Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v6n2p149

A marginalized model for zero-inflated, overdispersed and correlated count data
By Iddi, Molenberghs

Published: 14 October 2013

Electronic Journal of Applied Statistical Analysis
Vol. 6, Issue 2, 2013, 149-165
DOI: 10.1285/i20705948v6n2p149

A marginalized model for zero-inflated, overdispersed and correlated count data

Samuel Iddi[a] and Geert Molenberghs *[a,b]

[a] *I-BioStat, KU Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium*
[b] *I-BioStat, Universiteit Hasselt, Agoralaan 1, 3590 Diepenbeek, Belgium*

Published: 14 October 2013

Iddi and Molenberghs (2012) merged the attractive features of the so-called combined model of Molenberghs et al. (2010) and the marginalized model of Heagerty (1999) for hierarchical non-Gaussian data with overdispersion. In this model, the fixed-effect parameters retain their marginal interpretation. Lee et al. (2011) also developed an extension of Heagerty (1999) to handle zero-inflation from count data, using the hurdle model. To bring together all of these features, a marginalized, zero-inflated, overdispersed model for correlated count data is proposed. Using two empirical sets of data, it is shown that the proposed model leads to important improvements in model fit.

keywords: Marginal multilevel model, Maximum likelihood estimation, Random effects model, Negative binomial, Overdispersion, Partial Marginalization, Poisson model, Zero-Inflation.

1. Introduction

Count data are gathered in a multitude of settings. For their simplest, univariate form, a generalized linear model (GLM; Agresti, 2002; Nelder and Wedderburn, 1972) based on the Poisson distribution is regularly assumed, a well-known member of the exponential family. The mean response is then modeled linearly in unknown regression parameters, on the log scale. This frameworks easily admits maximum likelihood estimation techniques. In spite of this elegance, four various features have called for extension.

First, because in practice, the empirical data generally exhibit much more heterogeneity than that provided by the restrictive mean-variance relationship of the Poisson,

*Corresponding authors: geert.molenberghs@med.kuleuven.be

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

i.e., overdispersion, a large collection of extensions have been proposed. Note that underdispersion is equally well possible. Particularly, the negative binomial (NB) model (Breslow, 1984; Lawless, 1987) has been handy in addressing this issue.

Second, the occurrence of zeros beyond what is predicted by the Poisson distribution are encountered, for example, when the response of interest is rare. One model allowing for such excess zeros is the zero-inflated Poisson model (ZIP; Lambert, 1992). With this model, the process generating both the zeros and positive counts are distinguished from each other, each with their own set of parameters. Alternatively, a zero-inflated negative binomial model (ZINB; Ridout et al., 2001) simultaneously addresses overdispersion and extra-Poisson zeros.

We turn to the third issue. The target of inference may be geared towards an individual subject, e.g., the prediction of an outcome based on a covariate profile in a clinical context, on the one hand, or towards an entire (sub)population in a public-health context, on the other. For univariate outcomes, with or without overdispersion and/or zero inflation, the above models allow for both of these targets simultaneously. However, assuming measurements are taken repeatedly over time, or are otherwise clustered in compound units, these targets of inference generally cannot be addressed easily using a single model, apart from important exceptions in the setting of continuous outcomes. The additional feature present is within-unit association, which can be modeled, for example, through the introduction of individual-specific random effects, i.e., using the generalized linear mixed model (GLMM; Breslow, 1993; Engel, 1994; Laird and Ware, 1982). While this model is well established, further complication arises when overdispersion and zero inflation are also present.

Some progress has been made. For example, random effects were introduced into the NB, ZIP, and ZINB models (Hall, 2000; Min and Agresti, 2005; Yau et al., 2001). Further, to simultaneously address correlation and overdispersion, Molenberghs *et al* (2007, 2010) introduced the so-called combined model (CM) that decomposes the Poisson mean into two multiplicative components, each with its own random effect. The first random effect, usually chosen of a conjugate nature addresses overdispersion, while the second random effect, embedded into the linear predictor and often normally distributed, accounts for hierarchies in the data. The model has also been extended to address zero-inflation (ZICM; Kassahun et al., 2012).

Still, and this brings us to the fourth issue, by including individual-specific random effects into the linear predictor, the fixed-effect parameters no longer have a marginal interpretation. Rather, they are directly interpretable conditional upon the random effects. In other words, the models are natural for the individual-specific target of inference, but not for the population-averaged, or marginal, one. We therefore present a model that, while making use of the aforementioned random effects, still admits a marginal interpretation. This model was pioneered by Heagerty (1999). The marginalized multilevel model (MMM) defines separately a marginal mean model and a conditional mean model and the two models are held together by a so-called connector function. Iddi and Molenberghs (2012) extended this marginalized model to accommodate for overdispersion. Lee et al. (2011) also proposed an extension to zero-inflated clustered count data, using the hurdle model (Mullahy, 1986). We will show that further extension is possible to simulta-

neously account for overdispersion, zero-inflation, and data hierarchies, while retaining to the regression parameters their population-averaged interpretation. The proposed model is illustrated with two empirical datasets.

The remainder of the paper is organized as follows. Section 2 is devoted to the introduction of two datasets; these are analyzed in Section 5. The combined overdispersed marginalized multilevel model (COMMM) is reviewed, and the proposed zero-inflated version presented in Section 3. The maximum likelihood estimation strategy is the subject of Section 4.

2. Case Studies

2.1. A Clinical Trial in Epileptic Patients

A full description of the epilepsy dataset is provided in Faught et al. (1996). In summary, the data come from a randomized, double-blinded, parallel group multi-center study aimed at comparing placebo with a new anti-epileptic drug (AED), in combination with one or two other AED's. After a 12-week baseline stabilization period, during which the number of epileptic seizures were countered, 45 patients were randomly placed on the placebo, with a second group of 44 patients receiving the new active drug. For 16 weeks, patients were followed and the number of seizures counted. Thereafter, they entered a long term open-extension study during which some patients were followed for as long as 27 weeks after randomization. The key research interest was to investigate whether or not this new treatment helps to reduce the number of epileptic seizure.

2.2. The Whitefly Study

The whitefly dataset resulted from a horticultural experiment to examine the effect of six methods of applying the insecticide imidacloprid to poinsettia plants. These data have previously been reported by van Iersel et al. (2001) and also analyzed in Hall (2000) and Hall and Zhang (2004). Using a randomized complete block design, treatment (method) was applied to 18 experimental units that consisted of a trio of 18 poinsettia plants (54 plants in total); repeated measurements were taken over 12 consecutive weeks. The experimental units were randomly assigned to the 6 treatments in 3 complete blocks. One of the outcomes of this study, of interest here, was the number of immature whiteflies after treatment out of a number of insects caged in one leaf per plant, prior to measurement of the response. The study aimed at investigating the best method to control silverleaf whiteflies on the plants.

3. Methodology

3.1. Combined Overdispersed and Marginalized Multilevel Model (COMMM)

Let Y_{ij} denote the jth ($j = 1, 2, \ldots, n_i$) count outcome measured for cluster (subject) $i = 1, 2, \ldots, N$. Suppose the components in $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ follow Poisson distributions with mean number of events λ_{ij}. The combined overdispersed and marginal

multilevel model [15] brings together aspects of the combined model introduced by Molenberghs *et al* (2007, 2010) and the proposal of Heagerty (1999). The specific case of Poisson data takes the form: $Y_{ij} \sim \text{Poi}(\lambda_{ij}^c)$, with $\lambda_{ij}^c = \theta_{ij}\kappa_{ij}$ and $\kappa_{ij} = \exp(\Delta_{ij} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i)$ and marginal regression $\log(\lambda_{ij}^m) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}^m$, i.e., with $\lambda_{ij}^m = E(Y_{ij})$. The full design is $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ip})'$. The conditional mean also uses a log-link function and depends on the function Δ_{ij} and two random effects, $\theta_{ij} \sim \text{Gamma}(u_{ij}, v_{ij})$ and $\boldsymbol{b}_i \sim N(0, \boldsymbol{D})$ and another design matrix $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2}, \ldots, \boldsymbol{z}_{iq})'$ of known covariates associated with $\boldsymbol{b}_i$. The gamma and normal random effects are assumed to be independent. Here, θ_{ij} is used to handle the dispersion present in addition to what is already captured by the model. The normal random effects capture correlation between repeated measures. By integrating the conditional mean over the random effects, the marginal mean follows and hence the connector function Δ_{ij} can be obtained. Precisely, we can solve for Δ_{ij} from the equation:

$$\lambda_{ij}^m = \exp(\boldsymbol{x}_{ij}'\beta^m) = \int_b \int_\theta \theta_{ij} \exp(\Delta_{ij} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i) dG_\theta dF_b = \int_b E(\theta_{ij}) \exp(\Delta_{ij} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i) dF_b, \quad (1)$$

where $G_\theta(\cdot)$ and $F_b(\cdot)$ are the cumulative distribution functions of θ_{ij} and $\boldsymbol{b}_i$, respectively. Solving this integral leads to $\Delta_{ij} = -\log(u_{ij}v_{ij}) + \boldsymbol{x}_{ij}'\boldsymbol{\beta}^m - \frac{1}{2}\boldsymbol{z}_{ij}'\boldsymbol{D}\boldsymbol{z}_{ij}$. Evidently, the parameters, $\boldsymbol{\beta}^m$ have marginal interpretation. Furthermore, the joint marginal distribution is obtained from integrating out the two random effects from the conditional distribution.

3.2. Zero-Inflated, Overdispersed, and Marginalized Multilevel Model

When count data exhibit a high percentage of zero counts for some given covariates, the appropriate model to handle these excess zeros is the zero-inflated Poisson model or the extended zero-inflated negative binomial. Ideas can be transported to the context of the combined overdispersed and marginalized multilevel model presented in Section 3.1. To begin with, we will assume that the counts are generated from two processes. The first process generates the excess zeros with probability π_{ij}, while the full range of counts are generated with probability $(1 - \pi_{ij})$. We write these as:

$$Y_{ij} \sim \begin{cases} 0 & \text{with probability} \quad \pi_{ij}, \\ \text{Poisson}(\lambda_{ij}) & \text{with probability} \quad (1 - \pi_{ij}). \end{cases}$$

Further, assume that the zeros are generated from two sources based on probabilities of the two processes. This leads to the zero-inflated model with probabilities:

$$P(Y_{ij} = y_{ij}|\theta_{ij}, \boldsymbol{b}_i) = \begin{cases} \pi_{ij} + (1 - \pi_{ij}) f_i(0|\theta_{ij}, \boldsymbol{b}_{1i}, \lambda_{ij}) & \text{if} \quad y_{ij} = 0, \\ (1 - \pi_{ij}) f_i(y_{ij}|\theta_{ij}, \boldsymbol{b}_{1i}, \lambda_{ij}) & \text{if} \quad y_{ij} > 0, \end{cases}$$

where the mixing probability π_{ij} and the Poisson mean λ_{ij} are modeled with covariates and random effects. This yields the conditional zero-inflated combined overdispersed and correlated model (Kassahun et al., 2012). We propose a marginalized version referred

to as zero-inflated, combined overdispersed marginalized multilevel model (ZICOMMM) which can be spelled out as follows:

$$P(Y_{ij}=y_{ij})=\begin{cases}\pi_{ij}^m+(1-\pi_{ij}^m)f_i(0|\lambda_{ij}^m) & \text{if} \quad y_{ij}=0,\\ (1-\pi_{ij}^m)f_i(y_{ij}|\lambda_{ij}^m) & \text{if} \quad y_{ij}>0,\end{cases}$$

where the marginal mixing probability π_{ij}^m and marginal Poisson means λ_{ij}^m are related to only covariates through a logit-link and log-link function respectively. That is, $\text{logit}(\pi_{ij}^m)=\boldsymbol{x}'_{1ij}\boldsymbol{\beta}^m$ and $\log(\lambda_{ij}^m)=\boldsymbol{x}'_{2ij}\boldsymbol{\alpha}^m$. Next, the conditional specification is as follows;

$$P(Y_{ij}=y_{ij}|\theta_{ij},\boldsymbol{b}_i)=\begin{cases}\pi_{ij}^c+(1-\pi_{ij}^c)f_i(0|\theta_{ij},\boldsymbol{b}_{1i},\lambda_{ij}^c) & \text{if} \quad y_{ij}=0,\\ (1-\pi_{ij}^c)f_i(y_{ij}|\theta_{ij},\boldsymbol{b}_{1i},\lambda_{ij}^c) & \text{if} \quad y_{ij}>0,\end{cases}$$

where $\pi_{ij}^c=\Phi^{-1}(\Delta_{1ij}+\boldsymbol{z}'_{1ij}\boldsymbol{b}_{1i})$ and $\lambda_{ij}^c=\theta_{ij}\text{exp}(\Delta_{2ij}+\boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i})$. Here, $\Phi(\cdot)$ is the cumulative normal distribution function. Note that only non-zero count data exhibit overdispersion and so the overdispersion random effect θ_{ij} is introduced into the Poisson model. The binomial model for the mixing conditional probability has only one random effect, whereas two random effects are used in the positive-counts version, to address overdispersion and correlation. The functions, Δ_{1ij} and Δ_{2ij} can be calculated using iterated expectations, $E(Y_{ij})=E_b\{E_\theta[E_y(Y_{ij}|\theta_{ij},\boldsymbol{b}_i)]\}$. For $\boldsymbol{b}_i=(\boldsymbol{b}_{1i},\boldsymbol{b}_{2i})'\sim N(\boldsymbol{0},\boldsymbol{D})$ and based on (1) the closed-form expressions are:

$$\begin{aligned}\Delta_{1ij} &= \sqrt{1+\boldsymbol{z}'_{1ij}\boldsymbol{D}\boldsymbol{z}'_{ij}}\,\Phi^{-1}\left[\text{expit}(\boldsymbol{x}'_{1ij}\boldsymbol{\beta}^m)\right],\\ \Delta_{2ij} &= -\log(u_{ij}v_{ij})+\boldsymbol{x}'_{2ij}\boldsymbol{\alpha}^m-\tfrac{1}{2}\boldsymbol{z}'_{2ij}\boldsymbol{D}\boldsymbol{z}'_{ij}.\end{aligned}$$

We have used the probit link in the conditional mixing probability model so that analytical expressions for the integral in (1) can be obtained. The marginal mean model, however, still makes use of the logit link and thus, odds ratio interpretation can still be obtained from the fixed marginal parameters. Of course, should it be desirable, then this logit can be replaced by a probit as well. Assuming that it is adequate to use only random intercepts for both conditional models, that is, $\boldsymbol{z}'_{1ij}\boldsymbol{b}_{1i}=b_{1i}$ and $\boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i}=b_{2i}$, then the variance-covariance matrix in terms of correlation parameter ρ is given by

$$\boldsymbol{D}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2^2\\ \rho\sigma_1\sigma_2^2 & \sigma_2^2\end{pmatrix}. \tag{2}$$

The correlation parameter ρ measures the correlation between the binomial and count components. The variance parameters in the logistic part and the Poisson part could be estimated separately, but the covariance part needs to be estimated jointly.

4. Estimation

A number of estimation strategies are possible for fitting this type of models. We will pursue the maximum likelihood approach to obtain parameters and draw inferences.

Maximum likelihood estimation requires optimizing the full joint marginal likelihood. The likelihood is constructed by specifying the likelihood contribution from measurements $\boldsymbol{Y}_i$ conditioned on the random effects θ_{ij} and $\boldsymbol{b}_i$. The presence of the two random effects entails carrying out a double integration to find the joint marginal likelihood. Molenberghs *et al* (2007, 2010) proposed a so-called partial marginalization technique where the conjugate random effect is analytically integrated out while the normal random effect remains untouched. However, since statistical procedures, such as the NLMIXED procedure in SAS, allow for numerical integration of normal random effects by Gaussian and adaptive Gaussian quadrature methods (Pinheiro and Bates, 1995; Pinheiro and Bates, 2000), the conditional likelihood can be fed to the program to complete the full marginalization. The other advantage is that such a procedure further carries out the optimization and returns parameter estimates and standard errors. Molenberghs and Verbeke (2005) reviewed several of such estimation techniques.

The observed data likelihood of the ith subject conditioned on the two random effects is:

$$f_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{D}, \phi) = \int_b \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{b}_i) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i,$$

where

$$f(y_{ij}|\boldsymbol{b}_i) = \int_\theta f(y_{ij}|\theta_{ij}, \boldsymbol{b}_i) f(\theta_{ij}) d\theta_{ij},$$

from which we can derive the likelihood for $\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{D}, \phi$ as

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{D}, \phi) = \prod_{i=1}^{N} f_i(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{D}, \phi).$$

The parameters of the Poisson model, the zero-inflated model, the components of the variance-covariance matrix of the normal random effect and the parameters of the conjugate random effect are contained in the vectors $\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{D}$ and ϕ respectively.

For our count case, suppose that the response distribution is Poisson with mean composing of a normal and a conjugate gamma distributed random effect terms, then

$$f(y_{ij}|\boldsymbol{b}_i) = \left(\begin{array}{c} u_j + y_{ij} - 1 \\ u_j - 1 \end{array} \right) \left(\frac{v_j}{1 + \kappa_{ij} v_j} \right)^{y_{ij}} \left(\frac{1}{1 + \kappa_{ij} v_j} \right)^{u_j} \kappa_{ij}^{y_{ij}}.$$

For the zero-inflated version of the combined model:

$$f(y_{ij}|\boldsymbol{b}_i) = I(y_{ij} = 0)\pi_{ij} + (1 - \pi_{ij}) \left(\begin{array}{c} u_j + y_{ij} - 1 \\ u_j - 1 \end{array} \right) \left(\frac{v_j}{1 + \kappa_{ij} v_j} \right)^{y_{ij}} \left(\frac{1}{1 + \kappa_{ij} v_j} \right)^{u_j} \kappa_{ij}^{y_{ij}}.$$

In fitting the marginalized multilevel models, the conditional distributions are specified by replacing the terms $\boldsymbol{x}'_{1ij}\boldsymbol{\beta}$ and $\boldsymbol{x}'_{2ij}\boldsymbol{\alpha}$ in the zero-inflated version of the combined model with the analytical expressions for Δ_{1ij} and Δ_{2ij} respectively as the mean models relate separately, through link functions, to these terms.

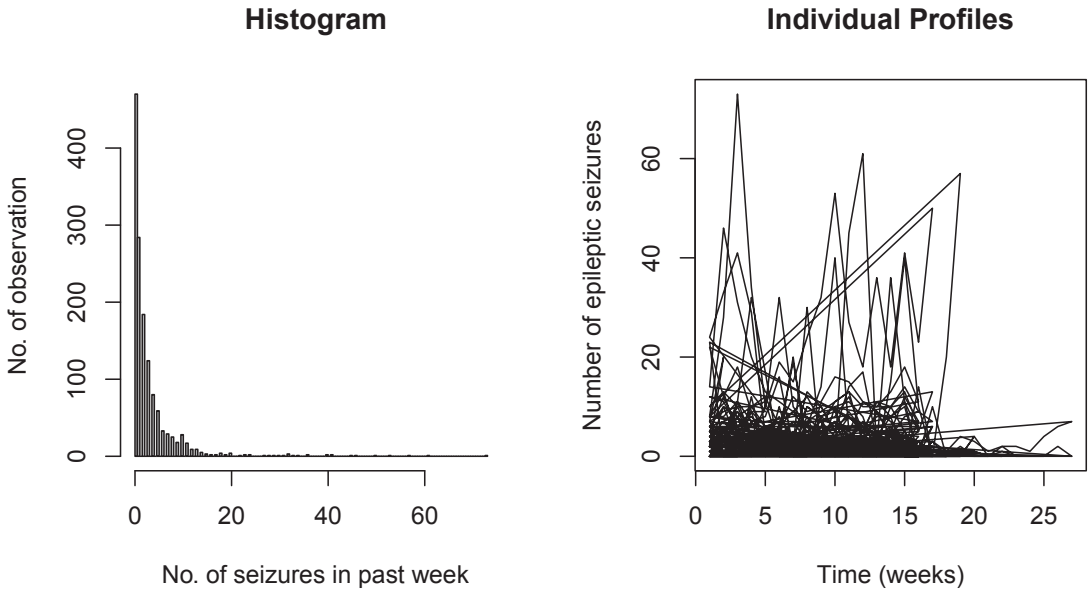

Figure 1: *Epilepsy Data. Histogram and individual profiles.*

Other estimation strategies may include, but are not limited to, pseudo-likelihood (Aerts et al., 2002; Molenberghs and Verbeke, 2005), generalized estimating equations (Zeger et al., 1988), and Bayesian methodology (Aregay et al., 2012).

To assess the fit of the models, we made use of Akaike Information Criterion (AIC; Akaike, 1994) and the Bayesian Information Criterion (BIC; Burnham, 2004). AIC is calculated as follows: AIC $= -2$Log-likelihood$+2k$ where k is the number of parameters in the model. The model with the minimum AIC value is usually the preferred model. Also, the formula to calculate BIC is given by: BIC $= -2$Log-likelihood$+k\log(n)$, where k number of model parameters and n is the sample size. This criterion also selects the 'best' model based on the minimum BIC value.

5. Analysis of Case Studies

5.1. Analysis of the Epilepsy Data

We begin by some data exploration. In Figure 1, the histogram of the number of epileptic seizures shows a higher proportion of excess zeros accounting for about 33% of the data. Also, a simple descriptive statistics shows a very high variance of 37.70, as compared to the empirical mean of 3.18, an indication of overdispersion. Finally, a plot of individual profiles reveals a higher between variability than variability within subjects, indicative of within-subject correlation. To investigate the impact of these features on inferences a number of models are fitted to account for each or a combination of zero-inflation,

overdispersion,and correlation. We denote the number of epileptic seizures experienced by the ith patient at the jth occasion by Y_{ij} and the occasion on which Y_{ij} was measured by t_{ij}. Assuming that Y_{ij} follows a Poisson distribution with mean parameter λ_{ij}^c, then the marginal mean model for the Poisson process is as follows:

$$\ln(\lambda_{ij}^m) = \begin{cases} \alpha_{00} + \alpha_{01} t_{ij} & \text{if placebo} \\ \alpha_{10} + \alpha_{11} t_{ij} & \text{if treated.} \end{cases}$$

For the combined model, $\lambda_{ij}^c = \theta_{ij}\kappa_{ij}$ with $\theta_{ij} \sim \text{Gamma}(u, v)$ and the corresponding marginal quantity:

$$\ln(\kappa_{ij}) = \begin{cases} \alpha_{00} + \alpha_{01} t_{ij} + b_i & \text{if placebo} \\ \alpha_{10} + \alpha_{11} t_{ij} + b_i & \text{if treated.} \end{cases}$$

The marginal model for the zero-inflated probabilities is given by

$$\ln(\pi_{ij}^m) = \beta_0 + \beta_1 t_{ij}.$$

The corresponding conditional models are specified by introducing a normally distributed random intercept, b_{1i} in the Poisson model and b_{2i} and the binomial model with D matrix as in (2).

Results of these models are presented in Table 1. Generally, the fixed-effect parameters are close to each other in the classical models. Their interpretations are not just subject-specific but can be extended to the whole population. We observe further that the standard errors of the parameter estimates in the Poisson model are underestimated as compared to the ZI Poisson. The same comparison can be made between the NB and ZI-NB models. These observations are crucial and point to the need of properly accommodating zero inflation and overdispersion. Furthermore, in terms of both AIC and BIC values, we observe that the ZI-NB model tends to perform better than the other models (in the top part of Table 1). It should be noted that, up to this point, correlation has been ignored. This leads us to the marginalized multilevel models where random intercepts are introduced to accommodate correlation. Comparing the MMM and zero-inflated MMM (ZIMMM) to the NB and ZINB models, we see improvement in the model fit owing to the normal random effect, but above and beyond these, still from overdispersion and zero-inflation. The normal random effects allow for correlated repeated measures and capture some but not all of the overdispersion. This explains why the model fit improves further if the normal random effects are supplemented with zero-inflation and the gamma random effects. Also here,the parameter estimates are rather close to their counterparts from the classical models, but the same is not true for the standard errors. In this particular dataset, the Poisson and ZI Poisson models excepting, all models yielded a significant slope difference between the placebo and the treated group. Finally, it is key that the more complex model results in a considerably improvement in the model. This is essential for inferences and for prediction.

Table 1: *Epilepsy Trial. Parameter estimates (standard errors) for the conventional marginal models (top) and the marginalized models (bottom). RE: random effect*

			Zero-Inflated	Negative	Zero-Inflated
		Poisson	Poisson	Binomial	Negative Binomial
Effect	Par.	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)
			Poisson Part		
Intercept placebo	α_{00}	1.2662(0.0424)	1.4205(0.0439)	1.2594(0.1119)	1.2361(0.1100)
Slope placebo	α_{01}	-0.0134(0.0043)	0.0061(0.0045)	-0.0126(0.0111)	-0.0072(0.0113)
Intercept treatment	α_{10}	1.4531(0.0383)	1.7608(0.0402)	1.4750(0.1093)	1.3974(0.1098)
Slope treatment	α_{11}	-0.0328(0.0038)	-0.01531(0.0041)	-0.0352(0.0101)	-0.0219(0.0112)
Slope difference	$\alpha_{01}-\alpha_{11}$	-0.0195(0.0058)	-0.0214(0.0061)	-0.0227(0.0150)	-0.0147(0.0153)
			Zero-Inflated Part		
Intercept	β_0		-1.2879(0.1203)		-7.1064(1.3344)
Slope	β_1		0.0593(0.0109)		0.2921(0.0655)
Overdispersion	$v=\frac{1}{u}$			0.5274(0.02553)	0.5595(0.03142)
-2Log-likelihood		-1492	-3321	-6755	-6763
AIC		-1484	-3309	-6745	-6749
BIC		-1463	-3278	-6719	-6712

			Zero-Inflated	Combined	Zero-Inflated
		MMM	MMM	MMM	Combined MMM
Effect	Par.	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)
			Poisson Part		
Intercept placebo	α_{00}	1.3960(0.1887)	1.3748(0.1695)	1.4757(0.1962)	1.4282(0.1831)
Slope placebo	α_{01}	-0.0143(0.0044)	-0.0041(0.0047)	-0.0248(0.0077)	-0.0124(0.0070)
Intercept treatment	α_{10}	1.2256(0.1901)	1.3777(0.1719)	1.2200(0.1970)	1.3373(0.1858)
Slope treatment	α_{11}	-0.0120(0.0043)	-0.0072(0.0045)	-0.0118(0.0075)	-0.0045(0.0069)
Slope diff.	$\alpha_{01}-\alpha_{11}$	0.0023(0.0062)	-0.0031(0.0065)	0.0130(0.0107)	0.0080(0.0096)
Variance of RE	σ_1^2	1.1567(0.1844)	0.9459(0.1602)	1.1290(0.1850)	1.0190(0.1742)
			Zero-Inflated Part		
Intercept	β_0		-2.2957(0.2963)		-2.4278(0.3206)
Slope	β_1		0.0657(0.0166)		0.0662(0.0183)
Variance of RE	σ_2^2		1.5728(0.4823)		1.6680(0.5361)
Overdispersion	$v=\frac{1}{u}$			0.4059(0.03481)	0.1792(0.0175)
Correlation	ρ		-0.1382(0.1601)		-0.0795(0.1669)
-2Log-likelihood		-6810	-7240	-7664	-7702
AIC		-6800	-7222	-7652	-7682
BIC		-6787	-7200	-7637	-7658

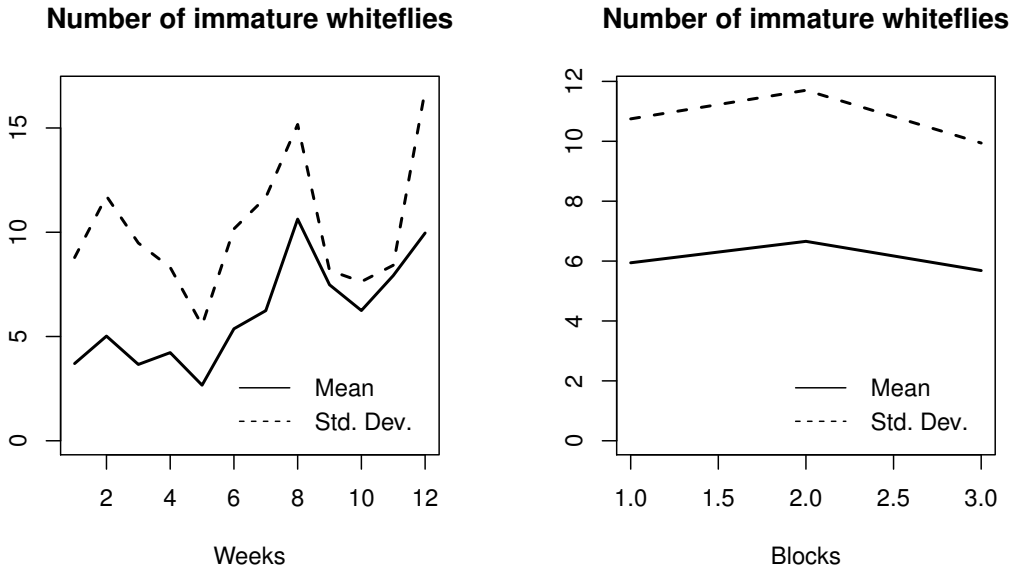

Figure 2: *Whitefly Data. Means and standard deviations by time (panel 1) and block (panel 2).*

5.2. Analysis of the Whitefly Data

Figure 2 shows that, at every level of treatment and block, the variance is always above the mean, reflecting overdispersion, with repetition inducing correlation. In Figure 3, the histogram reveals higher occurrences of zero immature whiteflies, which cannot be accounted for by the variance function of a Poisson or negative binomial distribution. It therefore seems sensible to apply our model.

Denote the number of immature whiteflies for the ith treatment in the jth block measured at the kth week by Y_{ijk}. Assume that $\boldsymbol{Y}_i$ follows a Poisson distribution with expected mean count $\lambda^c_{ijk} = \theta_{ijk}\kappa_{ijk}$. The marginal mean model for the Poisson model is given by:

$$\ln(\kappa_{ijk}) = \mu + \text{block}_j + \text{treatment}_i + \beta\text{week}_k + \log(n_{ijk})$$

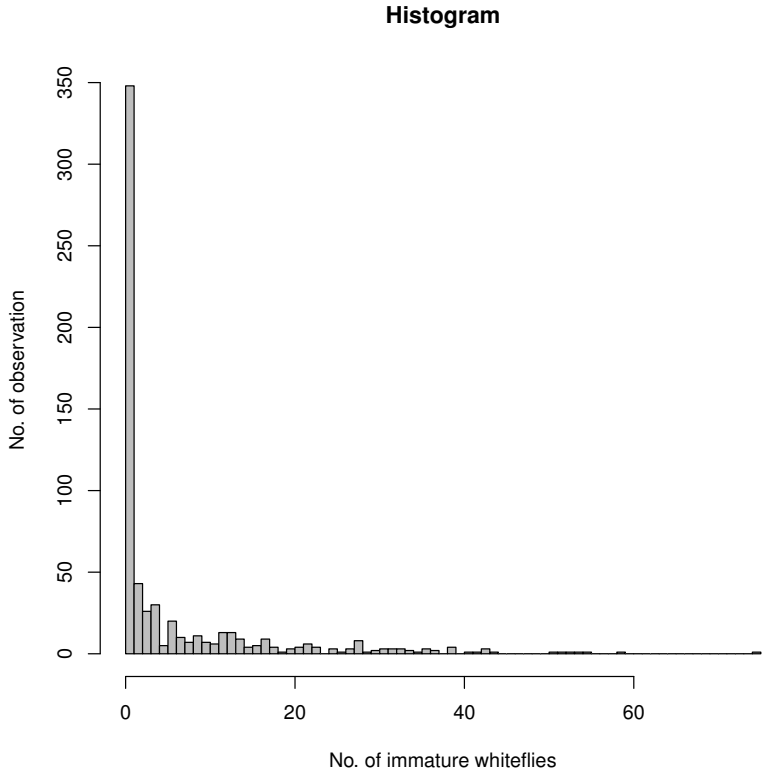

Figure 3: *Whitefly Data. Histogram of the number of immature whiteflies.*

and the probability of zeros π_{ijk} is modeled by $\text{logit}(\pi_{ijk}) = \gamma_0 + \gamma_1 \text{week}_k$, where n represents the number of adult insects placed on the leaf prior to measurement of the response. Week is treated as continuous and the other terms represent factor effects. Results of the fitted models are presented in Table 2. Only the best fitting models have been presented here. We observed that the parameter estimates and standard errors for the logistic part, for all the zero-inflated models seem identical. For the Poisson part, not all the parameters are similar, unlike in the previous example. This is not unexpected because the models are not all nested. Two versions of the ZICOMMM model have been presented here. The first one allows for random effects in the Poisson part, while the second enters them in the logistic part. One would expect that the random effects will be most important in the Poisson part of the model, or perhaps in both. However, the analysis proves otherwise. The higher counts occur in a purely Poisson way but they may be overdispersed. The binary part is rather highly correlated even though in general we expect the two to be present. This means that, given a nonzero count, we get independent replicates, while they are correlated given a zero count. This scenario seems more plausible than to expect heavy correlation among the higher values for the

Table 2: *Whitefly Data. Parameter estimates (standard errors) for the marginalized models. RE: random effect*

		MMM	Zero-Inflated Negative Binomial	Zero-Inflated MMM	Zero-Inflated Comb. MMM(1)	Zero-Inflated Comb. MMM(2)
Effect	Par.	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)
				Poisson Part		
Intercept	μ	-0.9890(0.1678)	-0.6640(0.2107)	-0.0964(0.1467)	-0.3913(0.2058)	-0.3968(0.1588)
Block 1		-0.1337(0.1495)	-0.0418(0.1147)	-0.2997(0.1331)	-0.2345(0.1544)	-0.0709(0.0978)
Block 2		-0.1116(0.1501)	-0.0534(0.1143)	-0.2210(0.1345)	-0.1970(0.1542)	-0.0105(0.0975)
Treatment 1		-1.0592(0.2010)	-1.1088(0.1622)	-0.9047(0.1700)	-0.9893(0.1967)	-0.9896(0.1357)
Treatment 2		-1.4603(0.2075)	-1.2796(0.1802)	-1.0185(0.1845)	-1.1186(0.2276)	-0.9951(0.1495)
Treatment 3		-2.1357(0.2249)	-2.0317(0.1968)	-1.7949(0.2218)	-2.0628(0.2389)	-1.6217(0.2296)
Treatment 4		-1.7553(0.2122)	-1.7034(0.1857)	-1.3379(0.1938)	-1.5730(0.2292)	-1.3471(0.1799)
Treatment 5		1.3570(0.1905)	1.0772(0.1386)	0.9193(0.1567)	1.0241(0.1775)	0.9702(0.1044)
Week	β	0.0976(0.0048)	0.0936(0.0182)	0.0459(0.0052)	0.0788(0.0140)	0.0651(0.0135)
Variance of RE	σ^2	0.1539(0.0428)		0.1007(0.0336)	0.0913(0.0445)	
				Zero-Inflated Part		
Intercept	γ_0		1.5231(0.2919)	1.6694(0.2191)	1.5346(0.2346)	1.6213(0.2610)
Week	γ_1		-0.3667(0.0524)	-0.2994(0.0332)	-0.2942(0.0353)	-0.2789(0.0308)
Variance of RE	σ^2					1.0871(0.3201)
Overdispersion	$v = \frac{1}{u}$		0.5442(0.0816)		0.2887(0.0349)	0.3062(0.0339)
-2Log-likelihood		4011.3	2628.8	3223.8	2675.8	2590.4
AIC		4031.3	2652.8	3247.8	2701.8	2616.4
BIC		4051.2	2706.4	3271.6	2727.6	2642.2

count. Allowing the random effect in the logistic part of the model therefore tends to improve the model fit significantly (smallest AIC and BIC). The model failed to converge when different random effects are used for the different parts of the model. This is not unexpected for models that make use of higher dimension of normal random effects in standard procedures such as in joint models. The two ZICOMMM models also fit better compared to the MMM and ZIMMM, highlighting the importance of acknowledging overdispersion in the model. Also, the higher AIC value for the MMM model brings out the inadequacy resulting from overlooking the all-important inflation of zeros and overdispersion.

6. Concluding Remarks

We have proposed a flexible model to simultaneously address issues of zero-inflation, overdispersion, and data hierarchies, while retaining a population-averaged interpretation of fixed effect parameters like in classical Poisson models. This was achieved by beneficially combining key aspects of recent modeling concept in statistical literature, namely the combined multilevel modeling approach of Iddi and Molenberghs (2012) and the model of Lee et al. (2011). Both models also combined aspects of Heagerty (1999) and the combined model of Molenberghs et al. (2010) in the case of the former, and the

latter introduced zero inflation in the model of Heagerty (1999). These models previously showed that disregarding zero inflation or overdispersion may hamper the fit of the model. Through two different empirical studies, we have demonstrated that it is not sufficient to address either two of the three phenomena, while ignoring the remaining one. Our extension led to considerable improvement, thereby ensuring parameter interpretation is for the whole population, where a population may be defined in terms of fixed-effects profile. Marginal interpretation is often of interest to public health experts, who seek solutions or interventions for the population at large and therefore might find conditional models such as the GLMM or the combined model cumbersome.

This notwithstanding, these features taken together do increase model and fitting complexity. An extensive search over starting values may be in place. This is particularly the case when high-dimensional random effects are used, a problem that is well-known in joint models, in non-linear mixed-effects models, etc.

The models proposed are fully specified and hence enable likelihood inference. For example, expressions for the full probability distribution of the response are available (Fitzmaurice and Laird, 1993; Molenberghs and Lesaffre, 1994). A further advantage is that inferences remain valid for incomplete data, where missingness is of the missing at random type. Reportedly, inferences are more robust under random-effect misspecification as compared to GLMM (Heagerty and Zeger, 2000).

Of course, because likelihood inference is possible does not mean that it is the only route available. Next to Bayesian inference, semi-parametric methods such as estimating-equation or pseudo-likelihood based techniques are equally well possible. Exploring these routes further falls outside of the scope of this paper.

Note that the modeling framework could find use outside of the scope of biomedical applications. For example, in imaging, whether in nuclear medicine or other physical applications, the methodology would allow for extra-Poisson variation in reconstructed images, where the extra variability might stem from scatter, attenuation, etc.

Acknowledgment

The authors gratefully acknowledge the financial support from the IAP research Network P7/06 of the Belgian Government (Belgian Science Policy).

References

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002) *Topics in Modelling of Clustered Data*. London: Chapman & Hall.

Agresti, A. (2002) *Categorical Data Analysis*. New York: John Wiley & Sons.

Akaike, H. (1994). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Aregay, M., Shkedy, Z., and Molenberghs, G. (2012). A hierarchical Bayesian approach

for the analysis of longitudinal count data with overdispersion: a simulation study. *Computational Statistics and Data Analysis*, **57**, 233–245.

Breslow, N.E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**, 38–44.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Burnham, K.P. and Anderson, D.R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods and Research*, **33**, 261–304.

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, **48**, 1–22.

Faught, E., Wilder, B.J., Ramsay, R.E., Reife, R.A, Kramer, L.D., Pledger, G.W., and Karim, R.M. (1996). Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosage. *Neurology*, **46**, 1684–1690.

Fitzmaurice, G.M. and Laird, N.M. (1993). A likelihood based method for analysing longitudinal binary responses. *Biometrika*. **80**, 141–151.

Hall, D.B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, **56**, 1030–1039.

Hall, D.B. and Zhang, Z. (2004). Marginal models for zero inflated clustered data. *Statistical Modelling*, **4**, 161–180.

Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**, 688–698.

Heagerty, P.J. and Zeger, S.L. (2000). Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, **15**, 1–26.

Iddi, S. and Molenberghs. G. (2012). A combined overdispersed and marginalized multilevel model. *Computational Statistics and Data Analysis*, **56**, 1944–1951.

Kassahun, W., Neyens, T., Faes, C., Molenberghs, G. and Verbeke, G. (2012). A zero-inflated overdispersed hierarchical Poisson model. *Submitted for publication.*

Laird, N.M., and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lambert, D. (1992). Zero-Inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–13.

Lawless, J. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15**, 209–225.

Lee, K., Joo, Y., Song, J.J., and Harper, D.W. (2011). Analysis of zero-inflated clustered count data: a marginalized model approach. *Computational Statistics and Data Analysis*, **55**, 824–837.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**, 1–19.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Asso-*

ciation, **89**, 633–644.

Molenberghs, G., Verbeke, G., and Demétrio, C. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, **13**, 513–531.

Molenberghs, G., Verbeke, G., Demétrio, C., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 325–347.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341–365.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series B*, **135**, 370–384.

Pinheiro, J.C. and Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computation and Graphical Statistics.* **4**, 12–35.

Pinheiro, J.C., and Bates, D.M. (2000). *Mixed Effects Models in S and S-Plus.* New York: Springer-Verlag.

Ridout, M., Hinde, J. and Demetrio, C.G.B. (2001). A score test for a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics.* **57**, 219–233.

van Iersel, M., Oetting, R., and Hall, D. B. (2001). Imidicloprid applications by subirrigation for control of silverleaf whitefly on poinsettia. *Journal of Economic Entomology.* **94**, 666–672.

Yau, K.K.W. and Lee, A.H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine*, **20**, 2907–2920.

Zeger, S.L., Liang, K.-Y., and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

Appendix

A. SAS Code

A.1. Combined Marginal Multilevel Model for the Epilepsy Data

```
proc nlmixed data=test qpoints=50;
title 'Gamma-Log-Log-Normal COMMM-  alpha*beta=1';
parms int0=1.3 slope0=-0.02 int1=1 slope1=1.2 sigma=1.2 alpha=2.5;
if (trt = 0) then eta = int0 + slope0*time;
else if (trt = 1) then eta = int1 + slope1*time;
beta=1/alpha;
delta=-log(alpha*beta)+ eta-sigma*sigma/2;
lambda_c =exp(delta+b);
loglik=lgamma(alpha+y)-lgamma(alpha)+y*log(beta)
      -(y+alpha)*log(1+beta*lambda_c)+y*log(lambda_c);
model y ~ general(loglik);
random b ~ normal(0,sigma**2) subject = id;
estimate 'difference in slope' slope1-slope0;
estimate 'ratio of slopes' slope1/slope0;
estimate 'variance RIs' sigma**2;
estimate 'beta=1/alpha' 1/alpha;
run;
```

A.2. Zero-Inflated Combined Marginal Multilevel Model for the Epilepsy Data

```
proc nlmixed data=test qpoints=20 tech=newrap;
title 'ZI Gamma-Log-Normal CM - alpha*beta=1';
parms int0=1.9 slope0=-0.05 int1=1.8 slope1=1.2 a0=-3.7 a1=-0.29
 sigma1=0.192 rho=-0.13 sigma2=2.2 alpha=2.5;
if (trt = 0) then eta_p = int0 + slope0*time;
else if (trt = 1) then eta_p = int1 + slope1*time;
eta_0=a0+a1*time;
pi_m=1/(1+exp(-eta_0));
beta=1/alpha;
delta1=  sqrt(1+(sigma1*sigma1)) * probit(pi_m);
pi_c=probnorm(delta1+b1);
delta2=-log(alpha*beta)+ eta_p-sigma2*sigma2/2;
lambda_c =exp(delta2+b2);
if y=0 then loglik=log((pi_c)+(1-pi_c)*exp(-lambda_c));
else loglik=log(1-pi_c)+lgamma(alpha+y)-lgamma(alpha)+y*log(beta)
            -(y+alpha)*log(1+beta*lambda_c)+y*log(lambda_c);
model y ~ general(loglik);
```

```
random b1 b2 ~ normal([0, 0],[sigma1**2,rho*sigma1*sigma2,sigma2**2])
               subject = id;
estimate 'difference in slope' slope1-slope0;
estimate 'ratio of slopes' slope1/slope0;
estimate 'variance RI1' sigma1**2;
estimate 'variance RI2' sigma2**2;
estimate 'beta=1/alpha' 1/alpha;
run;
```