L18 Models for data with many zeros

BIOS6643

EJC

Department of Biostatistics & Informatics, CU Anschutz

1. Reading
2. Data with many zeros
3. Zero-inflated Models
4. Truncated Models
5. Data example
6. Longitudinal data with many zeros
7. Extensions of two-part or zero-inflated models
8. Summary

Reading

- Zeiles (2008, J of Stats Software)
- Chapter 11 of Zuur (2009). Mixed effects models and extensions in ecology
- Buu (2012, SIM)
- Ridout (1998)
- Lambert (1991)

Data with many zeros

- Data with many zeros are common in several research areas including ecology, and health care. For example in studies of:
  - hot flushes/flashes in women in menopause
  - seizures in patients with focal epilepsy
  - migraines
  - cancer tumors
- Standard distributions cannot accommodate the amount of zeros found in the data
- **Ignoring zero inflation can have issues**:
  i. the estimated parameters and standard errors may be biased;
  ii. the excessive number of zeros can cause overdispersion.

There are two main models to accommodate extra-zeros

1. Truncated/two-part models
    - These models **separate the zeros from the non-zeros**
    - Then we can model the zeros with a binary distribution (e.g. logistic), and the non-zeros with a truncated (without 0s) distribution
    - Classic example comes from propagation experiments (agriculture) where there was interest in
        a. the proportion of cuttings that rooted and
        b. mean number of roots per rooted cutting
2. Zero-inflated models
    - These models make the distinction between **structural** versus **sampling zeros**
    - Structural zeros are those that are *inevitable*. For example, if there are cuttings unable to root.
    - Sampling zeros *occur by chance*. For example, if in a specific lot there were no rooted cuttings but the soil was good (maybe the did not root on time to be counted).

We will focus on models that build from Poisson model, but other models are also possible (e.g. negative binomial).

Zero-inflated Models

Consider π the probability of a true/structural zero. Then the mixture model of extra-zeros and a count distribution may be written as

$$f(y) = (\pi)I_\pi + (1 - \pi)f(y; \mu)$$

where I_π is the denegerate distribution taking the value of 0 with probability 1; μ is the parameter of the distribution of counts.

Zero-inflated Poisson (ZIP) model

Recall the Poisson probability density function with mean μ_i is $\frac{\mu^{y_i} e^{-\mu_i}}{y_i!}$. Note the probability of 0 under Poisson distribution is $e^{-\mu_i}$. Let p_i be the probability of a structural zero (e.g. those cuttings that are unable to root). Then the probability density of a Zero-inflated Poisson may be written as follows.

$$f(y_i = 0) = p_i + (1 - p_i)e^{-\mu_i} \tag{1}$$

$$f(y_i|y_i > 0) = (1 - p_i)\frac{\mu^{y_i} e^{-\mu_i}}{y_i!} \tag{2}$$

- Regression parameters may be introduced in p_i and in μ_i.

Truncated Models

Truncated models are also known as two-part, hurdle or altered-zero models.

To develop a two-part model we need to specify the probability the probability of 0s (let's call it π_0), and a distribution for the number of 'events/counts' when crossing the hurdle (i.e. $y_i > 0$).

Truncated Poission model

If we use Poisson for counts when $y_i > 0$, then the distribution of y is

$$f(y_i = 0) = \pi_i \tag{3}$$

$$f(y_i | y_i > 0) = (1 - \pi_i) \frac{\mu^{y_i} e^{-\mu_i}}{(1 - e^{-\mu_i}) y_i!} \tag{4}$$

Note the Poisson process is excluding the 0 count values.

Data example

- A sample of 915 biochemistry graduate students was obtained and the count of articles produced during last 3 years of PhD was recorded
- Variables recorded were: sex/gender (fem: men/women), marital status (mar), number of children aged 5 or younrder (kid5), prestige of PhD department (phd), number of articles produced by PhD mentor during last 3 years (ment)
- Study reported in Long(1990, Social Forces)
- *bioChemists* dataset, part of the *pscl* package
- We will use the *pscl* package to fit truncated/hurdle and ZIP models

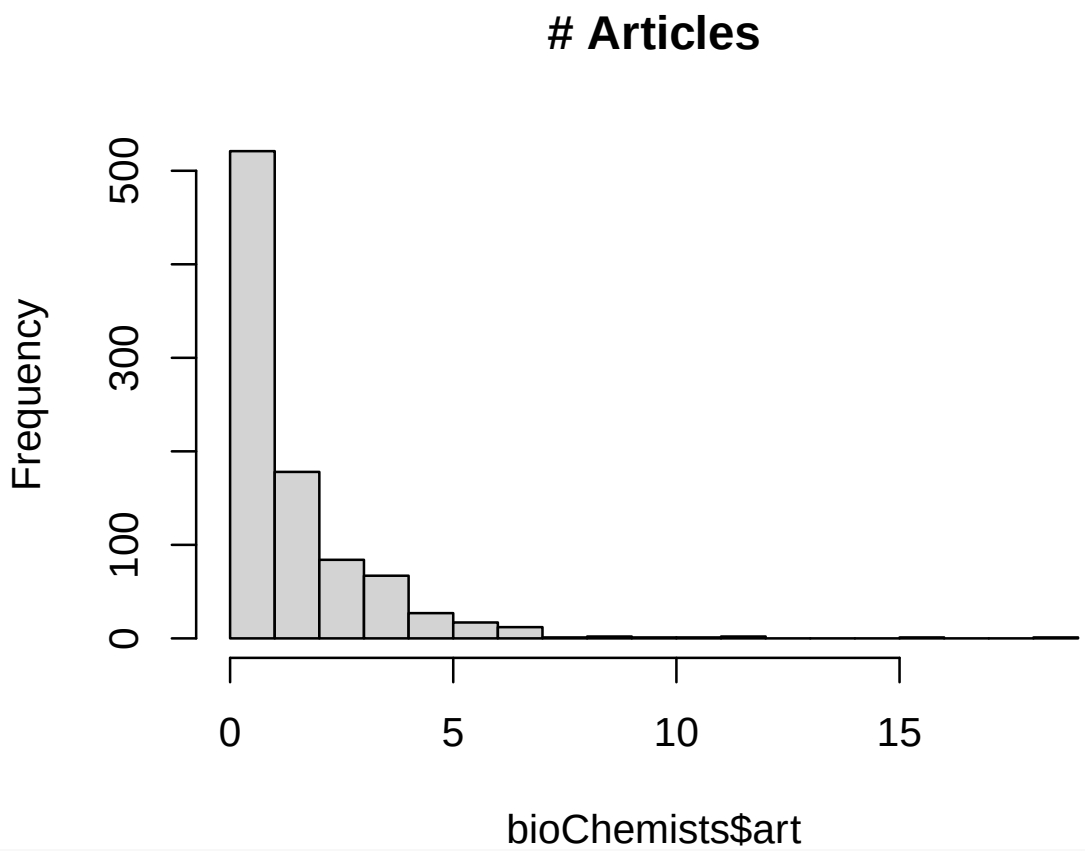
```
data("bioChemists", package = "pscl")

head(bioChemists, 2)
```

```
##   art   fem     mar kid5  phd ment
## 1   0   Men Married    0 2.52    7
## 2   0 Women  Single    0 2.05    6
```

```
##hist(bioChemists$art)
```

```
hist(bioChemists$art, breaks=20, main='# Articles')
```




```
table(bioChemists$art)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  16  19
## 275 246 178  84  67  27  17  12   1   2   1   1   2   1   1
```

Example of truncated/hurdle Poisson model in R

```r
fit.hurdle <- hurdle(art ~ fem + mar + kid5 + phd + ment | fem + mar + kid5 + phd + ment,
                     dist="poisson",
                     zero.dist = "binomial",
                     link = "logit", ## link function of the binomial zero hurdle (only used if zero.dist = "binomial")
                     data = bioChemists)
summary(fit.hurdle)
```

```
##
## Call:
## hurdle(formula = art ~ fem + mar + kid5 + phd + ment | fem + mar + kid5 +
##     phd + ment, data = bioChemists, dist = "poisson", zero.dist = "binomial",
##     link = "logit")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.4105 -0.8913 -0.2817  0.5530  7.0324
##
## Count model coefficients (truncated poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.67114    0.12246   5.481 4.24e-08 ***
## femWomen    -0.22858    0.06522  -3.505 0.000457 ***
## marMarried   0.09649    0.07283   1.325 0.185209
## kid5        -0.14219    0.04845  -2.934 0.003341 **
## phd         -0.01273    0.03130  -0.407 0.684343
## ment         0.01875    0.00228   8.222  < 2e-16 ***
## Zero hurdle model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.23680    0.29552   0.801   0.4230
## femWomen    -0.25115    0.15911  -1.579   0.1144
## marMarried   0.32623    0.18082   1.804   0.0712 .
## kid5        -0.28525    0.11113  -2.567   0.0103 *
## phd          0.02222    0.07956   0.279   0.7800
## ment         0.08012    0.01302   6.155 7.52e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 12
## Log-likelihood: -1605 on 12 Df
```

Example of ZIP model in R

```
fit.zip <- zeroinfl(art ~ fem + mar + kid5 + phd + ment | fem + mar + kid5 + phd + ment,
                    dist="poisson",
                    link = "logit", ## a binomial dist is always used
                    data = bioChemists)
summary(fit.zip)
```

```
##
## Call:
## zeroinfl(formula = art ~ fem + mar + kid5 + phd + ment | fem + mar +
##     kid5 + phd + ment, data = bioChemists, dist = "poisson", link = "logit")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.3253 -0.8652 -0.2826  0.5404  7.2976
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.640839   0.121307   5.283 1.27e-07 ***
## femWomen    -0.209144   0.063405  -3.299 0.000972 ***
## marMarried   0.103750   0.071111   1.459 0.144567
## kid5        -0.143320   0.047429  -3.022 0.002513 **
## phd         -0.006166   0.031008  -0.199 0.842376
## ment         0.018098   0.002294   7.888 3.07e-15 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.577060   0.509386  -1.133  0.25728
## femWomen     0.109752   0.280082   0.392  0.69517
## marMarried  -0.354018   0.317611  -1.115  0.26501
## kid5         0.217095   0.196483   1.105  0.26920
## phd          0.001275   0.145263   0.009  0.99300
## ment        -0.134114   0.045243  -2.964  0.00303 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 19
## Log-likelihood: -1605 on 12 Df
```

Longitudinal data with many zeros

- We can build extensions of hurdle and zero-inflated models to account for longitudinal data (see Buu (2012, SIM))
- For example, a **hurdle longitudinal model** for an outcome y_{ij} (time point j) could be specified as follows

$$f(y_{ij} = 0) = \pi_{ij} \tag{5}$$

$$f(y_{ij}|y_{ij} > 0) = (1 - \pi_{ij}) \frac{\mu^{y_{ij}} e^{-\mu_{ij}}}{(1 - e^{-\mu_{ij}}) y_{ij}!} \tag{6}$$

Let x_{ij} be the set of covariates that contribute to the 'intensity' of the events (e.g. number of articles in the data example), and z_{ij} the set of covariates that contribute to the probability of having no events (e.g. no articles published in the last 3 years of PhD). The parameters may be modeled by

$$log(\mu_{ij}) = x'_{ij}\boldsymbol{\beta} + f_m(t_{ij}) + a_i \tag{7}$$

$$logit(\pi_{ij}) = z'_{ij}\boldsymbol{\gamma} + f_0(t_{ij}) + b_i \tag{8}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are fixed effects for covariates x_{ij} and z_{ij}; a_i and b_i are random effects accounting for within and between heterogeneity; and f_m and f_0 are functions of time.

Often a normal distribution of random effects is assumed,
i.e. $(a_i, b_i)' \sim N(0, \Sigma)$.

Notes:

- If there is no correlation between random effects a_i and b_i, then the hurdle model can be fit by fitting two separate models. Otherwise, the sub-models needs to be fit, and this becomes a **joint model**.
- A longitudinal ZIP model may be similarly specified as above.

Extensions of two-part or zero-inflated models

- Other distributions may be considered to model the count (events) part of the models, e.g. negative binomial, geometric
- Joint models that share or model jointly some components of sub-models
- Models for continuous response with too many zeros; for example, cost
- Zero-deflated models are also possible

Discussion

- A model that accounts for overdispersion may accommodate some extra zeros. For example, the negative binomial accommodates more zeros than the standard Poisson.
- How do we decide between models:
  - Reasons for zeros; which model makes more sense?
  - Model validation, e.g. through analysis of residuals
  - Information criteria (AIC, BIC); compare parameter estimates too
  - Hypothesis testing, e.g. Poisson vs negative binomial (these are nested)
  - Compare observed and fitted values

Summary

- Data with many zeros are common. Ignoring zero inflation can biased parameters and standard errors, and cause overdispersion
- There are two main models to accommodate extra-zeros
  1. Truncated/two-part models: separate the zeros from the non-zeros. Then we can model the zeros with a binary distribution (e.g. logistic), and the non-zeros with a truncated (without 0s) distribution
  2. Zero-inflated models: make the distinction between *structural* versus *sampling zeros*. We model the probability of the structural zeros and the distribution of the distribution of events (including some sampling zeros).
- Package *pscl* may be used to fit truncated/hurdle and ZIP models for crossectional data
- For longitudinal data truncated and ZI-models may be extended (from models discussed here) using parameters that depend on time.
  - Random effects may be used to account for correlation between repeated measurements.
  - Joint models may be specified if association between random effects of sub-models are considered.
- How do we decide between models: reasons for zeros, model validation, information criteria, Poisson nested within NB, comparison of observed vs fitted values.