP

Models with negative binomial distribution: ZINB, ZANB, ZTNB

Summary. Which model is best?

Options for choosing the best model:

- Common sense, follow the schema: Poisson, overdispersed Poisson, NB, excess zeros (ZI, ZA models) or missing zeros (ZT models)
- Information criteria (AIC, BIC)
- Tests (Poisson vs NB, likelihood ratio test, Vuong's test)
- compare the predictions and actual values (RMSE, MAE)

Source: Zuur et al. (2009), p 291

Summary. Overview of count models

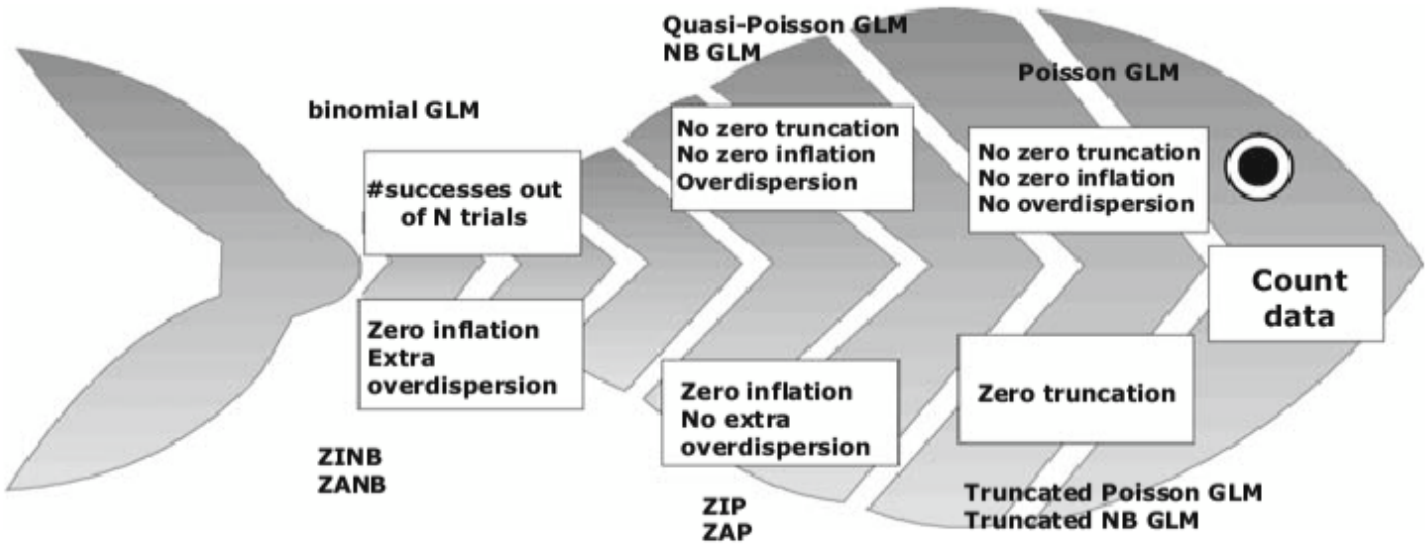

Fig. 11.8 GLMs for count data. Instead of GLM, you can also use GAM. Try sketching in the R functions for each box. If there is no zero truncation, no zero inflation and no overdispersion (*upper right box*), you can apply a Poisson GLM. If there is overdispersion (*upper middle box*), then consider quasi-Poisson or negative binomial GLM. The '#successes out of N trials' box refers to a logistic regression. The trials need to be independent and identical. For zero-truncated data (*lower right box*), you need to apply a zero-truncated Poisson GLM or a zero-truncated negative binomial GLM. If there is zero inflation, you are in the world of ZIP, ZAP, ZINB, and ZINB models. The difference between the P and NB is whether there is overdispersion in the non-zero data. It is a nice exercise to add the names of the corresponding R functions! You can also use the offset in the ZIP, ZAP, ZINB, and ZANB models

Source: Zuur et al. (2009), p 293

Generalized Poisson model

Joe & Zhu (2005), Consul (1989)

Pmf of generalized Poisson distribution

$$p(y;\lambda,\xi) = \frac{\lambda}{y!}[\lambda+\xi y]^{y-1}\exp(-\lambda-\xi y), \quad y = 0,1,\ldots$$

$\lambda > 0, \quad -1 \le \xi < 1$

Mean and variance:

$$\mathbf{E}Y = \mu = \frac{\lambda}{1-\xi}$$

$$\mathbf{D}Y = \frac{\lambda}{(1-\xi)^3} = \frac{\mu}{(1-\xi)^2} = \frac{1}{(1-\xi)^2}\mathbf{E}Y = \varphi\mathbf{E}Y,$$

where $\varphi = \frac{1}{(1-\xi)^2}$ is the variance function

Thus,

- $\xi = 0$ means $\varphi = 1$, i.e. we have Poisson distribution with parameter λ
- $\xi > 0$ means $\varphi > 1$, i.e. overdisperson
- $\xi < 0$ means $\varphi < 1$, i.e. underdisperson

Example. GP model. Elderly drivers

Data from 901 drivers with at least 65 years of age from Alabama 1991–1996
response variable – number of accidents within 5 years
arguments – several health related variables, driving habits, car type, final sample had 595 drivers

About data: mean number of accidents 0.76, variance 1.33 (overdispersion!)
Poisson model and GP model were estimated, GP model is better ($(\xi =)\alpha \neq 0$)
Both models have 7 significant arguments
NB model was also estimated, but GP is considered more flexible

Further developments
Famoye, Singh (2006): drivers who did not drive did not cause accidents $\Rightarrow$ ZIGP model
Zeros in data 47.2%, Poisson or GP model suggest 36%

Famoye, Wulu, Singh (2004). On the generalized Poisson Regression model with an application to accident data. *Journal of Data Science*, 2, 287–295

Example continued. Results of Poisson and GP models

Table 2: Determinants of elderly automobile accidents

Variable	Poisson Estimate±se	t-value	GPR Estimate±se	t-value
Intercept	$-0.5924 \pm .1849$	-3.20^*	$-0.6309 \pm .1996$	-3.16^*
Black	$0.1856 \pm .1138$	1.63	$0.2015 \pm .1226$	1.64
Ca_blo	$-0.4644 \pm .2010$	-2.31^*	$-0.4686 \pm .2098$	-2.23^*
Drivave	$0.2725 \pm .1245$	2.19^*	$0.2908 \pm .1348$	2.16^*
Everyday	$0.2250 \pm .0998$	2.25^*	$0.2167 \pm .1068$	2.03^*
Gender	$0.1735 \pm .0997$	1.74	$0.1689 \pm .1063$	1.59
Glaucmed	$-0.2288 \pm .2469$	-0.93	$-0.1883 \pm .2626$	-0.72
Walk	$0.6461 \pm .1232$	5.24^*	$0.5965 \pm .1359$	4.39^*
Vasodil	$-0.5904 \pm .3603$	-1.64	$-0.6075 \pm .3726$	-1.63
Hway	$0.4338 \pm .1404$	3.09^*	$0.4289 \pm .1487$	2.88^*
Objects	$-0.4582 \pm .1310$	-3.50^*	$-0.3977 \pm .1443$	-2.76^*
Work	$0.2828 \pm .1108$	2.55^*	$0.2450 \pm .1206$	2.03^*
Educ	$-0.1453 \pm .1048$	-1.39	$-0.1275 \pm .1119$	-1.14
α			$0.0794 \pm .0296$	2.68^*
Log-likelihood	-673.3		-667.0	

* means significant at 0.05 level, se = standard error

Further generalizations

- Zero inflated GP models (ZIGP)
- Zero truncated GP models (ZTGP)
- other generalizations to Poisson model (e.g. Conway-Maxwell-Poisson)

Example. Australian doctor visits (revisited)

Let us recall the Australian doctor visits dataset.

The dataset contains information for approximately 5,000 Australian individuals about the number and possible determinants of doctor visits that were made during a two-week interval.
Variables used for modelling:

- `doctorco` – response variable, the number of visits
- `sex` – 0/1 (male/female)
- `age` – age/100 (people over 72 are coded to age 72)
- `illness` - number of illnesses during 2 weeks (1, .., 5; over 5 coded to 5)
- `income` - income (in 1000AUD)
- `hscore` - health score (bigger score means worse health)

Example solution in R. GP model (1)

```
> library(VGAM)
> modelGP = vglm(doctorco ~ sex + age + illness + hscore,
                 family = "genpoisson", data = docvisit)
> summary(modelGP)
...
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  0.47618    0.03161  15.065  < 2e-16 ***
(Intercept):2 -2.63041    0.08154 -32.257  < 2e-16 ***
sex            0.21126    0.06530   3.235  0.00121 **
age            1.02032    0.15408   6.622 3.55e-11 ***
illness        0.24495    0.02093  11.701  < 2e-16 ***
hscore         0.09163    0.01136   8.067 7.18e-16 ***
...
Number of linear predictors:  2
Names of linear predictors: rhobit(lambda), loge(theta)
Log-likelihood: -3363.963 on 10374 degrees of freedom
Number of iterations: 5
Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):2'
```

Example solution in R. GP model (2)

How is R output related to GP model?

Recall the pmf of GP:

$$p(y_i; \lambda_i, \xi) = \frac{\lambda_i}{y_i!}[\lambda_i + \xi y_i]^{y_i - 1} \exp(-\lambda_i - \xi y_i), \quad y_i = 0, 1, \ldots$$

In R:

- `theta` corresponds to $\hat{\lambda}_i$
  $\Rightarrow \quad \ln(\hat{\lambda}_i) = \eta_i^{(2)} = -2.63041 + 0.21126 \cdot \texttt{sex} + 1.02032 \cdot \texttt{age} + \ldots$
- `lambda` corresponds to $\hat{\xi} \Rightarrow \quad \texttt{rhobit}(\hat{\xi}) = \eta^{(1)} = 0.47618$

What about `rhobit`?

- $\texttt{rhobit}(\xi) = \frac{1+\xi}{1-\xi}$, inverse function is $\texttt{rhobit}^{-1}(\eta) = \frac{\exp(\eta)-1}{\exp(\eta)+1}$
- no know relation to hobbit (or any other J.R.R.Tolkien's character)