Les Cahiers du GERAD ISSN: 0711–2440

Power and Sample Size Calculations for Poisson and Zero-Inflated Poisson Regression Models

N. Channouf, B. MacGibbon,
M. Fredette

G–2009–88

December 2009
Revised: July 2010

Power and Sample Size Calculations for Poisson and Zero-Inflated Poisson Regression Models

Nabil Channouf

GERAD & HEC Montréal
Montréal (Québec) Canada, H3T 2A7
nabil.channouf@gerad.ca

Brenda MacGibbon

GERAD & UQAM
Montréal (Québec) Canada, H3C 3P8
macgibbon.brenda@gmail.com

Marc Fredette

GERAD & HEC Montréal
Montréal (Québec) Canada, H3T 2A7
marc.fredette@hec.ca

December 2009
Revised: July 2010

Les Cahiers du GERAD

G–2009–88

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 covariates on the sample size except for the work of Shieh (2001) where several different values of the parameters of the multinomial distribution were considered. We propose to calculate the sample size for the Wald test in the Poisson regression model, assuming that the covariates may be correlated and have a multivariate normal distribution. We have chosen the two most popular correlation structures, that is, the exchangeable and the AR(1) correlation matrices, with different values for the correlation to illustrate our work. The method used here to calculate the sample size is based on a modification of the methodology proposed by Shieh (2001). Using Monte Carlo simulations, we conclude that the sample size depends on the number of covariates for the exchangeable correlation matrix, but much more so on the correlation structure. The AR(1) correlation matrix does not exhibit these types of changes. Our methodology is also extended for the case of the Zero-inflated Poisson regression model in order to obtain analogous results there.

Key Words: Wald test; Generalized linear models; Correlation structure; AR(1); Exchangeable; Monte Carlo Simulations.

Résumé

La détermination de la taille d'échantillon adéquate pour utiliser un modèle de régression de Poisson a été étudiée lorsqu'une ou deux variables explicatives sont disponibles. Par contre, à l'exception de Shieh (2001), très peu d'attention a été donnée sur l'effet de la corrélation entre les variables explicatives sur la taille de l'échantillon. Nous proposons de calculer la taille d'échantillon requise pour effectuer un test de Wald dans un modèle de régression de Poisson avec plusieurs variables explicatives afin d'étudier cet effet. Notre méthode est une modification de l'approche de Shieh (2001). Nous considérons le cas où les variables explicatives suivent une loi multinormale de différentes dimensions et ayant une structure de corrélation de type AR(1) ou échangeable. Nous illustrons la méthodologie avec des exemples numériques en utilisant des simulations de Monte Carlo pour conclure que la taille échantillonale diffère beaucoup selon les paramètres d'une structure de corrélation échangeable mais très peu pour la structure AR(1). Nous prolongeons également notre méthode dans le cas de modèles de régression de Poisson gonflé à zéro.

Mots clés : Test de Wald ; modèles linéaires généralisés ; matrice de corrélation; AR(1); échangeable; simulations de Monte Carlo.

Acknowledgments: The first author was supported by a GERAD postdoctoral fellowship and by the NSERC Discovery Grants of the second and third author. All three authors would like to thank Claude Gravel for some of the preliminary calculations for this research.

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.

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 approximation, originally proposed by Blom (1958) for transformed beta variables, in order to do power calculations but they did not include sample size calculations.

Although Lambert (1992), Hall (2003) 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 which allows for more than one covariate that is either discrete or continuous.

We consider here the case of testing whether one parameter is zero using the Wald test by introducing a modification to the method of Shieh (2001) for sample size calculations for the Poisson regression model with more than one covariate and then extend the methodology to the zero-inflated Poisson regression model. In particular, we study the influence of the correlation structure of the 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 were able to present an extension of the 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 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 generalized linear model, in particular, 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. 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 and with an accompanying discussion of these results. In Section 5 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:

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

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:

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

where g is the link function, $\lambda = \mathrm{E}[Y]$, $X = (1, X_1, \ldots, X_p)$ is a vector of $(p+1)$ covariates with density f_X and assumed independent of the parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$. 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:

$$\log\lambda_i = X_i\beta = \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p}, \quad \text{for} \quad i = 1, \ldots, N.$$

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

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

The score function is given by:

$$S_N(\beta) = \frac{\partial l(\beta; y, x)}{\partial \beta_k} = \sum_{i=1}^{N} x_{i,k} \left\{ y_i - e^{\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}} \right\},$$

for $k = 0, 1, \ldots, p$, with $x_{i,0} = 1$. The maximum likelihood estimates of β are given by the solution of the system $S_N(\beta) = 0$, 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 X. We first calculate the estimates of the parameters $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)|_{\beta_s=0}$ under H_0, by solving the following system of equations:

$$\mathbf{E}[S_N(\beta)|_{\beta_s=0}] = 0, \tag{2}$$

where the expectation is taken with respect to the true value of $\{\beta_0, \beta_1, \ldots, \beta_s\}$. It should be noted that we have changed the system of equations used by Shieh (2001) where he used the limit of the expectation, that is, $\lim_{N\to\infty} \mathbf{E}[S_N(\beta)|_{\beta_s=0}]/N$.

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

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

where Z_q is the $100(1-q)$th percentile of the standard normal distribution, α is the size of the test, and $V(\beta_0, \beta_1, \ldots, \beta_p)$ is computed with the pre-specified values and $V(\beta_0^*, 0, \beta_2^*, \ldots, \beta_p^*)$ is calculated using (2). We have modified Shieh's method here in (3) by using size α instead of $\alpha/2$ as in Shieh (2001). (For a discussion of the reasons for this modification, see Section 4.) It should also be noted that here the pre-specified value of β_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 the inverse augmented Hessian matrix of the moment generating function. In other words, let

$$m(t_0, t_1, \ldots, t_p) = \mathbf{E}[e^{t_0+t_1X_1+\ldots+t_pX_p}] \tag{4}$$

be the moment generating function. For example, if $(X_1, \ldots, X_p)$ is a random multivariate normal vector with vector of means $\theta = (\theta_1, \ldots, \theta_p)$ and covariance matrix $\Sigma = (\sigma_{ij})$, the moment generating function m of $(X_1, \ldots, X_p)$ 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 (4), that is,

$$M(t_0, t_1, \ldots, t_p) = \begin{pmatrix} m & \frac{\partial m}{\partial t_1} & \cdots & \frac{\partial m}{\partial t_j} & \cdots & \frac{\partial m}{\partial t_p} \\ \frac{\partial m}{\partial t_1} & \frac{\partial^2 m}{\partial t_1^2} & \cdots & \frac{\partial^2 m}{\partial t_1 \partial t_j} & \cdots & \frac{\partial^2 m}{\partial t_1 \partial t_p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \frac{\partial m}{\partial t_j} & \frac{\partial^2 m}{\partial t_j \partial t_1} & \vdots & \frac{\partial^2 m}{\partial t_j^2} & \vdots & \frac{\partial^2 m}{\partial t_j \partial t_p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial m}{\partial t_p} & \frac{\partial^2 m}{\partial t_p \partial t_1} & \cdots & \frac{\partial^2 m}{\partial t_p \partial t_j} & \cdots & \frac{\partial^2 m}{\partial t_p^2} \end{pmatrix}. \tag{5}$$

If we denote by $M_{i,j}^{-1}$ 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

$$V(\beta_0, \beta_1, \ldots, \beta_p)_{2,2} = M_{2,2}^{-1}(\beta_0, \beta_1, \ldots, \beta_p). \tag{6}$$

When testing $\mathrm{H}_0 : \beta_s = 0$, against the alternative $\mathrm{H}_1 : \beta_s > 0$, for $s > 1$, the formula for the variance function in (3) is defined in an analogous fashion.

We begin by finding the "estimates" of β by finding the solution to the system of equations given in (2) 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.

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, 2003), 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 the probability that Y_i is a zero-only random variable, with $0 \le \pi_i \le 1$. The random variable Y_i follows a ZIP distribution if:

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

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 β_s, the parameter which is being tested whether or not it is equal to 0).
- The distribution of the covariates 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 $\overline{\lambda} = E[e^{X\beta}]$. (The value $\overline{\lambda}$ was chosen equal to .05 as in (Shieh, 2001)).
- An initial value N for the sample size. Empirical studies suggest that the initial value of N has little effect on the results and that $N = 100$ suffices.
- The desired number of Monte Carlo replications B.

For multivariate normal covariates, the data must be standardized so that 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 N values from the distribution of X and compute the rates $\lambda_i(X)$, $i = 1. \ldots, N$.

2. **Compute** the values $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)|_{\beta_s=0}$, under H_0, by solving the system of nonlinear equations in (2);

3. **Compute** the inverse of the augmented Hessian matrix of the moment generation function of X and take its (s+1)-th diagonal elements using (5) and (6) to obtain the variance functions $V(\beta^*)$ and $V(\beta)$ of (3);

4. **Return** the sample size $N_{s,j}$ by formula (3).
5. **Repeat** steps 1 to 4 B times, $j = 1, \ldots, B$.
6. **Take** the average value over the outputs of the B replications, $N_s = \frac{\sum_{j=1}^{B} N_{s,j}}{B}$.

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 can come from two states, the zero state with probability π_i, or the Poisson state with a probability $(1-\pi_i)e^{-\lambda_i}$, for $i = 1, \ldots, N$. In other words, Y_i equals zero with probability π_i or follows a Poisson distribution with probability $(1-\pi_i)$. Both sets of parameters λ and π are modeled by link functions in order to have linear predictors as follows:

$$\mathrm{logit}(\pi) = G_i\gamma = \gamma_0 + \gamma_1 G_{i,1} + \ldots + \gamma_q G_{i,q}; \quad i = 1, \ldots, N,$$

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

$$\log(\lambda_i) = X_i\beta = \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p}; \quad i = 1, \ldots, N,$$

with $X = (1, X_1, \ldots, X_p)$ a vector of $(p+1)$ covariates and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ 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:

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

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 (3). However, the computations become more complex, since we have additional parameters for the logistic component, namely $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_q)$. In addition, the simplification by the moment generation function cannot be applied in this case.

The score function is given by:

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

for $l = 0, 1, \ldots, q$, and $k = 0, 1, \ldots, p$. The maximum likelihood estimates of γ and β are given by the solution of the system $S_N(\gamma, \beta) = 0$, and their variance-covariance matrix $V(\gamma, \beta)$ is given by the inverse of the Fisher information matrix $I(\gamma, \beta)$:

$$I(\gamma, \beta) = \begin{pmatrix} -\mathbf{E}\left[\frac{\partial^2 l(\gamma,\beta;y,g,x,z)}{\partial\gamma\partial\gamma^T}\right] & -\mathbf{E}\left[\frac{\partial^2 l(\gamma,\beta;y,g,x,z)}{\partial\gamma\partial\beta^T}\right] \\ -\mathbf{E}\left[\frac{\partial^2 l(\gamma,\beta;y,g,x,z)}{\partial\beta\partial\gamma^T}\right] & -\mathbf{E}\left[\frac{\partial^2 l(\gamma,\beta;y,g,x,z)}{\partial\beta\partial\beta^T}\right] \end{pmatrix}. \tag{7}$$

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 $\gamma^* = (\gamma_0^*, \gamma_1^*, \ldots, \gamma_q^*)$ and $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)|_{\beta_s=0}$ under H_0, are calculated by solving the modified system of equations:

$$\mathbf{E}[S_N(\gamma, \beta)|_{\beta_s=0}] = 0, \tag{8}$$

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

$$\begin{aligned} E[Z_i|y_i, \gamma, \beta] &= P(\text{zero state}|y_i, \gamma, \beta) \\ &= \frac{P(Y_i|\text{zero state})P(\text{zero state})}{P(Y_i|\text{zero state})P(\text{zero state}) + P(Y_i|\text{Poisson state})P(\text{Poisson state})} \\ &= \begin{cases} (1 + e^{-G_i\gamma - e^{X_i\beta}})^{-1} & \text{if } y_i = 0 \\ 0 & \text{if } y_i > 0, \end{cases} \end{aligned} \tag{9}$$

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 $\mathrm{H}_0 : \beta_1 = 0$, against the alternative $\mathrm{H}_1 : \beta_1 > 0$, the formula for estimating the minimum sample size N_s is given by

$$N_s = \left\lceil \left(\frac{V(\gamma^*, \beta_0^*, 0, \beta_2^*, \ldots, \beta_p^*)_{q+3,q+3}^{1/2} Z_\alpha + V(\gamma, \beta_0, \beta_1, \ldots, \beta_p)_{q+3,q+3}^{1/2} Z_{1-Power}}{\beta_1} \right)^2 \right\rceil, \tag{10}$$

where $V(\gamma, \beta_0, \beta_1, \ldots, \beta_p)$ 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 (8).

The implementation of the sample size calculations for a significance test on any parameter β_s, $s = 1, \ldots, p$, is summarized in Algorithm 2.

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 β_s, 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 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 values of β_0 and γ_0 are chosen to satisfy overall fixed means $\overline{\lambda} = E[e^{X\beta}]$ and $\overline{\pi} = E[e^{G\gamma}/(1+e^{G\gamma})]$, respectively.
- An initial value N for the sample size. Empirical studies suggest that the initial value of N has little effect on the results and that $N = 100$ 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 N values from the distributions of X and G, and compute the rates $\lambda_i(X)$, and $\pi_i(G)$, for $i = 1\ldots, N$.
2. **Compute** the maximum likelihood values $\gamma^* = (\gamma_0^*, \gamma_1^*, \ldots, \gamma_q^*)$ and $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)|_{\beta_s=0}$, by solving the system of $(q+s+1)$ nonlinear equations in (8);
3. **Compute** the inverse of the Fisher information matrix $I(\gamma, \beta)$, given in (7), 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 (10).
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 $X = (X_1, \ldots, X_p)$, 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:

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

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

We illustrated our methods for ρ ranging between $[-0.5, 0.5]$, and for four dimensions $p = 2, 3, 4, 5$. The lower bound of this range is used to ensure a positive semidefinite correlation matrix corresponding to the covariance matrix of type V_1, and the upper bound is to 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 X of covariates normally distributed, with zero mean and covariance matrix V_l, for $l = 1, 2$. The values for the β_i, for $i = 1, \ldots, p$, were chosen equal to those of Shieh (2001), that is, equal to log(2) and β_0 was chosen as indicated in Algorithm 1.

For the ZIP regression model, 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: $X\sim$Bernoulli(0.5), and $G\sim$Bernoulli(0.5); and then the case of $X\sim$N(0,1), and $G\sim$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 $(p_1, p_2, p_3, p_4) = (0.76, 0.0.19, 0.01, 0.04)$. As in Shieh (2001), the "true" parameters were taken as follows: β_0 is a chosen to satisfy an overall mean $\bar{\lambda} = 0.05 = \mathrm{E}[e^{X\beta}]$; γ_0 is chosen to satisfy an overall mean $\bar{\pi} = 0.05 = \mathrm{E}[e^{G\gamma}/(1 + e^{G\gamma})]$, while other parameters and the difference we wish to test are set, respectively, as $\gamma_1 = \beta_2 = \log(2)$ and $\Delta = \beta_1 = \log(2)$.

In all the numerical examples, we chose to test whether $\beta_1 = 0$ and we set the difference Δ equal to β_1. The estimates of the sample size N_1, 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 equation (3) 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.

The computations were based on independent Monte Carlo simulations, with 5000 replications. For each replication, a data set consisting of covariates of sizes $N = 100$ and 1000, were sampled from the specified distribution of X (and also of G for the ZIP regression model). 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. We observed that the two initial values chosen for N had little impact on the results. In general, we observed changes of less than 3% for the Poisson model and less than 0.2% for the ZIP regression model. For this reason, we present results only for $N = 100$ here.

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 p increases and it increases substantially for $\rho \leq 0.1$ as the dimension increases. For the AR(1) covariance matrix of type V_2, it is interesting to note that there is no substantial change in the sample size when we change ρ or the dimension.

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). This was done for the standard Poisson model and the ZIP model, but only the latter results were given in Tables 3 and 4 here. 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 all cases the results have high accuracy; the error between nominal and estimated power is relatively small. In general, the error is slightly higher for power=80%, but it never exceeds 0.07 in absolute value, which is about 8.7% in term of relative error. The estimated power usually tends to underestimate the nominal power, except for the extremely skewed case with the multinomial distribution, when $(p_1, p_2, p_3, p_4) = (0.76, 0.19, 0.01, 0.04)$. For this case, the power obtained is overestimated, but the error is no higher than the other cases that appear in Shieh (2001) 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 expect to have 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 distributed as a Bernoulli (for X_1) or a 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 a little impact on the increase. When $p = 2$, the sample size required when we augment 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%.

Table 1: Sample size for Poisson regression model for the test $H_0 : \beta_1 = 0$; $H_1 : \beta_1 > 0$, with significance level 2.5%. $X \sim$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 Poisson regression model for the test $H_0 : \beta_1 = 0$; $H_1 : \beta_1 > 0$, with significance level 2.5%. $X\sim$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

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.

	Power	N_s (ZIP)	N_s (Poisson)	Est. power with N_s (ZIP)	Error	Est. size with N_s (ZIP)
$Y\sim$ZIP(λ,π) $\lambda = e^{\beta_0+\beta_1 X}$	0.9	4261	4046	0.9312	0.0312	0.0256
$\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $X\sim$Bernoulli(0.1) $G\sim$Bernoulli(0.5)	0.8	3331	3164	0.8778	0.0778	0.0290
$Y\sim$ZIP(λ,π) $\lambda = e^{\beta_0+\beta_1 X}$	0.9	1933	1823	0.9022	0.0022	0.0264
$\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $X\sim$Bernoulli(0.5) $G\sim$Bernoulli(0.5)	0.8	1428	1355	0.7778	0.0222	0.0198
$Y\sim$ZIP(λ,π) $\lambda = e^{\beta_0+\beta_1 X}$	0.9	6599	6271	0.8592	0.0408	0.0218
$\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $X\sim$Bernoulli(0.9) $G\sim$Bernoulli(0.5)	0.8	4650	4418	0.6920	0.1080	0.0208
$Y\sim$ZIP(λ,π) $\lambda = e^{\beta_0+\beta_1 X}$	0.9	453	437	0.8550	0.0450	0.0282
$\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $X\sim$N(0,1) $G\sim$Bernoulli(0.5)	0.8	339	326	0.7362	0.0638	0.0326
$Y\sim$ZIP(λ,π) $\lambda = e^{\beta_0+\beta_1 X}$	0.9	462	439	0.8796	0.0204	0.0268
$\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$ $X\sim$N(0.5,1) $G\sim$Bernoulli(0.5)	0.8	345	328	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	N_s (ZIP)	N_s (Poisson)	Est. power with N_s (ZIP)	Error	Est. size with N_s (ZIP)
$Y \sim$ZIP(λ, π) $\lambda = e^{\beta_0+\beta_1 X_1+\beta_2 X_2}$ $\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$	0.9	1932	1835	0.8940	0.0060	0.0228
$(X_1, X_2)\sim$Multinomial(p) $p = (0.25, 0.25, 0.25, 0.25)$ $G\sim$Bernoulli(0.5)	0.8	1428	1356	0.7860	0.0140	0.0244
$Y \sim$ZIP(λ, π) $\lambda = e^{\beta_0+\beta_1 X+\beta_2 X_2}$ $\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$	0.9	3138	2927	0.8748	0.0252	0.0236
$(X_1, X_2)\sim$Multinomial(p) $p = (0.4, 0.1, 0.1, 0.4)$ $G\sim$Bernoulli(0.5)	0.8	2297	2184	0.7578	0.0422	0.0226
$Y \sim$ZIP(λ, π) $\lambda = e^{\beta_0+\beta_1 X+\beta_2 X_2}$ $\pi = \frac{e^{\gamma_0+\gamma_1 G}}{1+e^{\gamma_0+\gamma_1 G}}$	0.9	5972	5661	0.9278	0.0278	0.0326
$(X_1, X_2)\sim$Multinomial(p) $p = (0.76, 0.19, 0.01, 0.04)$ $G \sim$Bernoulli(0.5)	0.8	4633	4390	0.8514	0.0514	0.0256

5 Conclusion

Our objective here was to study the effect of the correlation structure 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 adapted 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 covariates on the sample size for two different types of correlation matrices. There is clear evidence that the correlation structure plays an important role in sample size determination for the exchangeable correlation matrix. We have shown these effects using simple correlation structures 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 also observed differences in sample size, not only as the correlation changes but also as the number of additional covariates increases for the exchangeable correlation matrix studied here.. We conclude that all of these aspects 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 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.

References

Blom, G. (1958). *Statistical Estimates and Transformed Beta-Variables*. Wiley, New York.

Christiansen, C. L. and Morris, C. N. (1996). Fitting and checking a two-level Poisson model: Modeling patient mortality rates in heart transplant patients. In Berry, D. and Stangl, D., editors, *Bayesian Biostatistics*, pages 467–501. Marcel Dekker, New York.

Golub, G. H. and Van Loan, C. F. (1989). *Matrix Computations*. John Hopkins University Press, Baltimore, second edition.

Hall, D. B. (2003). Zero-inflated Poisson and binomial regression with random effects: a case study. *International Journal for Quality in Health Care*, 15(4):319–329.

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., and 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. and Nelder, J. A. (1983). *Generalized Linear Models*. Chapman and Hall, London.

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.

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., and Hightower, A. W. (2007). Power calculations for ZIP and ZINB models. *Journal of Data Science*, 5:519–534.

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.

Yau, K. K. W., Lee, A. H., and 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.