ZERO INFLATED NEGATIVE BINOMIAL MODELS IN SMALL AREA ESTIMATION

Irene Muflikh Nadhiroh[1], Khairil Anwar Notodiputro[2], Indahwati[2]

[1]Mahasiswa S1 Departemen Statistika FMIPA IPB
[2]Dosen Departemen Statistika FMIPA IPB

Abstract

The problem of over-dispersion in Poisson data is usually solved by introducing prior distributions which lead to negative binomial models. Poisson data sometime is also suffered by excess zero problems, a condition when data contains too many zero or exceeds the distribution's expectation. Zero Inflated Negative Binomial (ZINB) method can be utilized to solve such problems. This paper demonstrates the adoption of ZINB methods in Small Area Estimation with excess zero data. It is shown that the excess zero problem has substantially influenced the Empirical Bayes (EB) estimates, and the adoption of ZINB methods has improved the precision and reliability of the estimates.

Key Words: Small Area Estimation, Zero-Inflation, Poisson-Gamma, Negative Binomial Regression, Empirical Bayes

INTRODUCTION

Background

Direct estimation is very difficult to adopt in some cases, which the sample is drawn from big scale survey such as National Economics and Social Survey. It is because the number of sample in region level is too small. Regarding the problem, indirect estimation, which adds covariates to estimate the parameter, is usually used. This type of estimation is broadly known as Small Area Estimation.

Kismiantini (2007) conducted a research in Small Area Estimation based on Poisson-Gamma models. She used Maximum Likelihood Estimator with Negative Binomial Regression technique to estimate prior parameter. Moreover, Negative Binomial Regression was used to resolve over-dispersion problem in the data.

In reality, count data is not characterized by over-dispersion only, but sometimes by excess-zero. Excess-zero is a condition when data contains too many zero or exceeds

the distribution's expectation. In this paper, Zero-Inflated models were used to take care of this type of situation.

Objectives

The research objectives are:

1. To investigate the performance of Negative Binomial Regression on Small Area Estimation in case of excess-zero.
2. To apply Zero-Inflated Count Models on Small Area Estimation in case of excess-zero.
3. To observe the performance of Zero-Inflated Count Models in estimating prior parameter toward the result of Small Area Estimation.

LITERATURE REVIEW

Small Area Estimation

In general, small area is used to denote any domain which the direct estimation with adequate precision can not be produced (Rao 2003). Small Area Estimation is used in effort to make estimation with adequate level of precision. It works as indirect estimation that lend the strength of variable interest values from related areas through the use of supplementary information related to variable interest such as, recent census count and current administrative records (Rao 2003).

Small Area Models

There are two link models in indirect estimation (Rao 2003):

1. Basic area level (type A) model.

Basic area level model or aggregate model includes all models that relate small area with area-specific auxiliary variables. These models are essential if unit (element) level data are not available. Assuming, parameter estimators, θ_i, is related to area specific auxiliary data or covariate variables, x_i, by a linear model:

$\theta_i = x_i^T \beta + b_i v_i$, with i=1,…,m, $v_i \sim N(0, \sigma_v^2)$ are area-specific random effect and β is $p \times 1$ vector of regression coefficients. Therefore, b_i are known as positive constants.

For making inferences about θ_i, direct estimators y_i are assumed available. Accordingly, assuming:

$y_i = \theta_i + e_i$, where i=1,…,m, with sampling error e_i ~N$(0, \sigma^2{}_{ei})$ and $\sigma^2{}_{ei}$ are known. At the end, both models are combined and as a result is new model: $y_i = x_i{}^T \beta + b_i v_i + e_i$, where i=1,…,m. (Rao 2003).

2. Basic unit level (type B) model.

Unit level model includes all models that relate unit values of the study variable to unit-specific auxiliary variables. Assuming, unit-specific auxiliary variables, x_{ij}, and correspondingly, a nested regression model: $y_{ij} = x_{ij}{}^T \beta + v_i + e_{ij}$, where i=1,…,m, and j=1,…, n_i, with v_i ~N$(0, \sigma_v^2)$ and also e_i ~N$(0, \sigma^2{}_{ei})$.

Empirical Bayes Methods

Novick on Good (1980) said that Bayes method is difficult to adopt and sometimes very sensitive, because it needs prior probability information, which is very difficult to obtain. So Robbin (1955) for the first time introduced Empirical Bayes methods with assumed that the prior is undefined, and then the data used to estimate the prior parameter. (Rao 2003) EB methods in Small Area Estimation may be summarized as follows:

1. Obtain the posterior probability density function of the small area's parameter interest.
2. Estimate the model parameters from the marginal density function.
3. Use the estimated posterior density for making inferences about small area parameters interest.

Poisson-Gamma Models

Poisson model is standard model in dealing with count data. This model is limited on variance and means when utilized to estimate single parameter. Generally, count data experience over-dispersion. Because of that, a Poisson formula had been developed to accommodate extra variance from sample data. So, two-stage models have been introduced for count data, known as mixed model Poisson-Gamma. Wakefield (2006) introduced Poisson-Gamma model which was more easy to use, with SMR (Standard

Mortality Ratio) as direct estimator. This study used Wakefield model with alteration in direct estimator.

If y_i is number of specific individual at small area-*i*, which has specific characteristic interest, and written as follow:

$$y_i = \sum_j y_{ij}$$

And y_{ij} are j-object at i-small area where j=1,…,n and i=1,…,m.

At first stage $y_i \overset{ind}{\sim} Poisson(\mu_i \theta_i)$ is assumed where $\mu_i = \mu(\underline{x}_i, \underline{\beta})$ describes a regression model in area level, $\underline{x}_i$, covariates and $\underline{\beta}_i = (\beta_i, ..., \beta_p)^T$, vector of regression coefficients.

At second stage $\theta_i \overset{iid}{\sim} gamma(\alpha, 1/\alpha)$ is assumed as prior with mean 1 and variance $1/\alpha$. The marginal distribution $y_i \mid \underline{\beta}, \alpha$ is negative binomial.

Then Wakefield (2006) using Bayes Theorem acquired posterior as: $\theta_i \mid y_i, \underline{\beta}, \alpha \sim gamma(y_i + \alpha, 1/\alpha + \mu_i)$ and EB estimator as:

$$\hat{\theta}_i^{EB} = \hat{\theta}_i^{B}(\underline{\hat{\beta}}, \hat{\alpha}) = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i)\mu_i$$

with $\hat{\gamma}_i = \hat{\mu}_i / (\hat{\alpha} + \hat{\mu}_i)$, $\hat{\theta}_i = y_i$ are direct estimation from θ_i.

Negative Binomial Regression

The negative binomial distribution function is written as:

$$g(y \mid x) = \frac{\Gamma(y+k)}{\Gamma(\mu)\Gamma(y+1)} \left(\frac{k}{\mu+k}\right)^k \left(\frac{\mu}{\mu+k}\right)^y$$

where y=0,1,2,… ; k and μ are negative binomial parameter with $E(y) = \mu$ and $\text{var}(y) = \mu + \mu^2 k$; k mention as disperse parameter which is shown that the data consist of over-dispersed.

Zero-Inflated Models

Zero-Inflated models consider two distinct sources of zero outcomes. One source is generated from individuals who do not enter into the counting process, the other from those who do enter the count process but result in a zero outcome (Hardin & Hilbe 2007).

For the zero-inflated model, the probability of observing a zero outcome equals the probability that an individual is in the always-zero group plus the probability that individual is not that group times the probability that the counting process produces a zero. If $B(0)$ as the probability that the binary process result in a zero outcomes and $\Pr(0)$ as the probability that the counting of a zero outcomes, the probability of a zero outcome for the system is then given by (Hardin & Hilbe 2007):

$$\Pr(y = 0) = B(0) + (1 - Z)\Pr(0)$$

The probability of a nonzero count is:

$$\Pr(y = k; k > 0) = [1 - B(0)]\Pr(k)$$

Zero-Inflated Negative Binomial

This model is used in over-disperse and excess-zero data. As a result, among parameter estimators, there would be k parameters which indicate that over-disperse occur in data, just as disperse parameter in negative binomial regression.

The probability distribution of this model is as follow:

$$P(Y_i = y_i \mid x_i) = \begin{cases} \varphi(\phi' x_i) + \{1 - \varphi(\phi' x_i)\} g(0 \mid x_i) \\ \{1 - \varphi(\phi' x_i)\} g(y_i \mid x_i) \end{cases}$$

Where φ is a function of $z_i'\phi$; x_i are vector of zero-inflated covariate and ϕ is a vector of zero-inflated coefficient, which will be estimated. Meanwhile, $g(y_i \mid x_i)$ is probability distribution of negative binomial, written as:

$$g(y_i \mid x_i) = \frac{\Gamma(y_i + \alpha)}{\Gamma(\mu_i)\Gamma(y_i + 1)} \left(\frac{\alpha}{\mu_i + \alpha} \right)^{\alpha} \left(\frac{\mu_i}{\mu_i + \alpha} \right)^{y_i}$$

Mean and variance of ZINB are:

$$E(y_i \mid x_i) = \mu_i (1 - \varphi_i)$$
$$V(y_i \mid x_i) = \mu_i (1 - \varphi_i)(1 + \mu_i (\varphi_i + \alpha))$$

Jackknife Method of Estimating MSE($\hat{\theta}_i^{EB}$)

This methods have been known by Tukey (1958) and developed to be a method that capable to be bias corrected of estimator by remove observation-i for i=1,…,m and performs parameter estimation.

Rao (2003) the Jackknife step to estimate MSE**(**$\hat{\theta}_i^{EB}$**)** are :

1. Assume that $\hat{\theta}_i^{EB} = k_i(y_i, \underline{\hat{\beta}}, \hat{\alpha})$, $\hat{\theta}_{i-1}^{EB} = k_i(y_i, \underline{\hat{\beta}}_{-1}, \hat{\alpha}_{-1})$, then calculate:

$$\hat{M}_{2i} = \frac{m-1}{m}\sum_{l}^{m}(\hat{\theta}_i^{EB} - \hat{\theta}_{i-1}^{EB})^2$$

2. Calculate the delete-i estimator $\underline{\hat{\beta}}_{-1}$ dan α_{-1}, then calculate :

$$\hat{M}_{1i} = g_{1i}(\underline{\hat{\beta}}, \hat{\alpha}, y_i) - \frac{m-1}{m}\sum_{i=m}^{m}[g_{1i}(\underline{\hat{\beta}}_{-1}, \hat{\alpha}_{-1}, y_i) - g_{1i}(\underline{\hat{\beta}}, \hat{\alpha}, y_i)]$$

   The estimator $\hat{M}_{1i}$ correct the bias of $g_{1i}(\hat{\sigma}_v^2)$

3. Calculate the jackknife estimator of MSE($\hat{\theta}_i^{EB}$) as:

$$MSE_J(\hat{\theta}_i^{EB}) = \hat{M}_{1i} + \hat{M}_{2i}$$

SUBJECT AND METHOD

Subject

This research assumed that the available auxiliary data is on area level, so this research used basic area level model. The data were simulated with 30 small areas and one covariate. Every batch generated different conditions of excess-zeros data, start from 0.1 until 0.9 probability of zero in small area. This research assumed structure of relation between respond and covariate was linear.

Method

Steps of generating data in SAS 9.1 were:

1. Fix the value of X_i, where i =the number of area
2. Fix the expected probability of zero in each small area $(P(Y_i = 0))$, then calculate: $Lambda_i = -\log(P(Y_i = 0))$
3. Generate : $\theta_i \sim Gamma(1,1)$
4. Calculate : $\lambda^* = \log(Lambda_i / \theta_i)$
5. Do linear regression between λ^* and X_i to get predicted of β_0 and β_1
6. Calculate : $\mu_i = \exp(X_i`\beta)$
7. Count: $parmlambda = \mu_i \times \theta_i$

8. Generate : $y_i \sim Poisson(parmlambda)$

And steps of analyzing data were:

1. Generate the negative binomial regression with *genmod* procedure in SAS 9.1 and Zero-Inflated Negative binomial Regression with *countreg* procedure in SAS 9.2.
2. Estimate the prior parameter, which are β and α ,
3. Calculate the estimation using EB method.
4. Calculate MSE for indirect estimation.
5. Calculate RRMSE (Root Relative Mean Square Error):

$$RRMSE(\hat{\theta}_i) = \frac{\sqrt{MSE(\hat{\theta}_i)}}{\hat{\theta}_i}$$

RESULT AND DISCUSSION

EB Estimator with Negative Binomial Regression on Estimated Prior Parameter

In case of non-excess-zero data, EB estimator with Negative Binomial Regression produced small and consistent MSE. Meanwhile, if data had contained excess-zero, approximately 30% or more with expected probability of zero is 0.6, it would have performed poorly and unreliably. As a result, EB estimation would produce negative values, which were meaningless as count estimators. RRMSE of EB estimator with NBR is increasing simultaneously with increase number of zero in data.

Table 1 Mean of MSE for EB estimator with NBR

Probability of zero	Mean of MSE	Median of MSE
0.1	0.333613	0.159448
0.2	0.353033	0.196963
0.3	0.400207	0.231014
0.4	0.41875	0.271669
0.5	0.452955	0.306641
0.6	-128.75	0.330078
0.7	2536.714	0.402092
0.8	-584495	0.301527
0.9	39135606	0.159682

Table 2 Mean of RRMSE for EB estimator with NBR

Probability of zero	Mean of RRMSE	Median of RRMSE
0.1	0.18188	0.1356
0.2	0.262006	0.205182
0.3	0.361013	0.299247
0.4	0.503566	0.422426
0.5	0.717336	0.588924
0.6	-0.38057	0.813734
0.7	-12.1606	1.355566
0.8	309.4627	2.115163
0.9	1.16E+10	6.639631

From table 1 shows that iterative process produced negative values of MSE, one simplest way to solve this problem is change the negative value to zero. This is do to the fact that, minimum value of MSE is zero. And the result were:

Table 3 Mean of MSE for EB estimator (II) with NBR

Probability of zero	Mean of MSE	Median of MSE
0.1	0.333613	0.159448
0.2	0.353033	0.196963
0.3	0.400207	0.231014
0.4	0.419181	0.271669
0.5	0.456258	0.306641
0.6	261.9677	0.330078
0.7	9500.073	0.402092
0.8	1444250	0.301527
0.9	41595285	0.159682

Table 4 Mean of RRMSE for EB estimator (II) with NBR

Probability of zero	Mean of RRMSE	Median of RRMSE
0.1	0.18188	0.1356
0.2	0.262006	0.205182
0.3	0.361013	0.299247
0.4	0.502895	0.422209
0.5	0.712314	0.585765
0.6	-0.34911	0.750369
0.7	-10.0163	0.995659
0.8	220.5437	1.104113
0.9	6.77E+09	0.557622

EB estimator (II) with NBR when data had expected probability of zero by 0.6 to 0.9 produced huge MSE. Similar with the MSE number, EB estimator (II) with NBR

produced big RRMSE when data had 0.8 to 0.9 expected probability of zero. But when data had 0.6 to 0.7 expected probability of zero, the value of RRMSE is negative due to the negative value of EB estimator.

EB Estimator with Zero Inflated Negative Binomial Regression on Estimated Prior Parameter

ZINB method produced EB estimator that was quite similar to which NBR method had given, although it slightly outperformed NBR method when the data only contained small number of zeros. In particular, if data had expected probability of zero by 0.1 to 0.5, ZINB would have produced bigger MSE for EB estimator than which NBR would have produced.

Whereas, if data had expected probability of zero by 0.6 to 0.7, ZINB would have given better estimator. The estimator was also unbiased as it covered parameter values adequately. However, ZINB would have begun to produce inconsistent estimator if data had expected probability of zero by 0.8 or more due to enormous MSE.

Table 5 Mean of MSE for EB estimator with ZINB

Probability of zero	Mean of MSE	Median of MSE
0.1	0.449506	0.168797
0.2	0.425844	0.201463
0.3	0.707983	0.277017
0.4	0.540615	0.284338
0.5	0.859252	0.334706
0.6	0.60793	0.376514
0.7	0.584152	0.255331
0.8	-1.27819	-1.4E-07
0.9	2954790	-1E-06

Table 6 Mean of RRMSE for EB estimator with ZINB

Probability of zero	Mean of RRMSE	Median of RRMSE
0.1	0.238099	0.136424
0.2	0.326097	0.210692
0.3	0.519524	0.317195
0.4	0.630776	0.420396
0.5	73228.07	0.662251
0.6	298.1671	1.033578
0.7	2181.192	1.942286
0.8	16269.7	3.754705
0.9	3.5E+278	6095.076

Besides, when data had expected probability of zero by 0.5 or more, EB estimator with ZINB produced huge RRMSE. It is because ZINB produced a small estimator, which is quite closed with the parameter value.

Table 7 Mean of MSE for EB estimator (II) with ZINB

Probability of zero	Mean of MSE	Median of MSE
0.1	0.450626	0.168797
0.2	0.428006	0.201463
0.3	0.716543	0.277017
0.4	0.549835	0.284338
0.5	0.94973	0.334706
0.6	0.749436	0.376514
0.7	1.501268	0.255331
0.8	1.755486	0
0.9	2954908	0

Table 8 Mean of RRMSE for EB estimator (II) with ZINB

Probability of zero	Mean of RRMSE	Median of RRMSE
0.1	0.23675	0.135647
0.2	0.320663	0.20709
0.3	0.506882	0.311937
0.4	0.606317	0.405926
0.5	65612.35	0.576343
0.6	234.0612	0.698814
0.7	1346.552	0.688797
0.8	7335.062	0
0.9	1.2E+278	0

EB estimator (II) with ZINB produced bigger average of MSE than EB estimator with ZINB. It is because, when negative value of MSE changed to zero, it doesn't have reduction factor in the average calculation**.**

CONCLUSION

Excess-zero in data would give high influence to the result of EB estimation. Conventional method such as negative binomial regression in prior estimation would provide unbiased and unreliable EB estimator for data with expected probability of zero by 0.6. It could be seen from the big number of MSE and negative value of estimator.

Meanwhile, EB estimation by ZINB method produced better and reliable estimator even though data had expected probability of zero by 0.6 to 0.7.

The ZINB methods just provide a reliable estimator for data with less than 53.33% of zeros. It because perform of ZINB decline when data had expected probability of zero by 0.8 or more. It identify in the big MSE and inconsistent estimator.

SUGGESTION

This research is still based on many assumptions and boundaries. If the assumptions and boundaries can be decreased, it will give better result because it will be closer to the real application. These are:

1. The generating process is not equal with real sampling process.
2. If population is generated first, it will be better because it can include the correction factor or offset.
3. The limited number of areas. It will be more interesting if experiment take account of larger number of areas, because the number of areas will influence data modeling.
4. Do the theoretical research about ZINB and EB estimator. So that will include all parameter of ZINB to the EB process.

REFERENCES

Hardin JW, JM Hilbe. 2007. *Generalized Linear Models and Extensions.* Texas: A Stata Press Publication.

Kismiantini. 2007. Pendugaan Statistik Area Kecil Berbasis Model Poisson-Gamma [Tesis] Bogor: Institut Pertanian Bogor, Fakultas Matematika dan Pengetahuan Alam.

McCullagh, P, J. A. Nelder. 1983. *Generalized Linear Models.* London: Chapmann and Hall.

Rao JNK. 2003. *Small Area Estimation*. New York: John Wiley & Sons.

Wakefield J. 2006. *Disease mapping and spatial regression with count data*. http://www.bepress.com/uwbiostat/paper286.pdf [24 April 2008].