Score Tests for Zero-Inflation and Overdispersion in Two-level Count Data

Lim, Hwa-Kyung
Institute of Statistics, Korea University
Anam-dong, Seongbuk-gu
Seoul 136-701, Korea
E-mail: hklim@korea.ac.kr

Song, Juwon
Department of Statistics, Korea University
Anam-dong, Seongbuk-gu
Seoul 136-701, Korea
E-mail: jsong@korea.ac.kr

Jung, Byoung Cheol
Department of Statistics, University of Seoul
jeonnong-dong 90, Dongdaemun-gu
Seoul 130-743, Korea
E-mail: bcjung@uos.ac.kr

Kang, Kee-Hoon
Department of Statistics, Hankuk University of Foreign Studies
Mohyun, Cheoin-Koo
Yongin 449-791, Korea
E-mail: khkang@hufs.ac.kr

Abstract

In a Poisson regression model, where observations are either clustered or represented by repeated measurements of counts, the number of observed zero counts is sometimes greater than the expected frequency by the Poisson distribution and the non-zero part of count data may be overdispersed. The zero-inflated negative binomial (ZINB) mixed regression model is suggested to analyze such data. Previous studies have proposed score statistics for testing zero-inflation and overdispersion separately in correlated count data. Here, we also deal with simultaneous score tests for zero-inflation and overdispersion in two-level count data by using the ZINB mixed regression model. Score tests are suggested for 1) zero-inflation in the presence of overdispersion, 2) overdispersion in the presence of zero-inflation, and 3) zero-inflation and overdispersion simultaneously. The level and power of score test statistics are evaluated by a simulation study. The simulation results indicate that score test statistics may occasionally underestimate or overestimate the nominal significance level due to variations in random effects. This study proposes a parametric bootstrap method to overcome this problem. The simulation results of the bootstrap test indicate that score tests hold the nominal level and provide good power.

keyword: Zero-Inflation, Overdispersion, Generalized Linear Mixed Models, Zero-Inflated Negative Binomial, Score Test, Bootstrap

1 Introduction

In fields such as medicine, public health, epidemiology, sociology, psychology, engineering, and agriculture, among others, the analysis of count data is a topic of major interest. For count data, Poisson regression models have been widely used to explain the relationship between the outcome variable

of interest and a set of explanatory variables. However, there are often the cases that the number of observed zero counts is greater than the expected frequency by the Poisson distribution. In such cases, a standard Poisson model may not perform well. A fair number of statistical methods has been developed to address count data with extra zeros. Böhning (1998) reviewed the related literature and presented some examples from a wide variety of disciplines. A popular approach for analyzing count data with excess zeros is to use the zero-inflated Poisson (ZIP) regression model by Lambert (1992). The ZIP regression model is a mixture of the Poisson distribution and a degenerate component of the point mass at zero. Van den Broek (1995) proposed a score test for zero-inflation under a Poisson distribution. This was extended to ZIP and zero-inflated binomial (ZIB) regression models with covariates (Deng and Paul, 2000; Jansakul and Hinde, 2002).

Both zero-inflation and dependency can often be present in hierarchical count data in which observations are either clustered or repeatedly measured from individual subjects. For example, subjects sampled from a common habitat (called a cluster) such as families, schools, and communities are more likely to be similar to one another than those sampled across different habitats, resulting in correlated responses within the cluster. Dependency among responses can be explained by hierarchical structures using random effects. Hall (2000), Yau and Lee (2001), Hur et al. (2002), and Wang et al. (2002) considered ZIP regression models with cluster-specific random effects to address the heterogeneous variances between clusters. Xiang et al. (2006) proposed a score test for zero-inflation in correlated count data. Lee et al. (2006) extended the ZIP regression model to a multilevel ZIP regression model with random effects. Recently, Moghimbeigi et al. (2009) proposed a score test for zero-inflation in multilevel count data.

Although the ZIP regression model can handle zero-inflation for Poisson data, the non-zero part of count data may be overdispersed. Under the Poisson distribution, the mean and the variance should be the same. In some applications, however, the variance often exceeds the mean, causing overdispersion. ZIP parameter estimates can be severely biased if nonzero counts are substantially overdispersed compared with the Poisson distribution. In such a case, the use of a zero-inflated negative binomial (ZINB) distribution can be a good alternative. Ridout et al. (2001) considered overdispersion in count data and proposed a score test for testing the ZIP regression model against ZINB alternatives. For hierarchical or correlated count data, it is especially true that ZIP parameter estimates can be severely biased when nonzero counts are overdispersed. Xiang et al. (2007) proposed a score test for assessing overdispersion based on the ZINB mixed model, while those of Xie et al. (2009) and Yang et al. (2010) focused on the zero-inflated generalized Poisson (ZIGP) mixed model. However, a simultaneous score test for zero-inflation and overdispersion in the ZINB mixed model or the ZIGP mixed model has not been proposed. Deng and Paul (2005) considered simultaneous score tests for zero-inflation and overdispersion in the ZINB regression model, but their model does not involve random effects for clustered count data.

In this paper, we deal with score tests for zero-inflation and/or overdispersion in two-level count data fitted by the ZINB mixed regression model. We propose score tests for 1) zero-inflation in the presence of overdispersion, 2) overdispersion in the presence of zero-inflation, and 3) zero-inflation and overdispersion simultaneously. Section 2 describes the ZINB mixed regression model. Section 3 suggests score tests for zero-inflation and/or overdispersion in the ZINB mixed regression model. We carried out a simple simulation study to check the adequacy of approximation of score test statistic. It indicates that the asymptotic null distribution works less well for larger σ_u because it would lead to more variations on the parameter estimates and consequently a larger variance for the score statistic. Thus, the score test statistics may occasionally underestimate or overestimate the nominal significance level due to variations in random effects. To solve this problem, Section 4 proposes a parametric bootstrap method.

2 ZINB Mixed Regression Model

Let Y_{ij} be the j^{th} response of a count variable from the i^{th} cluster. Then the ZINB distribution can be written as

$$P(Y_{ij}=y_{ij})=\begin{cases}\phi_{ij}+(1-\phi_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha} & \text{if } y_{ij}=0\\ (1-\phi_{ij})\frac{\Gamma(y_{ij}+1/\alpha)}{y_{ij}!\,\Gamma(1/\alpha)}(1+\alpha\lambda_{ij})^{-1/\alpha}\left(1+\frac{1}{\alpha\lambda_{ij}}\right)^{-y_{ij}} & \text{if } y_{ij}>0\end{cases}$$

for $i=1,\ldots,m$ and $j=1,\ldots,n_i$, where m is the number of clusters, n_i is the number of observations for cluster i, and $\alpha>0$ is an overdispersion parameter. Here ϕ_{ij} and λ_{ij} indicate the proportion of zero-inflation and the mean of the Poisson distribution, respectively. The mean and the variance of the ZINB response variable are given by

$$E(Y_{ij})=(1-\phi_{ij})\,\lambda_{ij},$$

$$Var(Y_{ij})=(1-\phi_{ij})\lambda_{ij}(1+\phi_{ij}\lambda_{ij}+\alpha\lambda_{ij}).$$

The parameters ϕ_{ij} and λ_{ij} can be modeled by linking linear predictors as follows:

$$\log\left(\frac{\phi_{ij}}{1-\phi_{ij}}\right)=\xi_{ij}=w_{ij}'\gamma+u_i,$$

$$\log(\lambda_{ij})=\eta_{ij}=x_{ij}'\beta+v_i,$$

where γ and β are the corresponding $p\times 1$ and $q\times 1$ vectors of regression coefficients for the logistic and Poisson models, respectively. The same explanatory variables can be used, and then w_{ij} would be equal to x_{ij}, and p would be equal to q. In these models, responses in different clusters are assumed to be independent, whereas those within the same cluster are likely to be correlated. To accommodate inherent correlations within clusters, random effects u_i and v_i are incorporated into the linear predictor ξ_{ij} for the zero-inflation model and η_{ij} for the Poisson model. The random effects $u=(u_1,\cdots,u_m)'$ and $v=(v_1,\cdots,v_m)'$ are assumed to be independently distributed as $N(0,\sigma_u^2 I_m)$ and $N(0,\sigma_v^2 I_m)$, respectively, where I_m denotes the $m\times m$ identity matrix.

Parameter estimation can be achieved by the restricted maximum likelihood (REML) approach of McGilchrist(1994). The penalized log-likelihood is given by $l=l_1+l_2$, with l_1 being the log-likelihood function with conditionally fixed random effects and l_2 being the log-likelihood of random effects.

$$\begin{aligned}l_1&=\sum_{i,j}I_{(y_{ij}=0)}\log(\phi_{ij}+(1-\phi_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha})+\sum_{i,j}I_{(y_{ij}>0)}[\log(1-\phi_{ij})\\&-\log(y_{ij}!)+\log(\tfrac{\Gamma(y_{ij}+1/\alpha)}{\Gamma(1/\alpha)})-(y_{ij}+1/\alpha)\log(1+\alpha\lambda_{ij})+y_{ij}\log(\alpha\lambda_{ij})],\\l_2&=-\tfrac{1}{2}\left[m\log(2\pi\sigma_u^2)+\sigma_u^{-2}u'u+m\log(2\pi\sigma_v^2)+\sigma_v^{-2}v'v\right].\end{aligned}$$

That is, l_1 is the log-likelihood of the ZINB variable expressed as a function of the linear predictors ξ_{ij} and η_{ij} and the overdispersion parameter α of the negative binomial distribution with conditionally fixed u and v. l_2 is the log-likelihood of independently and normally distributed u and v. Let the parameter vector of interest be α,γ,β,u,v. With suitable initial values, the REML estimates of α,γ,β,u,v can be obtained iteratively by maximizing l via an EM algorithm to ensure convergence by assuming that σ_u^2 and σ_v^2 are fixed. The variance component estimates for σ_u^2 and σ_v^2 are then computed from estimating equations involving REML estimates. For more details on the estimation procedure, the reader is referred to Yau et al.(2003).

3 Score Tests

3.1 Joint Score Test for Zero-Inflation and Overdispersion

A test of $H_0 : \theta = \alpha = 0$ is equivalent to a simultaneous test for zero-inflation and overdispersion in the zero-inflated negative binomial mixed model. Let $\tau = \sigma_v^2$. Taking the first and second derivatives of l with respect to β, v, τ, θ, and α, the score function $U_{\theta\alpha}$ and the Fisher information matrix $\Im_{\theta\alpha}(\beta, v, \tau, \theta, \alpha)$ can be obtained. Details of the derivatives are given in Appendix C. Under the null hypothesis $H_0 : \theta = \alpha = 0$, the reduced model is the Poisson mixed regression model. The score $U_{\theta\alpha}$ is obtained by evaluating the derivatives of l with respective to θ and α at the REML estimates $\hat{\beta}, \hat{v}$, and $\hat{\tau}$ of the Poisson mixed regression model:

$$U_{\theta\alpha} = [U_\theta,\ U_\alpha] = \left[\sum_{i,j} \Big\{ I_{\{y_{ij}=0\}} \exp(\hat{\lambda}_{ij}) - 1 \Big\},\quad \frac{1}{2} \sum_{i,j} \Big\{ (y_{ij} - \hat{\lambda}_{ij})^2 - y_{ij} \Big\} \right].$$

The expected Fisher information matrix is then

$$\Im_{\theta\alpha}(\beta, v, \tau, \theta, \alpha) = \begin{pmatrix} J_{\beta\beta} & J_{\beta v} & J_{\beta\tau} & J_{\beta\theta} & J_{\beta\alpha} \\ & J_{vv} & J_{v\tau} & J_{v\theta} & J_{v\alpha} \\ & & J_{\tau\tau} & J_{\tau\theta} & J_{\tau\alpha} \\ & & & J_{\theta\theta} & J_{\theta\alpha} \\ & & & & J_{\alpha\alpha} \end{pmatrix},$$

where the entries of $\Im_{\theta\alpha}(\beta, v, \tau, \theta, \alpha)$ under H_0 are obtained by evaluating the second derivatives of l at $\theta = 0$ and $\alpha = 0$. The formula is given in Appendix C. The matrix $\Im_{\theta\alpha}(\beta, v, \tau, \theta, \alpha)$ may be partitioned as follows:

$$\begin{pmatrix} \Im_{11} & \Im_{12} \\ \Im_{12}' & \Im_{22} \end{pmatrix},$$

where
$$\Im_{11} = \begin{pmatrix} -T'BT & -T'BP & 0 \\ -P'BT & -P'BP + \hat{\tau}^{-1} I_m & -\hat{\tau}^{-2}\hat{v} \\ 0 & -\hat{\tau}^{-2}\hat{v}' & -\hat{\tau}^{-2} m/2 + \hat{\tau}^{-3}\hat{v}'\hat{v} \end{pmatrix},$$
$$\Im_{12}' = \begin{pmatrix} 1_N'BT & 1_N'BP & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \Im_{22} = \begin{pmatrix} J_{\theta\theta} & J_{\theta\alpha} \\ J_{\theta\alpha}' & J_{\alpha\alpha} \end{pmatrix},$$
$$J_{\theta\theta} = \sum_{i,j} \Big(\exp(\hat{\lambda}_{ij}) - 1 \Big), \quad J_{\theta\alpha} = \tfrac{1}{2} \sum_{i,j} \hat{\lambda}_{ij}^2, \quad J_{\alpha\alpha} = \tfrac{1}{2} \sum_{i,j} \hat{\lambda}_{ij}^2,$$

with the matrices T, P, and B defined in Appendix C. The score statistic for jointly testing for zero-inflation and overdispersion in the ZINB mixed regression model is then

$$S_{\theta\alpha} = U_{\theta\alpha}'\, \Im^{\theta\alpha}\, U_{\theta\alpha},$$

where $\Im^{\theta\alpha}$ is the lower right-hand of the 2×2 matrices of the inverse information matrix $\Im_{\theta\alpha}^{-1}$, which is evaluated at the REML estimates of the Poisson mixed regression model. Under the null hypothesis, the test statistic $S_{\theta\alpha}$ asymptotically follows a χ_2^2 distribution.

3.2 Inadequacy of approximation of score test statistics

To check the appropriateness of χ^2 approximation of score test statistic, we carried out a simple simulation study. We assessed the effect of varying σ_v on the sampling distribution of S by simulating

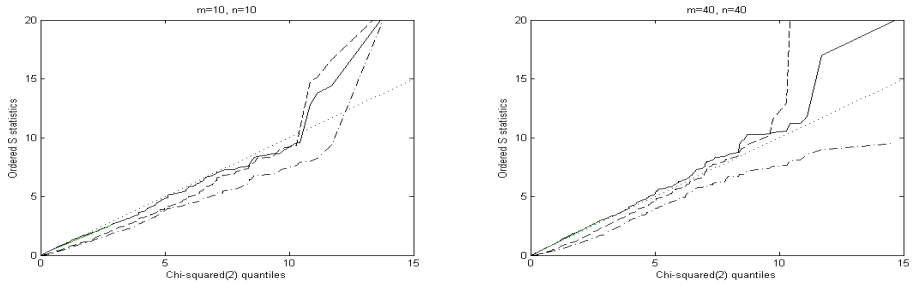

Figure 1. Q-Q plots of ordered score statistics against χ^2_2 quantiles based on 1000 replications generated from the Poisson mixed model under $H_0 : \phi = \alpha = 0$ with $\sigma_v = 0.1$(broken line), 0.5(solid line), 1.0(broken-dotted line)

1000 samples for $\sigma_v = 0.1$, 0.5, and 1 on two sample sizes ($m = 10$, $n = 10$) and ($m = 40$, $n = 40$). We present results for only joint test, but the other tests show similar results. The resulting Q-Q plots, given in Figure 1, confirm that the asymptotic null distribution is satisfactory for smaller σ_v value because of less variation in the random component. For larger σ_v, however, the asymptotic null distribution works less well. This result is consistent with Xiang et al.(2006) and Moghimbeigi et al.(2009). They mentioned that larger σ_v would lead to more variations on the parameter estimates and consequently a larger variance for the score statistic. On the other hand, the asymptotic distribution of the score statistic is often approached to χ^2 more slowly than that of the likelihood ratio statistic. Thus, significance levels derived from the score statistic may be misleading, particularly in small sample sizes. Due to this reason, the score test might underestimate the nominal significance level in small sample cases for testing a ZIP regression model against ZINB alternatives Ridout et al.(2001). Jung et al.(2005) proposed a parametric bootstrap method to overcome the underestimation of the nominal level. We consider the bootstrap method to avoid these problems and to obtain more accurate inference.

4 Bootstrap Method

Proceed with the following steps:

- Step 1.
  For the given data $(Y_{ij},\ x_{ij}),\ \ i = 1, \cdots, m,\ \ j = 1, \cdots, n_i$, obtain the REML estimates $\hat{\beta}^*,\ \ \hat{v}_i^*,\ \ \hat{\sigma}_v^{2*}$ under the Poisson mixed regression model and compute $\hat{\lambda}_{ij} = \exp(x'_{ij}\hat{\beta} + \hat{v}_i)$ and the score statistic $S_{\theta\alpha}$.

- Step 2.
  Generate a bootstrap sample Y_{ij}^* from the Poisson $(\hat{\lambda}_{ij})$ distribution.

- Step 3.
  For each bootstrap sample $(Y_{ij}^*,\ x_{ij}),\ \ i = 1, \cdots, m,\ \ j = 1, \cdots, n_i$, obtain the REML estimates $\hat{\beta}^*,\ \ \hat{v}_i^*,\ \ \hat{\sigma}_v^{2*}$ and compute the bootstrap score test statistic $S_{\theta\alpha}^*$.

- Step 4.
  Repeat Steps 2 and 3 independently B times. From the B possibly different values of $S_{\theta\alpha}^*$, obtain the $100(1-c)th$ percentile of $S_{\theta\alpha}^*$, $S_{\theta\alpha}^*(1-c)$.

- Step 5.
  If the score test statistic $S_{\theta\alpha}$ is greater than $S_{\theta\alpha}^*(1-c)$, then reject the null hypothesis at the significance level c.

REFERENCES (RÉFERENCES)

[1] Böhning, D., 1998. Zero-inflated Poisson models and C.A.MAN: A tutorial collection of evidence. Biometrical J. 40, 833-843.

[2] Deng, D., Paul, S.R., 2000. Score tests for zero-inflation in generalized linear models. Canad. J. Statist. 27, 563-570.

[3] Deng, D., Paul, S.R., 2005. Score tests for zero-inflation and over-dispersion in generalized linear models. Statist. Sinica 15, 257-276.

[4] Hall, D.B., 2000. Zero-inflated Poisson and binomial regression with random effects: A case study. Biometrics 56, 1030-1039.

[5] Hur, K., Hedeker, D., Henderson, W., Khuri, S., Daley, J., 2002. Modeling clustered count data with excess zeros in health care outcomes research. Health Serv. Outcomes Res. Method 3, 5-20.

[6] Jansakul, N., Hinde, J.P., 2002. Score tests for zero-inflated Poisson models. Comput. Statist. Data Anal. 40, 75-96.

[7] Jung, B.C., Jhun, M., Lee, J.W., 2005. Bootstrap Tests for Overdispersion in a Zero-Inflated Poisson Regression Model. Biometrics 61, 626-629.

[8] Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1-14.

[9] Lee, A.H., Wang, K., Scott, J.A., Yau, K.K.W., McLachlan, G.J., 2006. Multilevel zero-inflated Poisson regression modelling of correlated count data with excess zeros. Stat. Methods Med. Res. 15, 47-61.

[10] McGilchrist, C.A., 1994. Estimation in generalized linear mixed models. J. Roy. Statist. Soc. B 56, 61-69.

[11] Moghimbeigi, A., Eshraghian, M.R., Mohammad, K., McArdle, B., 2009. A score test for zero-inflation in multilevel count data. Comput. Statist. Data Anal. 53, 1239-1248.

[12] Ridout, M., Hinde, J., Demétrio, C.G.B., 2001. A score test for testing zero inflated Poisson regression model against zero inflated negative binomial alternatives. Biometrics 57, 219-223.

[13] Schall, R., 1991. Estimation in generalized linear models with random effects. Biometrika 78, 719-727.

[14] Silvapulle, M.J., 1994. On tests against one-sided hypotheses in some generalized linear models. Biometrics 50, 853-858.

[15] Van den Broek, J., 1995. A score test for zero-inflation in a Poisson distribution. Biometrics 51, 738-743.

[16] Wang, K., Yau, K.K.W., Lee, A.H., 2002. A zero-inflated Poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays. Comput. Methods Programs Biomed. 68, 195-203.

[17] Xiang, L., Lee, A.H., Yau, K.K.W., McLachlan, G.J., 2006. A score test for zero-inflation in correlated count data. Stat. Med. 25, 1660-1671.

[18] Xiang, L., Lee, A.H., Yau, K.K.W., McLachlan, G.J., 2007. A score test for overdispersion in zero-inflated Poisson mixed regression model. Stat. Med. 26, 1608-1622.

[19] Xie, F.C., Wei, B.C., Lin, J.G., 2009. Score tests for zero-inflated generalized Poisson mixed regression Models. Comput. Statist. Data Anal. 53, 3478-3489.

[20] Yang, Zhao., Hardin, J.W., Addy, C.L., 2010. Score tests for overdispersion in zero-inflated Poisson mixed models. Comput. Statist. Data Anal. 54, 1234-1246.

[21] Yau, K.K.W., Lee, A.H., 2001. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. Stat. Med. 20, 2907-2920.

[22] Yau, K.K.W., Wang, K., Lee, A.H., 2003. Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros. Biometrical J. 45, 437-452.