INTRODUCTION

Background

Direct estimation is usually applied in big scale survey but it is sometime difficult to utilize such estimator in a smaller region especially the sample size is too small. In this case, 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. Maximum Likelihood Estimation was used with Negative Binomial Regression techniques to estimate the respective prior parameter. Moreover, Negative Binomial Regression was used to resolve over-dispersion problem in the data.

In reality, count data is not only characterized by over-dispersion, but sometimes by excess-zero. Excess-zero is a condition when the data contains too many zero or exceeds the distribution’s expectation. 100 observations from Poisson model with response mean of 4, we could expect that there will be 2 zeros. If the data have 30 zeros, it should be obvious that the distributional assumptions have been violated. Therefore, the estimated parameter and standard error will be biased (Hardin & Hilbe 2007). In this paper, Zero-Inflated models were adapted to solve this type of problem.

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 evaluate the performance of Zero-Inflated Count Models in estimating prior parameter for Small Area Estimation.

LITERATURE REVIEW

Direct Estimation

Direct estimates are generally “design based” in the sense that they make use of “survey weight” and associated inferences are based on the probability distribution by the sample design, with the population values held fixed (Rao 2003). In particular, direct estimates of a domain parameter are based only on the domain-specific sample data.

Data from sample survey have been used to be a reliable estimate of parameter. Ramsini, et al (2001) mentioned that direct estimates of small area are unbiased although it would have big variance cause it’s small sample size.

Small Area Estimation

The term of small area can be everything depending on our object of interest. It can be a city, age group, sex group, region and rural district. In general, small area is used to denote any domain which the direct estimation with adequate precision can not be produced (Rao 2003). It happens because the sample size in small area is too small. As a result, direct estimation based on sampling design is not capable to produce direct estimation with adequate precision.

Furthermore, small area estimation is developed as a statistic technique for estimating the parameter of small area. This technique 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). Indirect estimation is a process of estimating a domain’s parameter by connecting the information in that domain with another domain using an appropriate model. So, the estimator works by including other domain’s data (Kurnia & Notodiputro 2006).

Small Area Models

There are two link models in indirect estimation. First, traditional method based on implicit models that provide a link to relate small area through supplementary data. Second, explicit small area models that make specific allowance between area variations (Rao 2003). This research used the second model and it could be classified into two broad types of basic model:

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, \( \theta_i \), is related to area specific auxiliary data or covariate variables, \( x_i = (x_{i1}, ..., x_{ip})^T \), by a linear model:
\[
\theta_i = x_i^T \beta + b_i v_i, \quad \text{with } i = 1, \ldots, m,
\]

\[v_i \sim N(0, \sigma_v^2)\] are area-specific random effect and \[\beta = (\beta_1, \ldots, \beta_p)^T\] is \(p \times 1\) vector of regression coefficients. Therefore, \(b_i\) are known as positive constants. For making inferences about \(\theta_i\), direct estimators \(y_i\) are assumed available. Accordingly, assuming:

\[y_i = \theta_i + e_i, \quad \text{where } i = 1, \ldots, m,\]

with sampling error \(e_i \sim N(0, \sigma_e^2)\) and \(\sigma_e^2\) are known. At the end, both models are combined and as a result is a new model:

\[y_i = x_i^T \beta + b_i v_i + e_i, \quad \text{where } i = 1, \ldots, m.\]

(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,

\[x_y = (x_{y1}, \ldots, x_{yn})^T,\]

correspondingly, a nested regression model:

\[y_i = x_i^T \beta + v_i + e_i, \quad \text{where } i = 1, \ldots, m, \quad \text{and } j = 1, \ldots, n_i,\]

\[v_i \sim N(0, \sigma_v^2)\] and also \(e_i \sim N(0, \sigma_e^2)\).

**Empirical Bayes Methods**

The Bayesian approach is based on Bayes Law, which was found by Thomas Bayes. This law was introduced by Richard Proce in 1763; two years after Thomas Bayes passed away. In 1774 and 1781, Laplace gave the details and relevancies for modern Bayesian statistics. (Gill 2002 in Kismiantini 2007).

Novick in Good (1980) mentioned that Bayes method is difficult to adopt and sometimes is very sensitive due to the requirement of prior probability information which is usually difficult to obtain. Robbin (1955) introduced Empirical Bayes methods by assuming a particular prior distribution estimating based on the sample. Rao (2003) said that EB (Empirical Bayes) and HB (Hierarchical Bayes) are compatible for binary and count data in Small Area Estimation. Therefore, EB method was used in this research.

Rao (2003) summarized EB methods in Small Area Estimation as follows:

1. Obtain the posterior probability density function of the small area parameter.
2. Estimate the parameters from the marginal density function.
3. Use the estimated posterior density for inferences regarding the parameters of interest.

**Poisson-Gamma Models**

Poisson model is a standard model in dealing with count data. Generally, count data can be suffered by over-dispersion problem. Therefore, a Poisson formula had been developed to accommodate extra variance from sample data. Two-stage models have been introduced for count data, known as mixed model Poisson-Gamma. Wakefield (2006) introduced Poisson-Gamma model which was easier to use, with SMR (Standard Mortality Ratio) as a direct estimator. This study used Wakefield model with alteration in direct estimator.

Let \(y_i\) be a number of specific individual at small area-\(i\), which has specific characteristic of interest, and written as follow:

\[y_i = \sum_j y_{ij}\]

\(y_{ij}\) are the-\(jth\) object at the-\(ith\) small area where \(j = 1, \ldots, n\) and \(i = 1, \ldots, m\).

First stage \(y_{ij} \sim \text{Poisson}(\mu_i, \theta_i)\) is assumed where \(\mu_i = \mu(x_i, \beta)\) describes a regression model in area level, \(x_i\) is a vector of covariates and \(\beta = (\beta_1, \ldots, \beta_p)^T\) is a vector of regression coefficients.

Second stage, distribution \(\theta_i \sim \text{gamma}(\alpha/\mu_i)\) is assumed as a prior distribution with mean 1 and variance \(1/\alpha\). Then the marginal distribution \(y_i | \beta, \alpha\) is negative binomial.

Moreover Wakefield (2006) used Bayes Theorem and acquired posterior distribution as:

\[\theta_i | y_i, \mu \sim \text{gamma}(y_i + \alpha, 1 + \alpha/\mu_i)\]

and EB estimator as:

\[\hat{\theta}^{EB}_i = \frac{1}{\gamma_i} (\bar{\gamma} \hat{\theta} + (1 - \bar{\gamma}) \mu_i)\]

with \(\bar{\gamma} = \mu_i / (\alpha + \mu_i)\), \(\hat{\theta} = y_i\) are direct estimation from \(\theta_i\), and \(y_i\) are the number of observation.

**Negative Binomial Regression**

The negative binomial regression model seems have been first discussed by Anscombe (1972). Others have pointed out its success in dealing with over-dispersed count data.
Lawless (1987) elaborated the mixture model parameterization of the negative binomial, providing formulas for its log likelihood, mean, variance and moments. Later Breslow (1990) cited Lawless’ work and since its inception to the late 1980’s, the negative binomial regression model have been construed as a mixture model that is useful for accommodating otherwise over-dispersed Poisson data (Hardin & Hilbe 2007). The negative binomial distribution function is written as:

\[ g(y | x) = \frac{\Gamma(y + k)}{\Gamma(y + 1) \Gamma(k)} \left( \frac{\mu}{\mu + k} \right)^{y} \left( \frac{\mu + k}{\mu} \right)^{k} \]

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

**Over-disperse at Count Data**

Count data for Poisson regression including by over-disperse if variance bigger than mean or if the expected value of variance is smaller than expected. This phenomenon is written as:

\[ Var(y_i) > E(y_i) \]

(McCullagh & Nelder 1989).

**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).

Lambert (1992) first described this type of mixture model in the context of process control in manufacturing. It has since been used in many applications and is now found discussed in nearly every book or article dealing with count response models.

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 \( \text{Pr}(0) \) 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):

\[ \text{Pr}(y = 0) = B(0) + (1 - Z) \text{Pr}(0) \]

The probability of a nonzero count is:

\[ \text{Pr}(y = k; k > 0) = [1 - B(0)] \text{Pr}(k) \]

This model would produce two groups of parameter, one is zero-inflation parameter which shown that the covariate significantly contribute to having a zero outcomes. And the other parameter is negative binomial parameter which modeling the response with the covariate.

**Zero-Inflated Negative Binomial**

There are many kinds of zero-inflated model; each model has plus and minus and is used in different type of data. Zero-Inflated negative binomial is one kind of them. 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:

\[ P(Y_i = y_i | x_i) = \begin{cases} \phi(x_i)\phi(0) + (1 - \phi(x_i))g(0 | x_i) & \\
(1 - \phi(x_i))g(y_i | x_i) & \end{cases} \]

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

\[ g(y_i | x_i) = \frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1) \Gamma(\alpha)} \left( \frac{\alpha}{\mu + \alpha} \right)^{y_i} \left( \frac{\mu}{\mu + \alpha} \right)^{\alpha} \]

Mean and variance of ZINB are:

\[ E(y_i | x_i) = \mu(1 - \phi) \]

\[ V(y_i | x_i) = \mu(1 - \phi)(1 + \mu(\phi + \alpha)) \]

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

Jackknife methods is one of general methods used in survey, because it’s unpretentious concept (Liang, Lahiri and Wan 2002). 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 \( i = 1, \ldots, m \) and performs parameter estimation.

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

1. Assume that \( \hat{\theta^{EB}} = k(y_i, \hat{\beta}, \hat{\alpha}) \), \( \hat{\theta}_{EB,i} = k(y_i, \hat{\beta}_{EB,i}, \hat{\alpha}_{EB,i}) \), then calculate:

   \[ \hat{M}_i = \frac{m - 1}{m} \sum_{i=1}^{m} (\hat{\theta}_{EB,i} - \hat{\theta}_{EB,\cdot})^2 \]

2. Calculate the delete-i estimator \( \hat{\beta}_{EB,i} \) and \( \alpha_{EB,i} \), then calculate:
\[ \hat{M}_i = g_i(\hat{\beta}, \hat{\alpha}; y_i) \frac{m-1}{m} \sum_{j=1}^m [g_j(\hat{\beta}, \hat{\alpha}; y_j) - g_i(\hat{\beta}, \hat{\alpha}; y_i)] \]

And \( g_i(\hat{\sigma}^2) \) is the variance estimator of posterior distribution which is used to measure the variability associated with \( \theta_i \).

The use of \( g_i(\hat{\sigma}^2) \) is leads to severe of underestimation of \( )\hat{\sigma}(1 \)

related with estimation in prior parameter. Therefore, the estimator \( \hat{M}_i \) correct the bias of \( g_i(\hat{\sigma}^2) \).

3. Calculate the jackknife estimator of \( \text{MSE}(\hat{\theta}) \) as:

\[ \text{MSE}_j(\hat{\theta}) = \hat{M}_{ji} + \hat{M}_{2i} \]

**METHODOLOGY**

**Data**

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.

**Methods**

The following steps in generating data using SAS 9.1 were used:

1. Fix the value of \( X_i \), for the-\( i \)-th area
2. Define the expected probability of zero in each small area \( P(Y_i = 0) \), then calculate: \( \Lambda_i = -\log(P(Y_i = 0)) \)
3. Generate: \( \theta_i \sim \text{Gamma}(1,1) \)
4. Calculate: \( \lambda^* = \log(\Lambda_i/\theta_i) \)
5. Fit linear regression between \( \lambda^* \) and \( X_i \) to obtain \( \hat{\beta}_i \) and \( \hat{\gamma}_i \)
6. Calculate: \( \mu_i = \exp(X_i/\beta) \)
7. Calculate: \( \text{parmlambda} = \mu_i \times \theta_i \)
8. Generate: \( y_i \sim \text{Poisson}(\text{parmlambda}) \)

Moreover, in analyzing data the following steps were applied:


2. Estimate the prior parameter, which are \( \beta \) and \( \alpha \)
3. Estimate using EB method.
4. Calculate \( \text{MSE} \) for indirect estimation.
5. Calculate \( \text{RRMSE} \) (Root Relative Mean Square Error):

\[ \text{RRMSE}(\hat{\theta}) = \sqrt{\frac{\text{MSE}(\hat{\theta})}{\theta_i}} \]

**RESULT AND DISCUSSION**

**Estimation of Prior Parameter is Based on EB Method with Negative Binomial Regression**

In case of non-excess-zero data, the estimator produced small and consistent MSE. Meanwhile, if the number of excess-zero is approximately 30% or more with expected probability of zero 0.6, the performance of estimates tends to be unreliable. As a result, EB estimation produced negative values.

RRMSE of the estimator increases simultaneously along with the increase of number of zero in the data. Furthermore, if the data contain excess zero at least 30%, the estimator is unreliable.

**Table 1** MSE and RRMSE of EB Estimator with NBR

<table>
<thead>
<tr>
<th>Probability of zero</th>
<th>Mean of MSE</th>
<th>Median of MSE</th>
<th>Mean of RRMSE</th>
<th>Median of RRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.33</td>
<td>0.16</td>
<td>0.18</td>
<td>0.13</td>
</tr>
<tr>
<td>0.2</td>
<td>0.35</td>
<td>0.20</td>
<td>0.26</td>
<td>0.20</td>
</tr>
<tr>
<td>0.3</td>
<td>0.40</td>
<td>0.23</td>
<td>0.36</td>
<td>0.30</td>
</tr>
<tr>
<td>0.4</td>
<td>0.42</td>
<td>0.27</td>
<td>0.50</td>
<td>0.42</td>
</tr>
<tr>
<td>0.5</td>
<td>0.45</td>
<td>0.31</td>
<td>0.72</td>
<td>0.59</td>
</tr>
<tr>
<td>0.6</td>
<td>-128.75</td>
<td>0.33</td>
<td>-0.38</td>
<td>0.81</td>
</tr>
<tr>
<td>0.7</td>
<td>2536.71</td>
<td>0.40</td>
<td>-12.16</td>
<td>1.35</td>
</tr>
<tr>
<td>0.8</td>
<td>-584495</td>
<td>0.30</td>
<td>309.46</td>
<td>2.11</td>
</tr>
<tr>
<td>0.9</td>
<td>39135606</td>
<td>0.16</td>
<td>1.16E+10</td>
<td>6.64</td>
</tr>
</tbody>
</table>

**Table 2** MSE (II) and RRMSE (II) of EB Estimator with NBR

<table>
<thead>
<tr>
<th>Probability of zero</th>
<th>Mean of MSE</th>
<th>Median of MSE</th>
<th>Mean of RRMSE</th>
<th>Median of RRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.33</td>
<td>0.16</td>
<td>0.18</td>
<td>0.13</td>
</tr>
<tr>
<td>0.2</td>
<td>0.35</td>
<td>0.20</td>
<td>0.26</td>
<td>0.20</td>
</tr>
<tr>
<td>0.3</td>
<td>0.40</td>
<td>0.23</td>
<td>0.36</td>
<td>0.30</td>
</tr>
<tr>
<td>0.4</td>
<td>0.42</td>
<td>0.27</td>
<td>0.50</td>
<td>0.42</td>
</tr>
<tr>
<td>0.5</td>
<td>0.46</td>
<td>0.31</td>
<td>0.71</td>
<td>0.58</td>
</tr>
<tr>
<td>0.6</td>
<td>261.97</td>
<td>0.33</td>
<td>-0.35</td>
<td>0.75</td>
</tr>
<tr>
<td>0.7</td>
<td>9500.07</td>
<td>0.40</td>
<td>-10.02</td>
<td>0.99</td>
</tr>
<tr>
<td>0.8</td>
<td>1444250</td>
<td>0.30</td>
<td>220.54</td>
<td>1.10</td>
</tr>
<tr>
<td>0.9</td>
<td>41595285</td>
<td>0.16</td>
<td>6.77E+09</td>
<td>0.56</td>
</tr>
</tbody>
</table>