Communications for Statistical Applications and Methods
2024, Vol. 31, No. 1, 55–64

https://doi.org/10.29220/CSAM.2024.31.1.055
Print ISSN 2287-7843 / Online ISSN 2383-4757

Sample size calculations for clustered count data based on zero-inflated discrete Weibull regression models

Hanna Yoo[1,a]

[a]Department of Applied Statistics, Hanshin University, Korea

Abstract

In this study, we consider the sample size determination problem for clustered count data with many zeros. In general, zero-inflated Poisson and binomial models are commonly used for zero-inflated data; however, in real data the assumptions that should be satisfied when using each model might be violated. We calculate the required sample size based on a discrete Weibull regression model that can handle both underdispersed and overdispersed data types. We use the Monte Carlo simulation to compute the required sample size. With our proposed method, a unified model with a low failure risk can be used to cope with the dispersed data type and handle data with many zeros, which appear in groups or clusters sharing a common variation source. A simulation study shows that our proposed method provides accurate results, revealing that the sample size is affected by the distribution skewness, covariance structure of covariates, and amount of zeros. We apply our method to the pancreas disorder length of the stay data collected from Western Australia.

Keywords: covariance structure, clustered count data, discrete Weibull regression, Monte Carlo simulations, sample size determination

1. Introduction

Clustered count data with an excess number of zeros are very common in the biomedical field, clinical studies, and health research. Zero-inflated models are a natural choice for this data type, with some studies utilizing these models. Hall (2000) first introduced random effects into a portion of the zero-inflated Poisson and binomial models. Tapak *et al*. (2019) extended the zero-inflated exponentiated-exponential geometric regression in the presence of a random effect. Furthermore, Choo-Wosoba *et al*. (2018) proposed a Bayesian approach by combining the Conway–Maxwell–Poisson distribution with a hurdle component.

Aside from proper data analysis, sample size determination before planning any statistical study is also a crucial part. Studies considering the sample size calculation for clustered count data with excessive zeros are scarce. Channouf *et al*. (2021) recently proposed sample size calculation methods for hierarchical Poisson and zero-inflated Poisson regression models by extending the work of Shieh (2001) and using the *h*-likelihood method with the Monte Carlo simulation to calculate the required sample size. In the present study, we calculate the required sample size based on a discrete Weibull model first introduced by Nakagawa and Osaki (1975). The discrete Weibull distribution can handle both overdispersion and underdispersion; thus, there is a low risk of failing to cope with the features of the dispersed data type. Yoo (2023) proposed a sample size calculation method for the clustered

[1]Department of Applied Statistics, Hanshin University, 137 Hanshindae-gil, Osan-si, Gyeonggi-do 18101, Korea. Email: yoohann@hs.ac.kr.

Published 31 January 2024 / journal homepage: http://csam.or.kr

discrete Weibull regression model, which we expand herein to the zero-inflated data. To date, no method for the sample size calculation dealing with clustered count data with the zero-inflated discrete Weibull regression model has yet been reported. In this study, we modify the method of Channouf *et al*. (2021) to calculate the required sample size for the zero-inflated discrete Weibull regression model in the presence of clustered data. Perumean-Chaney *et al*. (2013) showed that ignoring the overdispersion within the zero-inflated data inflates the Type 1 error and results in poor estimates. Thus, considering both dispersion and zero inflation when calculating the sample size is meaningful.

The remainder of this paper is organized as follows: Section 2 presents the proposed sample size calculation method for clustered zero-inflated discrete Weibull (ZI-DW) model; Section 3 elaborates on the simulation study conducted to examine the method performance; Section 4 depicts an illustrative example; and Section 5 discusses the results and concludes this work.

2. Sample size for the clustered zero-inflated discrete Weibull regression model

For the clustered data with an excessive number of zeros, we expanded the ZI-DW model with a random effect. Let Y_{ij} denote the j^{th} subject measured for cluster $i, i = 1, \ldots, n$; $j = 1, \ldots, n_i$ with the total number of subjects in the data, $N = \sum_{i=1}^{n} n_i$. We assumed Y_{ij} to follow the DW distribution:

$$Y_{ij} \sim \mathrm{DW}\left(q(X_{ij}), \beta\right), \tag{2.1}$$

where $\mathbf{X}'_{ij} = (1, X_{1,ij}, \ldots, \mathbf{X}_{p,ij})$ is a covariate vector. The probability mass function for the DW is

$$f_{DW}\left(Y_{ij}; q(\mathbf{X}_{ij}), \beta\right) = q(\mathbf{X}_{ij})^{Y_{ij}{}^{\beta}} - q(\mathbf{X}_{ij})^{(Y_{ij}+1)^{\beta}}, \tag{2.2}$$

for $Y_{ij} = 0, 1, 2, 3, \ldots, 0 < q(X_{ij}) < 1, \beta > 0$. Parameter $q(X_{ij})$ is the probability of the outcome random variable Y_{ij} having a value greater than 0 and parameter β controlling the value range of Y_{ij}. The distribution became highly skewed as β converged to 0. The DW distribution approached the Bernoulli distribution as β converged to ∞. We incorporated a random intercept into the model to consider the subject correlation in the same cluster. The clustered DW regression model is presented as follows:

$$\log\left(-\log(q(X_{ij}))\right) = \boldsymbol{\theta}'\mathbf{X}_{ij} + U_{0,i} = \theta_0 + \theta_1 X_{1,ij} + \theta_2 X_{2,ij} + \cdots + \theta_p X_{p,ij} + U_{0,i}, \tag{2.3}$$

where $\boldsymbol{\theta}' = (\theta_0, \theta_1, \theta_2, \ldots, \theta_p)$ denotes the corresponding $(p + 1)$ regression coefficients associated with $\mathbf{X}'_{ij} = (1, X_{1,ij}, \ldots, X_{p,ij})$, and $U_{0,i}$ is a random effect following an independently and identically distributed normal distribution with mean 0 and a common dispersion parameter σ_0^2.

In the presence of clustered data with an excessive number of zeros, we extended the DW model with a zero-inflation parameter. The clustered ZI-DW had the following probability mass function:

$$f_{ZI-DW}\left(Y_{ij}, \boldsymbol{\gamma}, q(\mathbf{X}_{ij}), \beta\right) = \begin{cases} \left(\pi(\mathbf{Z}_{ij}) + \left(1 - \pi(\mathbf{Z}_{ij})\right) f_{DW}\left(Y_{ij}, q(\mathbf{X}_{ij}), \beta\right)\right) & \text{for } y = 0, \\ \left(1 - \pi(\mathbf{Z}_{ij})\right) f_{DW}\left(Y_{ij}, q(\mathbf{X}_{ij}), \beta\right) & \text{for } y = 1, 2, 3, \ldots, \end{cases} \tag{2.4}$$

where $0 < \pi < 1$ is the zero-inflation parameter, and $f_{DW}(Y_{ij}, q(\mathbf{X}_{ij}), \beta)$ is the probability mass function shown in Equation (2.2). The logit link for π is usually used:

$$\mathrm{logit}\left(\pi(\mathbf{Z}_{ij})\right) = \boldsymbol{\gamma}'\mathbf{Z}_{ij} = \gamma_0 + \gamma_1 Z_{1,ij} + \cdots + \gamma_q Z_{q,ij}, \tag{2.5}$$

where $Z_{ij}{}' = (1, Z_{1,ij}, Z_{2,ij}, \ldots, Z_{q,ij})'$ is a vector with q covariates, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_q)'$ is the corresponding coefficient vector. A binary indicator variable δ_{ij}, which is $\delta_{ij} = 1$ if $Y_{ij} = 0$ and $\delta_{ij} = 0$ if $Y_{ij} > 0$, is used for the likelihood of the ZI-DW regression model. We estimated the parameters by using the *h*-likelihood first introduced by Lee and Nelder (1996). It extended the penalized quasi-likelihood to estimate the parameters and was based on the joint distribution of the outcome variable and the random effect. Unlike the classical likelihood, the *h*-likelihood is constructed for both fixed parameters and unobserved frailties (Lee *et al*., 2017). The *h*-likelihood avoids such integration itself and provides the inference for unobservable random effects (Ha *et al*., 2001). A big merit of *h*-likelihood is that once the statistical model is specified, the required inference procedures can be made: See (Jin and Lee, 2020) for recent review of *h*-likelihood.

Given a random sample $(Y_{ij}, X_{ij}, Z_{ij}, \delta_{ij})$, the log likelihood was obtained as follows:

$$
\begin{aligned}
l = & \sum_{i=1}^{n}\sum_{j=1}^{n_i} \delta_{ij} \log\left[\left(e^{-z'_{ij}\gamma}+1\right)^{-1} + \left[1-\left(e^{-z'_{ij}\gamma}+1\right)^{-1}\right]\left(1-e^{-e^{x'_{ij}\theta+U_{0,i}}}\right)\right] + \\
& \sum_{i=1}^{n}\sum_{j=1}^{n_i} \left(1-\delta_{ij}\right) \log\left[\left[1-\left(e^{-z'_{ij}\gamma}+1\right)^{-1}\right] + \left(e^{-y_{ij}{}^{\beta}e^{e^{x'_{ij}\theta+U_{0,i}}}} - e^{-(y_{ij}+1)^{\beta}e^{e^{x'_{ij}\theta+U_{0,i}}}}\right)\right] + \\
& \sum_{i=1}^{n}\sum_{j=1}^{n_i} \log f_{X,Z}\left(X_{ij}, z_{ij}\right) + \sum_{i=1}^{n}\left\{-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_0^2\right) - \frac{1}{2}u_{0,i}^2\right\}. \qquad (2.6)
\end{aligned}
$$

The score function is given as

$$
S_{n,k,l}\left(\boldsymbol{\theta}, \boldsymbol{\gamma}, \beta, \mathbf{u_0}\right) = \begin{cases} \frac{\partial l\left(y,x,\boldsymbol{\theta},\boldsymbol{\gamma},\beta\right)}{\partial\theta_0} \\ \frac{\partial l\left(y,x,\boldsymbol{\theta},\boldsymbol{\gamma},\beta\right)}{\partial\theta_k} \\ \frac{\partial l\left(y,x,\boldsymbol{\theta},\boldsymbol{\gamma},\beta\right)}{\partial\gamma_0} \\ \frac{\partial l\left(y,x,\boldsymbol{\theta},\boldsymbol{\gamma},\beta\right)}{\partial\gamma_l} \\ \frac{\partial l\left(y,x,\boldsymbol{\theta},\boldsymbol{\gamma},\beta\right)}{\partial\beta} \\ \frac{\partial l\left(y,x,\boldsymbol{\theta},\boldsymbol{\gamma},\beta\right)}{\partial u_{0,i}} \end{cases} \quad \text{for } k = 1,\ldots,p; l = 1,\ldots,q; i = 1,\ldots,n, \qquad (2.7)
$$

where

$$
\begin{aligned}
\frac{\partial l\left(y,x,\boldsymbol{\theta},\boldsymbol{\gamma},\beta\right)}{\partial\theta_0} = & \sum_{i=1}^{n}\sum_{j=1}^{n_i} \frac{\delta_{ij}\left[1-\left(e^{-z'_{ij}\gamma}+1\right)^{-1}\right]}{wzl_{ij}\left(\boldsymbol{\gamma},\theta_k\right)}\left[e^{x_{ij}{}'\theta+u_{o,i}-e^{x_{ij}{}'\theta+u_{0,i}}}\right] + \\
& \sum_{i=1}^{n}\sum_{j=1}^{n_i} \frac{\left(\delta_{ij}-1\right)\left[y_{ij}^{\beta}e^{-y_{ij}^{\beta}e^{x_{ij}{}'\theta+u_{0,i}}} - (y_{ij}+1)^{\beta}e^{-(y_{ij}+1)^{\beta}e^{x_{ij}{}'\theta+u_{o,i}}}\right]}{wzl_{ij}\left(\boldsymbol{\gamma},\theta_k\right)},
\end{aligned}
$$

$$\frac{\partial l\left(y,x,\boldsymbol{\theta},\boldsymbol{\gamma},\beta\right)}{\partial\theta_k}=\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\delta_{ij}\left[1-\left(e^{-z'_{ij}\gamma}+1\right)^{-1}\right]}{wzl_{ij}\left(\boldsymbol{\gamma},\theta_k\right)}\left[x_{ik}e^{x_{ij}{}'\theta+u_{0,i}-e^{x_{ij}{}'\theta+u_{0,i}}}\right]+$$
$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\left(\delta_{ij}-1\right)x_{ik}\left[y_{ij}^{\beta}e^{-y_{ij}^{\beta}e^{x_{ij}{}'\theta+u_{0,i}}}-(y_{ij}+1)^{\beta}e^{-(y_{ij}+1)^{\beta}e^{x_{ij}{}'\theta+u_{0,i}}}\right]}{wzl_{ij}\left(\boldsymbol{\gamma},\theta_k\right)}$$

$$\text{with } wzl_{ij}\left(\boldsymbol{\gamma},\boldsymbol{\theta}\right)=\left(e^{-z'_{ij}\gamma}+1\right)^{-1}+\left[1-\left(e^{-z'_{ij}\gamma}+1\right)^{-1}\right]\left(1-e^{-e^{x'_{ij}\theta+u_{0,i}}}\right)\quad\text{and}$$

$$\frac{\partial l\left(y,x,\boldsymbol{\theta},\boldsymbol{\gamma},\beta\right)}{\partial\boldsymbol{\gamma}_0}=\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{e^{z'_{ij}\gamma_l}\left[\left(\delta_{ij}-1\right)e^{e^{x'_{ij}\theta+u_{0,i}}}\left(e^{z'_{ij}\gamma_l}+1\right)+1\right]}{\left(e^{z'_{ij}\gamma_l}+1\right)\left(e^{z'_{ij}\gamma_l+e^{e^{x'_{ij}\theta+u_{0,i}}}}+e^{e^{x'_{ij}\theta+u_{0,i}}}-1\right)},$$

$$\frac{\partial l\left(y,x,\boldsymbol{\theta},\boldsymbol{\gamma},\beta\right)}{\partial\boldsymbol{\gamma}_l}=\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\gamma_l e^{z'_{ij}\gamma_l}\left[\left(\delta_{ij}-1\right)e^{e^{x'_{ij}\theta+u_{0,i}}}\left(e^{e^{z'_{ij}\gamma_l}+1}\right)+1\right]}{\left(e^{z'_{ij}\gamma_l}+1\right)\left(e^{z'_{ij}\gamma_l+e^{x'_{ij}\theta+u_{0,i}}}+e^{e^{x'_{ij}\theta+u_{0,i}}}-1\right)},$$

$$\frac{\partial l\left(y,x,\boldsymbol{\theta},\boldsymbol{\gamma},\beta\right)}{\partial\beta}=\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\left(\delta_{ij}-1\right)e^{x_{ij}{}'\theta}\left[y_{ij}^{\beta}\log(y_{ij})e^{-(y_{ij}+1)^{\beta}e^{x'_{ij}\theta+u_{0,i}}}-y_{ij}^{\beta}\log(y_{ij}+1)e^{-(y_{ij})^{\beta}e^{x'_{ij}\theta+u_{0,i}}}\right]}{e^{-\left(y_{ij}+1\right)^{\beta}e^{x'_{ij}\theta+u_{0,i}}}-e^{-y_{ij}^{\beta}e^{x'_{ij}\theta+u_{0,i}}}}$$

$$\frac{\partial l\left(y,x,\boldsymbol{\theta},,\boldsymbol{\gamma},\beta\right)}{\partial u_{0,i}}=\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\left(\delta_{ij}-1\right)\left[y_{ij}^{\beta}e^{-y_{ij}{}^{\beta}e^{x'_{ij}\theta+u_{0,i}}}-\left(y_{ij}+1\right)^{\beta}e^{-(y_{ij}+1)^{\beta}e^{x'_{ij}\theta+u_{0,i}}}\right]}{e^{-y_{ij}^{\beta}e^{x'_{ij}\theta+u_{0,i}}}-e^{-(y_{ij}+1)^{\beta}e^{x'_{ij}\theta+u_{0,i}}}}+$$
$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\delta_{ij}\left[1-\left(e^{-z'_{ij}\gamma}+1\right)^{-1}\right]\left(e^{-e^{x_{ij}{}'\theta+u_{0,i}}}\right)}{wzl_{ij}\left(\boldsymbol{\gamma},\theta_k\right)}-u_{0,i}.$$

The parameters $\boldsymbol{\theta},\boldsymbol{\gamma},\beta$ and $\mathbf{u_0}$ can be obtained by solving $S_{n,k,l}(\boldsymbol{\theta},\boldsymbol{\gamma},\beta,\mathbf{u_0})=0$. Parameter σ_0^2 was estimated through the second-stage procedure (Lee and Nelder, 1996). The variance–covariance matrix $V(\boldsymbol{\theta},\boldsymbol{\gamma},\beta,\mathbf{u_0})$ of the maximum h-likelihood estimators was obtained as follows from the expected inverse of the Fisher information matrix:

$$V\left(\boldsymbol{\theta},\boldsymbol{\gamma},\beta,\boldsymbol{u_0}\right)=E\left(-\frac{\partial^2 l}{\partial(\boldsymbol{\theta},\boldsymbol{\gamma},\beta,\boldsymbol{u}_o)}\right).\tag{2.8}$$

We restricted our scope to the case with the number of clusters fixed in advance and with an equal cluster size. We calculated the required sample size by modifying the methods of Channouf *et al.* (2014), which used a conditional expectation on the covariates using a Monte Carlo simulation for sample generation. The following hypothesis was tested:

$$H_0:\theta_1=0\quad\text{vs.}\quad H_1:\theta_1\neq 0.\tag{2.9}$$

Table 1: Calculated sample size for testing $H_0 : \theta_1 = 0$ vs. $H_1 : \theta_1 = -0.7$ at various excess zero probabilities and parameter settings under nominated powers of 80 and 90% with five clusters

π	β	σ	80%		90%	
			N	Est power	N	Est power
10%	0.5	0.3	51	0806	69	0.900
		2	58	0.809	77	0.911
	2	0.3	59	0.805	78	0.900
		2	63	0.795	84	0.910
30%	0.5	0.3	70	0.795	93	0.897
		2	78	0.791	104	0.900
	2	0.3	78	0.806	104	0.882
		2	85	0.788	114	0.900
50%	0.5	0.3	98	0.791	132	0.913
		2	108	0.811	144	0.910
	2	0.3	111	0.814	149	0.902
		2	113	0.787	158	0.887

The sample size was calculated based on four steps. In the first step, the maximum likelihood estimates of parameters $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)', \boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_q)', \beta$, and $\mathbf{u_0} = (u_{0,1}, \ldots, u_{0,n})'$ under the null hypothesis were obtained by solving the score function expectation as follows:

$$E\left[\left(\theta_0, 0, \ldots, \theta_p, \boldsymbol{\gamma}, \beta, \mathbf{u_0}\right) | \; \boldsymbol{X} = x, \boldsymbol{Z} = z\right] = 0. \tag{2.10}$$

The parameters obtained from Equation (2.10) are denoted as $\boldsymbol{\theta}^* = (\theta_0^*, 0, \theta_2^*, \ldots, \theta_p^*)', \boldsymbol{\gamma}^* = (\gamma_0^*, \gamma_1^*, \ldots, \gamma_q^*)', \beta^*$, and $\mathbf{u_0}^* = (u_{0,1}^*, \ldots, u_{0,n}^*)$. In the second step, the parameter variance was calculated under the null and alternative hypothesis. These variances, which were secondary diagonal elements, were obtained through the inverse of the Fisher information matrix and denoted as $V(\theta_0^*, 0, \theta_2^*, \ldots, \theta_p^*, \gamma_0^*, \gamma_1^*, \ldots, \gamma_q^*, \beta^*, u_{0,1}^*, \ldots, u_{0,n}^*)_{(2,2)}$ and $V(\theta_0, \theta_1, \theta_2, \ldots, \theta_{p,}, \gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_q, \beta, u_{0,1}, \ldots, u_{0,n})_{(2,2)}$.

In the third step, we calculated the sample size at each j^{th} $(j = 1, \ldots, B)$ Monte Carlo replication as follows:

$$N_j' = \left[\left(\frac{\sqrt{V\left(\theta_0^*, 0, \theta_2^*, \ldots, \theta_p^*, \gamma_0^*, \gamma_1^*, \ldots, \gamma_q^*, \beta^*, u_{0,1}^*, \ldots, u_{0,n}^*\right)_{(2,2)}} Z_{\alpha/2} + \sqrt{V\left(\theta_0, \theta_1, \theta_2, \ldots, \theta_{p,}, \gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_q, \beta, u_{0,1}, \ldots, u_{0,n}\right)_{(2,2)}} Z_r}{\theta_1}\right)^2\right], \tag{2.11}$$

where $Z_{\alpha/2}$ is the $100(1 - (\alpha/2))^{th}$ percentile of the standard normal distribution, and $1 - \gamma$ is power.

In the final step, the following sample size was used to test the hypothesis $H_0 : \theta_1 = 0$ against the two-sided alternative hypothesis $H_1 : \theta_1 \neq 0$ with a specified significance level α and power $1 - \gamma$:

$$N_{ZI-DW} = \frac{\sum_{j=1}^{B} N_{j'}}{B}. \tag{2.12}$$

3. Simulation study

We conducted a simulation study to assess the performance of our proposed formula for the sample size calculations. We calculated the required sample size for the 80% and 90% powers for each parameter setting with various cluster numbers. We then compared the true nominal power with the estimated power computed through a Monte Carlo simulation based on 1,000 independent datasets given the sample size. We considered the case of one and two covariates. In the one covariate case,

Table 2: Calculated sample size for testing $H_0 : \theta_1 = 0$ vs. $H_1 : \theta_1 = -0.7$ at various excess zero probabilities and parameter settings under nominated powers of 80 and 90% with 10 clusters

π	β	σ	80%		90%	
			N	Est power	N	Est power
10%	0.5	0.3	26	0.787	34	0.900
		2	28	0.800	36	0.900
	2	0.3	29	0.793	38	0.908
		2	30	0.806	39	0.894
30%	0.5	0.3	35	0.800	46	0.887
		2	37	0.805	49	0.890
	2	0.3	39	0.814	52	0.913
		2	41	0.789	54	0.888
50%	0.5	0.3	46	0.805	62	0.905
		2	51	0.808	66	0.895
	2	0.3	54	0.800	72	0.911
		2	56	0.797	74	0.891

Table 3: Calculated sample size for testing $H_0 : \theta_1 = 0$ vs. $H_1 : \theta_1 = -0.7$ at various excess zero probabilities and parameter settings under nominated powers of 80 and 90% with 20 clusters

π	β	σ	80%		90%	
			N	Est power	N	Est power
10%	0.5	0.3	13	0.804	17	0.905
		2	14	0.800	18	0.895
	2	0.3	14	0.810	19	0.900
		2	15	0.813	20	0.913
30%	0.5	0.3	18	0.800	23	0.906
		2	19	0.813	24	0.900
	2	0.3	20	0.809	26	0.900
		2	21	0.804	27	0.910
50%	0.5	0.3	24	0.812	32	0.916
		2	25	0.808	33	0.900
	2	0.3	26	0.792	35	0.910
		2	27	0.805	36	0.902

we assumed a uniform distribution $X \sim U(0, 1)$ and set the parameters to $\theta_0 = -1, \theta_1 = -0.7$ each to model the $q(\boldsymbol{X}_{ij})$ function. The random effect was assumed to follow a normal distribution with 0 mean and $\sigma_0^2, U_{0,i} \sim N(0\ \sigma_0^2)$ variance. We observed the effect of the random-effect variance by considering two different values of $\sigma_0 = 0.3, 2$. Moreover, we considered three different values for the zero-inflated parameter $\pi = 10\%, 30\%, 50\%$ for each representing the case of a rather small amount of zeros, a moderate amount of zeros, and a large amount of zeros in zero-state data. The Bernoulli distribution $Z \sim B(0.5)$ was assumed for the corresponding covariate of the zero-inflated model. Parameters γ_0 and γ_1 satisfied each corresponding π value. For the dispersion parameter, we set two different values as $\beta = 0.5, 2$ to determine how the sample size is affected by the distribution skewness. Table 1 presents the required sample sizes for testing $H_0 : \theta_1 = 0$ vs. $H_1 : \theta_1 = -0.7$ with a significance level of $\alpha = 0.05$ and powers of $1 - \gamma = 0.8$ and 0.9. Along with Table 1, Tables 2 and 3 show the results for the 5, 10, and 20 cluster numbers, respectively. The estimated power calculated based on the sample size formula for each covariate X distribution was close to the true nominated power, showing that our proposed sample size methodology was very accurate. The sample generally required increases as the dispersion parameter β increased. The increased dispersion parameter decreased the conditional mean value of the response variable, thereby requiring a larger sample size. The bigger variance of the random effect demonstrated a sample size increase. In each

Table 4: Calculated sample size for testing $H_0 : \theta_1 = 0$ vs. $H_1 : \theta_1 = 0.2$ at various correlations of the two covariates and parameter settings under the nominated powers of 80% and 90% with five clusters

ρ	π	β	σ	80%		90%	
				N	Est power	N	Est power
0.3	10%	0.5	0.3	58	0.793	78	0.888
			2	66	0.808	88	0.895
		2	0.3	68	0.812	92	0.904
			2	72	0.797	96	0.900
	30%	0.5	0.3	80	0.788	107	0.890
			2	88	0.805	118	0.900
		2	0.3	94	0.810	125	0.900
			2	98	0.794	131	0.900
	50%	0.5	0.3	111	0.782	148	0.893
			2	127	0.800	169	0.913
		2	0.3	136	0.813	183	0.909
			2	142	0.816	190	0.905
0.7	10%	0.5	0.3	101	0.813	139	0.913
			2	116	0.796	157	0.882
		2	0.3	115	0.800	160	0.887
			2	127	0.810	174	0.888
	30%	0.5	0.3	140	0.805	190	0.911
			2	157	0.811	216	0.903
		2	0.3	161	0.792	225	0.913
			2	173	0.790	240	0.900
	50%	0.5	0.3	199	0.790	263	0.912
			2	225	0.788	299	0.890
		2	0.3	223	0.820	313	0.900
			2	233	0.800	335	0.893

table, the required sample size for the fixed β and σ_0 increased by 30% when $\pi = 30\%$ and up to 100% when $\pi = 50\%$ compared to when $\pi = 10\%$ in average. The simulation study revealed that the required sample size was affected not only by the data skewness, but also by the zero-inflated parameter.

We held a simulation study with two covariates because the real data may contain more than one covariate. We assumed two covariates X_1 and X_2 following a bivariate normal distribution with mean $\mu' = (0, 0)$ and a variance–covariance matrix as $\sum = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ with $\rho = 0.3$ and 0.7, respectively. We set parameters $\theta_0 = -1, \theta_1 = 0.2$, and $\theta_2 = 0.2$. Tables 4–6 present the required sample sizes for testing $H_0 : \theta_1 = 0$ vs. $H_1 : \theta_1 = 0.2$ with a significance level of $\alpha = 0.05$ and powers of $1 - \gamma = 0.8$ and 0.9. Similar to the case with one covariate, the calculated power based on the calculated sample size was close to the true power. The effects of the parameters on the sample size were similar to those of one covariate. As regards the correlation effect between the covariates, a larger correlation ρ of the two covariates increased the required samples size. The sample size increased by 70% when $\rho = 0.7$ compared to when $\rho = 0.3$ in average.

4. Illustrative example

We applied our sample size calculation method to a real dataset. Yau *et al*. (2003) analyzed the pancreas disorder length of stay (LOS) data collected from 261 patients in Western Australia. The patients were collected from 36 different public hospitals. Yau *et al*. (2003) studied the factors associated with the patient LOS. The LOS distribution was highly skewed. The number of zeros was 17%. Among many risk factors affecting the LOS, the admission status (elective or emergency) was one of the factors associated with the LOS. We referred to their results to set the parameter

Table 5: Calculated sample size for testing $H_0 : \theta_1 = 0$ vs. $H_1 : \theta_1 = 0.2$ at various correlations of the two covariates and parameter settings under the nominated powers of 80 and 90% with 10 clusters

ρ	π	β	σ	80%		90%	
				N	Est power	N	Est power
0.3	10%	0.5	0.3	30	0.800	40	0.917
			2	33	0.787	43	0.897
		2	0.3	34	0.800	46	0.913
			2	36	0.805	48	0.886
	30%	0.5	0.3	41	0.810	53	0.912
			2	44	0.793	58	0.900
		2	0.3	48	0.805	62	0.906
			2	49	0.797	64	0.897
	50%	0.5	0.3	56	0.794	72	0.905
			2	60	0.800	79	0.912
		2	0.3	63	0.812	85	0.906
			2	67	0.790	88	0.905
0.7	10%	0.5	0.3	51	0.788	68	0.907
			2	58	0.806	77	0.915
		2	0.3	59	0.800	79	0.908
			2	64	0.795	84	0.894
	30%	0.5	0.3	71	0.806	92	0.912
			2	79	0.800	103	0.888
		2	0.3	83	0.816	108	0.884
			2	87	0.800	114	0.890
	50%	0.5	0.3	97	0.813	124	0.915
			2	109	0.796	141	0.912
		2	0.3	109	0.800	146	0.895
			2	120	0.805	156	0.900

values needed to calculate the required sample size. Furthermore, we considered a single covariate $Z, X \sim B(0.5)$ assumed to be the admission status to model the ZI-DW regression model. The DW parameters were set as $\theta_0 = -1.00, \theta_1 = -0.52$, and $\beta = 1.1$. The zero-inflation part parameters were set as $\gamma_0 = -2.3, \gamma_1 = -6.0$. The standard deviation of the random intercept was set as 0.9. Based on our sample size calculation, four children were required in each cluster. As shown in Table 7, based on our proposed sample size method 144 patients were required under 80% power (i.e., 180 patients were required to acquire 90% power), which was only 55% of the data (261) that were actually analyzed.

5. Conclusion

In this study, we proposed a sample size calculation method for the clustered ZI-DW regression model. A random intercept was incorporated into a regression model. A Monte Carlo simulation was used to calculate the conditional expectation of the score function and the parameter variances under the null and alternative hypothesis. Our proposed sample size calculation method showed accurate results in the conducted simulation study. The sample size was affected by the distribution skewness, cluster number, and random-effect variance. In addition, the correlation between the variables and the amount of zeros in the data was found to affect the required sample size.

We believe that the proposed sample size method based on the clustered ZI-DW regression model is a suitable choice because it can handle both overdispersed and underdispersed data types at the presence of an excessive number of zeros, which may have been caused by a subpopulation with only zero counts. In this work, our method was restricted to the case with an equal cluster size. Thus, the future study should perform a sample size calculation that can handle an unequal cluster size and

Table 6: Calculated sample size for testing $H_0 : \theta_1 = 0$ vs. $H_1 : \theta_1 = 0.2$ at various correlations of the two covariates and parameter settings under the nominated powers of 80 and 90% with 20 clusters

ρ	π	β	σ	80%		90%	
				N	Est power	N	Est power
0.3	10%	0.5	0.3	14	0.812	19	0.900
			2	16	0.800	21	0.895
		2	0.3	17	0.806	23	0.913
			2	18	0.800	23	0.895
	30%	0.5	0.3	20	0.795	27	0.900
			2	22	0.803	29	0.903
		2	0.3	24	0.815	31	0.893
			2	28	0.812	32	0.906
	50%	0.5	0.3	27	0.813	36	0.908
			2	29	0.806	39	0.890
		2	0.3	31	0.800	42	0.904
			2	32	0.804	43	0.900
0.7	10%	0.5	0.3	25	0.788	33	0.913
			2	28	0.812	38	0.905
		2	0.3	29	0.813	38	0.908
			2	31	0.795	41	0.890
	30%	0.5	0.3	35	0.794	46	0.912
			2	38	0.803	51	0.890
		2	0.3	40	0.800	53	0.895
			2	42	0.797	56	0.903
	50%	0.5	0.3	47	0.816	62	0.910
			2	52	0.813	69	0.906
		2	0.3	55	0.806	72	0.910
			2	58	0.794	76	0.900

Table 7: Sample size results of the pancreas disorder length of stay (LOS) data

Actual sample size	Required sample size	
	80% power	90% power
261	144	180

extend the model to a Bayesian approach, which offers more flexibility in model specification and uses prior information.

Funding

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1F1A1053119).

References

Channouf N, Fredette M, and MacGibbon B (2014). Power and sample size calculations for poisson and zero inflated poisson regression models, *Computational Statistics & Data Analysis*, **72**, 241–251.

Channouf N, Fredette M, and MacGibbon B (2021). Sample size calculations for hierarchical Poisson and zero-inflated Poisson regression models, *Communications in Statistics – Simulation and Computation*, **50**, 937–956.

Choo-Wosoba H, Gaskins J, Levy S, and Datta S (2018). A Bayesian approach for analyzing zero-

inflated clustered count data with dispersion, *Statistics in Medicine*, **37**, 801–812.

Ha ID, Lee Y, and Song JK (2001). Hierarchical likelihood approach for frailty models, *Biometrika*, **88**, 233–243.

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

Jin S and Lee Y (2020). A review of *h*-likelihood and hierarchical generalized linear model, *WIREs Computational Statistics*, **13**, e1527.

Lee Y and Nelder JA (1996). Hierarchical generalized linear models, *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 619–678.

Lee M, Ha ID, and Lee Y (2017). Frailty modeling for clustered competing risks data with missing cause of failure, *Statistical Methods in Medical Research*, **26**, 356–373.

Nakagawa T and Osaki S (1975). The discrete Weibull distribution, *IEEE Transactions on Reliability*, **24**, 300–301.

Perumean-Chaney SE, Morgan C, McDowall D, and Aban I (2013). Zero-inflated and overdispersed: What's one to do?, *Journal of Statistical Computation and Simulation*, **83**, 1671–1683.

Shieh G (2001). Sample size calculations for logistic and Poisson regression models, *Biometrika*, **88**, 1193–1199.

Tapak L, Hamidi O, Amini P, and Verbeke G (2019). Random effect exponentiated-exponential geometric model for clustered/longitudinal zero-inflated count data, *Journal of Applied Statistics*, **47**, 2272–2288.

Yau KWK, Wang K, and Lee AH (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros, *Biometrical Journal*, **45**, 437–452.

Yoo H (2023). Sample size for clustered count data based on discrete Weibull regression model, *Communications in Statistics - Simulation and Computation*, **52**, 5850–5856.

Received August 6, 2023; Revised October 18, 2023; Accepted October 18, 2023