Zero-Inflated Poisson Regression Mixture Model

Hwa Kyung Lim[a], Wai Keung Li[b], Philip L.H. Yu[b]

[a]*Department of Statistics, University of Michigan, MI, U.S.A.*
[b]*Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong, China*
Corresponding author: Hwa Kyung Lim, e-mail: hwlim@umich.edu

Abstract

Excess zeros and overdispersion are commonly encountered phenomena that limit the use of traditional Poisson regression models for modeling count data. The focus of this paper is on modeling count data in the case that a population has excess zero counts and also consists of several sub-populations in the non-zero counts. The proposed zero-inflated Poisson regression mixture model accounts for both excess zeros and heterogeneity. The performance of parameter estimation for the proposed model is evaluated through simulation studies.

Keywords: Zero-Inflation, Heterogeneity, Finite Mixture Models, Poisson

1. Introduction

Count data are common in many research areas (sociology, engineering, medical studies and others), which are usually modeled using the Poisson distribution. However, in various applications dispersion of the variable of interest exceeds dispersion. This phenomenon, called overdispersion, often results from unobserved heterogeneity, i.e., the sample of responses is drawn from a population consisting of several sub-populations. Mixtures of Poisson distributions have been widely used to deal with this problem. A finite Poisson mixture model with K components explains the population by giving weights π_k to sub-populations with means λ_k, $k = 1, \cdots, K$. This approach also provides a natural framework to classify observations into the components of the mixture model. The finite mixtures of the Poisson regression model with constant weight parameters have been developed by Brännaäs and Rosenqvist (1994), Wedel et al. (1993), Wang et al. (1996) and Alfò and Trovato (2004). Wang et al. (1998) discuss the finite mixed Poisson regression models that incorporate covariates in the weight parameters.

In addition, the count variable of interest may contain more zeros than what is to be expected under a Poisson model which is commonly observed in many applications. A popular approach to model excess zeros is to use a zero-inflated Poisson (ZIP) regression model discussed by Lambert (1992). The ZIP distribution is a mixture of a Poisson distribution and a degenerate distribution at zero. This regression setting allows for covariates in both the Poisson mean and weight parameters. Böhning (1998) and Ridout et al. (1998) provide reviews of related literature and present examples from a wide variety of disciplines.

If some overdispersion related to the counts may remain even after modeling excess zeros, a zero-inflated negative binomial (ZINB) model can be a good alternative. However, if a population has excess zero counts and also consists of several sub-populations in the non-zero counts, the ZINB model may not be sufficient for such data. The focus of this paper is on modeling heterogeneous count data with excess zero counts.

The paper is organized as follows. We describe the zero-inflated Poisson regression mixture model in Section 2. Several simulation studies to assess the performance and sensitivity of parameter estimation are presented in Section 3. Finally, we conclude by discussing the findings in Section 4.

2. ZIP regression mixture model

A popular approach to analyze count data with excess zeros is to use a zero-inflated Poisson (ZIP) regression model. If the non-zero counts consist of several sub-populations (unobserved heterogeneity), a single component of the ZIP regression model may be insufficient to describe the non-zero counts. We propose a model that accounts for the excess zeros and the heterogeneous non-zero counts simultaneously.

Suppose a count response variable Y follows a ZIP mixture distribution:

$$P(Y=y) = \begin{cases} \pi_1 + \pi_2 e^{-\lambda_2} + \cdots + \pi_K e^{-\lambda_K}, & y = 0 \\ \pi_2 \frac{e^{-\lambda_2}\lambda_2^y}{y!} + \cdots + \pi_K \frac{e^{-\lambda_K}\lambda_K^y}{y!}, & y > 0 \end{cases} \tag{1}$$

where K is the number of mixing components, λ_k is the mean and π_k is the mixing weight of component k such that $0 < \pi_k < 1$, $k = 1, \cdots, K$, and $\sum_{k=1}^{K} \pi_k = 1$. The weight π_1 determines the proportion of excess zeros compared with an ordinary Poisson mixture model. If K is equal to two, the ZIP mixture distribution in Eq. (1) is reduced to the ZIP distribution (Lambert (1992)).

To incorporate covariate information, we model the means $\{\lambda_k\}_{k=1}^K$ and the mixing weights $\{\pi_k\}_{k=1}^K$ using the following regression models that parameterize $\log(\lambda_k)$ and the multinomial logit transform of π_k as linear functions of covariates:

$$\log(\lambda_{ik}) = \mathbf{x}_i \beta_k, \quad i = 1, \cdots, N, \quad k = 2, \cdots, K \tag{2}$$

$$\pi_{ik}(\mathbf{w}_i, \gamma) = \frac{\exp(\mathbf{w}_i \gamma_k)}{1 + \sum_{k=2}^{K} \exp(\mathbf{w}_i \gamma_k)}, \quad \pi_{i1}(\mathbf{w}_i, \gamma) = 1 - \sum_{k=2}^{K} \pi_{ik}(\mathbf{w}_i, \gamma), \tag{3}$$

where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})$ and $\mathbf{w}_i = (w_{i1}, \cdots, w_{iq})$ are $1 \times p$ and $1 \times q$ row vectors of covariates (including an intercept), β_k and γ_k are the corresponding $p \times 1$ and $q \times 1$ vectors of regression coefficients for the kth component, respectively. Note that the mixing probability of the first component $\pi_{i1}(\mathbf{w}_i, \gamma)$ is the probability of excess zeros and is taken as the baseline for the multinomial logit. That is, the logit of the other components relative to π_{i1} is $\log(\pi_{ik}/\pi_{i1}) = \mathbf{w}_i \gamma_k$, $k = 2, \cdots, K$.

The generalized ZIP (GZIP) regression mixture model can be formulated as follows:

$$P(Y = y_i) = \pi_{i1}(\mathbf{w}_i, \gamma) I_{(y_i=0)} + \sum_{k=2}^{K} \pi_{ik}(\mathbf{w}_i, \gamma) Pois(y_i \mid \mathbf{x}_i, \beta_k), \tag{4}$$

where $I_{(\cdot)}$ is the indicator function of the argument. A special case of the above model will be obtained if the mixing weights π_{ik} are assumed to be constant functions of the covariates $\mathbf{w}_i$. In that case, the ZIP with fixed weights (FZIP) regression mixture model can be formulated as follows:

$$P(Y = y_i) = \pi_1 I_{(y_i=0)} + \sum_{k=2}^{K} \pi_k Pois(y_i \mid \mathbf{x}_i, \beta_k). \tag{5}$$

If both π_{ik} and λ_{ik} are constant functions, the GZIP mixture model reduces to the standard Poisson mixture model denoted by

$$P(Y = y_i) = \sum_{k=1}^{K} \pi_k Pois(y_i \mid \lambda_k). \tag{6}$$

Note that the first component (a degenerate distribution with all mass π_1 at $y_i = 0$) in Eq. (4) can be regarded as a Poisson distribution with mean $\lambda_1 = 0$ because of $Pois(y_i = 0 \mid \lambda_1 = 0) = 1$ and $Pois(y_i \neq 0 \mid \lambda_1 = 0) = 0$.

3. Results

A simulation study is conducted for evaluating the performance of the proposed EM estimation algorithm. We generate N samples from the following GZIP model with three components.

$$\pi_{i1}(\mathbf{w}_i, \gamma) I_{(y_i=0)} + \pi_{i2}(\mathbf{w}_i, \gamma) Pois(y_i \mid \lambda_2(\mathbf{x}_i, \beta_2)) + \pi_{i3}(\mathbf{w}_i, \gamma) Pois(y_i \mid \lambda_3(\mathbf{x}_i, \beta_3)).$$

The log-link for the Poisson mean λ_{ik} and the multinomial logit-link for weight π_{ik} used are as follows:

$$\begin{cases} \log(\lambda_2(\mathbf{x}_i, \beta_2)) = \beta_{20} + \beta_{21}\mathbf{x}_i \\ \log(\lambda_3(\mathbf{x}_i, \beta_3)) = \beta_{30} + \beta_{31}\mathbf{x}_i \end{cases} \quad \text{and} \quad \begin{cases} \log(\pi_{i2}/\pi_{i1}) = \gamma_{20} + \gamma_{21}\mathbf{w}_i \\ \log(\pi_{i3}/\pi_{i1}) = \gamma_{30} + \gamma_{31}\mathbf{w}_i, \end{cases}$$

where $\mathbf{x}_i$ and $\mathbf{w}_i$ are generated from Uniform(0,1), respectively. True values for the Poisson regression coefficients are assumed as $\beta_2' = (\beta_{20}, \beta_{21}) = (1.2, -0.4)$, $\beta_3' = (\beta_{30}, \beta_{31}) = (1.5, 0.8)$. To consider zero-inflation among the mixing weights, we highly set the probability of excess zeros as $\gamma_2' = (\gamma_{20}, \gamma_{21}) = (-1.2, 1.4)$, $\gamma_3' = (\gamma_{30}, \gamma_{31}) = (-1.3, 0.8)$ which can be reparameterized as π_k by Eq. (3). As a result, $\pi_1 \approx 0.497, \pi_2 \approx 0.301, \pi_3 \approx 0.202$ are expected on average.

To generate samples from the above model, for each subject i $(1, \cdots, N)$, a random number u is generated from Uniform(0,1). If u is less than π_{i1}, Y_i takes the value 0, or if u is between π_{i1} and $\pi_{i1}+\pi_{i2}$ then Y_i is a draw from Poisson(λ_{i2}), otherwise Y_i is generated from Poisson(λ_{i3}).

The results concerning the evaluation of parameter estimation are presented in Table 1. The results are based on 1000 replications for each of the three sample sizes ($N = 300, 500, 1000$). Bias, mean square error (MSE) and coverage probability are used to evaluate the estimation performance. Bias is calculated as the difference between the average estimate and the true value which should ideally be close to zero. MSE measures the average squared distance from the estimate and the true value which is a useful measure of overall accuracy of estimation. Coverage probability is the proportion of times the confidence interval obtained contains the true value. We consider bootstrapping for computing the coverage probability of confidence interval because the EM algorithm used in parameter estimation does not produce standard errors. The following formulas are used to obtain these measures.

- Bias$(\hat{\theta}) = \frac{1}{r} \sum_{j=1}^{r} \hat{\theta}_j - \theta$,
- MSE$(\hat{\theta}) = \frac{1}{r} \sum_{j=1}^{r} (\hat{\theta}_j - \theta)^2$,
- Bootstrap coverage : Proportion of times the $100(1-\alpha)\%$ bootstrap confidence interval $[2\hat{\theta}_j - \theta^*_{j,1-\alpha/2}, 2\hat{\theta}_j - \theta^*_{j,\alpha/2}]$ includes θ for $j = 1, \cdots, r$,

where $\theta \in \Theta = \{\beta_{k0}, \beta_{k1}, \gamma_{k0}, \gamma_{k1}\}_{k=2}^{3}$ is the true value for estimate of interest, r is the number of replications performed, $\hat{\theta}_j$ is the estimate of interest within each of the $j = 1, \cdots, r$ replication, and $\theta^*_{\alpha/2}$ is the $100(\alpha/2)th$ percentile of $\theta^* = (\theta^*_1, \cdots, \theta^*_B)$, where θ^*_b $(b = 1, \cdots, B)$ is computed for a bootstrap sample Y^*_i generated from $\hat{\pi}_{i1} I_{(y_i=0)} + \sum_{k=2}^{3} \hat{\pi}_{ik} Pois(\hat{\lambda}_{ik})$ and this process is independently repeated B=1000 times.

The results in Table 1 indicate that the EM algorithm indeed performs well in estimating the true coefficients. The bias and MSE for all the parameters decrease as sample size increases from 300 to 1000. Also, the bootstrap coverage probabilities are closer to the nominal confidence level as sample size increases.

Table 1: Bias, MSE, Coverage probability

			N=300		N=500		N=1000	
Evaluation criteria		$\hat{\theta}$	$k=2$	$k=3$	$k=2$	$k=3$	$k=2$	$k=3$
Bias		β_{k0}	-0.001	0.009	0.001	-0.002	0.001	-0.010
		β_{k1}	-0.016	-0.008	-0.019	0.006	-0.005	0.011
		γ_{k0}	0.000	-0.046	-0.010	-0.013	-0.003	-0.011
		γ_{k1}	0.048	0.003	0.021	-0.025	0.004	0.013
MSE		β_{k0}	0.042	0.021	0.023	0.016	0.013	0.008
		β_{k1}	0.127	0.045	0.071	0.031	0.038	0.017
		γ_{k0}	0.138	0.190	0.078	0.097	0.040	0.048
		γ_{k1}	0.374	0.505	0.196	0.277	0.104	0.131
Bootstrap Coverage	99%	β_{k0}	0.984	0.994	0.989	0.991	0.995	0.996
		β_{k1}	0.997	0.996	0.993	0.993	0.992	0.994
		γ_{k0}	0.987	0.990	0.992	0.992	0.990	0.991
		γ_{k1}	0.999	1.000	0.999	0.998	0.994	0.994
	95%	β_{k0}	0.964	0.967	0.970	0.963	0.970	0.963
		β_{k1}	0.971	0.977	0.974	0.972	0.954	0.964
		γ_{k0}	0.958	0.969	0.962	0.959	0.960	0.956
		γ_{k1}	0.986	0.994	0.979	0.977	0.956	0.957
	90%	β_{k0}	0.935	0.939	0.933	0.922	0.918	0.915
		β_{k1}	0.934	0.952	0.926	0.939	0.907	0.917
		γ_{k0}	0.932	0.933	0.931	0.918	0.913	0.901
		γ_{k1}	0.952	0.956	0.931	0.935	0.914	0.909

3.1. Sensitivity of the GZIP mixture model

We conduct simulations to examine sensitivity of the GZIP mixture model. Our first simulation is to examine the performance of the GZIP model for classification. We use the same setting as before to generate data from the GZIP mixture model with three components. A GZIP model with three components is fitted to the simulated data and used for classifying the data into three groups. We compute the misclassification rate by counting the number of observations classified into a group that is different from its original group. Results of this simulation based on 1000 repetitions are presented in Table 2. The overall misclassification rate is 11.3%. Moreover, it can be seen that the majority of the misclassification is between groups two and three which is not surprising because the means of these subgroups in some simulated datasets are close to each other making the boundary between groups slightly ambiguous (Poisson means $\lambda_k \approx (2.7, 6.8)$ for $k = (2, 3)$).

In the second simulation, we examine if the number of mixing components of the GZIP model can be identified using AIC and BIC model selection criteria. We generate data from the GZIP mixture with 3-components under the same setting as above. We now fit four GZIP mixture models with 2, 3, 4 and 5 components. Table 3 shows the result of model selection using AIC and BIC. We can see that BIC picks exactly the true model in all the cases but AIC sometimes choses more components than is necessary. This is consistent with the fact that BIC is

Table 2: Misclassification for 3-component model

Real group	Classified group 1	2	3	Misclassification rate
1	497	0	0	0.0%
2	22	242	40	20.5%
3	0	50	148	25.4%
Overall misclassification rate				11.3%

Data : GZIP mixture with 3-component,
Model : GZIP mixture with 3-component.

model selection consistent and indicates that it is more suitable for identifying the number of components.

Table 3: Model selection based on AIC and BIC

	GZIP mixture model k=2	k=3	k=4	k=5	Correct rate
AIC	0	912	81	7	91.2 %
BIC	0	1000	0	0	100.0 %

AIC = -2 log-likelihood + 2 M,
BIC = -2 log-likelihood + $M \log(N)$,
M is the number of estimated parameters,
N is the sample size.

The third experiment is performed to study the impact of using the GZIP model when the data is generated from a FZIP model that is simpler. The motivation for conducting this experiment is to see if the estimation of the GZIP model is less efficient (i.e., larger standard errors) than the FZIP model. A sample dataset (N=1000) is generated from the FZIP mixture model with 3-components given by Eq. (5).

True values for the regression coefficients are taken as $\beta_2' = (\beta_{20}, \beta_{21}) = (2.4, -1.8)$, $\beta_3' = (\beta_{30}, \beta_{31}) = (1.2, 0.6)$ and $\mathbf{x}_i$ is set at intervals of 0.2, i.e. 0.2 ($i = 1, \cdots, 100$), 0.4 ($i = 101, \cdots, 200$), $\cdots$, 2.0 ($i = 901, \cdots, 1000$). The corresponding weights are set to $(\pi_1, \pi_2, \pi_3) = (0.5, 0.3, 0.2)$. Based on 1000 replications, Table 4 shows that the estimates from the FZIP and GZIP mixture models. We can see that the GZIP model performs as well as the FZIP model for estimation both in terms of the estimates and their standard errors. The estimated coefficients for weights in the GZIP model are $(\hat{\gamma}_{20}, \hat{\gamma}_{21}) = (-0.509, 0.0005)$, $(\hat{\gamma}_{30}, \hat{\gamma}_{31}) = (-0.938, 0.016)$, which are reparameterized as $(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) = (0.499, 0.301, 0.199)$ by a multinomial logit transform.

Table 4: Parameter estimates (standard errors)

Component k	β_{k0} True	FZIP	GZIP	β_{k1} True	FZIP	GZIP	π_k True	FZIP	GZIP
1	-	-	-	-	-	-	0.5	0.500	0.499
2	2.4	2.399 (0.076)	2.398 (0.092)	-1.8	-1.802 (0.107)	-1.799 (0.154)	0.3	0.300	0.301
3	1.2	1.202 (0.096)	1.201 (0.099)	0.6	0.600 (0.063)	0.600 (0.065)	0.2	0.200	0.199

- indicates that the parameter is not estimated from the Poisson model.
Data : FZIP mixture with 3-components
Model : FZIP mixture with 3-components, GZIP mixture with 3-components

4. Conclusions

We have proposed the ZIP regression mixture model for the zero-inflated heterogeneous count data. Our simulation studies show that the proposed model works satisfactorily and estimation techniques perform well.

References

Alfò, M., Trovato, G., 2004. Semiparametric mixture models for multivariate count data, with application. *The Econometrics J.* 7(2), 426-454.

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

Brännaäs, K., Rosenqvist, G., 1994. Semiparametric estimation of heterogenous count data models. *European J. of Operational Research* 76, 247-258.

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

Ridout, M., Demétrio, C.G.B., Hinde, J. 1998. Models for count data with many zeros. International Biometric Conference, Cape Town.

Wang, P., Puterman, M.L., Cockburn, I., Le, N., 1996. Mixed poisson regression models with covariate dependent rates. *Biometrics* 52(2), 381-400.

Wang, P., Cockburn, I.M., Puterman, M.L., 1998. Analysis of patent data: a mixed Poisson regression model approach. *J. of Business and Economic Statistics* 16(1), 27-41.

Wedel, M., Desarbo, W.S., Bult, J.R., Ramaswamy, V., 1993. A latent class Poisson regression model for heterogeneous count data. *J. of Applied Econometrics* 8, 397-411.