Study of Small Area Estimation on Overdispersion Data with the Zero-Inflated Poisson Regression

Dian Christien Arisona\textsuperscript{1}, Anang Kurnia\textsuperscript{2}, Kusman Sadik\textsuperscript{3}\\\textsuperscript{1,2,3}Department of Statistics, Faculty of Mathematics and Natural Science, Bogor Agricultural University, INDONESIA

ABSTRACT

Small area estimation method is a method that is widely used to estimate the parameters of sub populations based on available survey data with very small sample sizes. Infant mortality data is a rare occurrence in a certain period of time so that small area estimation model is used with Poisson distribution. However, infant mortality data has a large number of zero value in each survey unit, so that the data indicating the existence of overdispersion that will certainly have an effect on the estimation process, then the proposed model is small area estimation with zero-inflated Poisson. This study compares the small area estimation with Poisson regression and small area estimation with zero-inflated Poisson regression. The goodness of the model is tested through simulation and revealed that small area estimation with zero-inflated Poisson regression is better than small area estimation with Poisson regression. Implementation on infant mortality data in West Java also shows that small area estimation with zero-inflated Poisson regression is better than small area estimate with Poisson regression. It is indicated by smaller standard error of small area estimation with zero-inflated Poisson regression than small area estimation with Poisson regression model.

Keywords— Overdispersion, small area estimation, zero-inflated Poisson

I. INTRODUCTION

Sample surveys have long been recognized as cost-effective means of obtaining data on a wide-ranging topics of interest at frequent intervals over time. In analysis on the indicators obtained from surveys often found data with variance smaller than average. This condition is known as overdispersion. Overdispersion can be caused by excess zeros value in the response variable. Zero-inflated Poisson (ZIP) model is a method that can be used to overcome the overdispersion caused by excess zeros in binary response. Previous research has begun by developing a zero-inflated Poisson regression with an application to defects in manufacturing by Lambert (1992). Then, Buu et al. (2013) developed a statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. Samuel and Kwabena (2015) developed a statistical model for overdispersed count outcome with many zeros. In the same year Baghban et al. (2015) also develops ZIP in the health sciences that is the application of zero-inflated Poisson mixed models in prognostic factors of hepatitis C.

Another problem from survey data is the size of the sample that is too small to estimate the parameters of the sub population. Small area estimation (SAE) has been used to estimate data with very small sample size. Small area estimation can be applied in health surveys to analyze infant mortality data. Proportion of infant mortality at the level level can be used as an indicator of health development and life quality of the people in those areas (IDHS, 2012). The indicator is also used to monitor and evaluate population growth, health, programs, and government policies.

Several studies about infant mortality rates with small area approaches have been discussed, such as Yadav and Ladusingh (2013) estimating infant mortality with synthetic models. Hajarisman (2013) estimates infant mortality with small area modeling through two levels model Poisson Bayes regression approach. Based on those studies, ignoring the large proportion of zeros in the data and modeling with linear mixed model (LMM) will result in biased estimators, unstable estimates, and other problems that may be caused. Research on SAE on data with many zero values has been done previously by Krieg (2015), which developed SAE with a zero-inflated model without ignoring the value of zero but dividing the data into two parts: zero value and non-zero value. The study of this research is SAE with Poisson regression and ZIP regression in overdispersion data and how to applied it to infant mortality data in West Java in 2012.
II. METHODOLOGY

This research uses simulation to measure the model fitness between Poisson regression and ZIP regression by relative root mean square error (RRMSE) and relative bias of the proportion estimates.

In this simulation we fixed area size is district/regency = 27, unit size is village = 300. In each simulation we generated n population values of auxiliary variable x from the normal distribution with mean 30 and standard deviation 5. Zero-inflated population values for y were generated by first generating Bernoulli (0/1) random variable with fixed probability \( p \) (proportion of non-zero values in the data) and multiplying them by for each area. A sample of size \( n=135 \) was then selected from the simulation population. Sampling was via stratified random sampling, with the strata defined by the small areas with average population. Sampling was via stratified random sampling, with the strata defined by the small areas with average population. Simulation was generated n population values of auxiliary variable x from the normal distribution with mean 30 and standard deviation 5. Zero-inflated population values for y were generated by first generating Bernoulli (0/1) random variable with fixed probability \( p \) (proportion of non-zero values in the data) and multiplying them by for each area.

A total of 1000 simulations were carried out and calculated RRMSE and RB \( \text{prop}_{i,q} \) of the formula:

\[
\text{RRMSE} = \sqrt{\frac{1}{m} \sum_{q=1}^{m} (\text{prop}_{i,q} - \text{prop}_{i,\text{mean}})^2 / m}
\]

\[
\text{RB}_i = \frac{\sum_{q=1}^{m} (\text{prop}_{i,\text{mean}} - \text{prop}_{i,q})}{\text{prop}_{i,\text{mean}}}
\]

where,

- \( \text{prop}_{i,q} \) : proportion estimation in i-th area by using SAE ZIP on the q-th simulation.
- \( \text{prop}_{i,\text{mean}} \) : The proportion of parameters from the i-th region.
- \( m \) : The number of simulations.

Then, comparing RRMSE and RB from Poisson regression and ZIP regression.

III. RESULT AND DISCUSSION

The simulation results are presented in Table 1, Figure 1, and Figure 2. Small area model used in the simulation is model unit level in 27 areas. Simulation is designed to calculate RRMSE and relative bias of the estimates.

Table 1: Summary of the Simulation using SAE ZIP and SAE Poisson Method

<table>
<thead>
<tr>
<th>Area</th>
<th>RRMSE (%)</th>
<th>RB (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZIP</td>
<td>Poisson</td>
<td>ZIP</td>
</tr>
</tbody>
</table>

The simulation result in Table 1 showed that the proposed method produced smaller RRMSE (average of RRMSE = 22.607 %) when compared to Poisson models (average of RRMSE = 26.757 %).

- **Figure 1**: Line chart of RRMSE (%) of the proportion estimate between SAE ZIP and SAE Poisson.
The relative bias of SAE ZIP method is closer to zero than SAE Poisson method (Figure 2). It showed that the proposed SAE ZIP method has better fit than SAE Poisson method to overcome overdispersed Poisson data.

IV. CONCLUSION

The results of data simulation on data with similar characteristics with infant mortality data in West Java in 2012 that is RRMSE for the estimated number of infant mortality and the estimated proportion of infant mortality and RB for the estimated number of infant mortality and estimated proportion of infant mortality, showed SAE ZIP better than SAE Poisson because SAE ZIP has a smaller RRMSE for each estimated number of infant mortality and an estimated proportion of infant mortality as well as near-zero RB for each estimated infant mortality and estimated proportion of infant mortality.

REFERENCES

[8] [IDHS]. Indonesia Demographic and Health Survey. Jakarta, Indonesia. 2012. (Book)