Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2022, 2022:22
https://doi.org/10.28919/cmbn/7117
ISSN: 2052-2541

THE DEVELOPMENT OF CLUSTERWISE REGRESSION MODEL ON GAMMA-NORMAL MIXED DISTRIBUTION WITH GENETIC ALGORITHM

MELLY AMELIA, AGUS M. SOLEH*, ERFIANI

Department of Statistics, IPB University, Indonesia

Abstract: Clusterwise regression is a statistical technique that combines clustering process and regression analysis. Cluster optimization was carried out using genetic algorithm (GA). GA is an optimization technique that adapts the theory of natural evolution starting from the formation of the initial population to produce the best generation. GA provides an optimal fitness value that describes the optimal clustering. This study aims to construct a clusterwise regression algorithm on Gamma-Normal mixed distribution using GA. Simulations were carried out to evaluate the results of the GA construction by generating data for various distributions. The simulation results on the Gamma distribution give an accuracy value 85% with cluster proportion 37% : 67%. Normal distribution with cluster proportion 51% : 49% yielded 95% accuracy. The highest accuracy is 98% in mixed distribution with cluster proportion 52% : 48%. Based on high accuracy value of the simulation results, it indicates the construction of an appropriate algorithm. This indicates that clusterwise regression Gamma-Normal mixed distribution with GA is able to cluster and model well.

Keywords: clusterwise regression; genetic algorithm; gamma-normal mixed distribution.

2010 AMS Subject Classification: 62J12, 65C60, 90C99.

*Corresponding author
E-mail address: agusms@apps.ipb.ac.id
Received December 28, 2021

1. INTRODUCTION

Research data sets generally come from a homogeneous population but sometimes cannot be modeled in one model. In fact, the data obtained consists of subpopulations or groups where each subpopulation has a different model [1]. This indicates that modeling in a data set is not effective if only one model is used. Therefore, the data set must be grouped to reduce the variability in the data that causes the modeling to be biased [2]. One of the analytical techniques that prioritizes grouping and modeling each group is clusterwise regression analysis. Clusterwise regression analysis is a combination of optimal data clustering techniques as well as regression analysis in each cluster [3].

The problem that often arises in cluster regression analysis is the search for the optimal cluster. Genetic algorithm is used as a technique for finding the optimal solution of a problem. In this case, genetic algorithms are used to group each observation according to its cluster. The genetic algorithm introduced by Holland [4] adopts the principles of the theory of evolution where the goal is to find the best generation of each process carried out [5]. Genetic algorithm is a metaheuristic strategy that is included in a large class of evolutionary algorithms [6].Genetic algorithms focus on finding solutions that have high quality by considering operators such as the selection process, convergent achievement and mutation processes [7]. For each generation that is processed in the genetic algorithm, the fitness value for each individual will be calculated. Therefore, the fitness value is the value of the objective function to be solved [8].The results of the study by Sartono et al. [1] stated that cluster linear regression with a genetic algorithm approach gave better results with an optimal fitness value of 99%. Cluster regression analysis with this genetic algorithm approach, will be applied in statistical downscaling modeling.

Based on the results of research Faladiba et al. [9] related to the development of cluster regression on the Gamma distribution, it gave quite good results with a smaller root mean squared error prediction (RMSEP) value than without clustering. However, the clustering stage with k-means was considered less suitable because the clustering was not done randomly. Therefore, in this study, we will develop a clusterwise regression Gamma-Normal mixed distribution with genetic algorithm.

2. Literature Review

2.1 Clusterwise Regression

Clusterwise regression is a clustering technique and regression analysis performed simultaneously [10]. The basic idea of clusterwise regression is the possibility that a data set has more than one regression model. So that linear regression is not able to overcome this.

Cluster regression will produce an optimal grouping of observations in k clusters and also a regression function in clusters so that simultaneously the best regression model will be obtained for each cluster [3]. In general, the cluster regression equation is as follows [11]:

$$y_i = \sum_{k=1}^{k} a_{ik} E(y|x) \tag{1}$$

where

y_i = i-th observation response variable,

a_{ik} = 1 if the i-th observation is in the k-th cluster, and 0 if the i-th observation is not included in the k-th cluster.

$E(y|x)$= regression model from the i-th observation.

Clusterwise regression will start with the initial stages, namely determining the number of clusters to be formed, randomizing the observations that will enter each cluster, estimating the regression parameters for each cluster using a linear model, until a new cluster is formed with optimal observations. Optimizing the observations that go into each cluster using a genetic algorithm approach.

2.2 Genetic Algorithm

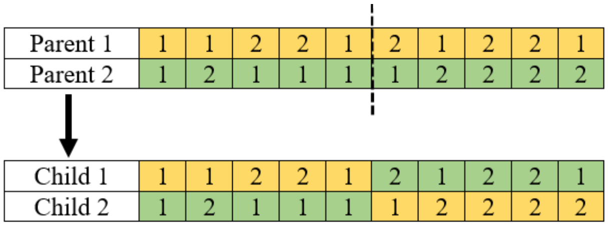
Genetic algorithm (GA) was first developed by John Holland (1975) in the book "Adaption in Natural and Artificial Systems" in New York, United States [4]. GA is an optimization technique for finding the optimal solution to a problem. The basic idea of GA is to adopt the process of genetics and natural selection [12]. The GA technique searches from several available solutions to obtain the optimal best solution based on specified criteria called the fitness function. In general, the principle of GA which adapts the theory of natural evolution consists of a selection process from the initial population, crossing over, mutation, natural selection to produce the best generation.

The terms in GA used in this study are

1. Gene, is a value that represents an information that plays an important role in the process of generation. In this study, the gene is a binary number that indicates the cluster code.
2. Chromosomes, are collection of several genes.
3. Individuals, in the form of values that represent solutions.
4. Population, is a collection of individuals that will be processed in the search for the best solution.
5. Children, are new individuals obtained from crossing over
6. Fitness function, is a function which optimal value will be sought which will produce a fitness value.

Furthermore, GA uses several operators in its process, namely [13]:

1. Initialization of the initial population, is the formation of the initial population needed in the process of finding the best solution. The population formed is the result of random generation.
2. Selection, is a process in GA like natural selection which selects each individual. Individuals with optimal fitness values are selected to be parents in the crossover stage.
3. Cross over, is the process of forming new children from 2 selected parental individuals. The two parents are exchanged for genes in their chromosomes. The crossover technique used is a one-point technique as illustrated in Figure 1.



Parent 1	1	1	2	2	1	2	1	2	2	1
Parent 2	1	2	1	1	1	1	2	2	2	2

Child 1	1	1	2	2	1	2	1	2	2	1
Child 2	1	2	1	1	1	1	2	2	2	2

Figure 1: One-point crossover technique

4. Mutation, is the process of exchanging the value of a gene which was initially valued at 1

to be changed to 2 or vice versa with the aim of maintaining a wider solution. In general, the probability of mutation is very small, $\leq 1\%$. According to Sartono [14], the rate of mutation cannot be adjusted too large because it is related to the length of time required to reach convergence.

3. ANALYTICAL PROCEDURE

The analytical procedure carried out in this study uses simulation data. Simulations were carried out to see the ability of cluster regression with genetic algorithms to group each observation according to the generated distribution. The following are the stages of cluster regression with a genetic algorithm approach.

1. Generating predictor variables that follow a uniform distribution (0.10).
2. Generating response variables consisting of Normal, Gamma, and mixed (Gamma – Normal) distributions. The number of data generated for each variable is 100 observations. The simulation scenarios for the generated data are in accordance with Table 1.

Table 1: Scenario of Clusterwise Regression

Simulation of Respon	1st Response Variable	2nd Response Variable	Predictor Variable
Gamma	$Y \sim \text{Gamma}\ (\xi, v)$ $\xi = 0.5\ ;\ \beta = 0.5$	$Y \sim \text{Gamma}\ (\xi, v)$ $\xi = 100; \beta = 0.75$	$X \sim \text{Uniform}\ (0,10)$
Normal	$Y \sim \text{Normal}(\mu, \sigma^2)$ $\beta = 4$	$Y \sim \text{Normal}(\mu, \sigma^2)$ $\beta = 6$	$X \sim \text{Uniform}\ (0,10)$
Mix	$Y \sim \text{Normal}(\mu, \sigma^2)$ $\beta = 4$	$Y \sim \text{Gamma}\ (\xi, v)$ $\xi = 100; \beta = 0.5$	$X \sim \text{Uniform}\ (0,10)$

3. Clusterwise regression analysis with genetic algorithm approach.

   The following are the stages in clusterwise regression with a genetic algorithm approach :

   a) Initial population initialization.

      In this case, 100 individuals are randomly generated. Because the number of clusters is 2, then each individual consists of the numeric numbers 1 and 2 as the code for the cluster.

b) Enter random generation data into each cluster. If the first observation has a code of 1, then the observation will enter the first cluster. Likewise with observations that have code 2.

c) Modeling each population according to cluster regression of Gamma distribution, Normal and Mix (Gamma-Normal).

d) Calculating fitness value

   The optimal fitness value in this study is the population that has the smallest root mean squared error (RMSE) value.

e) Population selection

   Ten populations with optimal fitness values will be selected to be parents in the crossover process. In this study, the selection was made based on the ranking of the RMSE values starting from the smallest.

f) Crossover process

   Cross over is done by combining every 2 parents from the population selection results. This process is carried out to find new children (population). In this study, the crossover technique used is one point, where the proportion of the population being crossed over is $50\%: 50\%$.

g) Mutation

   Each new population from the crossover process, will be mutated with probability 1%. This new population is then modeled, then the RMSE value is calculated.

h) Repeat steps (b) to (g) until the fitness value converges or reaches a maximum iteration of 500 iterations.

i) The best population selection is the population that has the smallest RMSE value of 10 individuals who have converged.

4. Calculating an accuracy value that shows how accurate the cluster regression model with a genetic algorithm approach is in grouping each observation according to its distribution.

4. Results and Discussion

4.1 Genetic Algorithm Performance

The first discussion in this study is about the performance of the genetic algorithm in cluster regression. The performance of the genetic algorithm can be seen based on the RMSE value and the level of accuracy in clustering. The best population or the best generation is the population that produces the highest accuracy value. As explained in the clusterwise regression stage, the genetic algorithm will find the best solution or the best population by selecting populations that produce optimal fitness values. Each new generation generated in the genetic algorithm stage will have a fitness value that is better or the same as the fitness value of the previous generation. Therefore, the fitness value in this case is the smallest RMSE value that provides the highest accuracy and has an increasing pattern from generation to generation.

Figure 2 shows the performance of the genetic algorithm in cluster regression which can be seen based on the increasing accuracy value in each iteration. The use of genetic algorithms produces a low accuracy value at the beginning and then increases with the number of iterations carried out. Figure 2 shows the level of accuracy in clustering. The cluster regression accuracy with the genetic algorithm approach is very high, reaching 98%.

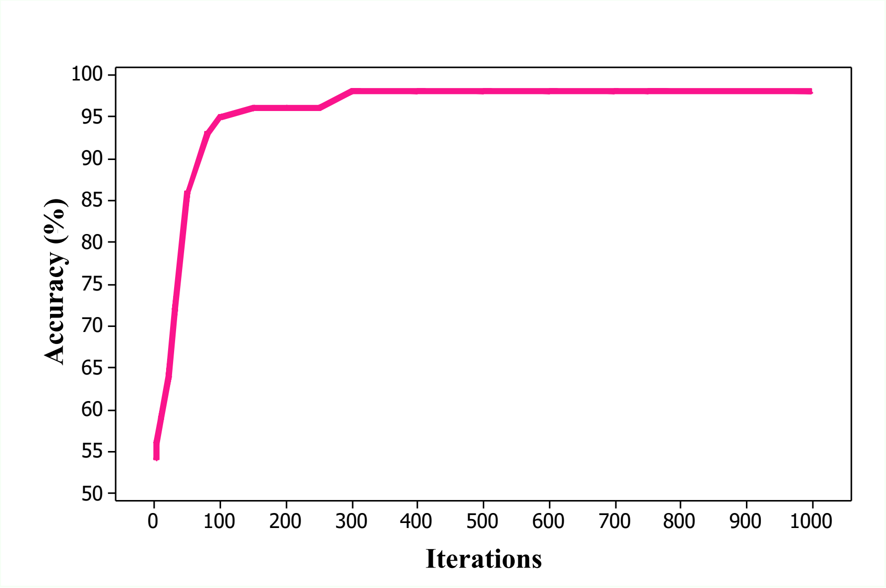
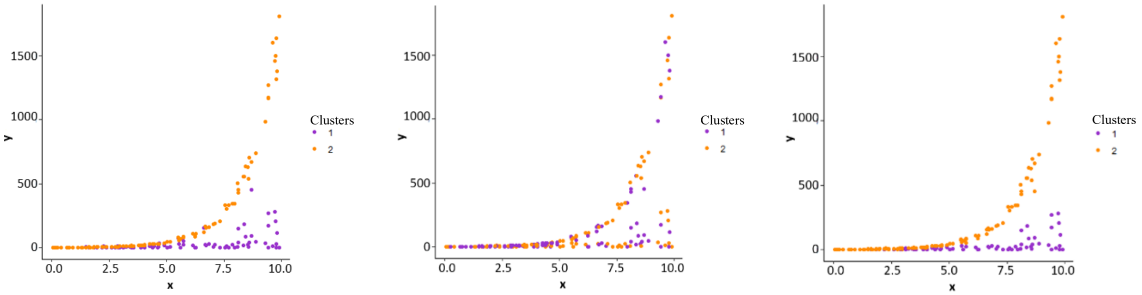


Figure 2: Performance of Genetic Algorithm in Clusterwise Regression

4.2 Clusterwise Regression Simulation with Genetic Algorithm

4.2.1 Gamma – Gamma Distribution

The cluster regression simulation with the response variable Gamma distribution using a genetic algorithm consists of two clusters. A total of 2 Gamma response variables were generated with different parameters. The first distribution with shape parameter $(\xi) = 0.5, \beta = 0.5$, and the second distribution with $\xi = 100$, $\beta = 0.75$. The predictor variables generated followed a uniform distribution (0,10) of 100 observations for each cluster. Plot (a) is the original generated data. After the data is generated, randomization is carried out at random and then enter all the observation data according to the random group as in plot (b). The next process is to perform cluster regression analysis with a genetic algorithm approach for each model, namely Gamma, Normal and mixed (Gamma-Normal). Plot (c) is the result of cluster regression simulation with Gamma distribution. The simulation results with this genetic algorithm approach seem to be able to separate the data well, where the highest accuracy is in the Gamma model of 85%, the accuracy of the Normal model is 62%, and the mixed model accuracy is 75%.



(a) (b) (c)

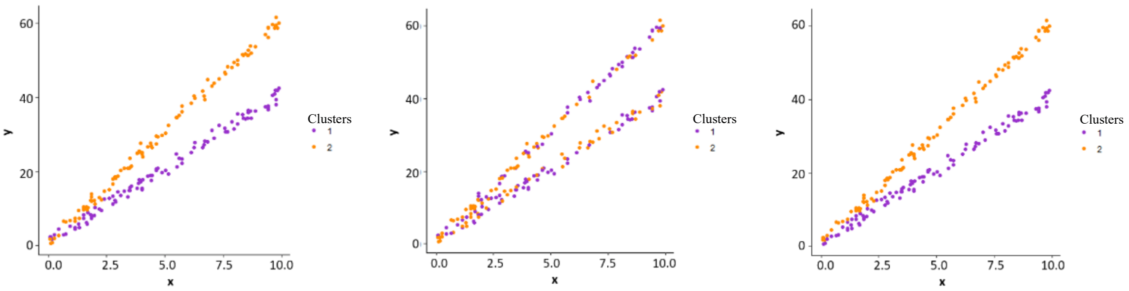
Figure 3: Plot of clusterwise regression simulation with Gamma distribution

(a) The original data of the generation, (b) The results of randomization, and (c) The results of clustering with GA

4.2.2 Normal - Normal Distribution

The cluster regression simulation with a Normal distribution is started by generating two response variables with different . The first distribution is generated with the value $\beta = 4$

and the second distribution with the value $\beta = 6$. As in the cluster regression simulation with Gamma distribution, the predictor variables generated are 100 observations that follow a uniform distribution (0, 10) and normally distributed error (0, 1). Based on the results of the generation data, two actual clusters are formed which can be seen in plot (a). Furthermore, the observation data that has been generated will be entered into the two clusters randomly as in plot (b). The cluster regression results with a Normal distribution are shown in plot (c).



(a) (b) (c)

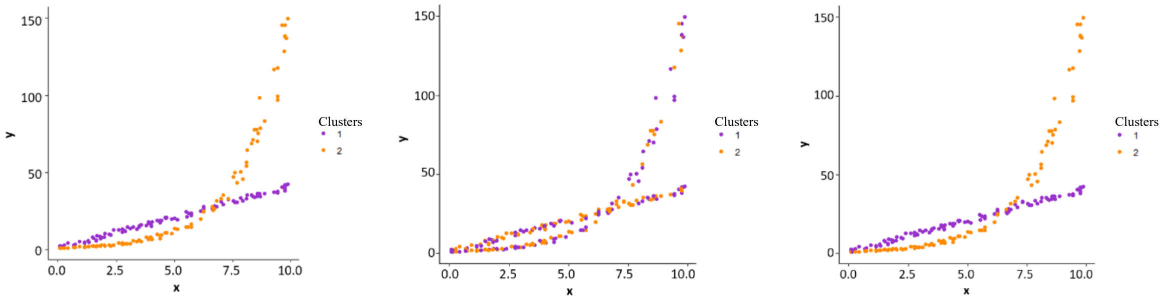
Figure 4: Plot of cluster regression simulation with Normal distribution

(a) The original data of the generation, (b) The results of randomization, and (c) The results of clustering with AG

Based on the cluster regression results in plot (c), it can be seen that cluster regression using a genetic algorithm approach gives very similar results to the initial generation data. This can also be seen with the highest accuracy value in the Normal model of 95%. This high accuracy value indicates that cluster regression with genetic algorithms is able to classify data according to the characteristics of the generated response variables.

4.2.3 Gamma – Normal Mixed Distribution

The last clusterwise regression model that was simulated was a model with a mixed distribution of Gamma - Normal. The first response variable that is generated follows a Normal distribution with a value of $\beta = 4$ and an error that follows a Normal distribution. Then the second response variable follows the Gamma distribution with the parameter $\xi = 100$ and the value $\beta = 0.5$.



(a) (b) (c)

Figure 5: Plot of cluster regression simulation with Gamma – Normal distribution

(a) The original data of the generation, (b) The results of randomization, and (c) The results of clustering with AG

Plot (a) shows the number of clusters formed, namely 2 clusters. Randomly, observations will enter the cluster as in plot (b). The results of the cluster regression simulation with mixed distribution are shown in plot (c). Based on the plot, it can be seen that there are similarities between the generation data and the cluster regression results with a Normal – Gamma distribution. This is evidenced by the highest accuracy value in the mixed model of 98%. Meanwhile, the accuracy values of the Gamma and Normal models are 85% and 65%, respectively. This shows that cluster regression with mixed models is a model that is very suitable for the distribution of the generated response variable data.

The initial generation data from each cluster has the same number of 100 observations. Table 2 shows the proportion of observations that enter each cluster after cluster regression with the three models is performed. The proportion ratio is almost the same in each group. This proves that cluster regression analysis with genetic algorithms is able to classify data well.

Table 2 : Summary of accuracy and proportion of the best cluster

Clusterwise Regression	Accuracy	Cluster	Proportion
Gamma	85%	1	37%
		2	63%
Normal	95%	1	51%
		2	49%
Mix	98%	1	52%
		2	48%

5. CONCLUSION

The construction of the genetic algorithm for the cluster regression that was carried out was appropriate. This can be seen based on the results of simulations carried out to see the performance of cluster regression with the genetic algorithm approach showing high accuracy reaching 98% in mixed models with the distribution of Gamma and Normal response variables. So from the results of this study, it can be concluded that the clusterwise regression Gamma-Normal method with a genetic algorithm approach is able to cluster and form models simultaneously well when compared to the regression model without clustering.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] B. Sartono, A. Hidayatuloh, A.M. Soleh, An implementation of genetic algorithm on clusterwise regression analysis, In: The 5th Annual Basic Science International Conference, Indonesia, (2015).

[2] R. Syafruddin, A.M. Soleh, A.H. Wigena, Clusterwise regression model development with gamma distribution In: Proceedings of the 1st International Conference on Statistics and Analytics, ICSA, 2019.

[3] A.M. Bagirov, J. Ugon, H. Mirzayeva, Nonsmooth nonconvex optimization approach to clusterwise linear regression problems, Eur. J. Oper. Res. 229 (2013), 132–142.

[4] J.H. Holland, Adaptation in natural and artificial systems, University of Michigan Press, Ann Arbor, (1975).

[5] K. Deb, An introduction to genetic algorithms, Sadhana. 24 (1999), 293–315.

[6] M. Mitchell, An introduction to genetic algorithms, MIT Press, MA, (1996).

[7] D. Whitley, A genetic algorithm tutorial, Stat. Comput. 4 (1994), 65–85.

[8] O. Kramer, Genetic Algorithm Essentials, Springer, Cham, 2017.

[9] F. Muthia Nadhira, A.M. Soleh, A. Djuraidah, Clusterwise regression model for statistical downscaling to predict daily rainfall using gamma distribution, J. Phys.: Conf. Ser. 1863 (2021), 012051.

[10] C. Hennig, Models and methods for clusterwise linear regression, in: W. Gaul, H. Locarek-Junge (Eds.), Classification in the Information Age, Springer Berlin Heidelberg, Berlin, Heidelberg, 1999: pp. 179–187.

[11] B. Grün, F. Leisch, Finite mixtures of generalized linear regression models, in: Recent Advances in Linear Models and Related Areas, Physica-Verlag HD, Heidelberg, 2008: pp. 205–230.

[12] M.G. Sahab, V.V. Toropov, A.H. Gandomi, Optimum design of composite concrete floors using a hybrid genetic algorithm, in: Handbook of Neural Computation, Elsevier, 2017: pp. 581–589.

[13] S.N. Sivanandam, S.N. Deepa, Genetic algorithms, in: Introduction to Genetic Algorithms, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008: pp. 15–37.

[14] B. Sartono, Pengenalan algoritma genetik untuk pemilihan peubah penjelas dalam model regresi menggunakan SAS/IML, Forum Stat. Komput. 15 (2010), 10-15.