Computational Statistics and Data Analysis 72 (2014) 241–251

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Power and sample size calculations for Poisson and zero-inflated Poisson regression models

Nabil Channouf [a], Marc Fredette [b,*], Brenda MacGibbon [c]

[a] Department of Operations Management & Business Statistics, College of Economics & Political Sciences, Sultan Qaboos University, Muscat, Sultanate of Oman

[b] Department of Management Sciences, HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal (Qc) H3T 2A7, Canada

[c] Department of Mathematics, Université du Québec à Montréal, Montréal, Canada

ARTICLE INFO

Article history:
Received 20 December 2011
Received in revised form 29 September 2013
Accepted 29 September 2013
Available online 7 October 2013

Keywords:
Wald test
Generalized linear models
Correlation structure
AR(1)
Exchangeable
Monte Carlo simulations

ABSTRACT

Although sample size calculations for testing a parameter in the Poisson regression model have been previously done, very little attention has been given to the effect of the correlation structure of the explanatory covariates on the sample size. A method to calculate the sample size for the Wald test in the Poisson regression model is proposed, assuming that the covariates may be correlated and have a multivariate normal distribution. Although this method of calculation works with any pre-specified correlation structure, the exchangeable and the AR(1) correlation matrices with different values for the correlation are used to illustrate the approach. The method used here to calculate the sample size is based on a modification of a methodology already proposed in the literature. Rather than using a discrete approximation to the normal distribution which may be much more problematic in higher dimensions, Monte Carlo simulations are used. It is observed that the sample size depends on the number of covariates for the exchangeable correlation matrix, but much more so on the correlation structure of the covariates. The sample size for the AR(1) correlation matrix changes less substantially as the dimension increases, and it also depends on the correlation structure of the covariates, but to a much lesser extent. The methodology is also extended to the case of the zero-inflated Poisson regression model in order to obtain analogous results.

1. Introduction

For scientific studies in diverse fields, calculation of the sample size needed for a test with pre-specified power and size is usually a necessary part of the design. This is especially true in biomedical research, where Poisson regression models are widely used. It can often happen that the observed count data may have many 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. In general, when the model involves multiple parameters, the inference becomes more complex, even more so for the zero-inflated Poisson models, when additional covariates are considered for the excess zero process.

* Corresponding author. Tel.: +1 514 3407108; fax: +1 514 3405634.
E-mail address: marc.fredette@hec.ca (M. Fredette).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.09.029

Sample size determination and power calculations for generalized linear models are usually based on the Wald test (Whittemore, 1981; Signorini, 1991; Shieh, 2001, 2005), the score test (Self and Mauritsen, 1988) or the likelihood ratio test (Self et al., 1992; Shieh, 2000). However, Matsui (2005) used a nonparametric test for sample size determination and Lyles et al. (2007) used a method of discrete approximation, originally proposed by Blom (1958) for transformed beta variables, in order to do power calculations but they did not include sample size calculations. Many of these authors concentrated on the univariate case. However, in most applications there are several explanatory covariates that are usually correlated. We propose to study the effect of this correlation structure of those covariates on the power and sample size calculations. It should be noted that we are only assuming that the covariates are correlated but that the responses are independent.

In our calculations, we propose a more flexible method based on Monte Carlo simulations. We do not need to proceed by using the type of discrete approximation as in Shieh (2001) or Lyles et al. (2007) when considering continuous random variables or discrete random variables with infinite support for the covariates. For example these authors used 10 or 11 points to approximate the Poisson distribution although Lyles et al. (2007) increased this number by adding pseudo-observations at each of the points.

Although Lambert (1992), Hall (2000) and Yau and Lee (2001) present complete and detailed descriptions of the Poisson (ZIP) regression model and its appropriate estimation procedures, there does not seem to be any previous research on sample size calculations for this model. However, Williamson et al. (2007) used the approximations of Blom (1958) to perform power calculations in this case.

We have chosen to study the Wald test here because it is frequently used due to its accessibility and its intuitiveness and because of several major recent contributions in sample size calculations for this test. For the logistic regression model, Whittemore (1981) approximated the Fisher information matrix in order to do sample size calculations. Signorini (1991) extended her technique to the Poisson regression model with one covariate in order to perform sample size calculations for the Wald test. The method presented by Shieh (2001) is a modification and a generalization of the methods of Signorini (1991) for testing one parameter in a Poisson regression model.

We consider here the case of testing in the Poisson regression model with more than one covariate whether one parameter is zero using the Wald test by introducing our modification to the method of Shieh (2001) for sample size calculations. We then extend the methodology to the zero-inflated Poisson regression model. In particular, we study the influence of the correlation structure of the explanatory covariates on the sample size using the two most popular types of correlation matrices. We also study the difference in sample size, not only as the correlation changes but also as the number of additional covariates increases.

This paper has two major contributions. We are able to modify the method of Shieh (2001) in order to study the effect of the correlation structure of the covariates in the higher dimensional case for the Poisson regression model by the use of Monte Carlo simulations and we present an extension of our methodology for sample size calculations to the ZIP regression model. We also studied the effect of the correlation between the covariates and the number of covariates on sample size calculations for both models. This latter contribution was motivated by two important features. First, although studies often have control variables, the impact on the sample size of having additional covariates is often neglected. Also, it is motivated by the importance of the correlation structure between covariates, especially in clinical trials, where the exchangeable and the AR correlation matrices are widely used. In some situations, the covariates may be equally correlated and the order has no importance, so the exchangeable structure can be considered for the correlation matrix. In other cases, when there is a time dependence and the order is important, the autoregressive structure can be more appropriate, and the matrix is build from an autoregressive process with a specific order.

The remainder of the paper is organized as follows. In the next section we introduce the Poisson regression model and describe our approach to sample size calculations for this model, with covariates having a multivariate normal distribution with different forms for the correlation matrix. It is important to emphasize that the method works with any type of correlation structure. In Section 3, we present the extension of our approach to the zero-inflated Poisson regression models. Section 4 contains numerical examples with different correlation structures for the covariates and with an accompanying discussion of these results. Section 5 contains an illustrative example. Finally, in Section 6, we draw some conclusions and discuss future directions of research related to this work.

2. Poisson regression model

The Poisson regression model is an important member of the family of generalized linear models (McCullagh and Nelder, 1983). In these models, the density of the response random variable Y takes the form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{1}$$

where a, b and c are specified functions, θ is the canonical parameter, ϕ is the dispersion parameter and the linear predictor takes the following form:

$$g(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where g is the link function, $\lambda = \mathrm{E}[Y]$, $\mathbf{X} = (1, X_1, \ldots, X_p)^T$ is a vector of $(p+1)$ covariates with density f_X and assumed independent of the parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$. We have $\mathrm{E}[Y] = b'(\theta)$ and $\mathrm{Var}(Y) = b''(\theta)a(\phi)$.

For the Poisson regression model with a random sample of size N, the density of Y takes the form given in (1) with $\theta = \log\lambda$, $b(\theta) = e^{\theta}$, $c(y, \phi) = -\log(y!)$, $\phi = 1$, the link function $g = \log$, and the transformation:

$$\log\lambda_i = \beta^T\mathbf{X}_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_p X_{p,i}, \quad \text{for } i = 1, \ldots, N.$$

Following Shieh (2001), we calculate the log-likelihood function associated with the data and the parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ for a random sample $(y_i, x_{i,1}, \ldots x_{i,p})$, $i = 1, \ldots, N$, as follows:

$$l(\beta; y, x) = \log L(\beta; y, x) = \sum_{i=1}^{N} \left\{ y_i\beta^T x_i - e^{\beta^T x_i} - \log(y_i!) + \log f_X(x_i) \right\}.$$

The score function for N observations is given by:

$$S_{N,k}(\beta) = \frac{\partial l(\beta; y, x)}{\partial \beta_k} = \sum_{i=1}^{N} x_{k,i} \left\{ y_i - e^{\beta^T x_i} \right\}, \tag{2}$$

for $k = 0, 1, \ldots, p$, with $x_{0,i} = 1$. The maximum likelihood estimates of β are given by the solution of the system $S_{N,k}(\beta) = 0$, and their variance–covariance matrix $V(\beta)$ is given by the inverse of the Fisher information matrix $I(\beta)$ whose elements are given by:

$$I(\beta)_{k,l} = -\mathbf{E}\left[\frac{\partial^2 l(\beta; y, x)}{\partial\beta_k \partial\beta_l} \right],$$

for $k, l = 0, 1, \ldots, p$.

We want to find the minimum required sample size with the Wald test for the hypothesis $H_0 : \beta_s = 0$, against $H_1 : \beta_s > 0$. To achieve a given power, we consider the approach given by Shieh (2001) with different forms of the covariance matrix of $\mathbf{X}$. Shieh (2001) chose to solve $\lim_{N\to\infty} \mathbf{E}[S_{N,k}(\beta_0, \beta_1, \ldots, \beta_{s-1}, 0, \beta_{s+1}, \ldots, \beta_p)]/N$, where the full expectation is always taken with respect to the true value of β. In his case, this expectation is possible to calculate because it is always expressed as a finite sum because of the discrete nature of the distribution of the covariates $\mathbf{X}$. Even for a continuous random variable such as the normal distribution, it is only necessary for him to evaluate the univariate normal density at 10 points according to his discrete approximation. We, however, choose to avoid using such discrete approximation in the higher dimensional case. Thus we use instead the conditional expectation, conditional on the covariates $\mathbf{X}$, using Monte Carlo simulation to generate samples from these covariates rather than trying to make potentially hazardous discrete approximations in higher dimensional spaces. In Eq. (2) we choose $N = K$ sufficiently large as a first step and calculate $S_{N,k}(\beta) = S_{K,k}(\beta)$ directly by substituting the values of the covariates from our Monte Carlo simulations and henceforth we use K as the subscript for the score function. Empirical evidence indicates that K =100 suffices. We first calculate the estimates of the parameters $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_{s-1}^*, 0, \beta_{s+1}^*, \ldots, \beta_p^*)^T$ under H_0, by solving the following system of equations:

$$\mathbf{E}[S_{K,k}(\beta_0, \beta_1, \ldots, \beta_{s-1}, 0, \beta_{s+1}, \ldots, \beta_p)|\mathbf{X} = x] = 0, \tag{3}$$

where the conditional expectation (given the simulated covariates $\mathbf{X}$) is taken with respect to the true value of β.

For testing the null hypothesis $\mathrm{H}_0 : \beta_1 = 0$, against the alternative $\mathrm{H}_1 : \beta_1 > 0$, the formula for estimating the minimum sample size N_s, for $s = 1$, proposed by Shieh (2001), is given by

$$N_s = \left\lceil \left(\frac{V(\beta_0^*, 0, \beta_2^*, \ldots, \beta_p^*)_{2,2}^{1/2} Z_\alpha + V(\beta)_{2,2}^{1/2} Z_{1-Power}}{\beta_1} \right)^2 \right\rceil, \tag{4}$$

where Z_q is the $100(1-q)$th percentile of the standard normal distribution, α is the size of the test, and $V(\beta)$ is computed with the pre-specified values and $V(\beta_0^*, 0, \beta_2^*, \ldots, \beta_p^*)$ is calculated using (3). We have modified Shieh's method here in (4) by using size α instead of $\alpha/2$ (for a discussion of the reasons for this modification, see Section 4). It should also be noted that here the pre-specified value of $\exp\beta_1$ represents the difference we want to detect by the test for a given power and size.

For the Poisson regression model, there is an exact relationship between the Fisher information matrix and the moment generating function (m.g.f.). The variance V is the inverse augmented Hessian matrix of the moment generating function. In other words, let

$$m(t_0, t_1, \ldots, t_p) = \mathbf{E}[e^{t_0 + t_1 X_1 + \cdots + t_p X_p}], \tag{5}$$

be the moment generating function. For example, if $\mathbf{X} = (X_1, \ldots, X_p)^T$ is a random multivariate normal vector with vector of means $\theta = (\theta_1, \ldots, \theta_p)^T$ and covariance matrix $\Sigma = (\sigma_{ij})$, the moment generating function m of $\mathbf{X}$ is given by

$$m(t_1, \ldots, t_p) = e^{\sum_{i=1}^{p} t_i\theta_j + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{ij}t_it_j}.$$

Let M be the $(p+1) \times (p+1)$ augmented Hessian matrix of m in (5), that is,

$$M(t_0, t_1, \ldots, t_p) = \begin{pmatrix} m & \frac{\partial m}{\partial t_1} & \cdots & \frac{\partial m}{\partial t_j} & \cdots & \frac{\partial m}{\partial t_p} \\ \frac{\partial m}{\partial t_1} & \frac{\partial^2 m}{\partial t_1^2} & \cdots & \frac{\partial^2 m}{\partial t_1 \partial t_j} & \cdots & \frac{\partial^2 m}{\partial t_1 \partial t_p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \frac{\partial m}{\partial t_j} & \frac{\partial^2 m}{\partial t_j \partial t_1} & \vdots & \frac{\partial^2 m}{\partial t_j^2} & \vdots & \frac{\partial^2 m}{\partial t_j \partial t_p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial m}{\partial t_p} & \frac{\partial^2 m}{\partial t_p \partial t_1} & \cdots & \frac{\partial^2 m}{\partial t_p \partial t_j} & \cdots & \frac{\partial^2 m}{\partial t_p^2} \end{pmatrix}. \tag{6}$$

If we denote by $M^{-1}_{i,j}$ the element of the inverse of M at the i^{th} row and the j^{th} column, then the variance of β_1 is given by

$$V(\beta)_{2,2} = M^{-1}_{2,2}(\beta). \tag{7}$$

When testing $H_0 : \beta_s = 0$, against the alternative $H_1 : \beta_s > 0$, for $s > 1$, the formula for the variance function in (4) is defined in an analogous fashion.

We begin by finding the "estimates" of β by solving the system of equations given in (3) under H_0, and then we compute the variances using the moment generating function. The calculation of the sample size necessary for a test with specified power and size on any one parameter β_s, $s = 1, \ldots, p$, is summarized in Algorithm 1. Note that all the programs (written in R) presented in this paper are available at http://neumann.hec.ca/pages/marc.fredette/CSDA1.html.

Algorithm 1 Sample size calculation for Poisson regression model.

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 which is being tested whether or not it is equal to 0).
- The distribution of the covariates $\mathbf{X}$.
- Some reasonable values for the other β_k, $k = 1, \ldots, p$, 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 $\overline{\lambda} = E[e^{\beta^T\mathbf{X}}]$. (The value $\overline{\lambda}$ was chosen equal to .05 as in Shieh (2001)).
- 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 desired number of Monte Carlo replications B.

For multivariate normal covariates, the data must be standardized so that the covariance matrix will be of the form V_1 or V_2 (as defined in Section 4) and the correlation must be specified. (Different values of the correlation could be tried.)
The algorithm returns the minimum sample size needed for the significance test.
The algorithm has the following steps:

1. **Generate** a sample of K values from the distribution of $\mathbf{X}$ and compute the rates $\lambda_i(X)$, $i = 1. \ldots, K$, under H_1.
2. **Compute** the values $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_{s-1}^*, 0, \beta_{s+1}^*, \ldots, \beta_p^*)$, under H_0, by solving the system of nonlinear equations in (3) with $N = K$;
3. **Compute** the inverse of the augmented Hessian matrix of the moment generating function of X and take its (s+1)-th diagonal elements using (6) and (7) to obtain the variance functions $V(\beta^*)$ and $V(\beta)$ of (4);
4. **Return** the sample size $N_{s,j}$ by formula (4).
5. **Repeat** steps 1 to 4 B times, $j = 1, \ldots, B$.
6. **Take** the average value over the outputs of the B replications, $N_s = \frac{\sum_{j=1}^{B} N_{s,j}}{B}$.

3. Zero-inflated Poisson model

It is well known that outcomes which consist of counts in biomedical research may have many zeros and the Poisson model may not be an adequate model in such a situation. Several approaches are introduced for these zero-inflated models. We cite the zero-inflated Poisson model (Lambert, 1992), the zero-inflated negative binomial model and zero-inflated binomial model (Hall, 2000), and the zero-inflated gamma model (Yau et al., 2002).

We concentrate here on the zero-inflated Poisson regression model (ZIP). Let the responses be denoted by $Y_i, i = 1, \ldots, N$, and let π_i be the probability that Y_i is a zero-only random variable, with $0 \leq \pi_i \leq 1$. The random variable Y_i follows a ZIP distribution if:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i} & \text{if } y_i = 0, \\ (1-\pi_i)\dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} & \text{if } y_i > 0, \end{cases}$$

with $\mathrm{E}[Y_i] = (1-\pi_i)\lambda_i$ and $\mathrm{Var}(Y_i) = (1-\pi_i)\lambda_i(1+\pi_i\lambda_i)$, for $i = 1, \ldots, N$. In this model the zeros for the outcomes Y_i can come from two states, the zero state with probability π_i, or the Poisson state with a probability $(1-\pi_i)$, for $i = 1, \ldots, N$. Both sets of parameters λ and π are modeled by link functions in order to have linear predictors as follows:

$$\mathrm{logit}(\pi_i) = \gamma^T \mathbf{G}_i = \gamma_0 + \gamma_1 G_{1,i} + \cdots + \gamma_q G_{q,i}; \quad i = 1, \ldots, N, \tag{8}$$

with $\mathbf{G} = (1, G_1, \ldots, G_q)^T$ a vector of $(q+1)$ covariates and $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_q)^T$ a vector of $(q+1)$ parameters. For the counts, we have

$$\log(\lambda_i) = \beta^T \mathbf{X}_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_p X_{p,i}; \quad i = 1, \ldots, N,$$

with $\mathbf{X} = (1, X_1, \ldots, X_p)^T$ a vector of $(p+1)$ covariates and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ a vector of $(p+1)$ parameters. The two models could have common covariates; however, in practice this is not usually recommended (cf. Lambert, 1992).

As shown in Lambert (1992), the log-likelihood function is given by:

$$\begin{aligned} l(\gamma, \beta; y, g, x, z) &= l(\gamma; y, g, z) + l(\beta; y, x, z) + \sum_{i=1}^{N}(1-z_i)\log(y_i!) \\ &= \sum_{i=1}^{N}\{z_i \gamma^T g_i - \log(1 + e^{\gamma^T g_i})\} + \sum_{i=1}^{N}(1-z_i)\{y_i \beta^T x_i - e^{\beta^T x_i} - \log(y_i!)\}, \end{aligned}$$

with

$$Z_i = \begin{cases} 1 & \text{if } Y_i \text{ comes from the zero state,} \\ 0 & \text{if } Y_i \text{ comes from the Poisson state,} \end{cases}$$

for $i = 1, \ldots, N$.

For the ZIP regression model, the sample size determination is analogous to the methodology for the Poisson regression model, since we use the same formula for the minimum sample size N_s, given by (4). However, the computations become more complex, since we have additional parameters for the logistic component, namely $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_q)^T$. In addition, the simplification by the moment generating function cannot be applied in this case.

The score function is given by:

$$S_{N,k}(\gamma, \beta) = \begin{cases} \dfrac{\partial l(\gamma, \beta; y, g, x, z)}{\partial \gamma_l} = \displaystyle\sum_{i=1}^{N} g_{l,i}\left\{z_i - \frac{e^{\gamma_0 + \gamma_1 g_{1,i} + \cdots + \gamma_q g_{q,i}}}{1 + e^{\gamma_0 + \gamma_1 g_{1,i} + \cdots + \gamma_q g_{q,i}}}\right\}; \quad l = 0, 1, \ldots, q, \\ \dfrac{\partial l(\gamma, \beta; y, g, x, z)}{\partial \beta_k} = \displaystyle\sum_{i=1}^{N}(1-z_i)x_{k,i}\left\{y_i - e^{\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}}\right\}; \quad k = 0, 1, \ldots, p. \end{cases}$$

The maximum likelihood estimates of γ and β are given by the solution of the system $S_{N,k}(\gamma, \beta) = 0$, and their variance–covariance matrix $V(\gamma, \beta)$ is given by the inverse of the Fisher information matrix $I(\gamma, \beta)$:

$$I(\gamma, \beta) = \begin{pmatrix} -\mathbf{E}\left[\dfrac{\partial^2 l(\gamma, \beta; y, g, x, z)}{\partial \gamma \partial \gamma^T}\right] & -\mathbf{E}\left[\dfrac{\partial^2 l(\gamma, \beta; y, g, x, z)}{\partial \gamma \partial \beta^T}\right] \\ -\mathbf{E}\left[\dfrac{\partial^2 l(\gamma, \beta; y, g, x, z)}{\partial \beta \partial \gamma^T}\right] & -\mathbf{E}\left[\dfrac{\partial^2 l(\gamma, \beta; y, g, x, z)}{\partial \beta \partial \beta^T}\right] \end{pmatrix}. \tag{9}$$

We can now calculate the minimum required sample size with the Wald test for the hypothesis $H_0 : \beta_s = 0$, against $H_1 : \beta_s > 0$. Analogous to the Poisson regression model, the estimates of the parameters $\gamma^* = (\gamma_0^*, \gamma_1^*, \ldots, \gamma_q^*)^T$ and

$\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_{s-1}^*, 0, \beta_{s+1}, \ldots, \beta_p^*)^T$ under H_0, are calculated by solving the modified system of equations:

$$\mathbf{E}[S_{N,k}(\gamma, \beta_0, \beta_1, \ldots, \beta_{s-1}, 0, \beta_{s+1}, \ldots, \beta_p)|\mathbf{X} = x] = 0, \tag{10}$$

where the Z_i are replaced by their conditional means given y_i, for $i = 1, \ldots, N$:

$$\begin{aligned} E[Z_i|y_i, \gamma, \beta] &= P(\text{zero state}|y_i, \gamma, \beta) \\ &= \frac{P(Y_i|\text{zero state})P(\text{zero state})}{P(Y_i|\text{zero state})P(\text{zero state}) + P(Y_i|\text{Poisson state})P(\text{Poisson state})} \\ &= \begin{cases} (1 + e^{-\gamma^T \mathbf{G}_i - e^{\beta^T \mathbf{X}_i}})^{-1} & \text{if } y_i = 0, \\ 0 & \text{if } y_i > 0, \end{cases} \end{aligned}$$

for $i = 1, \ldots, N$. It should be noted that we did not use the EM algorithm as proposed by Lambert (1992) to obtain these estimates.

For testing the null hypothesis $H_0 : \beta_1 = 0$, against the alternative $H_1 : \beta_1 > 0$, the formula for estimating the minimum sample size N_s is given by

$$N_s = \left\lceil \left(\frac{V(\gamma^*, \beta_0^*, 0, \beta_2^*, \ldots, \beta_p^*)_{q+3,q+3}^{1/2} Z_\alpha + V(\gamma, \beta)_{q+3,q+3}^{1/2} Z_{1-Power}}{\beta_1} \right)^2 \right\rceil, \tag{11}$$

where $V(\gamma, \beta)$ is computed with the pre-specified values and $V(\gamma, \beta_0^*, 0, \beta_2^*, \ldots, \beta_p^*)$ is calculated using the values obtained in (10).

The implementation of the sample size calculations for a significance test on any parameter $\beta_s, s = 1, \ldots, p$, is summarized in Algorithm 2. It is important to note that the same procedure for sample size calculation can be used when the test is about an element of $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_q)$, for the zero state model. In Eqs. (10) and (11), the maximum likelihood estimates of γ^* and β^* are calculated by taking the parameter concerned by the test, i.e. under H_0, equal to zero, and its variance estimation is the corresponding diagonal element of $V(\gamma, \beta)$. This approach can be also generalized to a test about any subset or combination of parameters simultaneously.

Algorithm 2 Sample size calculation for ZIP regression model.

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 which the user is testing whether it is equal to 0).
- The distribution of the covariates **X**.
- Some reasonable values for the other $\beta_k, k = 1, \ldots, p$, 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 overall fixed means $\overline{\lambda} = E[e^{\beta^T \mathbf{X}}]$ and $\overline{\pi} = E[e^{\gamma^T \mathbf{G}}/(1 + e^{\gamma^T \mathbf{G}})]$, respectively. (The values $\overline{\lambda}$ and $\overline{\pi}$ were chosen equal to 0.05).
- 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 desired number of Monte Carlo replications B.

The algorithm returns the minimum sample size needed for the significance test.

The algorithm has the following steps:

1. **Generate** a sample of K values from the distributions of **X** and **G**, and compute the rates $\lambda_i(\mathbf{X})$, and $\pi_i(G)$, for $i = 1. \ldots, K$, under H_1.
2. **Compute** the maximum likelihood values $\gamma^* = (\gamma_0^*, \gamma_1^*, \ldots, \gamma_q^*)$ and $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_{s-1}^*, 0, \beta_{s+1}^*, \ldots, \beta_p^*)$, by solving the system of $(q + s + 1)$ nonlinear equations in (10) with $N = K$;
3. **Compute** the inverse of the Fisher information matrix $I(\gamma, \beta)$, given in (9), to obtain $V(\gamma^*, \beta^*)$ and $V(\gamma, \beta)$, and take the (q+s+2)-th diagonal elements $V(\gamma^*, \beta^*)_{q+s+2,q+s+2}$ and $V(\gamma, \beta)_{q+s+2,q+s+2}$;
4. **Return** the sample size N_s by formula (11).
5. **Repeat** steps 1 to 4 B times, $j = 1, \ldots, B$.
6. **Take** the average value over the outputs of the B replications, $N_s = \frac{\sum_{j=1}^{B} N_{s,j}}{B}$.

4. Simulation study

In the study of the Poisson regression model, for the vector of covariates $\mathbf{X} = (X_1, \ldots, X_p)^T$, we considered two special cases for the parameterization of the covariance matrix V_l, $l = 1, 2$, which depend only on a single parameter ρ. The first

one represents the exchangeable case and the second is the AR(1) model. Each matrix is defined as follows:

$$V_1 : \mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} 1 & \rho & \dots & \rho & \cdots & \rho \\ \rho & 1 & & \rho & \cdots & \rho \\ \vdots & & \ddots & & & \vdots \\ \rho & \rho & & 1 & & \rho \\ \vdots & \vdots & & & \ddots & \\ \rho & \rho & \cdots & \rho & & 1 \end{pmatrix}.$$

$$V_2 : \mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} \frac{1}{1-\rho^2} & \frac{\rho}{1-\rho^2} & \cdots & \frac{\rho^{j-1}}{1-\rho^2} & \cdots & \frac{\rho^{p-1}}{1-\rho^2} \\ \frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & \cdots & \frac{\rho^{j-2}}{1-\rho^2} & \cdots & \frac{\rho^{p-2}}{1-\rho^2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \frac{\rho^{j-1}}{1-\rho^2} & \frac{\rho^{j-2}}{1-\rho^2} & \vdots & \frac{1}{1-\rho^2} & \vdots & \frac{\rho^{p-j}}{1-\rho^2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\rho^{p-1}}{1-\rho^2} & \frac{\rho^{p-2}}{1-\rho^2} & \cdots & \frac{\rho^{p-j}}{1-\rho^2} & \cdots & \frac{1}{1-\rho^2} \end{pmatrix}.$$

When choosing such covariance matrices, we had in mind the following types of data. A typical example of V_2 could be blood pressure measurements taken weekly for a certain period of time or results of certain blood tests measured on a continuous scale. A group of psychometric scores can sometimes be modeled as V_1.

We illustrated our methods for ρ ranging between $[-0.5, 0.5]$, and for four dimensions $p = 2, 3, 4, 5$. The lower bound for ρ is used to ensure a positive semidefinite correlation matrix for V_1, and the upper bound is used to avoid multicollinearity problems. The exchangeable correlation matrix is positive semidefinite for $-1/(p-1) < \rho < 1$. In fact, this condition ensures that the exchangeable correlation matrix is diagonally dominant. In general, a symmetric matrix with non negative diagonal entries is positive semidefinite if it is diagonally dominant (see Golub and Van Loan, 1989).

We considered the cases of a vector $\mathbf{X}$ of covariates normally distributed, with zero mean and covariance matrix V_l, for $l = 1, 2$. The values for the β_i, for $i = 1, \dots, p$, were chosen equal to those of Shieh (2001), that is, equal to $\log(2)$ and β_0 was chosen to satisfy an overall mean of 0.05 as indicated in Algorithm 1. An R program is available[1] to compute the sample size with the Poisson regression model when the covariates have a multivariate normal distribution with any type of correlation matrix, using Monte Carlo simulations.

For the ZIP regression model, the algorithm is more complex because we consider the conditional expectation for the score function, given in (10), so we need to generate the values of the covariates in each Monte Carlo replication. In order to be comparable with the examples in Shieh (2001) for the non-zero-inflated case, we considered similar distributions for the covariates. We began with a first example having a single covariate and then we considered the case of two covariates. For the single covariate models, we considered the following examples: $X \sim \mathrm{Bernoulli}(0.5)$, and $G \sim \mathrm{Bernoulli}(0.5)$; and then the case of $X \sim N(0, 1)$, and $G \sim \mathrm{Bernoulli}(0.5)$. For the case of two covariates, we assumed that (X_1, X_2) has a multinomial distribution with probabilities $(p_1, p_2, p_3, p_4 = 1 - p_1 - p_2 - p_3)$, corresponding to the values $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. In particular, we considered the three following cases corresponding to those in Shieh (2001): $(p_1, p_2, p_3, p_4) = (0.25, 0.25, 0.25, 0.25)$, $(p_1, p_2, p_3, p_4) = (0.4, 0.1, 0.1, 0.4)$, and $(p_1, p_2, p_3, p_4) = (0.76, 0.19, 0.01, 0.04)$. As in Shieh (2001), the "true" parameters were taken as follows: β_0 is chosen to satisfy an overall mean $\bar{\lambda} = 0.05 = \mathrm{E}[e^{\beta^T \mathbf{X}}]$; γ_0 is chosen to satisfy an overall mean $\bar{\pi} = 0.05 = \mathrm{E}[e^{\gamma^T \mathbf{G}}/(1 + e^{\gamma^T \mathbf{G}})]$, while other parameters and the difference we wish to test are set, respectively, as $\gamma_1 = \beta_2 = \log(2)$ and $\log(\Delta) = \beta_1 = \log(2)$.

The estimates of the sample size N_s, were calculated for a significance level $\alpha = 2.5\%$ and powers equal to 80% and 90%.

It is implicit in the work of Whittemore (1981), Signorini (1991) and Self and Mauritsen (1988) that this type of methodology to determine sample sizes described here in the previous sections is only useful for a unilateral test of the parameter. Thus, if Eq. (4) is used the size of the test is $\alpha/2$ and not α as reported in Shieh (2001). Because we wanted to compare our numerical results with those of Shieh (2001), we chose this rather unusual test size of .025. It should be noted that the sample size and power results for the test $H_0 : \beta_1 = 0$, against $H_1 : \beta_1 \neq 0$, for a significance level of α, are the same as for the sample size and power results for the test $H_0 : \beta_1 = 0$, against $H_1 : \beta_1 > 0$, for a significance level of $\alpha/2$, as shown in the tables here.

The computations were based on independent Monte Carlo simulations, with 5000 replications. For each replication, for both Poisson and ZIP regression models, a data set consisting of sets of covariates of sizes $K = 100$ and 1000, were sampled

[1] See http://neumann.hec.ca/pages/marc.fredette/CSDA1.html

Table 1
Sample size for the Poisson regression model for the test $H_0 : \beta_1 = 0; H_1 : \beta_1 > 0$, with significance level 2.5%. $\mathbf{X} \sim$ Multivariate normal$(0, V_1)$ (the exchangeable case).

Power	0.9				0.8			
Dimension	2	3	4	5	2	3	4	5
ρ								
−0.5	570	–	–	–	424	–	–	–
−0.4	514	799	–	–	382	580	–	–
−0.3	477	552	1095	–	356	409	791	–
−0.2	455	481	505	623	340	360	373	454
−0.1	443	453	444	441	332	339	331	327
0	439	445	435	428	328	333	325	319
0.1	443	451	449	443	332	339	336	331
0.2	455	472	473	467	339	353	352	348
0.3	474	500	496	481	355	373	369	355
0.4	505	530	513	483	376	392	376	349
0.5	548	562	522	473	406	412	373	333

Table 2
Sample size for the Poisson regression model for the test $H_0 : \beta_1 = 0; H_1 : \beta_1 > 0$, with significance level 2.5%. $\mathbf{X} \sim$ Multivariate normal$(0, V_2)$ (the AR(1) case).

Power	0.9				0.8			
Dimension	2	3	4	5	2	3	4	5
ρ								
−0.5	445	422	433	424	333	314	323	315
−0.4	442	424	432	426	331	315	323	318
−0.3	440	428	433	428	329	319	323	318
−0.2	440	434	434	429	329	324	323	319
−0.1	440	439	434	429	329	329	324	319
0	440	445	435	428	329	334	326	318
0.1	440	448	438	429	329	336	327	318
0.2	440	452	442	430	328	340	331	320
0.3	438	454	446	433	327	342	335	323
0.4	433	452	447	436	324	339	337	325
0.5	423	439	439	436	315	326	330	326

from the specified distribution of $\mathbf{X}$ (and also of G for the ZIP regression model). We experimented with both $K = 100$ and $K = 1000$, with negligible changes in the results (<3% of the sample size). For this reason we reported results only for $K = 100$. With these samples, we obtain the rates λ for the standard Poisson regression model and both the Poisson rates λ and the Bernoulli probabilities π for the ZIP regression model. We subsequently calculated the score and the log-likelihood functions, and then we used the method, described earlier in Algorithms 1 and 2, to find the required sample size N_s.

Tables 1 and 2 summarize the results for the Poisson regression model with the multivariate normal covariates. Results with the exchangeable covariance matrix V_1 indicated that the sample size increases when the absolute value of ρ increases; whereas, the sample size decreases slightly for positive ρ when the dimension increases and it increases substantially for $\rho < 0.1$ as the dimension increases. With the AR(1) covariance matrix V_2 the results indicated that for dimensions greater than two the sample size increases slightly when the value of ρ increases for $\rho > -0.2$. The sample size for the AR(1) covariance matrix does not change much when the dimension changes, although it seems to be slightly lower for odd dimensions for $\rho < -0.2$.

In Table 3, we calculated the necessary sample size for the two levels of power using our method, and we reported the sample size results for the standard Poisson regression model (with no excess zeros) from Shieh (2001) for comparison. We also calculated the estimated sizes of the tests in order to establish that the size is truly of the order of .025 and not of .05 as mentioned in Shieh (2001). We used the fixed nominal power for comparison with the estimated power obtained by Monte Carlo simulations based on 5000 independent replications. For each replicate, N_1 covariates X and G are generated from the initial distributions, and with these covariates we obtain the rates π and λ to generate the outcomes, and finally the test statistic is computed and compared to the critical value Z_α. Then, the estimated power will be the proportion of the replicates where the test is rejected (Shieh, 2001). We also reported the absolute error, |'Nominal power' − 'Estimated power'|, in the last column.

In general, the results are very accurate; the error between nominal and estimated power is relatively small. However, the error is higher for power = 80% and, in some asymmetric cases, e.g. when $X \sim$ Bernoulli(0.9) (Table 3), where the error equals 0.108. The estimated power usually tends to underestimate the nominal power, except for the extremely skewed case with the multinomial distribution (Table 4), when $(p_1, p_2, p_3, p_4) = (0.76, 0.19, 0.01, 0.04)$ and thus $P(X_1 = 0) = 0.95$. For this case, the power obtained is overestimated, but the error is no higher than the other cases that appear in Shieh (2001)

Table 3
Results with the zero-inflated Poisson regression model for the hypothesis test H_0: $\beta_1 = 0$, against H_1: $\beta_1 > 0$, with significance level 2.5%, and two covariates, X and G, or equivalently testing H_0: $\beta_1 = 0$, against H_1: $\beta_1 \neq 0$, with significance level 5%.

	Power	Poisson	ZIP			
		N_s	N_s	Est. power with N_s	Error	Est. size with N_s
$Y \sim \text{ZIP}(\lambda, \pi)$ $\lambda = e^{\beta_0+\beta_1 X}$ $\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $X \sim \text{Bernoulli}(0.1)$ $G \sim \text{Bernoulli}(0.5)$	0.9	4046	4261	0.9312	0.0312	0.0256
	0.8	3164	3331	0.8778	0.0778	0.0290
$Y \sim \text{ZIP}(\lambda, \pi)$ $\lambda = e^{\beta_0+\beta_1 X}$ $\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $X \sim \text{Bernoulli}(0.5)$ $G \sim \text{Bernoulli}(0.5)$	0.9	1823	1933	0.9022	0.0022	0.0264
	0.8	1355	1428	0.7778	0.0222	0.0198
$Y \sim \text{ZIP}(\lambda, \pi)$ $\lambda = e^{\beta_0+\beta_1 X}$ $\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $X \sim \text{Bernoulli}(0.9)$ $G \sim \text{Bernoulli}(0.5)$	0.9	6271	6599	0.8592	0.0408	0.0218
	0.8	4418	4650	0.6920	0.1080	0.0208
$Y \sim \text{ZIP}(\lambda, \pi)$ $\lambda = e^{\beta_0+\beta_1 X}$ $\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $X \sim N(0, 1)$ $G \sim \text{Bernoulli}(0.5)$	0.9	437	453	0.8550	0.0450	0.0282
	0.8	326	339	0.7362	0.0638	0.0326
$Y \sim \text{ZIP}(\lambda, \pi)$ $\lambda = e^{\beta_0+\beta_1 X}$ $\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $X \sim N(0.5, 1)$ $G \sim \text{Bernoulli}(0.5)$	0.9	439	462	0.8796	0.0204	0.0268
	0.8	328	345	0.7798	0.0202	0.0248

Table 4
Results with the zero-inflated Poisson regression model for the hypothesis test H_0: $\beta_1 = 0$, against H_1: $\beta_1 > 0$, with significance level 2.5%, and three covariates, X_1, X_2 and G.

	Power	Poisson	ZIP			
		N_s	N_s	Est. power with N_s	Error	Est. size with N_s
$Y \sim \text{ZIP}(\lambda, \pi)$ $\lambda = e^{\beta_0+\beta_1 X_1+\beta_2 X_2}$ $\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $(X_1, X_2) \sim \text{Multinomial}(p)$ $p = (0.25, 0.25, 0.25, 0.25)$ $G \sim \text{Bernoulli}(0.5)$	0.9	1835	1932	0.8940	0.0060	0.0228
	0.8	1356	1428	0.7860	0.0140	0.0244
$Y \sim \text{ZIP}(\lambda, \pi)$ $\lambda = e^{\beta_0+\beta_1 X+\beta_2 X_2}$ $\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $(X_1, X_2) \sim \text{Multinomial}(p)$ $p = (0.4, 0.1, 0.1, 0.4)$ $G \sim \text{Bernoulli}(0.5)$	0.9	2927	3138	0.8748	0.0252	0.0236
	0.8	2184	2297	0.7578	0.0422	0.0226
$Y \sim \text{ZIP}(\lambda, \pi)$ $\lambda = e^{\beta_0+\beta_1 X+\beta_2 X_2}$ $\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $(X_1, X_2) \sim \text{Multinomial}(p)$ $p = (0.76, 0.19, 0.01, 0.04)$ $G \sim \text{Bernoulli}(0.5)$	0.9	5661	5972	0.9278	0.0278	0.0326
	0.8	4390	4633	0.8514	0.0514	0.0256

with the standard Poisson regression model. As remarked by Shieh (2001), when the distribution is skewed, the nominal power is not accurate. For all types of distributions for the covariates, we were not surprised to observe that larger sample sizes are required with the ZIP regression model compared to the Poisson regression model.

In situations where the outcome variable contains excess zeros, we expected to obtain larger required sample size when we use the ZIP regression model instead of the standard Poisson model (Williamson et al., 2007). When the covariates are

Table 5
Results for testing H_0: $\beta_1 = 0$, against H_1: $\beta_1 \neq 0$, with significance level 5%. Poisson and ZIP regression models, with $X \sim$ Bernoulli(0.5) ($e^{\beta_0} = 0.85$ and $e^{\beta_1} = 1.3$).

Power	Poisson		ZIP		
	Signorini (1991)	Shieh (2001)	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.25$
0.8	406	370	389	411	493
0.9	555	513	540	570	684
0.95	697	649	683	721	865

Table 6
Results for testing H_0: $\beta_1 = 0$, against H_1: $\beta_1 \neq 0$, with significance level 5%. ZIP regression model, with $X \sim$ Bernoulli(0.5) ($e^{\beta_0} = 0.85$ and $e^{\beta_1} = 1.3$) and $G \sim$ Bernoulli(0.5). Two different probabilities of an excess zero for the two groups (π^0 and π^1).

Power	ZIP ($\pi^0 = 0.1$)			
	$\pi^1 = 0.05$	$\pi^1 = 0.1$	$\pi^1 = 0.15$	$\pi^1 = 0.2$
0.8	400	411	423	435
0.9	554	570	586	603
0.95	701	721	741	763

distributed as Bernoulli (for X_1) or multinomial (for (X_1, X_2)) random variables, the sample size needed is about 5.3% higher when we use the ZIP regression model compared to the standard Poisson model. When a single covariate has the standard normal distribution, the sample size is only 3.7% higher. In this latter case, we expect to have lower values for the required sample size than in the binomial case, as more zeros are generated by the Poisson process. In the multinomial case the sample size needed is between 5.3% and 7.2% higher when the ZIP regression model is used.

All our experiments were done for two levels of power, 80% and 90%. For the standard Poisson regression model, the sample size is, in general, 35% larger when the desired power increases from 80% to 90%. Of course, the increase in the required sample size depends on the type of covariance matrix used and the number of covariates in the model. With covariance matrix V_1, the sample size increases by 34% when $p = 2$ up to 42% when $p = 5$. With matrix V_2, as expected, the number of covariates has much less impact on the increase. When $p = 2$, the sample size required when we increase the power from 80% to 90%, increases by about 34%, as opposed to 35% for dimensions 3 and 4, and 36% when the dimension is 5. As for the ZIP regression model, the increase is comparable to what we observed for the standard Poisson regression model: the sample size needed to obtain a power of 90% is, on average, 34% higher than the one required to obtain a power of 80%.

5. Illustrative example

In order to illustrate our methodology with a real example, we reconsider here an application first presented in Signorini (1991). 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 they wanted a sample size large enough to detect an increase of 30% for ocean swimmers ($\Delta = e^{\beta_1} = 1.3$) for a given nominal power. It was also assumed that an equal number of ocean and non-ocean or infrequent swimmers would be sampled ($X_1 \sim$ Bernoulli(0.5)).

Using the new methodology we proposed for ZIP regression models, we could consider a more flexible model by accounting for a possible excess of zeros amongst swimmers than a Poisson model could allow. This could particularly be of interest for the group of infrequent swimmers. At first, we will consider that all swimmers have the same probability π of being measured from a zero-only random variable ($\pi = 0\%$ represents the usual Poisson regression model).

Table 5 shows the required sample size for both Poisson and ZIP regression for nominal powers 80%, 90% and 95% and where π could take the values 5%, 10% and 25%. For the Poisson regression results, we also added the results originally presented in Signorini (1991), we can see that our sample sizes are 5%–10% lower. Shieh (2001) also noted through a simulation study that the approximation developed by Signorini (1991) tends to give sample sizes that are too large.

Table 6 shows sample size results for new ZIP regression models. We are now supposing that the practitioners would want to consider that both groups of swimmers could have a different probability of being measured from a zero-only random variable. Using the notation in Eq. (8), let

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

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 $\pi^1 = 5\%$, 10%, 15% or 20%.

6. Conclusion

Our objective here was to study the effect of the correlation structure of the covariates and the number of covariates on the sample size required to attain certain levels of power and size for the Wald test when testing whether one parameter is zero in a multidimensional Poisson regression model and the zero-inflated Poisson regression model. We introduced Monte Carlo simulation techniques in order to adapt the approach of Shieh (2001) to do sample size calculations for the Poisson regression model with more than one covariate. We performed extensive simulation studies in order to determine the effect of the correlation structure of the explanatory covariates on the sample size for two different types of correlation matrices. There is clear evidence that the correlation structure of these covariates plays an important role in sample size determination for the exchangeable correlation matrix and a lesser one for the AR(1) correlation matrix. The dimension also has an effect on the sample size determination in the exchangeable case, but this effect is to a much lesser extent with the AR(1) correlation matrix. We have shown these effects using simple correlation structures for the covariates that are easily parameterized. In practice the correlation matrix is bound to be more complex and it will become important to obtain prior information about this matrix in order to do a correct assessment of the sample size required. Perhaps it will be necessary to do the calculations for several different correlation matrices before deciding exactly what sample size is needed. We conclude that both these aspects, the correlation structure of the covariates and the number of covariates, must be taken into account in sample size calculations.

There are several different directions in which we wish to continue our research. Although we did find some interesting differences when we did the sample size calculations for the zero-inflated Poisson models, we would like to pursue the study of these models in higher dimensions and in more detail. Another interesting avenue of research is to study the effect of the correlation structure of the covariates and the number of covariates on the sample size calculations for tests about multiple parameters. We would hope to use the method developed by Shieh (2005) which used the discrepancy between the non-central and the central chi-square approximations in our future study.

In the applied literature there is growing interest in hierarchical Poisson regression models such as in the study of longitudinal clustered data. In biomedical research the clusters could be hospitals and the patients could be studied over time and there would usually be several patient covariates such as sex, age, etc. Some interesting work on inference for these models has been done by Christiansen and Morris (1996), Tu and Piegorsch (2003), and Song (2007). For these hierarchical models, new methods of sample size calculations will be developed in order to accommodate the clustering effects.

Acknowledgments

The first author was supported by a GERAD postdoctoral fellowship and by the NSERC Discovery Grants of the second and third authors. All three authors would like to thank Claude Gravel for some of the preliminary calculations for this research. They would also like to thank the referees for their helpful comments which improved the original manuscript.

References

Blom, G., 1958. Statistical Estimates and Transformed Beta-Variables. Wiley, New York.
Christiansen, C.L., Morris, C.N., 1996. Fitting and checking a two-level Poisson model: modeling patient mortality rates in heart transplant patients. In: Berry, D., Stangl, D. (Eds.), Bayesian Biostatistics. Marcel Dekker, New York, pp. 467–501.
Golub, G.H., Van Loan, C.F., 1989. Matrix Computations, second ed. John Hopkins University Press, Baltimore.
Hall, D.B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 56 (4), 1030–1039.
Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34 (1), 1–14.
Lyles, R.H., Lin, H.-M., Williamson, J.M., 2007. A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. Statistics in Medicine 26, 1632–1648.
Matsui, S., 2005. Sample size calculations for comparative clinical trials with over-dispersed Poisson process data. Statistics in Medicine 24, 1339–1356.
McCullagh, P., Nelder, J.A., 1983. Generalized Linear Models. Chapman and Hall, London.
Self, S., Mauritsen, R.H., 1988. Power/sample size calculations for generalized linear models. Biometrics 44, 79–86.
Self, S., Mauritsen, R.H., 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., Correlated Data Analysis. In: Springer Series in Statistics, New York, 2007.
Tu, W., Piegorsch, W.W., 2003. Empirical Bayes analysis for a hierarchical Poisson generalized linear model. Journal of Statistical Planning and Inference 111, 235–248.
Whittemore, A.S., 1981. Sample size for logistic regression with small response probability. Journal of the American Statistical Association 76 (373), 27–32.
Williamson, J.M., Lin, H.-M., Lyles, R.H., Hightower, A.W., 2007. Power calculations for ZIP and ZINB models. Journal of Data Science 5, 519–534.
Yau, K.K.W., Lee, A.H., 2001. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. Statistics in Medicine 20, 2907–2920.
Yau, K.K.W., Lee, A.H., Angus, S.K.N., 2002. A zero-augmented gamma mixed model for longitudinal data with many zeros. Australian & New Zealand Journal of Statistics 44 (2), 177–183.