

# A Zero-Inflated and Overdispersed Marginalized Model for Correlated Counts

Samuel Iddi<sup>1</sup>, Geert Molenberghs<sup>2,1</sup>

<sup>1</sup> I-BioStat, KU Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium

<sup>2</sup> I-BioStat, Universiteit Hasselt, Agoralaan 1, 3500 Hasselt, Belgium

E-mail for correspondence: [Samuel.Iddi@med.kuleuven.be](mailto:Samuel.Iddi@med.kuleuven.be)

**Abstract:** 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 an empirical dataset, it is shown that the proposed model leads to important improvements in model fit.

**Keywords:** Marginal multilevel model; Random effects model; Overdispersion; Poisson model; Zero-Inflation

## 1 Introduction

Count data are gathered in a multitude of settings. For their univariate form, a generalized linear model (GLM) based on the Poisson distribution is regularly assumed, a member of the exponential family. Four features have called for extension. First, because empirical data generally exhibit more heterogeneity than that provided by the mean-variance relationship of the Poisson (overdispersion, but underdispersion is also possible), a collection of extensions has been proposed, such as the negative binomial (NB). Second, the occurrence of zeros beyond what is predicted by the Poisson are often encountered. Models addressing this are, for example, the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB). Third, assuming measurements are taken hierarchially, within-unit association is likely present. The generalized linear mixed model (GLMM) is a commonly used random-effects model to address this. While this model is well established, further complication arises when overdispersion and zero inflation are also present. To address this, overdispersion, Molenberghs *et al* (2010) introduced the combined model (CM) that decomposes the Poisson mean into two multiplicative components, one for each phenomenon.

Fourth, by including individual-specific random effects into the predictor, the fixed effects no longer have a marginal interpretation but are interpreted conditional upon the random effects. We present a model that, while making use of the aforementioned random effects, still admits a marginal interpretation. This multilevel marginal model (MMM) approach is based on Heagerty (1999). This model further simultaneously accounts for overdispersion and zero-inflation. The model is illustrated with real data.

## 2 Zero-Inflated, Overdispersed, Marginalized Multilevel Model

Let  $Y_{ij}$  denote count  $j = 1, \dots, n_i$  for cluster  $i = 1, \dots, N$ , following a Poisson distribution with mean number of events  $\lambda_{ij}$ . We formulate a model that allows for all four issues mentioned in the introduction (Iddi and Molenberghs 2012). The proposed model is:

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij}^m + (1 - \pi_{ij}^m)f_i(0|\lambda_{ij}^m) & \text{if } y_{ij} = 0, \\ (1 - \pi_{ij}^m)f_i(y_{ij}|\lambda_{ij}^m) & \text{if } y_{ij} = 1, 2, \dots \end{cases}$$

where the marginal mixing probability  $\pi_{ij}^m$  and marginal Poisson mean  $\lambda_{ij}^m = E(Y_{ij})$  are related to covariates:  $\text{logit}(\pi_{ij}^m) = x'_{1ij}\beta^m$  and  $\log(\lambda_{ij}^m) = x'_{2ij}\alpha^m$ . Next, a conditional specification follows:

$$P(Y_{ij} = y_{ij}|\theta_{ij}, b_i) = \begin{cases} \pi_{ij}^c + (1 - \pi_{ij}^c)f_i(0|\theta_{ij}, b_{1i}, \lambda_{ij}^c) & \text{if } y_{ij} = 0, \\ (1 - \pi_{ij}^c)f_i(y_{ij}|\theta_{ij}, b_{1i}, \lambda_{ij}^c) & \text{if } y_{ij} = 1, 2, \dots \end{cases}$$

where the probit  $\pi_{ij}^c = \Phi^{-1}(\Delta_{1ij} + z'_{1ij}b_{1i})$  and  $\lambda_{ij}^c = \theta_{ij}\exp(\Delta_{2ij} + z'_{2ij}b_{2i})$ . The overdispersion random effect,  $\theta_{ij} \sim \text{Gamma}(u_{ij}, v_{ij})$  is introduced in the Poisson model. For  $b_i = (b_{1i}, b_{2i})' \sim N(0, D)$  and based on

$$\lambda_{ij}^m = \int_b \int_\theta \theta_{ij} \exp(\Delta_{ij} + z'_{ij}b_i) dG_\theta dF_b = \int_b E(\theta_{ij}) \exp(\Delta_{ij} + z'_{ij}b_i) dF_b \quad (1)$$

where  $G_\theta(\cdot)$  and  $F_b(\cdot)$  are the cumulative distribution function of  $\theta_{ij}$  and  $b_i$  respectively, we derive:  $\Delta_{1ij} = \sqrt{1 + z'_{1ij}Dz'_{1ij}}\Phi^{-1}[\text{expit}(x'_{1ij}\beta^m)]$  and  $\Delta_{2ij} = -\log(u_{ij}v_{ij}) + x'_{2ij}\alpha^m - \frac{1}{2}z'_{2ij}Dz'_{2ij}$ . Thanks to the probit link, closed forms exist. The marginal mean still uses the logit, enabling an odds-ratio interpretation.

## 3 Estimation

We proceed via maximum likelihood. The observed data likelihood for subject  $i$ , conditional on the overdispersion random effect is:

$$f_i(\beta, \alpha, D, \phi) = \int_b \prod_{j=1}^{n_i} f(y_{ij}|b_i) f(b_i|D) db_i,$$

TABLE 1. *Epilepsy Trial. Parameter estimates (standard errors) for the marginalized models (bottom). RE: random effect.*

Effect	Par.	MMM	Zero-Inflated MMM	Combined MMM	Zero-Inflated Comb. MMM
		Est.(s.e.)	Est.(s.e.)	Est.(s.e.)	Est.(s.e.)
Poisson Part					
Interc. placebo	$\alpha_{00}$	1.396(0.189)	1.375(0.170)	1.476(0.196)	1.428(0.183)
Slope placebo	$\alpha_{01}$	-0.014(0.004)	-0.004(0.005)	-0.025(0.008)	-0.012(0.007)
Interc. treatment	$\alpha_{10}$	1.226(0.190)	1.378(0.172)	1.220(0.197)	1.337(0.186)
Slope treatment	$\alpha_{11}$	-0.012(0.004)	-0.007(0.005)	-0.019(0.008)	-0.005(0.007)
Slope diff.	$\alpha_{01} - \alpha_{11}$	0.002(0.006)	-0.003(0.007)	0.013(0.011)	0.008(0.010)
Std. Dev. RE	$\sigma_1$	1.076(0.086)	0.973(0.082)	1.063(0.087)	1.009(0.086)
Zero-Inflated Part					
Intercept	$\beta_0$		-2.296(0.296)		-2.428(0.321)
Slope	$\beta_1$		0.066(0.017)		0.066(0.018)
Std. Dev. of RE	$\sigma_2$		1.254(0.192)		1.292(0.208)
Overd. Par.	$v = \frac{1}{u}$			0.406(0.0348)	0.179(0.018)
Correlation	$\rho$		-0.138(0.1601)		-0.080(0.167)
AIC		-6810	-7222	-7664	-7682

from which the likelihood follows. The distribution of  $Y_i$  conditional on  $b_i$  and marginal over  $\theta_{ij}$  is given for the zero-inflated combined model by:

$$\begin{aligned}
 f(y_{ij}|b_i) = & I(y_{ij} = 0)\pi_{ij} + (1 - \pi_{ij}) \binom{u_j + y_{ij} - 1}{u_j - 1} \\
 & \times \left( \frac{v_j}{1 + \kappa_{ij}v_j} \right)^{y_{ij}} \left( \frac{1}{1 + \kappa_{ij}v_j} \right)^{u_j} \kappa_{ij}^{y_{ij}}.
 \end{aligned}$$

In fitting the MMM, the conditional distributions are specified by replacing the terms  $x'_{1ij}\beta$  and  $x'_{2ij}\alpha$  in the zero-inflated version of the combined model with the analytical expressions for  $\Delta_{1ij}$  and  $\Delta_{2ij}$ , respectively, as the mean models relate separately to these terms. Implementation is within SAS NLMIXED.

#### 4 Analysis of Epilepsy Data

A description of the data is provided in Molenberghs *et al* (2010). 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. Weekly seizure counts are available. We fit our model and several sub-models to the data. Denote the number of epileptic seizures for patient  $i$  at week  $j$  by  $Y_{ij}$  and the occasion on which  $Y_{ij}$  was measured by  $t_{ij}$ . Assuming that  $Y_{ij}$  follows a combined model with  $\lambda_{ij}^c = \theta_{ij}\kappa_{ij}$ , assume  $\theta_{ij} \sim \text{Gamma}(u, v)$ , and

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

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,  $b_{1i} \sim N(0, \sigma_1^2)$  in the Poisson model and  $b_{2i} \sim N(0, \sigma_2^2)$  in the binomial model and the correlation between the binomial and count components is represented by  $\rho$ .

Results of these models are presented in Table 1. Generally, the fixed-effect parameters are close to each other. Their interpretations are not just subject-specific but can be extended to the whole population. Use ‘CO’ for combined and ‘ZI’ for zero inflation. Comparing the MMM and ZIMMM to the COMMM and ZICOMMM models, we see improvement in the model fit owing to the gamma random effects. Also, model fit improves if the normal random effects are supplemented with zero-inflation. Therefore, it is key that the more complex model results in a considerably improvement in the model. This is essential for inferences and for prediction.

## 5 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. Through an empirical study, 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. A 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.

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

## References

- Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**, 688–698.
- Iddi, S. and Molenberghs, G. (2012). A combined overdispersed and marginalized multilevel model. *Computational Statistics and Data Analysis*, **56**, 1944–1951.
- 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.