

**Sample size calculations for hierarchical
Poisson and zero-inflated poisson
regression models**

N. Channouf, B. MacGibbon,
M. Fredette

G-2015-63

June 2015

Les textes publiés dans la série des rapports de recherche *Les Cahiers du GERAD* n'engagent que la responsabilité de leurs auteurs.

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2015.

The authors are exclusively responsible for the content of their research papers published in the series *Les Cahiers du GERAD*.

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2015.

Sample size calculations for hierarchical Poisson and zero-inflated poisson regression models

Nabil Channouf^a

Brenda MacGibbon^b

Marc Fredette^c

^a *Sultan Qaboos University, Muscat, Oman*

^b *GERAD & Université du Québec à Montréal, Montréal (Québec) Canada, H3C 3P8*

^c *GERAD & HEC Montréal, Montréal (Québec) Canada, H3T 2A7*

nabil.channouf@gmail.com
macgibbon.brenda@gmail.com
marc.fredette@hec.ca

June 2015

**Les Cahiers du GERAD
G–2015–63**

Copyright © 2015 GERAD

Abstract: In biomedical research there is a growing interest in the use of hierarchical Poisson regression models. Although sample size calculations for testing parameters in a Poisson regression model with pre-specified power and size have been previously done, very little attention has been paid to this problem for the hierarchical model. We propose to use Monte Carlo simulations to calculate the sample size necessary for the Wald test. The effect of the number of clusters and the covariance structure of the fixed effects is also studied in a simulation study. The method and the simulation study are also extended to the case of the hierarchical Zero-inflated Poisson regression model in order to obtain analogous results there.

Key Words: Wald test, hierarchical generalized linear models, Monte Carlo simulations.

Résumé : Dans la recherche biomédicale, il y a un intérêt croissant pour l'utilisation de modèles hiérarchiques de régression de Poisson. Nous proposons d'utiliser des simulations de Monte Carlo pour calculer la taille de l'échantillon nécessaire afin d'obtenir la puissance voulue avec le test de Wald. L'effet du nombre de grappes et de la structure de covariance des effets fixes est également étudié via simulations. Ensuite cette méthode sera adaptée pour les modèles de régression hiérarchiques pour les lois de Poisson gonflés à zéro.

1 Introduction

In the applied literature hierarchical Poisson regression models are used in interesting ways such as in the studies of clustered data of Christiansen and Morris (1997), Yau and Lee (2001) and Royle (2008). We also cite Hall (2000), Tu and Piegorsch (2003) and Roy et al. (2007). In biomedical research the clusters could be hospitals, and for the patients within hospitals there would usually be several explanatory covariates such as sex, age, etc. For such studies, calculation of the sample size needed for a test with pre-specified power and size is usually a necessary part of the design.

Sample size calculations for the standard non-hierarchical Poisson model are often based on the Wald test (Signorini, 1991; Shieh, 2001, 2005; Channouf et al., 2014) or the score test (Self and Mauritsen, 1988) or the likelihood ratio test (Self et al., 1992; Shieh, 2000).

Although statistical inferential methods for the hierarchical generalized linear model have been proposed by Lee and Nelder (1996, 2001), and Ha and Lee (2003), with few exceptions the problem of sample size determination has largely been ignored. The exceptions include, Liu and Liang (1997), who extended the method of Self and Mauritsen (1988) to the multivariate case with correlated data using the generalized estimating equation method and the quasi-score statistic (Liang and Zeger, 1986), and Matsui (2005), who used asymptotic approximations for a two-sample non-parametric test to develop a new formula for sample size calculations for clinical trials with Poisson or over-dispersed Poisson data.

We have chosen here to use the Wald test for sample size calculations for the hierarchical Poisson regression model, because it is frequently used due to its accessibility and its intuitiveness and because of recent major contributions in sample size calculations for this test for the standard non-hierarchical Poisson model (Shieh, 2001, 2005; Channouf et al., 2014).

Another problem with count data is that the observed data may have an excess number of zeros and that the Poisson model may not be an adequate model for the counts in such situations. The more complex zero-inflated Poisson (ZIP) model could be a better choice here. For the non-hierarchical model, Channouf et al. (2014) used the estimation approach of Lambert (1992), based on the complete-data likelihood function (Hall, 2000), to do sample size calculations for the ZIP model, while Williamson et al. (2007) used the incomplete log-likelihood function (equation (3) in Lambert (1992)) to do power calculations for ZIP and ZINB regression models. Hall (2000) presented a detailed study of the zero-inflated Poisson and binomial regression models and their extension to correlated data. The idea of using the zero-inflated Poisson model with correlated data in biomedical applications was first studied by Yau and Lee (2001), although sample size calculations were not considered.

Influenced by the original work of Shieh (2001) for the standard Poisson regression model when the covariates were assumed to have a finite support, the major contribution of this paper is the development of the methodology necessary to do sample size calculations using Monte Carlo simulations for the Wald test for different hierarchical Poisson regression models, and the extension of this methodology to the case of the hierarchical Zero-inflated Poisson model. It should be noted that our methodology is also used here for continuous covariates. Without loss of generality, we restrict our numerical examples to the case of equal cluster size. For parameter estimation we use the h -likelihood method proposed by Lee and Nelder (1996). We also study the effect of different correlation structures on the sample size. Our work is motivated by the fact that no existing method was proposed in the literature for sample size calculations for hierarchical Poisson regression models.

The remainder of the paper is organized as follows. In the next section, we introduce the hierarchical Poisson regression model and describe our use of the h -likelihood method and our approach to sample size calculations for this model. In Section 3, we present the extension of our approach to zero-inflated hierarchical Poisson regression models using Hall's (2000) estimation techniques. Section 4 contains numerical examples with an accompanying discussion of these results. Section 5 contains an illustrative data example. In Section 6, we draw some conclusions and discuss further directions of research related to this work.

2 Hierarchical Poisson regression model

The hierarchical Poisson regression model belongs to the class of hierarchical generalized linear models in which the distribution of the random effects is not restricted to be normal, but of general form. Often, this distribution is conjugate to that of the outcomes. For such a model having both random and fixed effects, we use a conditional approach as in Lee and Nelder (1996).

In general for the hierarchical Poisson model, we consider the case of clustered data, where each vector of responses $\mathbf{Y}_i$ denotes a cluster, and where the outcomes are $Y_{i,j}$, for $i = 1, \dots, n$; $j = 1, \dots, n_i$, with $N = \sum_{i=1}^n n_i$. Given a vector of unobserved random effects $\mathbf{U}_i = (U_{0,i}, \dots, U_{r,i})$ associated with the cluster $\mathbf{Y}_i$, the outcomes $Y_{i,1}, \dots, Y_{i,n_i}$ are mutually independent and have a Poisson distribution with mean $E[Y_{i,j}|\mathbf{u}_i] = \lambda_{i,j}$, where the linear predictor takes the form:

$$\log(\lambda_{i,j}) = \boldsymbol{\beta}^T \mathbf{X}_{i,j} + v(\mathbf{u}_i)^T \mathbf{Z}_{i,j}, \quad i = 1, \dots, n; \quad j = 1, \dots, n_i,$$

where $\mathbf{X}_{i,j}^T = (1, X_{1,i,j}, \dots, X_{p,i,j})$ represents the explanatory covariates associated with the fixed $(p+1)$ effects $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_p)$, and $\mathbf{Z}_{i,j}^T = (1, Z_{1,i,j}, \dots, Z_{r,i,j})$ the covariates associated with the $(r+1)$ random effects $\mathbf{v}_i = v(\mathbf{U}_i)$ for a strictly monotonic link function v . The random vectors $\mathbf{X}$ and $\mathbf{Z}$ are also assumed independent. The random variables $U_{0,i}, \dots, U_{r,i}$ are assumed independent and the distribution of each $U_{k,i}$, for $k = 0, \dots, r$, is assumed to belong to the exponential family and it is characterized by one dispersion parameter α_k .

The choice of the link function v is very important as well and affects the estimation of the parameters and the dispersion. Some special distributions for the responses $\mathbf{Y}_i$ and the random variables $U_{0,i}, \dots, U_{r,i}$, for $i = 1, \dots, n$, involve conjugate models. For such models, in addition to the joint likelihood function, as defined by Lee and Nelder (1996, p. 621), to estimate the random effect and the parameters of the model, an explicit form of the marginal likelihood can be obtained as well without any integration.

Our numerical examples include the usual conjugate hierarchical Poisson-Gamma model with different covariate distributions for the explanatory covariates. Analogous results could be obtained for other conjugate hierarchical models in the generalized linear exponential family, but they will not be discussed here.

2.1 Hierarchical likelihood estimation

The hierarchical likelihood method, called h -likelihood, was introduced by Lee and Nelder (1996), and it is based on the joint distribution of $\mathbf{Y}$ and $\mathbf{U}$. It should be noted that Kuk and Cheng (1999) gave a comparative study of different algorithms for maximum likelihood estimates in random effects models. They argued that the h -likelihood can be inappropriate, especially with binary data. Lee et al. (2007) gave a detailed study of the h -likelihood and effectively responded to the criticism of Kuk and Cheng (1999).

For hierarchical generalized linear models like the hierarchical Poisson model, both fixed effects and random effects are of interest, so they are estimated simultaneously. In general, for models where the joint distribution is used, some difficulties arise and complex integrals need to be evaluated, especially for high dimensions. Different methods have been presented in order to circumvent these analytical difficulties by using approximations, such as penalized quasi-likelihood estimation or restricted maximum likelihood estimation, where Laplace transforms are used to approximate the integrals. For more details about likelihood-based inference in generalized linear mixed models, see Song (2007, chapter 7). The h -likelihood estimation method can be regarded as an extension of penalized quasi-likelihood estimation, since it does not involve any integration (Jiang, 2007, chapter 3).

Assume we have a random sample from the joint distribution of $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$ with the conditional p.d.f. $f(\mathbf{Y}, \mathbf{X}, \mathbf{Z}|\mathbf{U}) = f(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \mathbf{U})f(\mathbf{X})f(\mathbf{Z})$, since $\mathbf{X}$ and $\mathbf{Z}$ are independent and do not depend on $\mathbf{U}$ or any of the unknown parameters of the model $(\beta_0, \beta_1, \dots, \beta_p, \alpha_0, \dots, \alpha_r)$. The joint log-likelihood function for the hierarchical Poisson model, called the h -likelihood, associated with the data, the parameters $\boldsymbol{\beta}$ and the random effects $\mathbf{U}$, is given by:

$$h(\boldsymbol{\beta}; \mathbf{y}, \mathbf{x}, \mathbf{z}, \mathbf{v}) = \log L(\boldsymbol{\beta}; \mathbf{y}, \mathbf{x}, \mathbf{z}, \mathbf{v}) = l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{x}, \mathbf{z}|\mathbf{v}) + l(\alpha_0, \dots, \alpha_r; \mathbf{v}) \quad (1)$$

where $l(\beta; \mathbf{y}, \mathbf{x}, \mathbf{z}|\mathbf{v}) = \sum_{i=1}^n \sum_{j=1}^{n_{i,j}} \log f_{\mathbf{Y}}(y_{i,j}|x_{i,j}, z_{i,j}, v_i)$, with $f_{\mathbf{Y}}$ having the p.d.f. of a Poisson distribution $\mathbf{Y}$ (conditional on $\mathbf{v}$) and $l(\alpha_0, \dots, \alpha_r; v)$ representing the log-likelihood of $\mathbf{v}$ with parameters $\alpha_0, \dots, \alpha_r$. Note that we can use both the density of $\mathbf{Y}|v(\mathbf{U})$ or $\mathbf{Y}|\mathbf{U}$, since v is a strictly monotonic function of $\mathbf{U}$. We also note that in the estimation of the parameters in the hierarchical generalized linear model, these parameters are conditioned on the random effects, which are not observed. However, they can be estimated jointly with the fixed effects.

The score function for this estimation problem of dimension $(p+q+2)$ is defined as follows

$$S_N(\beta; \mathbf{y}, \mathbf{x}, \mathbf{z}, \mathbf{v}) = \begin{cases} \frac{\partial h(\beta; \mathbf{y}, \mathbf{x}, \mathbf{z}, \mathbf{v})}{\partial \beta} = \frac{\partial l(\beta; \mathbf{y}, \mathbf{x}, \mathbf{z}|\mathbf{v})}{\partial \beta} \\ \frac{\partial h(\beta; \mathbf{y}, \mathbf{x}, \mathbf{z}, \mathbf{v})}{\partial \mathbf{v}} = \frac{\partial l(\beta; \mathbf{y}, \mathbf{x}, \mathbf{z}|\mathbf{v})}{\partial \mathbf{v}} + \frac{\partial l(\alpha_0, \dots, \alpha_r; \mathbf{v})}{\partial \mathbf{v}}. \end{cases} \quad (2)$$

The maximum h -likelihood estimates of β and $\mathbf{v}$ are then obtained by solving the system of equations $S_N(\beta; \mathbf{y}, \mathbf{x}, \mathbf{z}, \mathbf{v}) = 0$. For the estimation of the dispersion parameters $\alpha_0, \dots, \alpha_r$, a second stage procedure is used on the estimators of the fixed and random effects (Lee and Nelder, 1996).

2.2 Hierarchical Poisson-Gamma model with one random effect

Here we consider the special case of the hierarchical Poisson model with its conjugate gamma distribution, with the logarithm as the link function and with only one (intercept) random factor $\mathbf{U}_0$ for each cluster. This type of model is often called “the random-intercept model”.

For each group (cluster) $i = 1, \dots, n$, we have outcomes $Y_{i,1}, \dots, Y_{i,n_i}$, a vector of fixed effects $(\beta_0, \beta_1, \dots, \beta_p)^T$ associated with $X_{i,j}^T = (1, X_{1,i,j}, \dots, X_{p,i,j})$, and one random effect $\mathbf{U}_0$, where $U_{0,1}, \dots, U_{0,n}$ are i.i.d. having a gamma distribution with a common dispersion parameter α_0 and satisfying $E[U_{0,i}] = 1$. We have $Y_{i,j}|u_{0,i} \sim \text{Poisson}(\lambda_{i,j})$, where the conditional mean $E[Y_{i,j}|u_{0,i}] = \lambda_{i,j}$ satisfies:

$$\log(\lambda_{i,j}) = \beta^T \mathbf{X}_{i,j} + v_{0,i}, \quad (3)$$

for $i = 1, \dots, n$, and $j = 1, \dots, n_i$, with $v_{0,i} = \log(u_{0,i})$. From (1), the h -likelihood function associated with the parameters $\beta^T = (\beta_0, \beta_1, \dots, \beta_p)$ and the unobserved random effects $v_{0,i}$, for a random sample $(y_{i,j}, x_{1,i,j}, \dots, x_{p,i,j})$, is given by:

$$\begin{aligned} h(\beta; \mathbf{y}, \mathbf{x}, \mathbf{v}_0) &= l(\beta; \mathbf{y}, \mathbf{x}|\mathbf{v}_0) + l(\alpha_0; \mathbf{v}_0) \\ &= \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ y_{i,j}(\beta^T x_{i,j} + v_{0,i}) - e^{\beta^T x_{i,j} + v_{0,i}} - \log(y_{i,j}!) + \log f_X(x_{i,j}) \right\} \\ &\quad + \sum_{i=1}^n \{ \alpha_0 \log(\alpha_0) + \alpha_0 v_{0,i} - \alpha_0 e^{v_{0,i}} - \log \Gamma(\alpha_0) \}. \end{aligned} \quad (4)$$

Once α_0 is specified or fixed in the first step as in Lee and Nelder (1996), the estimates of the random effects can be easily obtained from the score function and the second equation of (2). We have:

$$\hat{u}_{0,i} = e^{\hat{v}_{0,i}} = \frac{y_{i,+} + \alpha_0}{\lambda_{i,+} + \alpha_0}, \quad (5)$$

where $y_{i,+} = \sum_{j=1}^{n_i} y_{i,j}$ and $\lambda_{i,+} = \sum_{j=1}^{n_i} \lambda_{i,j}$, with $\lambda_{i,j} = E[Y_{i,j}]$ equal to the unconditional mean of $Y_{i,j}$. It should be noted that $E[U_{0,i}] = 1$ for any given α_0 , and when $\alpha_0 \rightarrow \infty$, we recover the Poisson regression model with fixed effects.

2.3 Sample size determination

We want to find the minimum required sample size in order to achieve a given power with the Wald test for the hypothesis $H_0 : \beta_s = 0$, for any given $s \in \{0, 1, \dots, p\}$. To do this, we generalize the approach used by Signorini (1991) and Shieh (2001) for non-hierarchical models. In our case, we also prefer the conditional expectation, conditional on the covariates $\mathbf{X}$, using Monte Carlo simulation to generate samples

from these covariates instead of the full expectation as used by Shieh (2001). Shieh (2001) chose to solve $\lim_{N \rightarrow \infty} \mathbf{E}[S_N(\beta_0, \beta_1, \dots, \beta_{s-1}, 0, \beta_{s+1}, \dots, \beta_p)]/N$, where the full expectation is always taken with respect to the true value of β . In his case, this expectation can be calculated because it is always expressed as a finite sum since the support of the covariates $\mathbf{X}$ is treated as finite. Even for continuous random variables such as the normal distribution, he evaluates it at a finite number of points. We, however, choose to avoid using such a finite approximation in the more complex case of the hierarchical Poisson regression model. In equation (2) we choose $N = K$ sufficiently large as a first step and calculate $S_N(\beta) = S_K(\beta)$ directly by substituting the values of the covariates from our Monte Carlo simulations and then use K as the subscript for the score function. We first calculate the estimates of the parameters $(\beta_0^*, \beta_1^*, \dots, \beta_{s-1}^*, 0, \beta_{s+1}^*, \dots, \beta_p^*)$ and $\hat{\mathbf{v}}_0^T = (\hat{v}_{0,1}^*, \dots, \hat{v}_{0,n}^*)$ under H_0 , by solving the following system of equations:

$$\mathbf{E}[S_K(\beta_0, \beta_1, \dots, \beta_{s-1}, 0, \beta_{s+1}, \dots, \beta_p, \mathbf{v}_0) | \mathbf{X} = \mathbf{x}] = 0, \quad (6)$$

where the conditional expectation (given $\mathbf{X}$) is taken with respect to the true values of $\beta_0, \dots, \beta_p$. The score function is given by:

$$S_K(\beta, \mathbf{v}_0) |_{\beta_s=0} = \begin{cases} \frac{\partial h(\beta; \mathbf{y}, \mathbf{x}, \mathbf{v}_0)}{\partial \beta_0} = \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{i,j} - e^{\beta^T x_{i,j} + v_{0,i}}), \\ \frac{\partial h(\beta; \mathbf{y}, \mathbf{x}, \mathbf{v}_0)}{\partial \beta_l} = \sum_{i=1}^n \sum_{j=1}^{n_i} x_{l,i,j} (y_{i,j} - e^{\beta^T x_{i,j} + v_{0,i}}); \quad l = 1, \dots, p. \\ \frac{\partial h(\beta; \mathbf{y}, \mathbf{x}, \mathbf{v}_0)}{\partial v_{0,i}} = \alpha_0 - \alpha_0 e^{v_{0,i}} + \sum_{j=1}^{n_i} (y_{i,j} - e^{\beta^T x_{i,j} + v_{0,i}}); \quad i = 1, \dots, n. \end{cases}$$

In order to estimate the variance-covariance matrix of the maximum h -likelihood estimates of β and v_0 , we use the Fisher information matrix by considering

$$V(\beta, \mathbf{v}_0) = I(\beta, \mathbf{v}_0)^{-1}, \quad (7)$$

where

$$I(\beta, \mathbf{v}_0) = -\frac{1}{K} \begin{pmatrix} \mathbf{E} \left[\frac{\partial^2 h(\beta; \mathbf{y}, \mathbf{x}, \mathbf{v}_0)}{\partial \beta \partial \beta^T} \right] & \mathbf{E} \left[\frac{\partial^2 h(\beta; \mathbf{y}, \mathbf{x}, \mathbf{v}_0)}{\partial \beta \partial \mathbf{v}_0^T} \right] \\ \mathbf{E} \left[\frac{\partial^2 h(\beta; \mathbf{y}, \mathbf{x}, \mathbf{v}_0)}{\partial \beta \partial \mathbf{v}_0^T} \right]^T & \mathbf{E} \left[\frac{\partial^2 h(\beta; \mathbf{y}, \mathbf{x}, \mathbf{v}_0)}{\partial v_0 \partial \mathbf{v}_0^T} \right] \end{pmatrix}, \quad (8)$$

and where the asymptotic normality of the maximum h -likelihood estimates of β and $\mathbf{v}_0$ holds as the $n_i \rightarrow \infty$, at the same rate, whereas the number of clusters n remains constant (Lee and Nelder, 1996). Xia et al. (2006) also studied the asymptotic behavior of these estimators.

Without loss of generality, let us assume that the test concerns the parameter β_1 , and we test $H_0 : \beta_1 = 0$, against the alternative $H_1 : \beta_1 > 0$, for a given power. Since asymptotic normality of the estimates of the β parameters holds, we can use equations (7) and (8) to calculate the necessary variances in the formula for estimating the minimum sample size N_s , for $s = 1$, proposed by Signorini (1991) and Shieh (2001), is given by:

$$N_s = \left(\frac{V(\beta_0^*, 0, \beta_2^*, \dots, \beta_p^*, \mathbf{v}_0^*)^{1/2} Z_\alpha + V(\beta_0, \beta_1, \dots, \beta_p, \mathbf{v}_0)^{1/2} Z_{1-\text{Power}}}{\beta_1} \right)^2, \quad (9)$$

where Z_q is the $100(1 - q)$ th percentile of the standard normal distribution, α is the size of the test, $V(\beta_0, \beta_1, \dots, \beta_p, \mathbf{v}_0)$ is computed with the pre-specified values and $V(\beta_0^*, 0, \beta_2^*, \dots, \beta_p^*, \mathbf{v}_0^*)$ is computed with the values obtained in (6). We have modified Shieh's method here in (9) by using α instead of $\alpha/2$ as was originally intended by Signorini (1991). It should be noted that here the pre-specified value of the parameter β_1 represents the difference we want to detect by the test for a given power and size. The detailed calculations of the sample size for a significance test with specified power and size on any parameter β_s , $s = 1, \dots, p$, are described in Algorithm 1.

3 Extension to the hierarchical zero-inflated Poisson regression model

For this type of model, we modify the methods of Hall (2000) who studied data having excess zeros using the generalized linear mixed model. We adapt his inference methodology for the regression parameters to the

Algorithm 1 Sample size calculation for the hierarchical Poisson-gamma model with one random effect (intercept).

The algorithm has the following input:

- The desired *Power*.
- The significance level α .
- The value of the difference Δ that is to be detected. This difference is set equal to e^{β_s} , where β_s is the parameter being tested.
- The distribution of the covariates $\mathbf{X}$.
- The parameters of U_0 are chosen so that $\mathbb{E}[U_0] = 1$; $U_0 \sim \Gamma(\alpha_0, 1/\alpha_0)$.
- Some reasonable values for the other β_k , $k = 1, \dots, p$, also have to be specified. (These could come from previous studies, a pilot study or an educated guess). The value of β_0 is chosen to satisfy an overall fixed mean $\bar{\lambda} = E[e^{\beta^T \mathbf{X}}]$.
- An initial number K of simulated sets of covariates. Empirical studies suggest that the value of K has little effect on the results and that $K = 100$ suffices.
- The number of clusters n is specified.
- The desired number of Monte Carlo replications B .

For normal covariates, the Quasi Monte Carlo method is used for integration, whereas analytical formula can be obtained for other distributions (e.g. Bernoulli or multinomial).

The algorithm then returns the minimum sample size needed for the significance test. It has the following steps:

1. **Generate** a sample of K values from the distribution of $\mathbf{X}$ and compute the rates $\lambda_i(X)$, $i = 1, \dots, K$, under H_1 .
 2. **Compute** the values $\beta^* = (\beta_0^*, \beta_1^*, \dots, \beta_{s-1}^*, 0, \beta_{s+1}^*, \dots, \beta_p^*)$, and $u_{0,1}, \dots, u_{0,n}$, under H_0 , by solving the system of nonlinear equations in (6) with $N = K$, using the relation in (5);
 3. **Compute** the Fisher information matrix using (8) to obtain the variance functions $V(\beta^*, v_0^*)$ and $V(\beta, v_0)$ in (7) ($v_0 = \log(u_0)$), and then take its $(s+1)$ -th diagonal elements;
 4. **Repeat** steps (1) to (3) B times and each time return the sample size $N_{s,j}$ obtained from equation (9), for $j = 1, \dots, B$.
 5. **Take** the average value over the outputs of the B replications, $N_s = \frac{\sum_{j=1}^B N_{s,j}}{B}$.
-

hierarchical zero-inflated Poisson model by using what he called the complete-data log-likelihood with one random intercept. However, when calculating the log-likelihood function, we consider the gamma distribution for the intercept instead of the normal one used by Hall (2000).

We again consider responses and covariates that come from clustered data. We assume that for the random effects $\mathbf{U}_i = (U_{0,i}, \dots, U_{r,i})$, and link functions $v_i = v(U_i)$, the outcomes $Y_{i,j}$, for $i = 1, \dots, n$, and $j = 1, \dots, n_i$, have the following conditional probability density function:

$$P(Y_{i,j} = y_{i,j} | u_i) = \begin{cases} \pi_{i,j} + (1 - \pi_{i,j})e^{-\lambda_{i,j}} & \text{if } y_{i,j} = 0 \\ (1 - \pi_{i,j}) \frac{e^{-\lambda_{i,j}} \lambda_{i,j}^{y_{i,j}}}{y_{i,j}!} & \text{if } y_{i,j} > 0, \end{cases}$$

with $E[Y_{i,j} | u_i] = (1 - \pi_{i,j})\lambda_{i,j}$ and $\text{Var}(Y_{i,j} | u_i) = (1 - \pi_{i,j})\lambda_{i,j}(1 + \pi_{i,j}\lambda_{i,j})$. In this model, the zeros for the outcomes can come from two states, the zero state with probability $\pi_{i,j}$, or the Poisson state with a probability $(1 - \pi_{i,j})e^{-\lambda_{i,j}}$. In other words, $Y_{i,j}$ equals zero with probability $\pi_{i,j}$ or follows a Poisson distribution with probability $(1 - \pi_{i,j})$. Both sets of parameters $\lambda_{i,j}$ and $\pi_{i,j}$ are modeled by link functions in order to have linear predictors as follows:

$$\text{logit}(\pi_{i,j}) = \gamma^T \mathbf{G}_{i,j} = \gamma_0 + \gamma_1 G_{1,i,j} + \dots + \gamma_q G_{q,i,j},$$

for $i = 1, \dots, n$, and $j = 1, \dots, n_i$, with $\mathbf{G}_{i,j}^T = (1, G_{1,i,j}, \dots, G_{q,i,j})$ the $(q+1)$ elements of the covariates and $\gamma^T = (\gamma_0, \gamma_1, \dots, \gamma_q)$, a vector of $(q+1)$ parameters. For the counts, we have

$$\begin{aligned} \log(\lambda_{i,j}) &= \beta^T \mathbf{X}_{i,j} + v_i^T \mathbf{Z}_{i,j} \\ &= \beta_0 + \beta_1 X_{1,i,j} + \dots + \beta_p X_{p,i,j} + v_{0,i} + v_{1,i} Z_{1,i,j} + \dots + v_{r,i} Z_{r,i,j}, \end{aligned}$$

with $\mathbf{X}_{i,j}^T = (1, X_{1,i,j}, \dots, X_{p,i,j})$, the $(p+1)$ elements of the covariates, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_p)$, a vector of $(p+1)$ parameters, $\mathbf{Z}_{i,j}^T = (1, Z_{1,i,j}, \dots, Z_{r,i,j})$, the elements of the $(r+1)$ covariates associated to the random effects $\mathbf{v}_i^T = (v_{0,i}, v_{1,i}, \dots, v_{r,i}) = (\log u_{0,i}, \log u_{1,i}, \dots, \log u_{r,i})$.

For a hierarchical ZIP model with one random gamma effect $\mathbf{U}_0$ (“the random-intercept model”), the joint log-likelihood combines the logarithm of the density of the outcomes conditional on v_0 with parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, and the logarithm of the density function of v_0 with parameter α_0 . We have:

$$h(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x}, \mathbf{v}_0) = l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x} | \mathbf{v}_0) + l(\alpha_0; \mathbf{v}_0),$$

where $l(\alpha_0; \mathbf{v}_0)$ is computed as for the hierarchical Poisson regression model given in (4), and $l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x} | \mathbf{v}_0)$ is calculated following Hall (2000). Then for the h -likelihood, we obtain:

$$\begin{aligned} h(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x}, \mathbf{v}_0) &= \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ T_{i,j} \boldsymbol{\gamma}^T \mathbf{g}_{i,j} - \log(1 + e^{\boldsymbol{\gamma}^T \mathbf{g}_{i,j}}) \right\} \\ &+ \sum_{i=1}^n \sum_{j=1}^{n_i} (1 - T_{i,j}) \left\{ y_{i,j} (\boldsymbol{\beta}^T \mathbf{x}_{i,j} + v_{0,i}) - e^{\boldsymbol{\beta}^T \mathbf{x}_{i,j} + v_{0,i}} - \log(y_{i,j}!) \right\} \\ &+ \sum_{i=1}^n \left\{ \alpha_0 \log(\alpha_0) + \alpha_0 v_{0,i} - \alpha_0 e^{v_{0,i}} - \log \Gamma(\alpha_0) \right\}, \end{aligned} \quad (10)$$

with

$$T_{i,j} = \begin{cases} 1 & \text{if } Y_{i,j} \text{ comes from the zero state,} \\ 0 & \text{if } Y_{i,j} \text{ comes from the Poisson state,} \end{cases}$$

for $i = 1, \dots, n$, and $j = 1, \dots, n_i$.

The sample size determination is then analogous to the methodology for the hierarchical Poisson regression model, since we use the same formula for the minimum sample size N_s , given by (9). However, the computations become more complex, since we have additional parameters for the logistic component, namely $\boldsymbol{\gamma}^T = (\gamma_0, \gamma_1, \dots, \gamma_q)$.

To find the minimum required sample size for the Wald test, we have the following score function:

$$S_N(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{v}_0) = \begin{cases} \frac{\partial l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x}, \mathbf{v}_0)}{\partial \gamma_0} = \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ T_{i,j} - \frac{e^{\boldsymbol{\gamma}^T \mathbf{g}_{i,j}}}{1 + e^{\boldsymbol{\gamma}^T \mathbf{g}_{i,j}}} \right\} \\ \frac{\partial l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x}, \mathbf{v}_0)}{\partial \gamma_l} = \sum_{i=1}^n \sum_{j=1}^{n_i} g_{l,i,j} \left\{ T_{i,j} - \frac{e^{\boldsymbol{\gamma}^T \mathbf{g}_{i,j}}}{1 + e^{\boldsymbol{\gamma}^T \mathbf{g}_{i,j}}} \right\}; \quad l = 1, \dots, q. \\ \frac{\partial l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x}, \mathbf{v}_0)}{\partial \beta_0} = \sum_{i=1}^n \sum_{j=1}^{n_i} (1 - T_{i,j}) \left\{ y_{i,j} - e^{\boldsymbol{\beta}^T \mathbf{x}_{i,j}} \right\}, \\ \frac{\partial l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x}, \mathbf{v}_0)}{\partial \beta_k} = \sum_{i=1}^n \sum_{j=1}^{n_i} (1 - T_{i,j}) \left\{ x_{k,i,j} (y_{i,j} - e^{\boldsymbol{\beta}^T \mathbf{x}_{i,j}}) \right\}; \quad k = 1, \dots, p. \\ \frac{\partial l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x}, \mathbf{v}_0)}{\partial v_{0,i}} = \alpha_0 - \alpha_0 e^{v_{0,i}} + \sum_{j=1}^{n_i} (1 - T_{i,j}) \left\{ y_{i,j} - e^{\boldsymbol{\beta}^T \mathbf{x}_{i,j}} \right\}; \quad i = 1, \dots, n. \end{cases}$$

The maximum likelihood estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are given by the solution of the system $S_N(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{v}_0) = 0$, and their variance-covariance matrix $V(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{v}_0)$ is estimated by the inverse of the Fisher information matrix $I(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{v}_0)$, given by:

$$I(\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{v}_0) = -\frac{1}{N} \begin{pmatrix} \mathbf{E} \left[\frac{\partial^2 l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x}, \mathbf{v}_0)}{\partial(\boldsymbol{\gamma}, \boldsymbol{\beta})^T \partial(\boldsymbol{\gamma}, \boldsymbol{\beta})} \right] & \mathbf{E} \left[\frac{\partial^2 l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x}, \mathbf{v}_0)}{\partial(\boldsymbol{\gamma}, \boldsymbol{\beta})^T \partial \mathbf{v}_0^T} \right] \\ \mathbf{E} \left[\frac{\partial^2 l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x}, \mathbf{v}_0)}{\partial(\boldsymbol{\gamma}, \boldsymbol{\beta})^T \partial \mathbf{v}_0^T} \right]^T & \mathbf{E} \left[\frac{\partial^2 l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{g}, \mathbf{x}, \mathbf{v}_0)}{\partial \mathbf{v}_0 \partial \boldsymbol{\beta}^T} \right] \end{pmatrix}. \quad (11)$$

Without loss of generality, let us assume that the test concerns the parameter β_1 . We can now calculate the minimum required sample size with the Wald test for the hypothesis $H_0 : \beta_1 = 0$, against $H_1 : \beta_1 > 0$. Analogous to the Poisson regression model, the estimates of the parameters $\boldsymbol{\gamma}^{*T} = (\gamma_0^*, \dots, \gamma_q^*)$ and $(\beta_0^*, 0, \beta_2^*, \dots, \beta_p^*)$ under H_0 , are calculated by solving the modified system of equations:

$$\mathbf{E}[S_N(\gamma_0, \dots, \gamma_q, \beta_0, 0, \beta_2, \dots, \beta_p; \mathbf{v}_0) | \mathbf{X} = \mathbf{x}] = 0, \quad (12)$$

where the conditional expectation is taken with respect to the true values of γ and β , and the $T_{i,j}$ are replaced by their conditional means obtained from the EM algorithm used for maximum likelihood estimation of parameters in ZIP models with random effects (Hall, 2000). We have

$$\begin{aligned}
 E[T_{i,j}|y_{i,j}, \gamma, \beta] &= P(\text{zero state}|y_{i,j}, \gamma, \beta) \\
 &= \frac{P(Y_{i,j}|\text{zero state})P(\text{zero state})}{P(Y_{i,j}|\text{zero state})P(\text{zero state}) + P(Y_{i,j}|\text{Poisson state})P(\text{Poisson state})} \\
 &= \begin{cases} (1 + e^{-G_{i,j}\gamma - e^{\beta^T \mathbf{x}_{i,j} + v_{0,i}}})^{-1} & \text{if } y_{i,j} = 0 \\ 0 & \text{if } y_{i,j} > 0, \end{cases} \quad (13)
 \end{aligned}$$

for $i = 1, \dots, n$, and $j = 1, \dots, n_i$.

As Hall (2000) has shown that the maximum likelihood estimates of β are asymptotically normal, the analogue of formula (9) can be used to estimate the minimum sample size N_s , which is given by:

$$N_s = \left\lceil \left(\frac{V(\gamma^*, \beta_0^*, 0, \beta_2^*, \dots, \beta_p^*, \mathbf{v}_0^*)^{1/2}_{q+3, q+3} Z_\alpha + V(\gamma, \beta, \mathbf{v}_0)^{1/2}_{q+3, q+3} Z_{1-Power}}{\beta_1} \right)^2 \right\rceil, \quad (14)$$

where the variance $V(\gamma, \beta, \mathbf{v}_0)$ is computed with the pre-specified values of the parameters and $V(\gamma^*, \beta_0^*, 0, \beta_2^*, \dots, \beta_p^*, \mathbf{v}_0^*)$ is calculated using the values obtained in (12). All the steps for the calculations of the sample size for a significance test with specified power and size on any parameter β_s , $s = 1, \dots, p$, are described in Algorithm 2.

Algorithm 2 Sample size calculation for the hierarchical ZIP-gamma model with one random effect (intercept).

The algorithm has the following input:

- The desired *Power*.
- The significance level α .
- The value of the difference Δ that is to be detected. This difference is set equal to e^{β_s} , where β_s is the parameter being tested.
- The distribution of the covariates $\mathbf{X}$ and $\mathbf{G}$.
- The parameters of U_0 are chosen so that $\mathbb{E}[U_0] = 1$; $U_0 \sim \Gamma(\alpha_0, 1/\alpha_0)$.
- Some reasonable values for the other β_k , $k = 1, \dots, p$, and γ_l , $l = 1, \dots, q$, also have to be specified. (These could come from previous studies, a pilot study or an educated guess). The values of β_0 and γ_0 are chosen to satisfy an overall fixed means $\bar{\lambda} = E[e^{\beta^T \mathbf{x}}]$ and $\bar{\pi} = E[e^{\gamma^T \mathbf{G}} / (1 + e^{\gamma^T \mathbf{G}})]$, respectively.
- An initial number K of simulated sets of covariates. Empirical studies suggest that the value of K has little effect on the results and that $K = 100$ suffices.
- The number of clusters n is specified.
- The desired number of Monte Carlo replications B .

For normal covariates, the Quasi Monte Carlo method is used for integration, whereas analytic formula can be obtained for other distributions (e.g. Bernoulli or multinomial).

The algorithm then returns the minimum sample size needed for the significance test. It has the following steps:

1. **Generate** a sample of K values from the distributions of $\mathbf{X}$ and $\mathbf{G}$, and compute the rates $\lambda_i(X)$, and $\pi_i(G)$, for $i = 1, \dots, K$, under H_1 .
 2. **Compute** the values $\gamma^* = (\gamma_0^*, \gamma_1^*, \dots, \gamma_q^*)$ and $\beta^* = (\beta_0^*, \beta_1^*, \dots, \beta_{s-1}^*, 0, \beta_{s+1}^*, \dots, \beta_p^*)$, and $u_{0,1}, \dots, u_{0,n}$, under H_0 , by solving the system of nonlinear equations in (12) with $N = K$, and using the relation in (13);
 3. **Compute** the Fisher information matrix using (11) to obtain the variance functions $V(\gamma^*, \beta^*, v_0^*)$ and $V(\gamma, \beta, v_0)$ ($v_0 = \log(u_0)$), and then take its $(q+s+2)$ -th diagonal elements;
 4. **Repeat** steps (1) to (3) B times and each time return the sample size $N_{s,j}$ obtained from equation (14), for $j = 1, \dots, B$.
 5. **Take** the average value over the outputs of the B replications, $N_s = \frac{\sum_{j=1}^B N_{s,j}}{B}$.
-

4 Numerical examples

All our simulation studies are illustrated using the conjugate Poisson-gamma model and clusters of equal size with different distributions for the covariates. The required sample sizes for our numerical examples are computed through Monte Carlo simulations. For each replication the values of the covariates are generated from the specified distributions, given we use the conditional expectation of the score function (equations (6) and (12)). We consider first an example with a single covariate and then we extend this to the case of two covariates. It should also be noted that here the pre-specified value of β_1 also represents the difference we want to detect by the test for a given power and size. We consider the cases of $X \sim \text{Bernoulli}(p)$ with $p = 0.1, 0.5, 0.9$, and $X \sim N(\mu, 1)$ with $\mu = 0, 0.5$. The case of two covariates, (X_1, X_2) , corresponding to the values $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$, is considered as a multinomial with probabilities $(p_1, p_2, p_3, p_4 = 1 - p_1 - p_2 - p_3)$. The three following cases from Shieh (2001) are considered: $(p_1, p_2, p_3, p_4) = (0.25, 0.25, 0.25, 0.25)$, $(p_1, p_2, p_3, p_4) = (0.4, 0.1, 0.1, 0.4)$, and the extreme case $(p_1, p_2, p_3, p_4) = (0.76, 0.19, 0.01, 0.04)$. The “true” parameters are chosen as follows (Shieh, 2001): β_0 is chosen to satisfy $\bar{\lambda} = E[e^{\beta^T \mathbf{X}}] = 0.05$, $\beta_1 = \beta_2 = \log(2)$, and we arbitrarily take $\alpha_0 = 2$. We also consider equal size clusters, i.e. $n_i = N/n$, for $i = 1, \dots, n$.

For the hierarchical ZIP model, we use only one covariate for the logit part of the model and the same types of covariate distributions for the counts as were used with the hierarchical Poisson model. We also add the case of two covariates having the bivariate normal distribution $N(0, \Sigma_\rho)$, with $\rho = -0.5, 0, 0.5$, where $\Sigma_\rho = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ is the covariance matrix. The “true” parameters of the logit model are taken as follows: γ_0 is chosen to satisfy $\bar{\pi} = E[\frac{e^{\gamma^T \mathbf{G}}}{1 + e^{\gamma^T \mathbf{G}}}] = 0.05$ and $\gamma_1 = \log 2$.

The estimates of the sample size N_s for testing $H_0 : \beta_1 = 0$, against $H_1 : \beta_1 > 0$, are based on independent Monte Carlo simulations with 5000 replications, using the software R. The significance level is taken to be $\alpha = 0.05$, and the powers considered are 80% and 90%. We also calculated the intraclass correlation (ICC) as defined by Commenges and Jacqmin (1994) for the various hierarchical Poisson models defined in Tables 1 to 6. They defined the ICC as the ratio of the “variation within classes” to the “total variation”. However, each of these models considered here satisfies the properties of the general intraclass correlation random effects model as defined by Commenges and Jacqmin (1994). Thus, the

$$\text{ICC} = \text{corr}(Y_{i,j}, Y_{i,k}),$$

which only depends on i . It should be noted that our chosen $\alpha_0 = 2$ yields intraclass correlations (ICC) varying between 1% and 2.5%.

The second and third columns of each table contain the initial number of clusters and the cluster size, respectively, where the total number of observations generated from the original distributions of $\mathbf{X}$ (and $\mathbf{G}$) is fixed at $K = 100$ for all the cases. This is done since simulations showed almost no change if we used $K = 1000$ or $K = 100$. The first case corresponding to a unique cluster is an extreme case where all the outcomes are correlated and we expect to have an extremely large sample size here. As the number of clusters increases, the sample sizes obtained converge toward the “independent” case of a non-hierarchical Poisson regression model and can thus be compared with the results of Shieh (2001) and Channouf et al. (2014). It is important to note that, to the best of our knowledge, no other existing studies with numerical results are presented in the literature for hierarchical Poisson regression models with more than two clusters, so we are unable to present a comparative study with the results of others.

In Table 1, we present examples with one covariate X having a binomial distribution with $p = 0.1, 0.5, 0.9$. In all cases the sample size is fairly large when we have one cluster, especially when $p = 0.9$, and then the sample size decreases when the number of individuals within a cluster decreases. This indicates the importance of using the hierarchical structure even when the ICC is very close to 0. Changing power has an important impact on the cluster size. In general, between 27% and 47% more individuals are needed when the power increases from 80% to 90%, and the largest difference is observed when $p = 0.9$. When the covariate X follows a normal distribution (Table 2), we observe the same tendency when the number of clusters increases. Increasing power from 80% to 90%, the sample size can increase by an amount between 28% and 51%, and it is more pronounced when the observations are sampled from a normal distribution with mean 0.5. For the numerical example with two covariates having the multinomial distribution, the results

Table 1: Results for the hierarchical Poisson regression model for testing $H_0 : \beta_1 = 0$, against $H_1 : \beta_1 \neq 0$, with significance level 5%. One covariate X has the Bernoulli distribution.

	Number of clusters	N_s	
		Power=0.9	Power=0.8
$Y u_0 \sim \text{Pois}(\lambda)$	1	5585	4535
$\lambda = e^{\beta_0 + \beta_1 X + \log(u_0)}$	2	3977	3118
$X \sim \text{Bernoulli}(0.1)$	4	3579	2719
$U_0 \sim \Gamma(2, 1/2)$	10	3337	2533
	25	3271	2484
	50	3253	2467
	100	3236	2456
$Y u_0 \sim \text{Pois}(\lambda)$	1	2459	1813
$\lambda = e^{\beta_0 + \beta_1 X + \log(u_0)}$	2	1840	1334
$X \sim \text{Bernoulli}(0.5)$	4	1646	1182
$U_0 \sim \Gamma(2, 1/2)$	10	1550	1112
	25	1523	1087
	50	1513	1080
	100	1508	1076
$Y u_0 \sim \text{Pois}(\lambda)$	1	8517	6049
$\lambda = e^{\beta_0 + \beta_1 X + \log(u_0)}$	2	6384	4358
$X \sim \text{Bernoulli}(0.9)$	4	5705	3937
$U_0 \sim \Gamma(2, 1/2)$	10	5388	3679
	25	5291	3613
	50	5269	3582
	100	5252	3572

Table 2: Results for the hierarchical Poisson regression model for testing $H_0 : \beta_1 = 0$, against $H_1 : \beta_1 \neq 0$, with significance level 5%. One covariate X has the standard normal distribution.

	Number of clusters	N_s	
		Power=0.9	Power=0.8
$Y u_0 \sim \text{Pois}(\lambda)$	1	589	459
$\lambda = e^{\beta_0 + \beta_1 X + \log(u_0)}$	2	441	322
$X \sim N(0,1)$	4	394	285
$U_0 \sim \Gamma(2, 1/2)$	10	371	268
	25	363	263
	50	360	261
	100	359	260
$Y u_0 \sim \text{Pois}(\lambda)$	1	894	604
$\lambda = e^{\beta_0 + \beta_1 X + \log(u_0)}$	2	679	456
$X \sim N(0.5,1)$	4	601	401
$U_0 \sim \Gamma(2, 1/2)$	10	574	380
	25	563	373
	50	559	370
	100	559	369

are quite similar to the model with a single covariate having a Bernoulli distribution in the well balanced case with $(p_1, p_2, p_3, p_4) = (0.25, 0.25, 0.25, 0.25)$. For the extremely skewed case, the sample size increases substantially for both levels of power. When we increase the power from 80% to 90%, the sample size increases from 27% to 42% for the balanced case, and from 20% to 32% for the unbalanced case. The results with the multinomial case are listed in Table 3. As the number of clusters increases, the sample sizes approach those of Shieh (2001) for the non hierarchical model.

For the hierarchical Zero-inflated Poisson regression model, we consider the same distributions cited previously and we also add the bivariate normal distribution. The results are summarized in Tables 4–7. Compared to the hierarchical Poisson model, calculation with the hierarchical ZIP model provides sample sizes which are between 7% and 10% larger. As the number of clusters increases, the sample sizes approach those of Channouf et al. (2014) obtained with the non-hierarchical ZIP regression model, with all the types

Table 3: Results for the hierarchical Poisson regression model for testing $H_0 : \beta_1 = 0$, against $H_1 : \beta_1 \neq 0$, with significance level 5%. Two covariates X_1 and X_2 have the multinomial distribution.

	Number of clusters	N_s	
		Power=0.9	Power=0.8
$Y u_0 \sim \text{Pois}(\lambda)$	1	2581	1814
$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \log(u_0)}$	2	1857	1348
$(X_1, X_2) \sim \text{Multinomial}(p)$	4	1654	1192
$p = (0.25, 0.25, 0.25, 0.25)$	10	1547	1107
$U_0 \sim \Gamma(2, 1/2)$	25	1521	1087
	50	1514	1079
	100	1507	1077
$Y u_0 \sim \text{Pois}(\lambda)$	1	3993	3126
$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \log(u_0)}$	2	3009	2150
$(X_1, X_2) \sim \text{Multinomial}(p)$	4	2716	1925
$p = (0.4, 0.1, 0.1, 0.4)$	10	2527	1795
$U_0 \sim \Gamma(2, 1/2)$	25	2482	1754
	50	2465	1748
	100	2459	1737
$Y u_0 \sim \text{Pois}(\lambda)$	1	9214	7657
$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \log(u_0)}$	2	5812	4540
$(X_1, X_2) \sim \text{Multinomial}(p)$	4	5028	3826
$p = (0.76, 0.19, 0.01, 0.04)$	10	4702	3578
$U_0 \sim \Gamma(2, 1/2)$	25	4624	3495
	50	4601	3462
	100	4593	3461

of distributions of the covariates studied. The results for the normal case with two covariates exhibit larger sample sizes with positive correlation when compared to the results with the non-hierarchical Poisson regression model. When $\rho = -0.5$, the sample size is larger by a percentage ranging between 14% and 17%, and for $\rho = 0$, it is between 41% and 46% larger. When $\rho = 0.5$, the sample size is larger than results with the non-hierarchical Poisson regression model by a percentage ranging from 91% to 94%. This dramatic increase in sample size must be taken into account when introducing random effects into a standard Poisson model.

For both types of Poisson and ZIP models, and with all types of covariate distributions considered, the results indicated, as expected, that the required total sample size always decreases when the number of clusters increases. This means fewer correlated observations and therefore fewer individuals are needed.

5 Illustrative example

In order to illustrate our methodology with a real example, we reconsider here an application first presented in Signorini (1991), and then studied by Channouf et al. (2014) using the ZIP regression model. During a study of water pollution around Sydney Australia, the number of illnesses and infections contracted during a swimming season was examined by the Sydney Water Board. The objective was to determine if there was a significant difference between non-ocean or infrequent swimmers ($X_1 = 0$) and ocean swimmers ($X_1 = 1$). Using a Poisson regression to model the number of illnesses and infections per swimmer, it was originally assumed that the infection rate per non-ocean or infrequent swimmer was 0.85 ($e^{\beta_0} = 0.85$) and that the sample size large enough to detect an increase of 30% for ocean swimmers ($\Delta = e^{\beta_1} = 1.3$) should be calculated for a given nominal power. It was also assumed that an equal number from both groups of swimmers would be sampled ($X_1 \sim \text{Bernoulli}(0.5)$).

We used the method described here for hierarchical ZIP models for more flexibility in order to account for a possible excess of zeros amongst swimmers and we can assume that the swimmers were selected from 5 different cities (clusters). In this case, swimmers within the same city should be correlated and we examine the effect of the ICC on the sample size for different values of the dispersion parameter α_0 . At first, we consider that all swimmers have the same probability π of being measured from a zero-only random variable ($\pi = 0\%$ represents the hierarchical Poisson regression model). In a second step, we suppose that the practitioners

Table 4: Results for the hierarchical zero-inflated Poisson regression model for testing $H_0 : \beta_1 = 0$, against $H_1 : \beta_1 \neq 0$, with significance level 5%. One covariate G with Bernoulli distribution and one covariate X has the Bernoulli distribution.

	Number of clusters	N_s	
		Power=0.9	Power=0.8
$Y u_0 \sim \text{ZIP}(\lambda, \pi)$	1	5989	4796
$\lambda = e^{\beta_0 + \beta_1 X + \log(u_0)}$	2	4387	3414
$\pi = \frac{e^{\gamma_0 + \gamma_1 G}}{1 + e^{\gamma_0 + \gamma_1 G}}$	4	3909	2977
$X \sim \text{Bernoulli}(0.1)$	10	3646	2781
$G \sim \text{Bernoulli}(0.5)$	25	3552	2710
$U_0 \sim \Gamma(2, 1/2)$	50	3530	2701
	100	3523	2684
$Y u_0 \sim \text{ZIP}(\lambda, \pi)$	1	2659	2087
$\lambda = e^{\beta_0 + \beta_1 X + \log(u_0)}$	2	2004	1471
$\pi = \frac{e^{\gamma_0 + \gamma_1 G}}{1 + e^{\gamma_0 + \gamma_1 G}}$	4	1787	1301
$X \sim \text{Bernoulli}(0.5)$	10	1685	1214
$G \sim \text{Bernoulli}(0.5)$	25	1646	1185
$U_0 \sim \Gamma(2, 1/2)$	50	1636	1176
	100	1630	1171
$Y u_0 \sim \text{ZIP}(\lambda, \pi)$	1	9581	6585
$\lambda = e^{\beta_0 + \beta_1 X + \log(u_0)}$	2	6936	4884
$\pi = \frac{e^{\gamma_0 + \gamma_1 G}}{1 + e^{\gamma_0 + \gamma_1 G}}$	4	6164	4235
$X \sim \text{Bernoulli}(0.9)$	10	5869	3999
$G \sim \text{Bernoulli}(0.5)$	25	5734	3914
$U_0 \sim \Gamma(2, 1/2)$	50	5697	3889
	100	5680	3880

Table 5: Results for the hierarchical zero-inflated Poisson regression model for testing $H_0 : \beta_1 = 0$, against $H_1 : \beta_1 \neq 0$, with significance level 5%. One covariate G with Bernoulli distribution and one covariate X has the normal distribution.

	Number of clusters	N_s	
		Power=0.9	Power=0.8
$Y u_0 \sim \text{ZIP}(\lambda, \pi)$	1	665	496
$\lambda = e^{\beta_0 + \beta_1 X + \log(u_0)}$	2	481	350
$\pi = \frac{e^{\gamma_0 + \gamma_1 G}}{1 + e^{\gamma_0 + \gamma_1 G}}$	4	427	310
$X \sim N(0, 1)$	10	401	294
$G \sim \text{Bernoulli}(0.5)$	25	394	286
$U_0 \sim \Gamma(2, 1/2)$	50	392	285
	100	390	283
$Y u_0 \sim \text{ZIP}(\lambda, \pi)$	1	658	486
$\lambda = e^{\beta_0 + \beta_1 X + \log(u_0)}$	2	482	353
$\pi = \frac{e^{\gamma_0 + \gamma_1 G}}{1 + e^{\gamma_0 + \gamma_1 G}}$	4	425	313
$X \sim N(0.5, 1)$	10	402	293
$G \sim \text{Bernoulli}(0.5)$	25	394	287
$U_0 \sim \Gamma(2, 1/2)$	50	391	284
	100	390	283

Table 6: Results for the hierarchical zero-inflated Poisson regression model for testing $H_0 : \beta_1 = 0$, against $H_1 : \beta_1 \neq 0$, with significance level 5%. One covariate G with Bernoulli distribution and two covariates X_1 and X_2 have the multinomial distribution.

	Number of clusters	N_s	
		Power=0.9	Power=0.8
$Y u_0 \sim \text{ZIP}(\lambda, \pi)$	1	2720	2022
$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \log(u_0)}$	2	2045	1458
$\pi = \frac{e^{\gamma_0 + \gamma_1 G}}{1 + e^{\gamma_0 + \gamma_1 G}}$	4	1790	1304
$(X_1, X_2) \sim \text{Multinomial}(p)$	10	1684	1218
$p = (0.25, 0.25, 0.25, 0.25)$	25	1650	1186
$G \sim \text{Bernoulli}(0.5)$	50	1639	1177
$U_0 \sim \Gamma(2, 1/2)$	100	1636	1172
$Y u_0 \sim \text{ZIP}(\lambda, \pi)$	1	4358	3267
$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \log(u_0)}$	2	3252	2363
$\pi = \frac{e^{\gamma_0 + \gamma_1 G}}{1 + e^{\gamma_0 + \gamma_1 G}}$	4	2920	2084
$(X_1, X_2) \sim \text{Multinomial}(p)$	10	2744	1951
$p = (0.4, 0.1, 0.1, 0.4)$	25	2684	1907
$G \sim \text{Bernoulli}(0.5)$	50	2670	1900
$U_0 \sim \Gamma(2, 1/2)$	100	2665	1890
$Y u_0 \sim \text{ZIP}(\lambda, \pi)$	1	8529	6591
$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \log(u_0)}$	2	6190	4726
$\pi = \frac{e^{\gamma_0 + \gamma_1 G}}{1 + e^{\gamma_0 + \gamma_1 G}}$	4	5503	4185
$(X_1, X_2) \sim \text{Multinomial}(p)$	10	5141	3937
$p = (0.76, 0.19, 0.01, 0.04)$	25	5034	3815
$G \sim \text{Bernoulli}(0.5)$	50	5008	3788
$U_0 \sim \Gamma(2, 1/2)$	100	4998	3775

Table 7: Results for the hierarchical zero-inflated Poisson regression model for testing $H_0 : \beta_1 = 0$, against $H_1 : \beta_1 \neq 0$, with significance level 5%. One covariate G with Bernoulli distribution and two covariates X_1 and X_2 have the bivariate normal distribution.

	Number of clusters	N_s	
		Power=0.9	Power=0.8
$Y u_0 \sim \text{ZIP}(\lambda, \pi)$	1	1109	824
$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \log(u_0)}$	2	826	604
$\pi = \frac{e^{\gamma_0 + \gamma_1 G}}{1 + e^{\gamma_0 + \gamma_1 G}}$	4	731	536
$(X_1, X_2) \sim N(0, \Sigma_\rho)$	10	688	499
$\rho = -0.5$	25	675	490
$G \sim \text{Bernoulli}(0.5)$	50	672	487
$U_0 \sim \Gamma(2, 1/2)$	100	668	485
$Y u_0 \sim \text{ZIP}(\lambda, \pi)$	1	1035	783
$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \log(u_0)}$	2	765	578
$\pi = \frac{e^{\gamma_0 + \gamma_1 G}}{1 + e^{\gamma_0 + \gamma_1 G}}$	4	694	512
$(X_1, X_2) \sim N(0, \Sigma_\rho)$	10	653	481
$\rho = 0$	25	641	469
$G \sim \text{Bernoulli}(0.5)$	50	638	466
$U_0 \sim \Gamma(2, 1/2)$	100	637	464
$Y u_0 \sim \text{ZIP}(\lambda, \pi)$	1	1699	1273
$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \log(u_0)}$	2	1294	960
$\pi = \frac{e^{\gamma_0 + \gamma_1 G}}{1 + e^{\gamma_0 + \gamma_1 G}}$	4	1160	849
$(X_1, X_2) \sim N(0, \Sigma_\rho)$	10	1093	800
$\rho = 0.5$	25	1072	782
$G \sim \text{Bernoulli}(0.5)$	50	1064	777
$U_0 \sim \Gamma(2, 1/2)$	100	1061	775

would want to consider that both groups of swimmers could have a different probability of being measured from a zero-only random variable. We have:

$$\pi_i = \begin{cases} \pi^0 & \text{if } G_{1,i} = 0, \\ \pi^1 & \text{if } G_{1,i} = 1. \end{cases} \quad (15)$$

We also have $G_{1,i} = X_{1,i}$ and the parameters γ_0 and γ_1 are chosen such that $\pi^0 = .10$ and $\pi^1 = .05, .15$ or $.20$.

Table 8 shows the required sample sizes for hierarchical ZIP regression for nominal powers 80%, 90% and 95%. with the two cases for π . For the single probability ZIP model as π increases from 0.05 to 0.25, the necessary sample size increases by 56%-58% for $\alpha_0 = 2$. For $\alpha_0 = 100$, this increase is of the order of 46%-51%. In the more complex model given by equation (15) with $\pi^0 = 0.1$ and π^1 varying between 0.05 and 0.2, the sample size increase varies between 14% and 17% for all α_0 considered. The necessary sample sizes increase as π increases in the simple model and as π^1 increases in the more complex model. Increasing the power while keeping the other parameters fixed increases the necessary sample size. It should be noted that the necessary sample size for $\pi = \pi^1 = 0.05$ are larger for the complex model.

Table 8: Results for testing $H_0: \beta_1 = 0$, against $H_1: \beta_1 \neq 0$, with significance level 5%. Hierarchical ZIP regression model, with $X \sim \text{Bernoulli}(0.5)$ ($e^{\beta_0} = 0.85$ and $e^{\beta_1} = 1.3$), $G \sim \text{Bernoulli}(0.5)$ and 5 clusters. Two cases: one case with a single probability of an excess zero ($\pi^0 = \pi^1 = \pi$), and a second case with two different probabilities of an excess zero for the two groups (π^0 and π^1).

	Power	$\pi^0 = \pi^1 = \pi$			$\pi^0 = 0.1$		
		$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.25$	$\pi^1 = 0.05$	$\pi^1 = 0.15$	$\pi^1 = 0.2$
$\alpha_0 = 2$	0.8	579	646	913	597	666	692
	0.9	769	845	1209	787	907	946
	0.95	946	1056	1476	998	1113	1189
	ICC	0.30	0.29	0.25	0.29	0.26	0.25
$\alpha_0 = 5$	0.8	537	598	847	571	561	664
	0.9	726	795	1127	752	753	880
	0.95	893	982	1370	938	934	1102
	ICC	0.15	0.13	0.10	0.14	0.13	0.12
$\alpha_0 = 10$	0.8	527	582	813	554	613	646
	0.9	707	780	1083	742	823	865
	0.95	871	950	1329	907	1008	1066
	ICC	0.08	0.07	0.05	0.08	0.07	0.07
$\alpha_0 = 100$	0.8	515	565	776	538	592	622
	0.9	686	749	1016	716	785	823
	0.95	847	924	1237	883	967	1010
	ICC	0.01	0.01	0.01	0.01	0.01	0.01

6 Conclusion

We have succeeded in developing Monte Carlo simulation methods for calculating the sample size necessary for the Wald test of one parameter for hierarchical Poisson models using the estimation procedure proposed by Lee and Nelder (1996, 2001) for the parameters, where the power and the significance levels, as well as the values of the parameters of interest are specified under the null and the alternative hypothesis. We show that these methods can be extended also to more complex models, such as Zero-inflated hierarchical Poisson models. Our simulations indicate that the sample size will be sensitive to the type of covariate distributions, as well as the size of the correlations.

Our results indicate that caution should be taken when adding random effects to Poisson regression models, because the required sample size for the same size and power can be substantially larger.

Further work should include the case of unequal sizes for the clusters which will result in more complex calculations. For these to be efficient, we will use quasi-Monte Carlo methods. This methodology could also be applied to other hierarchical generalized models.

References

- Channouf, N., Fredette, M., and MacGibbon, B. (2014). Power and sample size calculations for Poisson and Zero-inflated Poisson regression models. *Computational Statistics & Data Analysis*, 72:241–251.
- Christiansen, C.L. and Morris, C.N. (1997). Hierarchical Poisson regression modeling. *Journal of the American Statistical Association*, 92(438):618–632.
- Commenges, D. and Jacqmin, H. (1994). The intraclass correlation coefficient: distribution-free definition and test. *Biometrics*, 50(2):517–526.
- Ha, I.D. and Lee, Y. (2003). Estimating frailty models via Poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics*, 12(3):663–681.
- Hall, D.B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4):1030–1039.
- Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Series in Statistics.
- Kuk, A.Y.C. and Cheng, Y.W. (1999). Pointwise and functional approximations in monte carlo maximum likelihood estimation. *Statistics and Computing*, 9(2):91–99.
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.
- Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(4):619–678.
- Lee, Y. and Nelder, J.A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4):987–1006.
- Lee, Y., Nelder, J.A., and Noh, M. (2007). H-likelihood: problems and solutions. *Statistics and Computing*, 17(1):49–55.
- Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.
- Liu, G. and Liang, K.-Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics*, 53:937–947.
- Matsui, S. (2005). Sample size calculations for comparative clinical trials with over-dispersed Poisson process data. *Statistics in Medicine*, 24:1339–1356.
- Roy, A., Bhaumik, D.K., Aryal, S., and Gibbons, R.D. (2007). Sample size determination for hierarchical longitudinal designs with differential attrition rates. *Biometrics*, 63(3):699–707.
- Royle, J.A. (2008). Hierarchical modeling of cluster size in wildlife surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, 13(1):23–36.
- Self, S. and Mauritsen, R.H. (1988). Power/sample size calculations for generalized linear models. *Biometrics*, 44:79–86.
- Self, S., Mauritsen, R.H., and Ohara, J. (1992). Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 48:31–39.
- Shieh, G. (2000). On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 56:1192–1196.
- Shieh, G. (2001). Sample size calculations for logistic and Poisson regression models. *Biometrika*, 88(4):1193–1199.
- Shieh, G. (2005). On power and sample size calculations for Wald tests in generalized linear models. *Journal of Statistical Planning and Inference*, 128:43–59.
- Signorini, D.F. (1991). Sample size for Poisson regression. *Biometrika*, 78(2):446–450.
- Song, P.X.-K. (2007). *Correlated Data Analysis*. Springer Series in Statistics, New York.
- Tu, W. and Piegorsch, W.W. (2003). Empirical Bayes analysis for a hierarchical Poisson generalized linear model. *Journal of Statistical Planning and Inference*, 111:235–248.
- Williamson, J.M., Lin, H.-M., Lyles, R.H., and Hightower, A.W. (2007). Power calculations for ZIP and ZINB models. *Journal of Data Science*, 5:519–534.
- Xia, N., Zhang, Z.-Z., and Ying, Z.-L. (2006). Convergence rate of the L-N estimator in Poisson-gamma models. *Acta Mathematicae Applicatae Sinica, English Series*, 22(4):639–654.
- 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.