Journal of Economics and Administrative Sciences
2024; 30(140), pp. 434-449

P-ISSN 2518-5764
E-ISSN 2227-703X

Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq
DOI: https://doi.org/10.33095/xammnc51

Expectation Parameters in the Poisson Mixture Regression Model for Latent Class by Applying Genetic Algorithm and Maximization Algorithm

Ahmed Khuder Eleas*
Department of Statistics
College of Administration and Economics
University of Baghdad
Baghdad, Iraq
ahmed.khedr2101m@coadec.uobaghdad.edu.iq
***Corresponding author**

Emad Hazim Aboudi
Department of Statistics
College of Administration and Economics
University of Baghdad
Baghdad, Iraq
emadhazim@coadec.uobaghdad.edu.iq

Received:22/10/2023 Accepted:3/12/2023 Published Online First: 30 /4/ 2024

Abstract:

In a Poisson mixture regression model for latent class, observations come from different sub-sources or classes, and the observed data are assumed to be generated by a specific (finite) mixture of unobserved or latent classes. The problem lies in the optimal assignment of observations to their respective classes. This requires sophisticated methods for estimating the parameters in the model. Usually, the model parameters are estimated by the conventional EM algorithm. The research aims to compare the EM algorithm and the genetic algorithm GA. Using simulation, the two algorithms were compared based on the MSE criterion, with different sample sizes (n = 50, 90, 120) and three scenarios (S1, S2, S3) for default values of the parameters. The results showed the superiority of the GA genetic algorithm over the EM algorithm, as the GA genetic algorithm gave the lowest MSE values.

Paper type: *Research paper*
Keywords: Mixture Poisson Regression, Latent Class, Expectation Maximization (EM), Genetic Algorithm (GA).

P-ISSN 2518-5764
E-ISSN 2227-703X

1. Introduction:

In contemporary statistical analysis, sophisticated algorithms play a pivotal role in enhancing the accuracy and reliability of parameter estimation in regression models. This research paper embarks on a journey into the realm of regression analysis, specifically focusing on estimating parameters within the mixture Poisson regression model for the latent class. The amalgamation of two potent algorithms, namely the Expectation Maximization (EM) algorithm and the Genetic Algorithm (GA), forms the cornerstone of our analytical approach.

The regression analysis has witnessed an unparalleled surge in significance across various domains, due to its intrinsic capacity to elucidate relationships between dependent and independent variables. While conventional regression models have been extensively explored and utilized, the complexity and nuance of real-world data often necessitate the exploration of more intricate techniques. This paper seeks to bridge this gap by delving into the innovative fusion of EM and GA in the context of the mixture Poisson regression model.

Counting data, characterized by discrete and non-negative outcomes, is omnipresent in numerous fields, including but not limited to epidemiology, finance, and social sciences. The statistical modeling of counting data has profound implications for decision-making processes, policy formulation, and scientific discovery. The mixture Poisson regression model, with its capacity to capture unobserved latent classes within the data, presents a robust framework for modeling such data.

1.1 Literature Review:

Pernkopf and Bouchaffra (2005) proposed a genetic-based expectation-maximization (GA-EM) algorithm for learning Gaussian mixture models from multivariate data.The experiments on simulated and real data show that the GA-EM outperforms the EM method. Sundararajan and Mengshoel (2016) suggested a genetic algorithm for expectation maximization (GAEM) for learning parameters in Bayesian networks. It combined the global search property of the genetic algorithm with the local search property of EM. The global convergence of GAEM has been demonstrated theoretically, empirically, GAEM has been shown to provide significant speedups because it tends to select fitter individuals, who converge faster, as parents of the next. Papastamoulis et al (2016) used the EM algorithm to estimate the parameters of a zero-inflated bivariate Poisson mixture regression model, and the method was applied to a car insurance claims dataset, and the results showed that this algorithm significantly improved the modelling of the dataset. Tzougas (2020) presented an inverse Poisson-Gamma regression model used with data with a long tail and high dispersion, .The researcher developed an EM algorithm to estimate the parameters of the inverse Poisson-Gamma model, and the researcher applied it to car insurance data in order to verify the efficiency of the algorithm. AlKhafaji and AlBakri (2021) used the genetic algorithm (GA) and the iterative reweighting (IR) algorithm to estimate the parameters of the skewed normal distribution; the results proved, using Monte Carlo simulation, that the genetic algorithm is best when the sample size is small, and that the iterative reweighting algorithm is best when the sample size is large. Kareem and Hashim (2021) compared three methods (FlexMix, MixTLE, MixLP) for estimating the mixed linear regression model, and the simulation results proved that the (FlexMix) method is more efficient than other methods. Gonçalves et al (2022) presented a latent Poisson-Birnbaum-Saunders regression model in which observations within the same group are driven by the same latent random effect that follows a Birnbaum-Saunders distribution. The Expectation Maximization (EM) algorithm was used to estimate the model parameters and a simulation was conducted to evaluate the performance of the estimators. Radam and Hameed (2023) compared the Poisson regression model and the Conway-Maxwell-Poisson model using simulation and with different sample sizes, and the researchers demonstrated through the results the superiority of the Poisson model through the Akaike criterion and the mean square error criterion.

P-ISSN 2518-5764
E-ISSN 2227-703X

The problem of this research is that Poisson regression models often need to adequately capture the complex structure of data, especially when the data is heterogeneous, that is, when observations come from different subgroups or sources. Poisson mixture regression models have been used to address this problem, but determining the number of mixture components and appropriate assignment of observations to classes can be difficult due to the complex structure of the data. Particularly with the use of latent classes, sophisticated parameter estimation methods are required.

The research's objective is to compare the EM algorithm and the genetic algorithm used to estimate the parameters of the mixture Poisson regression model for latent class.

2. Material and Methods:

2.1 Poisson Regression Model:

The classical linear regression model assumes that the response variable depends on the explanatory variables, which can be either continuous or countable. However, when the response variable takes the form of countable data, the assumptions of linear regression are not met. The Poisson regression model was introduced as a suitable alternative for such cases (Algama and Abdalteef, 2019). The Poisson regression model is a form of regression analysis specifically designed for modeling countable data, making it well-suited for analyzing rare events. The Poisson regression model is mathematically expressed using the following formula (Cameron and Trivedi, 2013):

$$y = e^{X\beta+U} \qquad (1)$$

Where:

y **:** The response vector is a variable with dimensions of (n×1).
β : The matrix containing the explanatory variables has a size of (n×(L+1)), where n represents the number of observations, and (L+1) indicates the number of explanatory variables, including the constant term.
U : The random error vector has dimensions of (n×1).

The Poisson regression model assumes that the response variable (y) follows a Poisson distribution with a mean and variance of (λ) (Algama and Abdalteef, 2019). Additionally, it is assumed that the logarithm of the expected value of the response variable can be represented as a linear combination involving several unknown parameters. Because of this property, the Poisson regression model is often referred to as the log-linear model (McCullagh and Nelder, 1989):

$$E(y) = \lambda = x\beta \qquad (2)$$

According to this linear formula, the explanatory variables allow for any real value of the mean of response variable y, which contradicts the nature of the Poisson distribution parameter (λ) since it must take positive values. To overcome this issue, we employ the logarithmic link function to establish a relationship between the mean (λ) and the explanatory variables (x), thereby adopting the generalized linear model (McCullagh and Nelder, 1989):

$$Log\ \lambda = \beta_0 + \beta_1 x_1 + \cdots \beta_L x_L \qquad \text{Where: } x_0 = 1$$
$$\lambda = \exp(\beta_0 + \beta_1 x_1 + \cdots \beta_L x_L) \qquad (3)$$

P-ISSN 2518-5764
E-ISSN 2227-703X

2.2 Mixture Poisson Regression For Latent Class :

We will work with the dependent variable y_i which denotes the total number of events measured in a sample comprising (n) observations. The independent observations $y_{1\ldots} y_n$ are assumed to be organized into (c) classes, and each observation y_i belonging to class k follows a Poisson distribution with the parameter $\lambda_{i|k}$ as follows :

$$P_k(y_i|\lambda_{i|k}) = \frac{e^{-\lambda_{i|k}}(\lambda_{i|k})^{y_i}}{y_i!} . \quad i = 1. \ldots . n \, . k = 1. \ldots . c \qquad (5)$$

$$\ln P_k(y_i|\lambda_{i|k}) = y_i \ln(\lambda_{i|k}) - \lambda_{i|k} - \ln(y_i!)$$

As per the generalized linear model (McCullagh and Nelder, 1989), the conventional parameter is expressed as follows:

$$Q(\lambda_{i|k}) = \ln \lambda_{i|k}$$

$$\ln \lambda_{i|k} = \beta_{0k} + x_{i1}\beta_{1k} + \cdots + x_{iL}\beta_{Lk} \qquad (6)$$

The latent class model is commonly employed to analyse grouped discrete data, presumed to adhere to a mixture distribution (Clapperton, 2022). Consequently, the Poisson regression model for the latent class incorporates a blend of linear logarithmic (Poisson) regression distributions with a latent variable k (k = 1, 2, ..., c) as depicted below:

$$P(y_i|\alpha.\beta) = \sum_{k=1}^{c} \alpha_k \, P_k(y_i|\lambda_{i|k}) \qquad (7)$$

The parameter (α_k) can be understood as the unconditional probability of an individual belonging to class k. This probability is based on the assumption that, given an individual i belongs to class k, the number of events for that individual follows a Poisson distribution with parameter $\lambda_{i|k}$ (Yang and Lai, 2005).

To address heterogeneity across individuals, we employ two approaches (Lin and Tsai, 2022):
Firstly, we use a formula where the average event rate takes on a discrete patchwork distribution, which varies across a finite number of unobserved populations.
Secondly, the average event rate within each category also varies based on the explanatory variables.

2.3 Estimation of Model Parameters Using Expectation Maximization Algorithm:

The EM algorithm is an iterative algorithm for an estimation of the parameters. Instead of maximizing the log-likelihood function for dataset Y containing n observations, it maximizes the complete-data log-likelihood. The complete data are assembled from the observed data Y (sampled dataset) and the missing data (Panić at el., 2020) .The expectation-maximization (EM) algorithm is developed by (Yang and Lai, 2005) by considering $\{y_1. \ldots . y_n\}$ as an incomplete data set, with the latent class variable $z_{ik} = (z_{i1}. \ldots . z_{ic})'$ being treated as missing. The variable $\boldsymbol{z_{ik}}$ signifies whether the observation y_i belongs to latent class k, and it takes on two possible values (0, 1) , as follows:

$$z_{ki} = z_k(y_i) = \begin{cases} 1. & if \ y_i \in kth \ class \\ 0. & otherwise \end{cases}$$

It is assumed that z_{ki} is a polynomial i.i.d with probabilities α_k since with its distribution function, it is (Yang and Lai, 2005):

$$P(z_i) = \prod_{k=1}^{c} \alpha_k^{z_{ki}} \qquad (8)$$

Where $P(z_i)$ represents the probability of observing i belonging to one of the classes.

If z_{ki} equals 0, it signifies that the observation y_i does not belong to class k, resulting in a probability of $1 - \alpha_k$ for belonging to class k. Conversely, if z_{ki} equals 1, it indicates that the observation y_i belongs to class k, resulting in a probability of α_k for belonging to class k. As a

result, we calculate the probability of observing y_i belonging to class k by raising α_k to the power of z_{ki}.

$$P(y_i|z_i) = \prod_{k=1}^{c} (P_k(y_i|\beta_k))^{z_{ki}} \quad (9)$$

$P(y_i|z_i)$ represents the probability of observation y_i, given class membership z_i. And $P_k(y_i|\beta_k)$ represents the probability mass function of the Poisson distribution of observation y_i when it belongs to class k, whose parameters are determined by the class-specific parameters β_k .

$$P(y_i, \dots, y_n, z_1, \dots, z_n) = \prod_{i=1}^{n} P(y_i, z_i) = \prod_{i=1}^{n} P(y_i|z_i)P(z_i)$$

$$P(y_i, \dots, y_n, z_1, \dots, z_n) = \prod_{i=1}^{n} \left(\prod_{k=1}^{c} \left(P_k(y_i|\beta_k)\right)^{z_{ki}} \alpha_k^{z_{ki}} \right) \quad (10)$$

Therefore, the logarithmic maximum likelihood function will be as follows:

$$\ln L = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ki} \ln P_k(y_i|\beta_k) + \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ki} \ln \alpha_k \quad (11)$$

During the E -step, we utilized $E(z_{ki}|y_i.\alpha.\beta_k)$ to estimate z_{ki} this is because z_{ki} is missing data.

$$P(z_{ki}|y_i) = \frac{P(y_i \,.\, z_{ki})}{P(y_i)} = \frac{P(y_i \mid z_{ki})\, P(z_{ki})}{\sum_{s=1}^{c} \alpha_s \, P_s(y_i \mid \beta_s)}$$

$$P(z_{ki}|y_i) = \frac{(\alpha_k P_k(y_i \mid \beta_k))^{z_{ki}} \left(\sum_{\substack{s=1 \\ s \neq k}}^{c} \alpha_s P_s(y_i \mid \beta_s) \right)^{1-z_{ki}}}{\sum_{s=1}^{c} \alpha_s \, P_s(y_i \mid \beta_s)} \quad (12)$$

$$\hat{z}_{ki} = E[z_{ki}|y_i \,.\, \alpha \,.\, \beta_k] = 1 \times \frac{\alpha_k P_k(y_i \mid \beta_k)}{\sum_{s=1}^{c} \alpha_s \, P_s(y_i \mid \beta_s)} + 0$$

$$\hat{z}_{ki} = \frac{\alpha_k P_k(y_i \mid \beta_k)}{\sum_{s=1}^{c} \alpha_s \, P_s(y_i \mid \beta_s)} \quad (13)$$

Therefore:

$$E(\ln L) = \sum_{i=1}^{n} \sum_{k=1}^{c} \hat{z}_{ki} \ln P_k(y_i|\beta_k) + \sum_{i=1}^{n} \sum_{k=1}^{c} \hat{z}_{ki} \ln \alpha_k$$

During the M-step, we aim to maximize $E(\ln L)$ with the restriction $\sum_{k=1}^{c} \alpha_k = 1$; considering the Lagrange multiplier multiplication, we subsequently take the derivative with respect to α_k and set the resulting equation to zero.

$$\frac{\partial}{\partial \alpha_k} (E(\ln L) - \gamma(\sum_{s=1}^{c} \alpha_s - 1)) = 0$$

$$\hat{\alpha}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ki}}{n} \quad (14)$$

P-ISSN 2518-5764
E-ISSN 2227-703X

$$L(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{c} \hat{z}_{ki}\, lnP_k(y_i|\beta_k) + \sum_{i=1}^{n}\sum_{k=1}^{c} \hat{z}_{ki}\; ln\hat{\alpha}_k$$

$$\frac{\partial}{\partial \beta_{lk}} L(\beta) = \sum_{i=1}^{n} \hat{z}_{ki} \left(\frac{\partial \ln P_k(y_i|\lambda_{i|k})}{\partial \lambda_{i|k}}\right)\left(\frac{\partial \lambda_{i|k}}{\partial \beta_{lk}}\right)$$

Where:

$$\ln P_k(y_i|\lambda_{i|k}) = y_i \ln(\lambda_{i|k}) - \lambda_{i|k} - \ln(y_i!)$$

$$\ln \lambda_{i|k} = \sum_{l=1}^{L} x_{il}\beta_{lk}$$

$$\frac{\partial\, L(\beta)}{\partial \lambda_{i|k}} = \left(\frac{y_i}{\lambda_{i|k}} - 1\right) \quad , \quad \frac{\partial \lambda_{i|k}}{\partial \beta_{lk}} = x_{il}\lambda_{i|k}$$

$$\frac{\partial}{\partial \beta_{lk}} L(\beta) = \sum_{i=1}^{n} \hat{z}_{ki} \left(\frac{y_i}{\lambda_{i|k}} - 1\right) x_{il}\lambda_{i|k} = \sum_{i=1}^{n} \hat{z}_{ki}\,(y_i - \lambda_{i|k}) x_{il}$$

Fisher Matrix is:

$$-E\left[\frac{\partial^2 L(\beta)}{\partial \beta_{lk}\partial \beta_{l'k}}\right] = \sum_{i=1}^{n} \hat{z}_{ki}\, \lambda_{i|k}\, x_{il'} x_{il} = \{X' W_k X\}_{ll'}$$

Where:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1L} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nL} \end{pmatrix} \quad and \quad W_k = \begin{pmatrix} \hat{z}_{k1}^0 \lambda_{1|k} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{z}_{kn}^0 \lambda_{n|k} \end{pmatrix} \qquad (15)$$

Typically, the Newton-Raphson process or method is employed to adjust or modify $\ln\lambda_{i|k}$ and subsequently estimate β_k.

$$\begin{aligned} \gamma_{i|k} &= \ln \lambda_{i|k} + (y_i - \lambda_{i|k}) \frac{d \ln \lambda_{i|k}}{d\, \lambda_{i|k}} \\ &= \ln \lambda_{i|k} + (y_i - \lambda_{i|k}) \frac{1}{\lambda_{i|k}} \end{aligned} \qquad (16)$$

$$X' W_k X \hat{\beta}_k = X' W_k\, \gamma_k$$

$$\hat{\beta}_k = (X' W_k X)^{-1} X' W_k\, \gamma_k \qquad (17)$$

P-ISSN 2518-5764
E-ISSN 2227-703X

Below is a diagram of the EM algorithm:

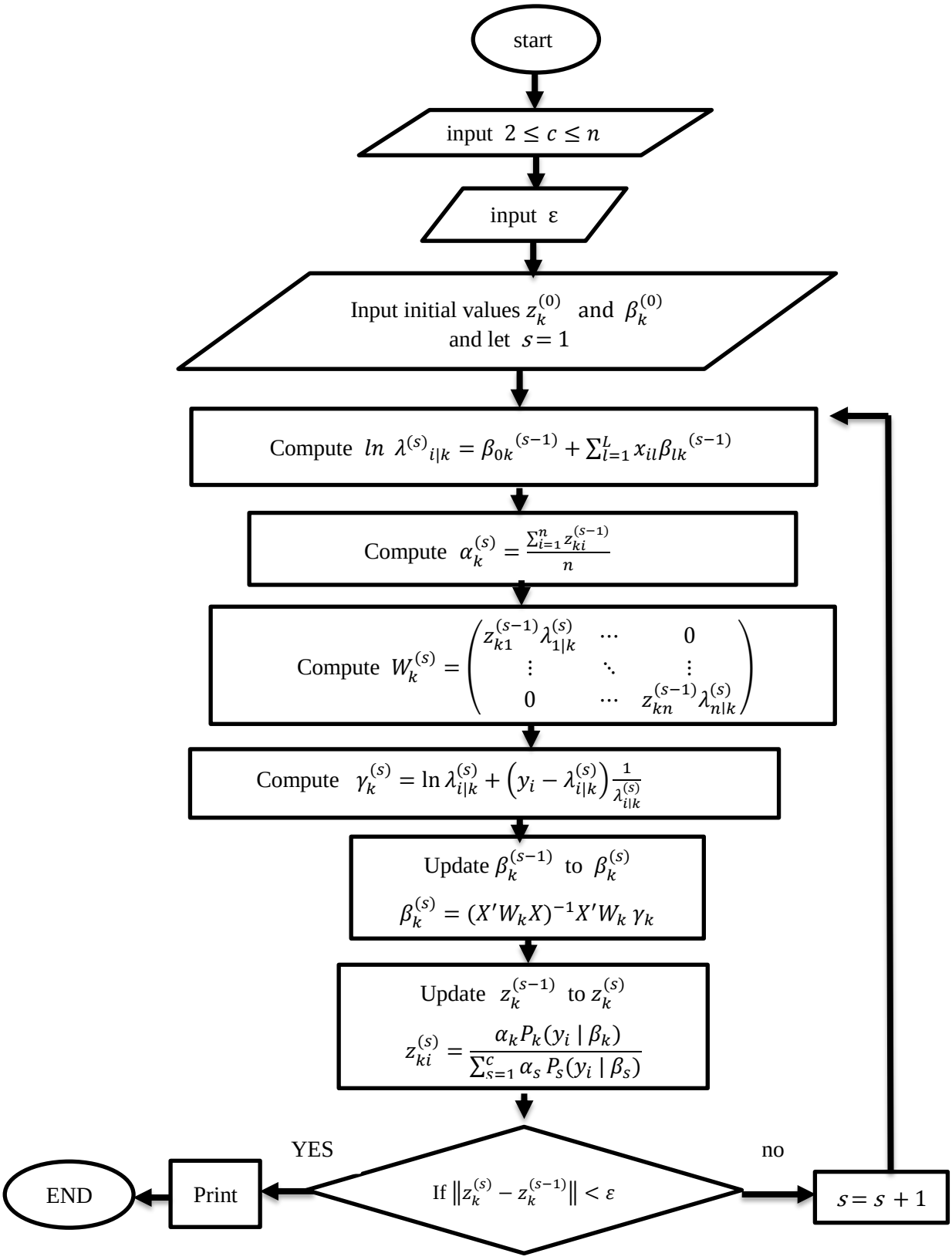

Figure 1:- Diagram of EM algorithm

P-ISSN 2518-5764
E-ISSN 2227-703X

2.4 Genetic algorithm:

Genetic algorithm is one of the most important tools in artificial intelligence and machine learning (Seretis et al, 2018). It is based on biological evolution and is used in many applications in various fields. One of these important applications is using genetic algorithms to improve the parameters of regression models in general. Latent class mixture Poisson regression is also one of the models in which the genetic algorithm can be applied to improve the values of its parameters, Below is a diagram of the genetic algorithm (Mahdavi et al, 2009) :

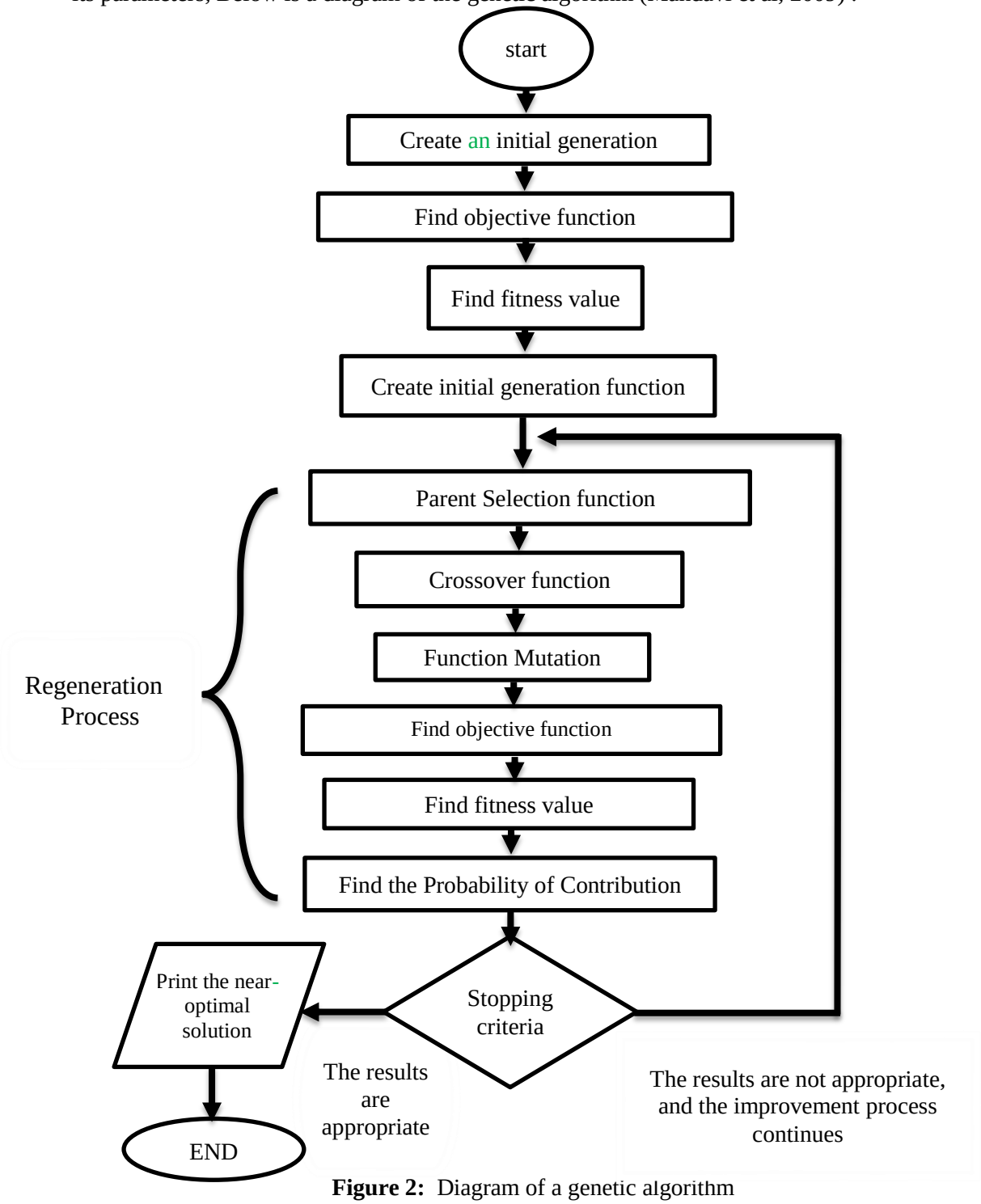

Figure 2: Diagram of a genetic algorithm

P-ISSN 2518-5764
E-ISSN 2227-703X

2.5 Applying the genetic algorithm to estimate the parameters of the Poisson mixture regression model for latent classes

We apply the steps of the genetic algorithm to the objective function equation of a Poisson mixture regression model for the latent class to estimate the model parameters :

1- Start: The chromosome is formed by the parameter values, where the genes represent the chromosome.

2- Initialization: Generating the first generation (or initial) by assigning initial values to the genes, generated randomly.

3- In the objective function, the chromosome is evaluated, and the one possessing a smaller objective function value corresponding to a higher likelihood is selected. Then, the evaluation function is determined using the following equation:

$$fitness\ function = \frac{1}{1 + objective\ function}$$

The probability of the evaluation function (best evaluation) can be calculated using the following formula:

$$C_i = \frac{f_{(i)}}{\sum_{i=1}^{N} f_{(i)}}$$

Where:

C_i : represents the probability of chromosome i.

f_i : represents the evaluation function for chromosome i.

N : is the population size.

And using one of the selection criteria known as the "roulette wheel," a random number, denoted as $r_{(c)}$, is generated within the interval [0,1]. This number is then compared to the first chromosome, denoted as $c_{(1)}$.–The first chromosome is selected if $r_{(c)}$ is less than $c_{(1)}$. This process is repeated for each iteration, and it determines one chromosome for the new population based on the evaluation function.

4- Selected chromosomes are hybridized through mating between two chromosomes, employing one of the hybridization criteria known as regulated hybridization. This is done based on the hybridization probability, denoted as P_c , and this probability value is determined by the researcher, typically falling within the range $P_c \geq 0 \cdot 25$. This value is then compared with the genetic values of the chromosomes (parents) to generate the new generation (offspring), and the exchange occurs when the gene value is greater than or equal to the specified P_c.

5- The mutation process, which depends on the probability value P_m for the parameters, and this probability value is calculated using the following formula:

$$P_m = \begin{cases} 0 \cdot 09 - \dfrac{Fitvalue - f_{mean}}{f_{max}} & if\ \ Fitvalue > f_{mean} \\ 0 \cdot 09 & otherwise \end{cases}$$

Where: Fitvalue represents the evaluation function value, f_{mean} represents the population mean. f_{max} represents the maximum value in the population.

By replacing randomly selected genes with new values also generated randomly, we obtain them using the following formula:

The sum of genes = (number of genes in the chromosome) × (population size).

6- We refer back to step three until the separate achievement criterion is met.

P-ISSN 2518-5764
E-ISSN 2227-703X

7- The evaluation of parameters is carried out based on the value of the objective function to estimate the parameters of a Poisson mixture regression model for the latent class.

3. Discussion of Results:

3.1 Simulation Preparation:

This paper's simulation experiments involved writing several MATLAB programs to generate simulated data to compare methods across different sample sizes. Three sample sizes were adopted for generating the data, namely ($n_1 = 50$, $n_2 = 90$, $n_3 = 130$). The independent variables were generated from a uniform distribution:

$$x_1 \sim U(0,2) \quad , \quad x_2 \sim U(0,32)$$

The model will include two independent variables, x_1 and x_1 , so we will use:

$$\ln \lambda_{i|k} = \beta_{0k} + \beta_{1k} x_{i1} + \beta_{2k} x_{i2}$$

The generation values are based on three scenarios for default parameter values:

Table 1: Represents the three scenarios, S1,S2, and S3, the default parameter values in the presence of x_1,x_2.

Scenario	β_{01}	β_{11}	β_{21}	β_{02}	β_{12}	β_{22}
S1	1	2	-2	0.5	3	-1
S2	2	-1	-3	0.8	1	-2
S3	0.3	1	-1	1.5	2	-3

Alpha (α) is selected to be 0.3 of the sample size (n), where data will be generated in the amount of 0.3n according to the first class:

$$\ln \lambda_{i|1} = \beta_{01} + \beta_{11} x_{i1} \qquad i = 1,2,3 \ldots (0.3\, n)$$

As for the observations of the second class, they are generated from the remaining 0.7n based on the model:

$$\ln \lambda_{i|2} = \beta_{02} + \beta_{12} x_{i1} \qquad i = (0.3\, n + 1) \ldots \; n$$

We calculate a parameter of the Poisson distribution from the equation:

$$\lambda_{i|k} = e^{\beta_{0k} + \beta_{1k} x_{i1}}$$

After that, y_i is generated from a Poisson distribution.

$$y_i \sim poisson(\lambda_{i|k})$$

Then, estimation methods, represented by the Expectation Maximization (EM) algorithm and the genetic (GA) algorithm, are applied to the generated data.

3.2 Results of Simulation :

Each experiment was repeated 500 times. The estimated parameter $\hat{z}_{ki}$ is used to determine the observation's membership. The number of observations correctly classified according to the Rate equation is calculated. The Mean Squared Error (MSE) is used as a statistical measure for comparison between the estimation methods, which are the Expectation Maximization (EM) algorithm and the Genetic Algorithm (GA).

P-ISSN 2518-5764
E-ISSN 2227-703X

Table 2: Represents the Mean Squared Error (MSE) for the estimators and the three scenarios (S1, S2, S3) with respect to sample size n=50 and $\alpha = 0.3$ when considering x_1 , x_2

scenario		α	$1-\alpha$	β_{01}	β_{11}	β_{21}	β_{02}	β_{12}	β_{22}	Rate
S1	EM	0.31	0.31	0.21	0.36	0.25	0.53	0.42	0.34	0.54
	GA	0.21	0.21	0.17	0.29	0.18	0.43	0.31	0.14	0.62
S2	EM	0.37	0.37	0.25	0.32	0.33	0.49	0.27	0.41	0.57
	GA	0.29	0.29	0.21	0.28	0.31	0.41	0.27	0.40	0.63
S3	EM	0.19	0.19	0.29	0.36	0.31	0.28	0.34	0.33	0.57
	GA	0.12	0.12	0.22	0.31	0.27	0.15	0.32	0.31	0.68

By looking at **Table 2**, we observe that for a sample size of n=50 and across all three scenarios (S1, S2, S3), the genetic algorithm outperformed the EM algorithm for all parameter values, as indicated by the Mean Squared Error (MSE) values for each parameter. Significantly, in the genetic algorithm, the MSE values for all parameters are lower than the MSE values in the EM algorithm. This indicates that the genetic algorithm has indeed improved parameter estimation at n=50.

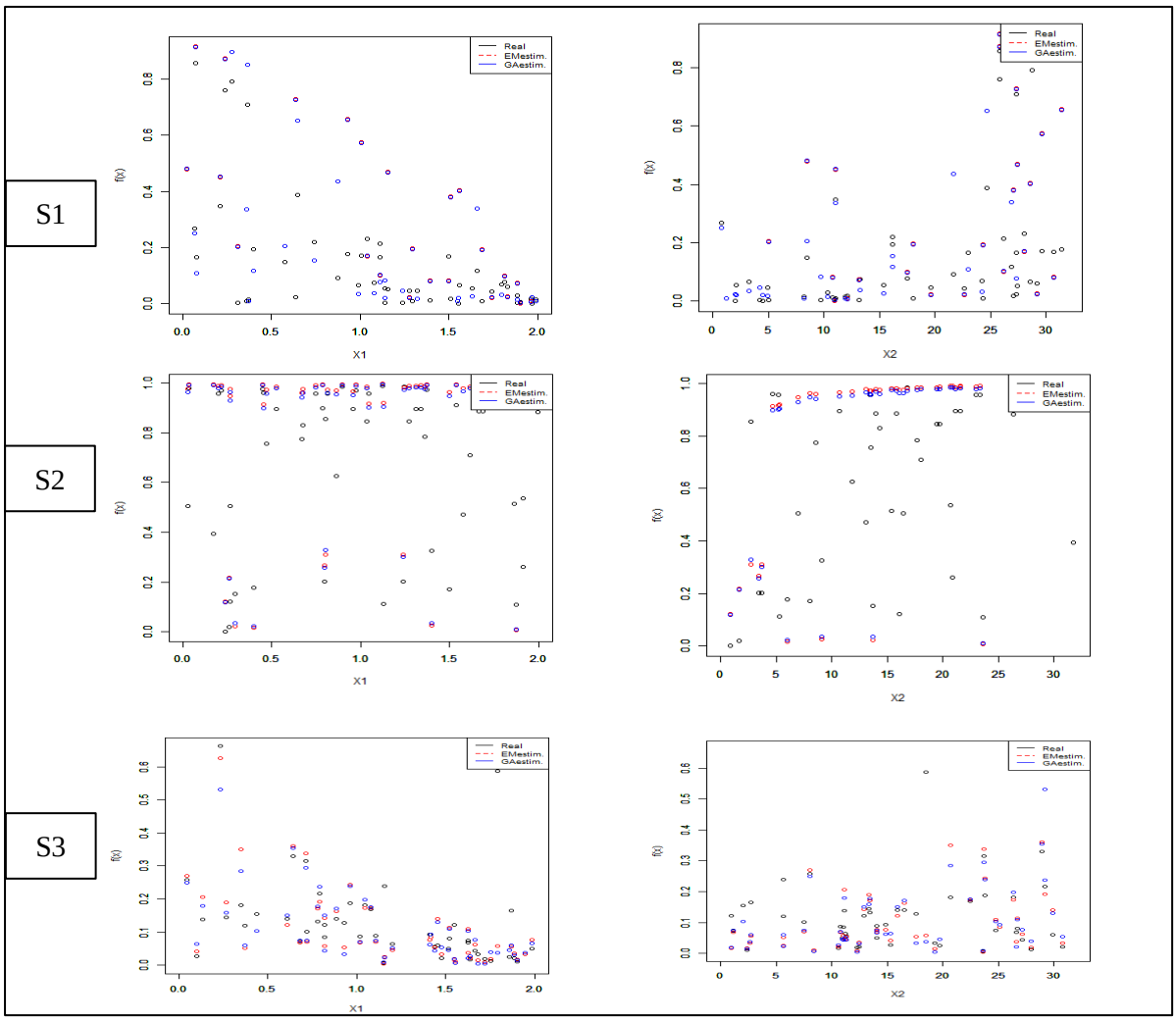

Figure 3 : represents the true function $f(x)$ and the estimated functions $f(x)$ obtained using the EM algorithm and the Genetic Algorithm (GA) for all three scenarios (S1, S2, S3) for x_1 *and* x_2 when the sample size is n=50.

P-ISSN 2518-5764
E-ISSN 2227-703X

Through Figure (3), we notice the closeness of the real values and the values estimated by the EM and genetic algorithms when the sample size is n=50.

Table 3: Represents the Mean Squared Error (MSE) for the estimators and the three scenarios (S1, S2, S3) with respect to sample size n=90 and $\alpha = 0.3$ when considering x_1, x_2

scenario		α	$1-\alpha$	β_{01}	β_{11}	β_{21}	β_{02}	β_{12}	β_{22}	Rate
S1	EM	0.034	0.034	0.068	0.048	0.036	0.042	0.037	0.031	0.67
	GA	0.027	0.027	0.059	0.041	0.031	0.039	0.032	0.028	0.79
S2	EM	0.027	0.027	0.054	0.032	0.037	0.027	0.022	0.017	0.66
	GA	0.021	0.021	0.043	0.026	0.030	0.025	0.018	0.016	0.81
S3	EM	0.033	0.024	0.055	0.042	0.033	0.048	0.039	0.037	0.68
	GA	0.028	0.024	0.050	0.041	0.031	0.039	0.031	0.032	0.83

By looking at **Table 3**, we observe that for a sample size of n=90 and across all three scenarios (S1, S2, S3), the genetic algorithm outperformed the EM algorithm for all parameter values, as indicated by the Mean Squared Error (MSE) values for each parameter. Significantly, in the genetic algorithm, the MSE values for all parameters are lower than the MSE values in the EM algorithm. This indicates that the genetic algorithm has indeed improved parameter estimation at n=90.

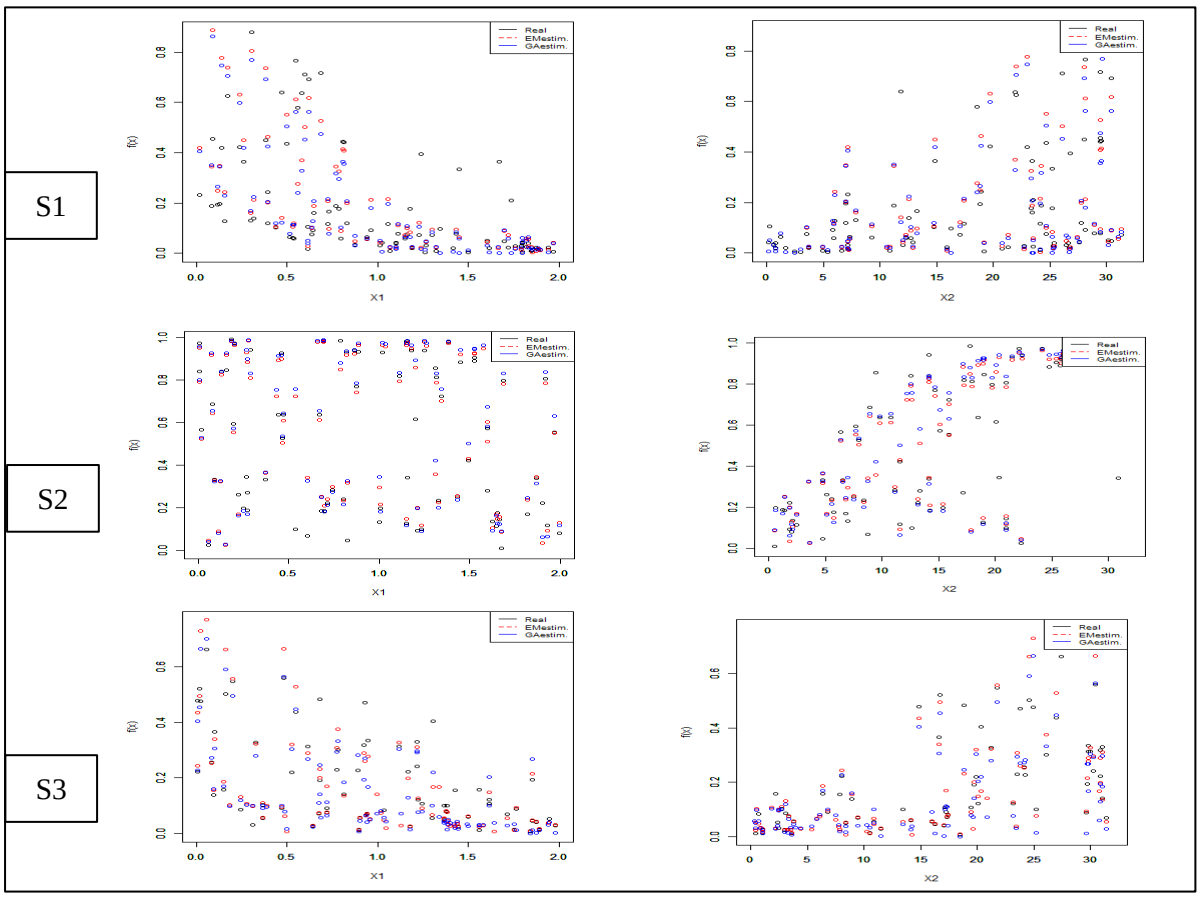

Figure 4: represents the true function $f(x)$ and the estimated functions $f(x)$ obtained using the EM algorithm and the Genetic Algorithm (GA) for all three scenarios (S1, S2, S3) for x_1 *and* x_2 when the sample size is n=90.

P-ISSN 2518-5764
E-ISSN 2227-703X

Through Figure (4), we notice the closeness of the real values and the values estimated by the EM algorithm and the genetic algorithm when the sample size is n=90.

Table 4: Represents the Mean Squared Error (MSE) for the estimators and the three scenarios (S1, S2, S3) with respect to sample size n=130 and $\alpha = 0.3$ when considering x_1 , x_2

scenario		α	$1-\alpha$	β_{01}	β_{11}	β_{21}	β_{02}	β_{12}	β_{22}	Rate
S1	EM	0.0087	0.0087	0.0076	0.0058	0.0053	0.0062	0.0066	0.0048	0.76
	GA	0.0079	0.0079	0.0068	0.0051	0.0051	0.0043	0.0038	0.0032	0.84
S2	EM	0.0062	0.0062	0.0058	0.0049	0.0036	0.0033	0.0032	0.0029	0.76
	GA	0.0058	0.0058	0.0053	0.0041	0.0031	0.0022	0.0029	0.0025	0.84
S3	EM	0.0031	0.0031	0.0053	0.0045	0.0038	0.0033	0.0032	0.0029	0.79
	GA	0.0029	0.0029	0.0048	0.0037	0.0032	0.0028	0.0029	0.0025	0.88

By looking at **Table 3**, we observe that for a sample size of n=130 and across all three scenarios (S1, S2, S3), the genetic algorithm outperformed the EM algorithm for all parameter values, as indicated by the Mean Squared Error (MSE) values for each parameter. Significantly, in the genetic algorithm, the MSE values for all parameters are lower than the MSE values in the EM algorithm. This indicates that the genetic algorithm has indeed improved parameter estimation at n=130.

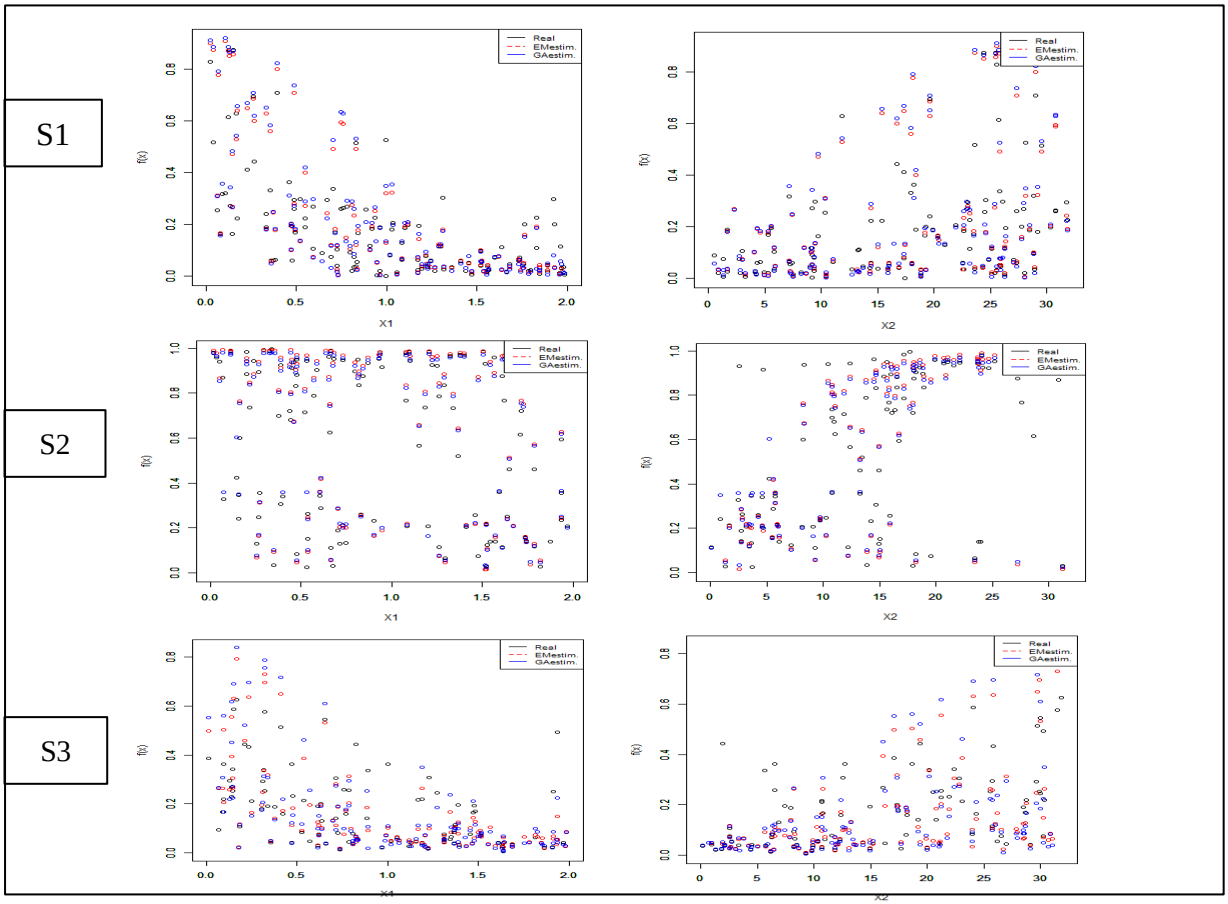

Figure 5: represents the true function $f(x)$ and the estimated functions $f(x)$ obtained using the EM algorithm and the Genetic Algorithm (GA) for all three scenarios (S1, S2, S3) for x_1 *and* x_2 when the sample size is n=130 .

P-ISSN 2518-5764
E-ISSN 2227-703X

Through Figure (5), we notice the closeness of the real values and the values estimated by the EM algorithm and the genetic algorithm when the sample size is n=130.

4.Conclusion:

1- Through simulation and for different sample sizes (50, 90, 130), we observe the superiority of the Genetic Algorithm over the Expectation Maximization (EM) algorithm, as the Genetic Algorithm (GA) yielded lower Mean Squared Error (MSE) values for all parameters at all sample sizes and for all three scenarios (S1, S2, S3), This indicates that the genetic algorithm has improved the parameter values.

2- As the sample size increases, there is an increase in the convergence of the real observations belonging to each class, and the observations are estimated using both the EM algorithm and the genetic algorithm (GA) for the first and second variables.

Authors Declaration:

Conflicts of Interest: None

-We Hereby Confirm That All The Figures and Tables In The Manuscript Are Mine and Ours. Besides, The Figures and Images, Which are Not Mine, Have Been Permitted Republication and Attached to The Manuscript.

- Ethical Clearance: The Research Was Approved By The Local Ethical Committee in The University.

References:

1.Algamal, Z.Y. and Abdalteef, A.M. (2019). "Variable Selection in Poisson Regression Model Using Penalized Likelihood Methods", Journal of Administration and Economics, vol. 37, No. 118, pp 285-294

2.AlKhafaji, M. A. and AlBakri, R. A. (2021) "Using Iterative Reweighting Algorithm and Genetic Algorithm to Calculate The Estimation of The Parameters Of The Maximum Likelihood of The Skew Normal Distribution", Journal of Economics and Administrative Sciences, Vol. 27, No127, pp. 253-264.

3.Cameron, A. C. and Trivedi, P. K. (2013), "Regression Analysis of Count Data", Cambridge University Press, New York USA.

4.Clapperton, B. (2022), "Applying Latent Class Cluster Analysis And Data Mining Methods to Identify Classes of Chronic Fatigue Syndrome Patients That Are Predictive of Treatment Success", Doctoral Dissertation, King's College, London.

5.Gonçalves, J. N. Barreto-Souza, W. and Ombao, H. (2022) "Poisson-Birnbaum-Saunders Regression Model for Clustered Count Data", Arxiv Preprint Arxiv, Vol. 1, pp 10.

6.Kareem, U. I. and Hashim, F. M. (2021) "Comparison of Some Methods for Estimating Mixture of Linear Regression Models with Application", Journal of Economics and Administrative Sciences, Vol. 27, No.129, pp. 171-184.

7.Lin, T. H. and Tsai, M. H. (2022) "Solving Unobserved Heterogeneity With Latent Class Inflated Poisson Regression Model", Journal of Applied Statistics,Vol. 49 , No.11, pp. 2953-2963.

8.Mahdavi, I. Paydar, M. M. Solimanpur, M. and Heidarzade, A. (2009) "Genetic Algorithm Approach for Solving a Cell Formation Problem in Cellular Manufacturing", Expert Systems with Applications, Vol 36, No. 3, pp 6598-6604.

McCullagh, P. and Nelder, J. A. (1989) "Generalized Linear Models", Routledge, New York USA.
https://www.routledge.com/Generalized-Linear-Models/McCullagh-Nelder/p/book/9780412317606

9.Papastamoulis, P. Martin-Magniette, M. L. and Maugis-Rabusseau, C. (2016) "On the Estimation of Mixtures of Poisson Regression Models with Large Number of Components", Computational Statistics and Data Analysis, Vol. 93, pp. 97-106.

10. Panić, B. Klemenc, J. and Nagode, M. (2020) "Improved Initialization of The EM Algorithm for Mixture Model Parameter Estimation", Mathematics, Vol. 8, No. 3, pp. 373.

P-ISSN 2518-5764
E-ISSN 2227-703X

11. Pernkopf, F. and Bouchaffra, D. (2005) "Genetic-based EM algorithm for learning Gaussian mixture models", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No 8, pp.1344-1348.

12. Radam, S. A. and Hameed, L. M. A. (2023) 'Comparison of Poisson Regression and Conway Maxwell Poisson Models Using Simulation", Journal of Economics and Administrative Sciences, Vol. 29, No.136, pp. 81-89

13. Seretis, G. Kouzilos, G. Manolakos, D. and Provatidis, C. (2018) "Multi-Objective Curing Cycle Optimization for Glass fabric/epoxy Composites Using Poisson Regression and Genetic Algorithm", Materials Research, Vol. 21, No. 2, pp.e20170815.

14. Sundararajan,P.K. and Mengshoel, O.J. (2016) "August. A genetic algorithm for learning parameters in Bayesian networks using expectation maximization", Conference on probabilistic graphical models, Vol 52, pp. 511-522.

15. Tzougas, G. (2020) "EM estimation for the Poisson-inverse Gamma Regression Model with Varying Dispersion: an Application to Insurance Ratemaking", Risks, Vol. 8, No. 3, pp. 97.

16. Yang, M.S. and Lai, C.Y. (2005) 'Mixture Poisson Regression Models for Heterogeneous Count Data Based on Latent and Fuzzy Class Analysis', Soft Computing, Vol. 9, No.7, pp.519-524

Journal of Economics and Administrative Sciences
2024; 30(140), pp. 434-449

P-ISSN 2518-5764
E-ISSN 2227-703X

استخدام الخوارزمية الجينية وخوارزمية تعظيم التوقع لتقدير المعلمات في انموذج انحدار خليط بواسون للفئة الكامنة

أحمد خضر الياس
جامعة بغداد / كلية الإدارة والاقتصاد / قسم الإحصاء بغداد ، العراق
ahmed.khedr2101m@coadec.uobaghdad.edu.iq

عماد حازم عبودي
جامعة بغداد / كلية الإدارة والاقتصاد / قسم الإحصاء بغداد ، العراق
emadhazim@coadec.uobaghdad.edu.iq

Received:22/10/2023 Accepted:3/12/2023 Published Online First: 30 /4/ 2024

مستخلص البحث:
في أنموذج انحدار خليط بواسون للفئة الكامنة تأتي المشاهدات من مصادر فرعية أو فئات مختلفة ، حيث يفترض أن البيانات التي تم مشاهدتها يتم إنشاؤها بواسطة خليط محدود من الفئات غير المرصودة أو الكامنة حيث تكون المشكلة في التخصيص المناسب للملاحظات الى كل فئة ويتطلب هذا الامر أساليب معقدة لتقدير المعلمات في الانموذج . عادة، يتم تقدير المعلمات في نموذج انحدار خليط بواسون للفئة الكامنة بواسطة خوارزمية EM التقليدية . تهدف الورقة البحثية إلى المقارنة بين خوارزمية EM و الخوارزمية الجينية GA . باستخدام المحاكاة تمت مقارنة الخوارزميتين بناءً على معيار MSE وبأحجام عينات مختلفة (n=50,90,120) وبثلاثة سيناريوهات (S1,S2,S3) للقيم الافتراضية للمعلمات . أثبتت النتائج تفوق الخوارزمية الجينية GA على خوارزمية EM، حيث كانت الخوارزمية الجينية GA تمتلك اقل قيم (MSE) .

نوع البحث : ورقة بحثية.

المصطلحات الرئيسة للبحث : انحدار خليط بواسون ، الفئة الكامنة ، خوارزمية تعظيم التوقعات ، الخوارزمية الجينية .

*البحث مستل من رسالة ماجستير