Clusterwise Regression Model Development with Gamma Distribution

Reski Syafruddin[1*], Agus M. Soleh[2], Aji H. Wigena[3]
{reski_syafruddin@apps.ipb.ac.id[1], agusms@apps.ipb.ac.id[2], ajiwigena@ymail.com[3]}

Student of Department of Statistics, IPB University, Indonesia[1],
Lecturer of Department of Statistics, IPB University,Indonesia[23]

Abstract. This paper present development of clusterwise regression with a data set that has gamma distribution. Clusterwise regression is a method that finds simultaneously an optimal member of data in *k* cluster and each cluster have the best regression model. Analysis of a simulated data set has also been presented for illustrative purposes. Gamma and normal distributions were used for distribution of responses scenario with different parameters. This simulation study is carried out by initializing the number of clusters, classify observations randomly as an initial partition, move observation to the cluster giving the smallest residual and re-estimate the regression model from final partition. This simulation showed that clusterwise regression is able to form partition according to the distribution of data, also to form the best generalized linear model with Gamma distribution and linear regression model.

***Keywords*:** Clusterwise regression, Gamma distribution, generalized linear model with Gamma distribution, *linear regression model*.

1 Introduction

There are many observations in the economics, social and science that arise with a high variance. When the data was estimated a single regression model there was a mistake in presenting the data structure. The data can be classified to reduce variance. One of the ways to classify the data is clusterwise regression techniques. The clusterwise regression is based on the combination of these two techniques that find simultaneously an optimal partition of data in k cluster and regression function within cluster [1]. It is assumed that samples come from a certain number of populations and consider the existence of subpopulations of heterogeneous populations. The proportion of subpopulations is unknown. A specific form for each sub-population can be determined and the purpose of clusterwise regression is to describe the sample into mixture components based on the subpopulation.

Estimating parameters in the cluster wise regression method is needed to estimate the regression coefficients for each cluster. Various kinds of algorithms are formed to overcome this problem, among others, based on exchange algorithm [9], statistical techniques [5] [6] [10], and optimization techniques [2] [3] [4]. The application of cluster regression to various fields of research has been carried out, such as business research, physics, and social studies. Data not only arise from normal distribution but also arise from exponential families. Statistical techniques from [6] [10] developed the cluster wise method for the approach for generalized linear models.

There are several studies that have observations that only have non-negative values. These observations can be seen as random variables with a range of values> 0. Estimation of

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290521

random variable values in the range of values> 0 with the linear model estimation approach will not naturally get an estimate > 0 because the estimation techniques are based on responses from a normal distribution with ranges (-∞, ∞). Gamma distribution is a good distribution that describes the case with reasons that are the gamma distribution is positively skewed, meaning that it has an extended tail to the right of the distribution [7]. The observation often suitable for data that are continuous, positive, right-skew and where variance is near-constant on the log-scale. Since we need μ>0 we need η<0, which gives restrictions on β. Thereforethe canonical link is not often used. Most often the log link is used.

This paper present the development of clusterwise regression with gamma distribution that has a log-link function. Simulated data set has been presented for illustrative purposes to make good data clustering.

2 Materials

The data usedinthisstudyaresimulation data generatedfromtwopopulationsthathavedifferentdistributions. Response data inthefirstpopulationcomesfrom a normal distributionwith a populationsizeof 100 which has the parameter β = 4 andwithoutusinganintercept.. The response data from a normal distribution is formed by calculating the value of **y** = **x**β + **e**. In this case, **e** is an error that is normally distributed. The response data inthesecondpopulationcomesfromtheGammadistributionwith a populationsize of 100. The response data with the distribution of gamma is formed by stages, namely determinedshape parameterξ = 100 then calculating the value of $\mu = \exp(\boldsymbol{X}\beta)$ with β = 0.5. The rate parameters for gamma distribution are obtained by calculating v = ξ / (μ) and then generating data on gamma distribution with two parameters(y ~ Gamma (ξ, v)). The response data generated from the normal distribution and Gamma distribution is a data setthat was processed by clusterwise regression method.

3 Methods

Data analysis was first performed to initialize the partition by determining the number of k clusters. Then each observation is grouped into one of the clusters randomly. After obtaining the initial partition, estimating the regression model is carried out for each cluster. The model formed is generalized linear models, in this case, the response variable comes from the exponential family [8]. The generalized linear model is formulated consisting of a random variable comes from an exponential family, a systematic component (η), and a function g (), which link the random components and systematic components [10].

A general form of the probability density function of the exponential family in the k-cluster is

$$f_{ik}(y_{ik}|\theta_{ik},\lambda_k) = \exp\left\{\frac{y_{ik}\theta_{ik} - b(\theta_{ik})}{a\lambda_k + c(y_{ik},\lambda_k)}\right\} \quad (1)$$

where a (.), b (.), and c (.) are certain functions, θ is a canonical parameter, and λ is a dispersion parameter and is assumed to be constant over observation in k cluster. The probability density function of the 2-parameter gamma distribution is

$$f(y|v,\xi) = \frac{v^{\xi}}{\Gamma(\xi)} y^{\xi-1}\exp(-vy) \quad (2)$$

where ν is the rate parameter and ξ is a shape parameter. The relationship of the parameters of the rate and form of the distribution of gamma is ν = ξ / μ with the parameter ξ is assumed to be constant [8].

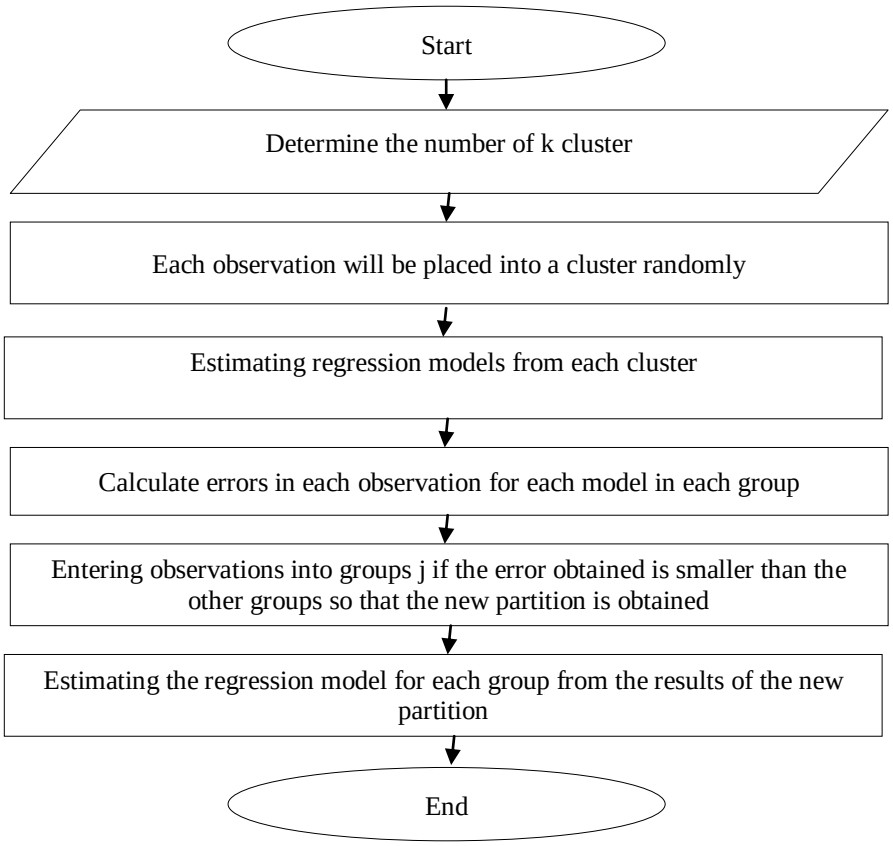


Fig. 1. Flowchart of clusterwise regression

The systematic component and link functions for each k-cluster are defined as follows:

$$\eta_{ijk} = g(\mu_{ijk}) \tag{3}$$

where$\eta_i = \beta_0 + \sum_{p=1}^{P} X_j \beta_j$ for each group of k. The parameter of the most linear model is assumed to use the maximum likelihood method, namely by maximizing the density function. The function parameter μ is assumed by the parameter β_j in a systematic component based on a link function. The canonical link for gamma distribution responses in this study uses logs so that they are obtained

$$\mu = \exp(\beta_0 + \sum_{p=1}^{P} X_j \beta_j) \tag{4}$$

whereas canonical link functions for normal distribution responses is identity so that they are obtained

$$\mu = \beta_0 + \sum_{p=1}^{P} X_j \beta_j \tag{5}$$

After the model in each cluster was obtained, then calculate the residual of each observation on each model formed. Each observation is moved to a cluster giving the smallest residual. This process was carried out from the first observation to the last observation. All observations have been reclassified so that a new partition is obtained. Furthermore, estimating the regression model is carried out on each new cluster. The flow chart of the algorithm used in this study is presented in Fig. 1.

The precision of observations clustering was done by calculating accuracy based on the actual classification of data generated with the final classification. Accuracy describes how accurately a model can classify a data or comparison between predicted data correctly with the whole data so that the accuracy value can be defined as follows:

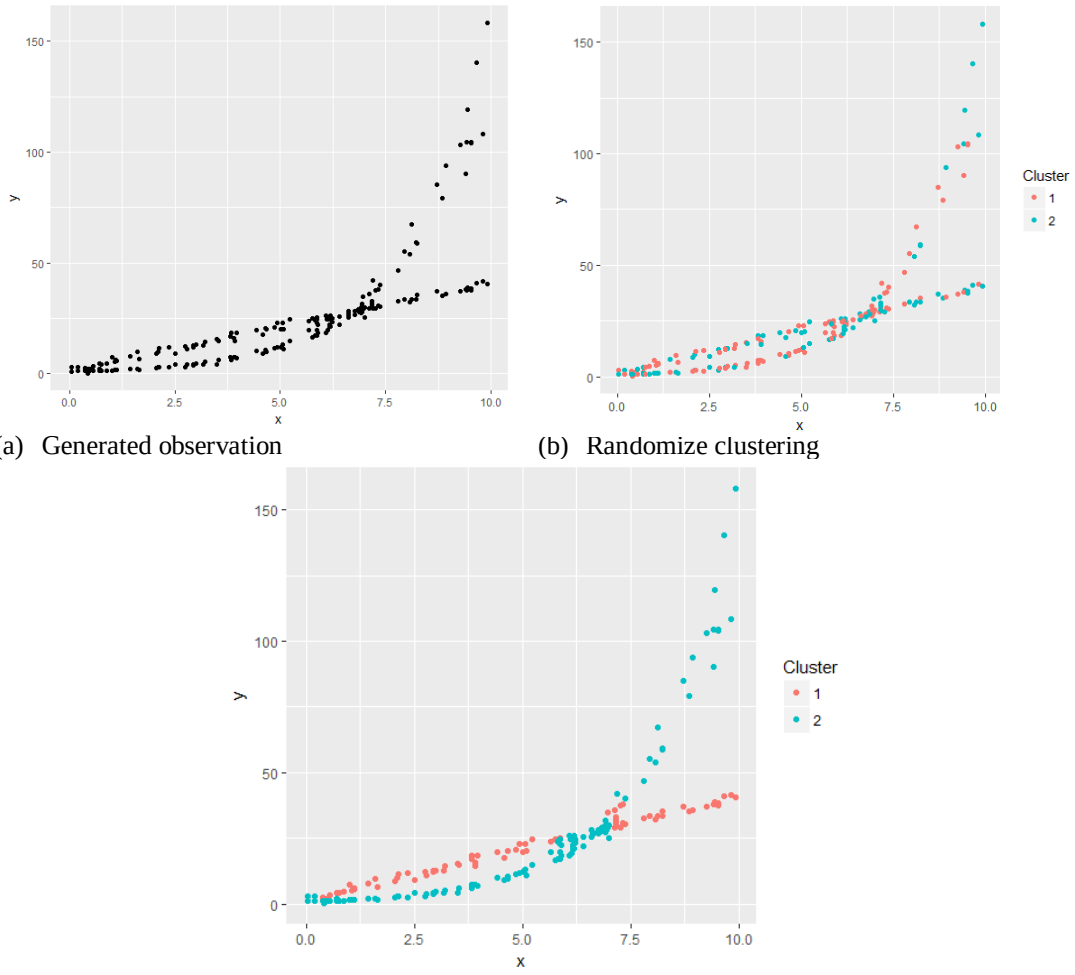
Table 1. Classification table between actual and prediction classes

Actual	Predicted	
	1	0
1	TP	FN
0	FP	TN

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

4 Results and Discussion

The cluster regression simulation was formed from generating a data set consisting of gamma distribution with a form parameter that is ξ = 100, β coefficient = 0.5 and normal distribution with the parameter β = 4. The cluster information formed from the cluster of data can be detected by scatter plot between the response variable (Y) and the independent variable (X). The scatter diagram presented in Fig. 2(a)provides information that observations form two clusters. Determining the number of clusters in certain cases can be easily identified, but in complex cases, it is difficult to detect the number of clusters formed.



(a) Generated observation (b) Randomize clustering

(c) Final result

Fig. 2. clusterwise regression simulation with Gamma and normal distribution

Initialization is done by determining the number of clusters and randomizing observations into one cluster of clusters formed. The results of the randomization of observations for each cluster are presented in Fig. 2(b). Each cluster has a gamma regression model and linear regression. The examination of each observation into one cluster is carried out by calculating the smallest error in each cluster so that there is a fixed observation or move to another group. The final results of examining all observations are presented in Fig. 1(c). The observations are clustered with observations that have the same characteristics so that the model is obtained according to the distribution of each cluster. The accuracy obtained is also quite good at 0.875. Therefore, the clusterwise regression algorithm that is formed is able to make good data clustering.The actual proportion of each group has the same proportion. The clusterwise regression with mixed distribution also has the ability to classify response data into clusters appropriately which has the proportion in each group is almost the same, cluster one is 41% and cluster two is 59%.

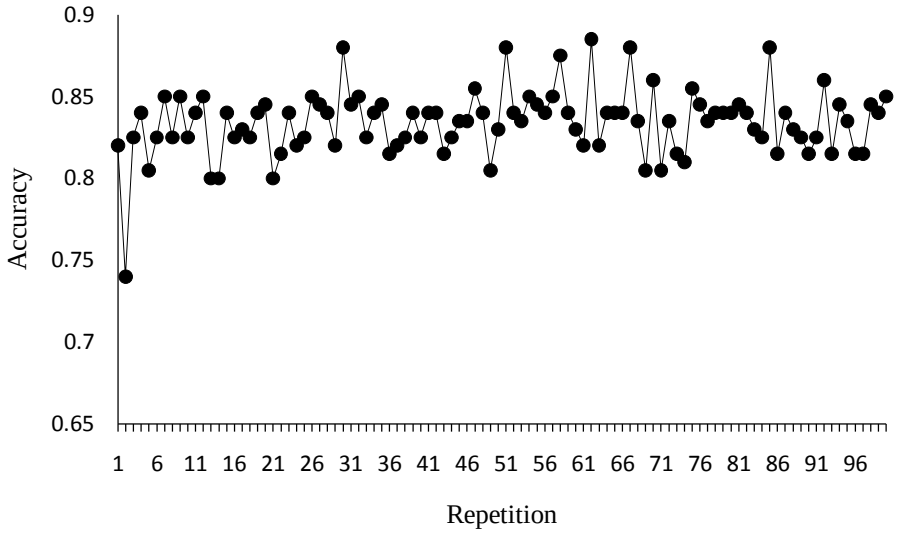


Figure 2accuracy of clusterwise regression model with 100 repetitions

The consistency of thealgorithmformedisdonebylookingattherepetitionaccuracyvalue of 100 times. theaccuracyvaluecanbeseeninFig. 2. Fig. 2 showstheaccuracyvaluearoundthevalue of 0.80 to 0.88, eventhoughthereisanaccuracyvaluethatindicates a valuebelow 0.75. Thisshowsthatthealgorithmformedhas a highaccuracyvalueandwasconsistent for eachrepetition. Therefore, thisalgorithmcanbeused on data that has a high varianceor data that hasmixeddistribution.

5 Conclusion

This study developed a clusterwise regression method for the observation that had gamma distribution and normal distribution. It has created a new partition that was almost similar to the actual distribution and regression function based on the distribution for each cluster simultaneously. The prediction performance of the models was evaluated by the accuracy that obtained between class actual based on distribution and a new class. The results presented in this paper demonstrate that simulation clusters model with gamma distribution that had a log-link function and normal distribution can make the right partition. These results also demonstrate that the clusterwise regression method was better to exist models for heterogeneous data.

References

[1] Bagirov AM, Mahmood A, Barton A.: Prediction of monthly rainfall in Victoria, Australia: clusterwise linear regression. J Atmosres. pp. 20-29(2017).

[2] Bagirov AM, Ugon J, Mirzayeva H.: An algorithm for clusterwise linear regression based on smoothing techniques. Optimization Letters 9. pp. 2: 375-390 (2015).

[3] Bagirov AM, Ugon J, Mirzayeva H.: Nonsmooth nonconvex optimization approach to clusterwise linear regression problems. Eropan journal of operational research. pp. 229: 132-142 (2013).
[4] Bagirov AM, Ugon J, Mirzayeva H.: Nonsmooth optimization algorithm for solving clusterwise linear regression problems. Journal of optimization. pp. 3: 375-390 (2015).
[5] Desarbo WS, Cron WL.: A Maximum Likelihood Methodology for Clusterwise Linear Regression. Journal of Classification. pp. 5: 249-282 (1988).
[6] Grun B, Leisch F.: Finite Mixtures of Generalized Linear Regression Model.University of Munich, DE (2007).
[7] Husak GJ, Michaelsen J, Funk C.: Use of The Gamma Distribution to Represent Monthly Rainfall in Africa for Drought Monitoring Applications. Int J Climatol. pp. 27: 935-944 (2007).
[8] Soleh AM.: Pemodelan Linear Sebaran Gamma dan Pareto Terampat dengan Regularisasi L1 pada Statistical Downscaling untuk Pendugaan Curah Hujan Bulanan. Institut Pertanian Bogor, ID (2015).
[9] Spath H.: Algorithm 39: Clusterwise Linear Regression. Computing. pp. 22: 367-373 (1979).
[10] Wedel M, DeSarbo WS.: A Mixture Likelihod Approach for Generalized Linear Models. Journal of Classification. pp. 12: 21-55 (1995).