



Hierarchical Hurdle Models for Zero-In(De)flated Count Data of Complex Designs

Marek Molas¹, Emmanuel Lesaffre^{1,2}

¹ Erasmus MC
Erasmus Universiteit - Rotterdam
The Netherlands

² L-Biostat
Katholieke Universiteit Leuven
Belgium

26th August 2009
International Society for Clinical Biostatistics 30
Prague, Czech Republic

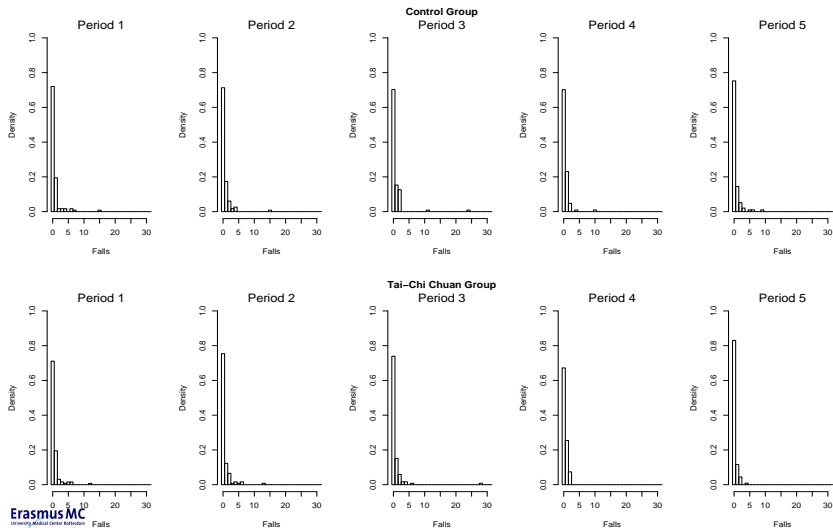
Outline

- Challenges in the analysis of the motivating example
- Hurdle models
- H-likelihood for hurdle models
- Practical application

Motivating Example

- Randomized clinical trial (General Practice Erasmus MC)
- Two physical exercises regimes: standard and Tai-Chi Chuan
- Outcome: **number of days** elderly patients experienced a fall
- Each patient is followed for about a year: **5 measurements**
- Baseline covariates: age, gender, BMI, alcohol use
- Is there a **reduction of number of falls in the Tai-Chi Chuan patients?**

Histogram - Tai-Chi Chuan



The Zero-Inflated Poisson and the Hurdle Model

- **The zero-inflated Poisson model** is a mixture:
 - Point mass at zero
 - Standard Poisson distribution
- **The hurdle model** has two components:
 - Point mass at zero
 - Truncated Poisson distribution

The Zero-Inflated Poisson and the Hurdle Model

Zero-Inflated Poisson Model

$$P(X = 0) = p_Z + (1 - p_Z)e^{-\mu_Z}$$

$$P(X = k) = (1 - p_Z)e^{-\mu_Z} \frac{\mu_Z^k}{k!}$$

Hurdle Model

$$P(X = 0) = p_H$$

$$P(X = k) = (1 - p_H) \frac{1}{1 - e^{-\mu_H}} e^{-\mu_H} \frac{\mu_H^k}{k!}$$

The Zero-Inflated Poisson and the Hurdle Model

- Relation between zero-value probabilities:

$$\{p_H = p_Z + (1 - p_Z)e^{-\mu_Z}\} > 0$$

- Inflation part probability:

$$p_Z = \frac{p_H - e^{-\mu_H}}{1 - e^{-\mu_H}}$$

- Relation between means of the distributional parts:

$$\mu_Z = \mu_H$$

- Hurdle model allows for **zero-deflation** and **zero-inflation**
ZIP model allows only **zero-inflation**

Hurdle Model for Clustered Data

- Add covariates and random effects as follows:

$$\begin{aligned}\text{logit}(p_H) &= \mathbf{X}_0^T \boldsymbol{\beta}_0 + b_0 \\ \log(\mu_H) &= \mathbf{X}_1^T \boldsymbol{\beta}_1 + b_1\end{aligned} \quad \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \sim \mathbf{G}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Likelihood factorizes if random effects are independent:
 - Bernoulli model
 - Truncated Poisson model

Hurdle Model for Clustered Data

- **Marginal Likelihood** - SAS PROC NLMIXED (Min & Agresti 2005)
 - 2-level data
 - Normal random effects
 - Different distribution for random effects (Liu & Yu 2007)
- **Marginal Likelihood** - Non-parametric Maximum Likelihood (Min & Agresti 2005)

Hurdle Model for Clustered Data

What are we looking for?

- Method to handle **multilevel** or **multi-membership** data
- Method to allow for **dispersion parameters** to depend on **covariates**
- **Efficient estimation**

Hurdle Model for Clustered Data

- **H - Likelihood** (Lee & Nelder, 1996, 2001; Noh & Lee 2007)
 - What is it?
 - **Computational framework** allowing estimation of models involving random effects
 - It offers:
 - Efficient estimation algorithm
 - REML type of inference for dispersion components
 - Implications:
 - Random effects can be not normal
 - Multilevel / multi-membership data is easily handled
 - Dispersion parameters can depend on covariates
 - Overdispersion can depend on covariates

H-likelihood

Three types of parameters:

- Fixed effects parameters - β
- Random effects - \mathbf{v}
- Dispersion parameters - λ

H-likelihood

- Extended likelihood

$$L_E(\beta, \lambda, \mathbf{v}|\mathbf{y}, \mathbf{v}) = \prod_{i=1}^N \prod_{j=1}^{n_i} f_{\beta, \lambda}(y_{ij}|\mathbf{v}_i) f_{\lambda}(\mathbf{v}_i)$$

- Extended log-likelihood

$$h = \log(L_E(\beta, \lambda, \mathbf{v}|\mathbf{y}, \mathbf{v}))$$

H-likelihood

Adjusted profile likelihood:

- Estimation of β :

$$p_{\mathbf{v}}(h) = h(\beta, \lambda, \mathbf{v})|_{\mathbf{v}=\hat{\mathbf{v}}} - 0.5 \log \left| \frac{D(h, \mathbf{v})}{2\pi} \right|_{\mathbf{v}=\hat{\mathbf{v}}}$$

- Laplace approximation to the integral:

$$L_M(\beta, \lambda | \mathbf{y}) = \prod_{i=1}^N \int \prod_{j=1}^{n_i} f_{\beta, \lambda}(y_{ij} | \mathbf{v}_i) f_{\lambda}(\mathbf{v}_i) d\mathbf{v}_i.$$

H-likelihood

Adjusted profile likelihood:

- Estimation of λ :

$$p_{\beta, \nu}(h) = h(\beta, \lambda, \nu) \Big|_{\beta=\hat{\beta}, \nu=\hat{\nu}} - 0.5 \log \left| \frac{D[h, (\beta, \nu)]}{2\pi} \right|_{\beta=\hat{\beta}, \nu=\hat{\nu}}$$

Application to the Hurdle Model - Truncated Poisson

Standard exponential family distribution

$$f_{\beta}(y_{ij} | \mathbf{v}_i) = \exp(y_{ij}\theta_{ij} - b(\theta_{ij}) + c(y_{ij}))$$
$$\theta_{ij} = \mathbf{x}_{ij}^T \beta + \mathbf{z}_{ij}^T \mathbf{v}_i$$

Truncated exponential family distribution

$$f_{\beta}(y_{ij} | \mathbf{v}_i) = \exp(y_{ij}\theta_{ij} - b(\theta_{ij}) - \log(M(\theta_{ij})) + c(y_{ij}))$$
$$\theta_{ij} = \mathbf{x}_{ij}^T \beta + \mathbf{z}_{ij}^T \mathbf{v}_i$$

Truncated Poisson Distribution

- Standard weight matrix

$$\mathbf{W} = \left(\frac{\partial \mu}{\partial \eta} \right)^2 V^{-1}(\mu)$$

- Modified weight matrix

$$\tilde{\mathbf{W}} = \mathbf{W} + \left(\frac{M''(\theta)}{M(\theta)} - \left(\frac{M'(\theta)}{M(\theta)} \right)^2 \right) V^{-1}(\mu) \mathbf{W}$$

- Standard adjusted dependent variable

$$\mathbf{z} = \eta + (\mathbf{y} - \mu)(\partial \eta / \partial \mu)$$

- Modified adjusted dependent variable

$$\mathbf{z} = \eta + (\mathbf{W} / \tilde{\mathbf{W}}) (\mathbf{y} - \mu - \frac{M'(\theta)}{M(\theta)}) (\partial \eta / \partial \mu)$$

Application - Hurdle Model

- Bernoulli model

$$\text{logit}[\rho(Y_{ij} > 0)] = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i$$

$$u_i = \frac{\exp(v_i)}{1 + \exp(v_i)}$$

$$u_i \sim \text{Beta}\left(\frac{1}{\lambda_{10}}, \frac{1}{\lambda_{10}}\right)$$

$$\log(\lambda_{10}) = \gamma_{100} + \gamma_{101} \times \text{Female}_i$$

- Truncated Poisson model

$$\log\left(\frac{\mu_{ij}}{\text{days}_{ij}}\right) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i + v_{ij}$$

$$v_i \sim \mathcal{N}(0, \lambda_{20})$$

$$u_{ij} = \exp(v_{ij})$$

$$u_{ij} \sim \text{Gamma}\left(\frac{1}{\lambda_{21}}, \lambda_{21}\right)$$

$$\log(\lambda_{21}) = \gamma_{210} + \gamma_{211} \times \text{Female}_i$$

Application

Bernoulli Model

Effect	Estimate	P-value
Intercept	-5.69	0.004
Female	0.96	0.007
Time	$-1.18 \cdot 10^{-3}$	0.132
Time*Trt(Tai-Chi)	$-0.64 \cdot 10^{-3}$	0.454
Age	0.05	0.042
γ_{100}	-0.42	0.096
γ_{101} (Female)	-1.22	0.001

Application

Truncated Poisson Model

Effect	Estimate	P-value
Intercept	-3.54	<0.001
Female	-0.66	0.035
Time	$-2.48 \cdot 10^{-3}$	0.006
Time*Trt(Tai-Chi)	$-2.17 \cdot 10^{-3}$	0.048
γ_{20}	0.11	0.670
γ_{210}	-0.33	0.431
γ_{211} (Female)	-1.50	0.032

Further Research

- Joint estimation of the binary and truncated Poisson model within H-likelihood framework
- Correlated random effects

Thank you for your attention!

m.molas@erasmusmc.nl