Journal of Physics: Conference Series

PAPER • OPEN ACCESS

AMAZonn (*A Multicollinearity-adjusted Adaptive LASSO for Zero-infated Count Regression*) with Weight of *Expectation Maximization Standard Error Adaptive LASSO* (SEAL AL) for Zero Inflated Poisson Data

To cite this article: Ismah *et al* 2021 *J. Phys.: Conf. Ser.* **1776** 012050

AMAZonn (*A Multicollinearity-adjusted Adaptive LASSO for Zero-infated Count Regression*) with Weight of *Expectation Maximization Standard Error Adaptive LASSO* (SEAL AL) for Zero Inflated Poisson Data

Ismah[1], Khairil Anwar Notodiputro[2], Bagus Sartono[2]

[1] Mathematics Education Department, Universitas Muhammadiyah Jakarta, Indonesia
[2] Institut Pertanian Bogor, Indonesia

Corresponding email: ismah.fr@gmail.com

Abstract. Lasso as selecting predictor variables continue to experience development like an adaptive Lasso which gives weight value in its formula. In a modelling case, variables selecting technique is needed in order to get a stable model, however, multicollinear cases are often found in several cases which caused the models obtained are unstable since the values of variance became large. Besides, data counting on response variable, with the presence of excess zero, causes the linear model cannot be applied, hence, using generalized linier modelling with zero inflated poisson (ZIP) model can become the solution. Thus, in this research, ZIP model will be applied after selecting the variables through AMAZoon (*A Multicollinearity-adjusted Adaptive LASSO for Zero-infated Count Regression*) with Weight of *Expectation Maximization Standard Error Adaptive LASSO* (SEAL AL), and the comparison towards the results gained by ZIP model without prior variables selection will be done. The comparison was seen based on the value of the smallest *Akaike Information Criterion* (AIC). The data analysis revealed that there was multicollinear case in the data, and also ZIP model, after conducting the variables selection by using AMAZoon with Weigth SEAL AL, reached smaller AIC than ZIP model which having no variable selection. Therefore, ZIP model after the selection of variables by using AMAZoon with Weight of SEAL AL was better to be used when multicollinear happened.

1. Introduction

In a statistical analysis, modelling is a technique which relates variables, i.e. between response variable and predictor variable. The method which generally used is the regression analysis with parameter estimation technique with *least square* method. Least Squares Method stricts with few assumptions including there is no large correlation (multicollinear) between predictor variables, and the variables involved in the modelling are continues data or distributed normally. These assumptions, oftentimes, cannot be fulfilled in the real data. Big number of predictor variables in modelling causes multicollinear happens, also, variables as discrete data or categorical data is frequently found. In the statistical analysis, response variables which are not normally distributed can be accomplished by *generalized linier modelling* if the response variables independent and identical distributed from exponential family.

In several cases, there are response variables with counting data which poisson distributed so that they can be accomplished by poisson model. However, in counting data, *excess zeros* is frequently found, which means there is no amount in that observed data. Thus, the presence *excess zeros* cannot be accomplished with poisson model because in its model the assumption that should be fulfilled is that the variants and averages of response variables are the same. While in the case of *excess zeros* in which variant value of response variable is bigger than the average value, thus it is called overdispersion which causes the linear model cannot be applied. This overdispersion case can be solved with various methods previously introduced such as *zero inflated poisson* (ZIP) which proposed by Lambert (1992), *zero inflated binomial* (ZIB), *zero inflated generalized poisson* (ZIGP) by Famoye (2003), amd *zero inflated negative binomial* (ZINB).

Multicollinear in modelling can influence the variant value becomes bigger and alleged value obtained with least squares method becomes unstable. Multicollinear can be solved by eliminating one of the variables which correlated big, and or by adding new predictor variables. Besides those two ways, multicollinear can also be solved by selecting the coefficient model or regression until zero. The method that can be used in selecting the coefficient model until zero is *Least Absolute shrinkage and Selection Operator* (LASSO) which was introduced by Tishibrani (1996). LASSO method is getting developed, its development was done by Zou (2006) by giving adaptive weight into the equation of LASSO so that this method is called adaptive LASSO.

Banerjee et,al. (2018) used adaptive LASSO for the regression of Zero Inflated Poisson. It was done since sometimes in the regression of ZIP, multicollinear cases were found in the variables with big dimensions. Therefore, model gained by ZIP is unstable. Banerjee used simulation data and real data to determine the best variable coefficient selections method in the regression of ZIP. The methods compared were AMAZonn (*A Multicollinearity-adjusted Adaptive LASSO for Zero-infated Count Regression*) with two kinds of adaptive weights, i.e. first, the opposite of the maximum estimator likelihood called EM AL which stands for *Expectation Maximization Adaptive LASSO*, and second, the opposite of the maximum estimator likelihood which was divided from standard error estimator called EM SEAL which stands for *Expectation Maximization Standard Error Adaptive LASSO*. The analysis results used simulation data conducted by Banerjee, et al. between AMAZonn with EM AL, AMAZonn with EM SEAL and EM LASSO showed that AMAZonn with EM SEAL was better compared to others based on the values of *bayesian information criterion* (BIC) gained from those three methods.

Based on the research conducted by Sulistyaningsih (2019), it was found that the results of ZIP was better than ZIGP in overcoming overdispersion problem on the data of maternal mortality rate in Bali Province. ZIGP is the development of ZIP model to overcome counting data with a lot of zero values. However, the length of short interval of the smallest counting data to the biggest one made this ZIGP model became inappropriate to be used so that ZIP was considered better. In a certain case, ZIGP was not better than another simpler models reffering to the research results conducted by Ismah (2019) who compared poisson regression model (POI), Generalized Poisson (GP), ZIP and ZIGP toward the data of overdispersion of number of child's deaths in the family which was taken from the data of Indonesian Demographic and Health Survey 2017, in which GP model was better compared to the other three models based on *Akaike Information Criterion* (AIC) obtained by each model.

Based on the previous research, this research will show data analysis results of overdispersion of the number of child's deaths in the family obtained from Indonesian Demographic and Health Survey in 2017, using ZIP model with no variable coefficient selection and ZIP model with variable coefficient selection by applying tehnique of adaptive lasso AMAZonn with Weight of EM SEAL. The analysis results of those two methods will be compared to determine which model wss better based on the value of AIC.

2. Materials

2.1. Zero Inflated Poisson Model

For example $\boldsymbol{y} = (y_1, y_2, y_3, \cdots, y_n)'$ and **y** has distribution poison, so the Probability Density Function and parameter estimated use maximum likelihood method and link function shown in the following table 1.

Table 1. Probability Density Function and Parameter Estimated (*Maximum Likelihood*)

Model	Criterion	
	Probability Density Function	**Mean and Variance**
ZIP	$P(Y_i = y_i) = \begin{cases} \varphi_i + (1-\varphi_i)e^{-\mu_i}, \varphi_i = 0 \\ (1-\varphi_i)\dfrac{e^{-\mu_i}{\mu_i}^{y_i}}{y_i!}, \ \ y_i = 1, 2, \ldots, ; 0 \le \varphi_i \le 1 \end{cases}$	$E(y) = (1-\varphi_i)\mu;$ $Var(y) = (1-\varphi_i)({\mu_i}^2 + \mu_i) - (1-\varphi_i)^2{\mu_i}^2$
	Parameter Estimated (*Maximum Likelihood*)	**Link Function**
	For $y_i = 0$ $\prod_{y_i=0} \dfrac{e^{(X_i^T\gamma)} + e^{\left(-e^{(X_i^T\beta)}\right)}}{\left(1 + e^{(X_i^T\gamma)}\right)}$ For $y_i > 0$ $\prod_{y_i>0} \dfrac{e^{\left(-e^{(X_i^T\beta)}\right)}\left(e^{(X_i^T\beta)}\right)^{y_i}}{\left(1 + e^{(X_i^T\gamma)}\right)y_i!}$	$\text{logit}\,\varphi_i = \gamma_0 + \sum_{j=1}^{k} z_{ij}\gamma_j \,; i = 1, 2, \ldots, n$ $\ln \mu_i = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j \,; i = 1, 2, \ldots, n$

x_{ij} and z_{ij} are covariate vector for model count and zero,$\boldsymbol{\beta_j} = (\beta_1, \beta_2, \cdots, \beta_k)'\gamma_j = (\gamma_1, \gamma_2, \cdots, \gamma_k)'$

2.2. AMAZonn Method

AMAZoon uses two weights of EM adaptive Lasso, first, the opposite of the maximum estimator likelihood and second, the opposite of the maximum estimator likelihood which was divided into standard error for each parameter such as presented in the following table 2.

Tabel 2. Weight of AMAZonn

Weight Scheme	Count	Zero
AMAZoon-EM AL	$\dfrac{1}{\lvert\hat{\beta}_{jML}\rvert}$	$\dfrac{1}{\lvert\hat{\gamma}_{jML}\rvert}$
AMAZoon-EM SEAL	$\dfrac{SE(\hat{\beta}_{jML})}{\lvert\hat{\beta}_{jML}\rvert}$	$\dfrac{SE(\hat{\gamma}_{jML})}{\lvert\hat{\gamma}_{jML}\rvert}$

The formula of ZIP regression with EM adaptive Lasso was defined by Tang et al (2014), namely:

$$\widehat{\boldsymbol{\theta}}^* = arg\min\{-L(\boldsymbol{\theta})\} + v_1\sum_{j=1}^{k} w_{1j}\lvert\beta_j\rvert + v_2\sum_{j=1}^{k} w_{2j}\lvert\gamma_j\rvert$$

In which $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$vector paramater with known weights of $w_{1j} = (w_{11}, w_{12}, \cdots, w_{1k})'$ and $w_{2j} = (w_{21}, w_{22}, \cdots, w_{2p})'$. Qian and Yang (2013) the opposite of estimator maximum likelihood as weight is not always stable, commonly when multicollinear happens. Therefore, Banarjee et al (2018) concluded that EM SEAL is more stable based on the analysis results used simulation data.

3. Methods

Data used in this research are: number of child's deaths in the family as dependent variable (Y), while the independent variables consist of mother's last education (X_1), father's last education (X_2), type of contraceptives used (X_3), place of delivery (X_4), wealth index combined (X_5) and marital status (X_6). The data were collected from Indonesian Demographic and Health Survey in 2017. The determination of both dependent and independent variables was based on the results of the research conducted by Baiye et, al. (2018) who overcame the overdispersion in semi parametric spatial model on ferlility data gained from National Demographic and Health Survey (NDHS) in Nigeria in 2013. The descriptions of the data used in this research can be seen in Table 3.:

Table 3. The Description of Variables of the Research

NO	Variabel	Description
1	Number of child deaths in the family (Y)	0 = None
2	Highest educational level (X_1)	0=noeducation 1=primary 2=secondary 3=higher
3	Husband/partner's education level (X_2)	0=no education 1=primary 2=secondary 3=higher 4=don't know
4	Contraceptive used and intention (X_3)	1=using modern method 2=using traditional methods 3=Non-user- intend to use later 4=does not intend to use 5=Never had sex
5	Place of delivery (X_4)	10=home 11=respondent's home 12=other home 20=public sector 21=goverment hospital 22=goverment clinic 23=goverment health centre 26=other public sector 30=private sector 31=private hospital/clinic 37=other private sector 96=other
6	Wealth index combined (X_5)	1=poorest 2=poorer 3=middle 4=richer 5=richest
7	Current marital status (X_6)	1=married 2= single

The method used in this research wss ZIP regression with and without variable coefficient selection used the algorithm AMAZonn with weights of SEAL AL. Algorithm EM is streamlining parameter estimators in optimizing the formula of ZIP regression with EM adaptive Lassp defined by Tang et, al (2014),$z_i = 1$ if y_i from zero state, and $z_i = 0$ if y_i from count state with i = 1, 2, ..., n, so that the following formula was obtained (Barnajee et all (2018)).

$$Q^*(\theta) = -L(\theta) + v_1 \sum_{j=1}^{k} w_{1j}|\beta_j| + v_2 \sum_{j=1}^{k} w_{2j}|\gamma_j|$$

with $L(\theta) = \sum_{i=1}^{n}[z_i X_i \gamma - \log(1 + \exp(X_i\gamma)) + (1 - z_i)\{y_i X_i \beta - (y_i + 1)\log(1 + X_i\beta\}]$. In order to accomplish $Q^*(\theta)$ iterated was done to get the convergent results. The minimum of Bayesian information criterion (BIC) in this case did not only show the performance of obtained model but also used to determine the parameter tunning (θ).

4. Results and Discussion

The descriptions of the research data are as follows:

```
      X1              X2              X3              X4              X5             X6              Y
 Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :11.00   Min.   :1.000   Min.   :1.00   Min.   :0.0000
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:11.00   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.0000
 Median :2.000   Median :2.000   Median :1.000   Median :23.00   Median :3.000   Median :1.00   Median :0.0000
 Mean   :1.896   Mean   :1.879   Mean   :1.756   Mean   :22.83   Mean   :2.683   Mean   :1.02   Mean   :0.1336
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:31.00   3rd Qu.:4.000   3rd Qu.:1.00   3rd Qu.:0.0000
 Max.   :3.000   Max.   :9.000   Max.   :4.000   Max.   :96.00   Max.   :5.000   Max.   :2.00   Max.   :6.0000
```

The changing response of Y is number of child's deaths gathered from 17.212 families, source of the data gained from Indonesian Demographic and Health Survey, in which minimum of 0 number of child's deaths in the family and maximum 6 children's deaths in the family. Bar chart frequency of number of child's deaths in the family of 0, 1, 2, 3, 4, 5, and 6 children shown in the following picture 1. A total number of 89% from total number of the family declared that they had no death child (Y=0), the presence of very high data count in the changing response was excess zero. Therefore, the data in the analysis used zero inflated method. The followings are the analysis results of ZIP method before and after the selection of variable coefficient, in which the coefficient variable selection applied adaptive lasso AMAZonn with weights of EM SEAL.

Based on the results of data analyisis used ZIP method before the selection of variables had been conducted by Ismah, et.al (2018) which gained AIC model value of 13589. While the analysis results which used ZIP method after the selection of variables by applying adaptive lasso AMAZonn technique with weight of EM SEAL shown in the following table 4.

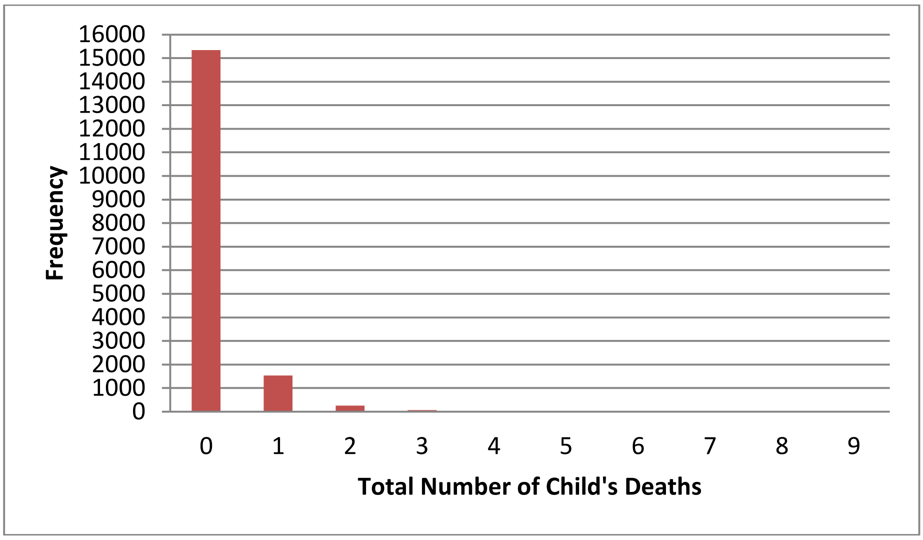

Figure 1. The Frequency of Total Number of Child's Deaths

Table 4. Estimating Model Parameter

Remarks	**Count**		**Zero**	
	Estimate	**SE**	**Estimate**	**SE**
Intercept	-0.2861	0.17744	-6.4571	6.40463
X1	-0.4521	0.03714	0	-
X2	0	-	0	-
X3	0.10141	0.02093	0	-
X4	-0.0199	0.00318	0	-
X5	-0.1488	0.02189	0	-
X6	-0.3436	0.15158	0	-
AIC	0.0019		13467.01	
BIC	0.0019		13529.04	
Theta	0.600702			
Log Likelihood	-6725.507			

Table 4 shows that in variable count model, only one which was selected, i.e. X2, namely husband/partner's education level, by removing excess zero in this model, it can be seen that gained value of AIC for this model was very small, i.e. 0.0019 in which Qian dan Yang (2013) explained that the opposite of estimator maximum likelihood as weight is not always stable, in which it commonly happens in multicollinear, so that it can be concluded that multicollinear between predictor variables happened in this case, it was shown by the value of AIC obtained became under-estimate.

In zero model in which excess zero involved in this model, it was seen that all variable of X1 until X6 was being selected and gained the AIC value which was more stable, i.e. 13467.01, with the value of parameter tuning (theta) optimum reached 0.600702 and the value of log likelihood -6724.507. The comparison of ZIP method before and after the selection of variables, through the selection of variable coefficient adaptive lasso AMAZonn technique with weight of EM SEAL seen based on the smallest value of AIC obtained from zero model, and it can be concluded that ZIP method after the selection of variables in zero model was better than ZIP method before the selection of variables.

The range values between minimum value with the smallest maximum value in the data count of variable response will influence the results analysis of ZIP model, and it has been known from the data used in this research which had very small reached value of 6 so that the result analysis became less optimum. Due to the limitation of the researchers in getting the data needed in implementing the method , thus, further research should apply ZIP modelling after conducting the selection of variables by using AMAZoon EM SEAL with suitable data such as big number of predictor variables, high multicollinear happens and big reached value in data count in response variables.

References

[1] Baiye, E.A., Oluwayemisi O. Alaba, Olusanya E. Olubusoye, J. O. Olaomi. Handling Overdispersion in Spatial Semi-parametric Modeling of Fertility Data

[2] Banerjee, P., Broti G., Himel M., Shrabanti C., Saptarshi C 2018 A Note on the Adaptive LASSO for Zero-Inflated Poisson Regression *Hindawi Journal of Probability and Statistics* Volume 2018, Article ID 2834183, 9 pages https://doi.org/10.1155/2018/2834183

[3] Lambert, D 1992 Zero-inflated Poisson regression, with an application torandom defects in manufacturing *Technometrics*, 34, 1-14

[4] Famoye, F and Singh, K, 2006 Zero-Inflated Generalized PoissonRegression Model with an Application to Domestic Violence Data *J of Data Science* 4, 2006, 117-130

[5] Tibshirani, R 1996 Regression Shrinkage and Selection via the Lasso *Journal of the Royal Statistical Society Series B* 58(1): 267-288

[6] W. Qian and Y. Yang 2013 Model selection via standard error adjusted adaptive lasso *Annals of the Institute of Statistical Mathematics*, vol. 65, no. 2, pp. 295–318, 2013

[7] Y. Tang, L. Xiang, and Z. Zhu, 2014 Risk Factor Selection in Rate Making: EM Adaptive LASSO for Zero-Infated Poisson Regression Models *Risk Analysis* vol. 34, no. 6, pp. 1112–1127

[8] Zou, H. 2006 The Adaptive Lasso and Its Oracle Properties *Journal of the American Statistical Association*, 101, 1418-1429