Zero inflation Poisson Regression estimation by using Genetic Algorithm (GA)

Tahir R. Dikheel
University of AL-Qadisiyah, Al Diwaniyah, Iraq

Huyam A. Jouda
University of AL-Qadisiyah, Al Diwaniyah, Iraq

Abstract

In this research, the phenomena that follow the models for dealing with numerical data, which take the form of time series, have been dealt with, the time series that take the numerical form often suffer from the problem of zero inflation, which represents a problem that cannot be overlooked, so a set of methods and methods have been proposed to deal with this problem, and one of the most important models in this case is Poisson's zero-inflated model. It should be noted here that there are a set of methods that are used for the purpose of estimating the parameters of the Poisson zero-amplified model, including the MLE method and the NLM method. In this paper, a genetic algorithm method was proposed for the purpose of improving the above estimates. The simulation and real data were used for the purpose of verifying the performance of the proposed method, where the research relied on the MSE standard for the purpose of comparison, which proved that the proposed method gave good results compared to other methods.

Keywords: *Time series; Hardel model; Zero inflation; Poisson regression; Genetic Algorithm.*

INTRODUCTION

Poisson distribution its a default starting point to modeling count data, its p.m.f. is provided as follows (For example, data that accepts only positive integer values).

$$\Pr[Y = y] = \frac{\exp(-\mu)\,\mu^{y}}{y!} \quad ; \quad y = 0,1,2,\ldots$$

as is well-known, ($\boldsymbol{\mu} > 0$) this distribution is said to be "equi-dispersed" because both its mean and variance are greater than zero. The Poisson distribution is less beneficial when the data is "over-dispersed," or when the variance of them is greater than the mean of them. The well-known Negative Binomial (NB II) distribution results from letting the variation be described by a gamma distribution in turn. The latter can detect data overdispersion.

It's Naturally would do the same thing here and includes covariates in model by assigning covariates to the dependent variable's (conditional) mean in linear regression as a function of parameters and covariates.

$$\mu = \exp(x'\beta)$$

where it is demonstrated using the exponential function that ($\mu > 0$), as is clearly necessary. Maximum likelihood can then be used to estimate the parameters., This is straightforward because the log-likelihood function is concave (So is the NB II model). The following discussion will be successful if it is understood that the Poisson model and conventional variations that allow for over-dispersion cannot describe multi-modal data.. (More accurately if μ is integer, then the

modes of the Poisson distribution are μ and (μ – 1), but never at values that are not contiguous. If μ is non-integer, the single mode takes place at [μ].) A variation of the well-known The Poisson regression model, which permits a surplus of zero counts in the data, zero-inflated Poisson (ZIP) regression model. This phenomena frequently occurs in real-world situations. For this model, typical references include Heilbron (1989), Additionally, excellent conversations are offered by Winkelmann (2000) and Cameron & Trivedi (1998).

The fact that the data comes from two regimes is the key concept. While counts in one regime (RI) always result in zero, counts in the other regime (RII) proceed according to a regular Poisson process.

Suppose

$$\Pr[y_i \in R_I] = \omega_i \,; \Pr[y_i \in R_{II}] = (1-\omega_i) \,; \quad i = 1,2,3,\dots,n$$

when

$$\Pr[y_i = 0] = \omega_i + (1-\omega_i)\exp(-\mu_i)$$

And

$$\Pr[y_i = r] = \frac{(1-\omega_i)\exp(-\mu_i)\,\mu_i^r}{r!} \quad ; \quad r = 1,2,3,\dots$$

As previously, the conditional mean serves as the model's entry point for covariates, μ_i, of Poisson distribution

$$\mu_i = \exp(x_i'\beta)$$

where x_i' is $(1\times k)$ vector of i^{th} observation on the covariates, and β is a $(k\times 1)$ vector of coefficients.

Clearly

$$E[y_i|x_i] = (1-\omega_i)\mu_i$$

And

$$Var[[y_i|x_i] = (1-\omega_i)(\mu_i + \omega_i\mu_i^2)$$

and as a result, this structure also allows for excessive data dispersion. ($if\ \omega_i>0$). When the Poisson model is applied to the Negative Binomial model, the over-dispersion does not result from heterogeneity. Rather, it results from the data's division into the two regimes. In actuality, one or both of these factors could contribute to the prevalence of over-dispersion. (Mullahy, 1986) (Greene, 2003, p. 750).

According to Lambert (1992), it is customary and practical, to modeled ω_i utilizing the Logit model, so:

$$\omega_i = [\exp(z_i'\gamma)]/[1+\exp(z_i'\gamma)]$$

Where z_i' is $(1\times p)$ the vector of i^{th} observing a few covariates, and γ is a $(p\times 1)$ added parameter vector. Of course, the elements of z_i may contain components of x_i, in addition to the possibility of replacing the Logit definition with a Probit (or other) specification.

If we has n It is clear from the sample's independent observations that the log-likelihood function can showed as

$$\begin{aligned}\log L(\beta,\gamma) = &\sum_{y_i=0} \log[\exp(z_i'\gamma) \\ &+ \exp(-\exp(x_i'\beta))] \\ &+ \sum_{y_i=0} [y_i x_i'\beta - exp(x_i'\beta) \\ &- \log(y_i!)] \\ &- \sum_{i=0}^{n} \log[1+\exp(z_i'\gamma)]\end{aligned}$$

(Cameron & Trivedi,1998,p.126.)

We must consider the various ranges of summing while coding the aforementioned log-likelihood function for usage in R (or any other program that needs just one observation on the log-density before adding all n, assuming that the observations are independent). The log-third likelihood's term doesn't need to be changed because the range of summation includes all n. We can create a dummy variable to cope using the summation ranges for the first two terms, that, D_i ,if it takes the value one $y_i = 0$, and else zero. The i^{th} Following that, the observation on the log-likelihood It will be encoded as

$$\begin{aligned}\log L_i(\beta,\gamma) = D_i \log[&\exp(z_i'\gamma) \\ &+ \exp(-\exp(x_i'\beta))] \\ &+ (1 - D_i)[y_i x_i'\beta - exp(x_i'\beta) \\ &- \log(y_i!)] - \log[1 + \exp(z_i'\gamma)]\end{aligned}$$

Time series:

time series is collection of random variables over an extended period of time. The measurements are typically taken at regular intervals. Numerous uses of time series analysis include forecasting of the economy, sales, the stock market, sports, and many more.

Fitting a model that describes the time series' structure and offers practical interpretations is the fundamental goal of time series modeling. An application for a fitted model is:

• To draw attention to the essential elements of the time series, such as change-points, seasonality, and trend

• To clarify the relationship between present and pasts occurrences so that future values of the series can be predicted.

The data aren't always independent, which is one way time series analysis differs from regression analysis. Let $(\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_\mathrm{n})$ be a time series with n elements, denoted as $\{X_t\}_{t=1}^n$ The mean composition of $\{X_t\}_{t=1}^n \, is \, \mu_t = E[X_t]$. The structure of covariance of $\{X_t\}_{t=1}^n$ can be showed by its autocovariance function (ACVF). ACVF with h lag at time can be written as:

$$\begin{aligned}\gamma_\mathrm{x}(\mathrm{t}, \mathrm{t}+\mathrm{h}) &= \mathrm{Cov}(\mathrm{X_t}, \mathrm{X_{t+h}}) \\ &= \mathrm{E}(\mathrm{X_t}, \mathrm{X_{t+h}}) - \mathrm{E}(\mathrm{X_t})\mathrm{E}(\mathrm{X_{t+h}})\end{aligned}$$

the time series $\{X_t\}$ if it meets the following requirements:

a) the mean $\mathrm{E}(\mathrm{X_t})$ its same for each t

b) for all t and every $h \in \{0,1,2,\cdots\}$ the covariance between X_t and X_{t+h} is same. Similarly, It is argued that a time series $\{X_t\}$ is strictly stationary if $(X_1, X_2, \ldots, X_n)$ & $(\mathrm{X_{1+h}}, \mathrm{X_{2+h}}, \ldots, \mathrm{X_{n+h}})$ for all integers, the joint distribution is the same $h > 0$ and $n > 0$. Clearly, Weakly stationary is implied by rigidly stationary.

In the case of a (weakly) stationary series $\{X_t\}$, In this setting, The lag h ACVF is independent of t for some of h

$$\gamma_x(\mathrm{t}, \mathrm{t}+\mathrm{h}) = \gamma_x(0, \mathrm{h})$$

For ease of notation, all ACVFs can utilize the same argument: $\gamma_x(h) = \gamma_x(0, h)$.

Sometimes it is simpler to examine correlations than covariances. A stationary time series' autocorrelation function (ACF) $\{X_t\}$ is defined as

$$\rho(h) = Corr(X_t, X_{t+h}) = \frac{\gamma(\mathrm{h})}{\gamma(0)}$$

The Cauchy-Schwarz inequality clearly shows that ACFs are between -1 and 1. The impact of

series dispersion is eliminated in ACFs. ACFs can be used to compare how dependent various series are.

For series, it's advantageous to pursue a partial autocorrelation function (PACF). In general, a conditional correlation is a partial correlation. The conditional correlation between X_t & X_{t+h}, $h > 0$, is what is used to (PACF) for a time series between X_t and X_{t+h}, conditional on $X_{t+1}, X_{t+2}, \dots, X_{t+h-1}$

$$\kappa(h) = Corr(X_t, X_{t-h} | X_{t+1}, \cdots, X_{t+h-1}),$$

after linear prediction for all variables between X_t and X_{t+h}, where the conditional correlation is taken between X_t and X_{t+h}.

The most popular model class in stationary time series analysis is an autoregressive moving average (ARMA) model class. The general (ARMA) model introduced by (Peter Whittle in the 1970). The most current observation in a series is linked to older observations and incorrect forecasts using the ARMA model class. the ARMA(p,q) model contains moving-average terms up to order q as well as autoregressive terms up to order p. It abides by recursion.

$$X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \cdots + \emptyset_p X_{t-p} + z_t - \theta_1 z_{t-1} - \theta_2 z_{t-2} - \cdots - \theta_q z_{t-q}$$

where p *and q* non-negative integers. White noises make up the series (z_t) which is frequently thought to have an independent, uniform distribution in time t. the ARMA(p,q) model is also known as a moving-average model (MA(q)) where $p = 0$. Likewise, when $q = 0$, This model is known as an autoregressive model of order p (AR(p)).

We can use model (ARIMA) Autoregressive integrated moving average to illustrate patterns in non-stationary series. The model's components, which consist of d difference operation, q moving average terms and p autoregressive terms, are defined in the form ARIMA (p, d, q). greater formality, operation { X_t } if $(1-B)^d X_t$ is ARMA(p,q), when(p,d,q) are positive integers is called to be ARIMA (p,d,q) , $(1-B)^d$ is the d_{th} operator for order differences.

The models that belong to the generalized linear family, they negative binomial (NB) regression model and poisson regression model, etc...

The Poisson Regression Model:

Poisson distribution models the probability of y, its formula:

$$Pr(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!} \qquad (y = 0,1,2, \dots)$$

Keep in mind that there is only one parameter used to define the Poisson distribution. This is a rare event's average incidence rate per unit of exposure. parameter $\boldsymbol{\mu}$ it can be explained as the risk of a new incident of the event during a given exposure period, t. The probability of y events is given by the relation:

$$Pr(Y = y|\mu, t) = \frac{e^{-\mu t}(\mu t)^y}{y!} \qquad (y = 0,1,2, \dots)$$

The likelihood that the mean, variance and Poisson distribution's are equal is.

Regular multiple regression is similar to Poisson regression, but the dependant variable (Y) in Poisson regression is a count that is observed and follows Poisson distribution. Thus, the possible values of (Y) are positive integers: (0, 1, 2,3,ect.). It's believed that high numbers are uncommon. In light of the fact

that logistic regression also involves a discrete response variable, Poisson regression is comparable to it. However, unlike in logistic regression, the answer is not constrained to particular values. The investigation of the relationships between the colony counts of bacteria and various environmental factors and dilutions is one example of an appropriate application of Poisson regression. Another illustration is the quantity of machine breakdowns under various operating circumstances. Another illustration would be crucial data on cancer incidence or newborn mortality in certain demographic groups.

In Poisson regression, we assume that of collection of k regressor variables (X's) determine the Poisson incidence ratio. A formula of this quantity is:

$$\mu = t\ exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

The regression coefficients $\beta_1, \beta_2, \dots, \beta_k$ a parameters who are unknown and are estimate using the set of data. Their estimates is labeled $b_1, b_2, \dots, b_k$. that notation is used For an observation, the basic Poisson regression model is expressed as

$$Pr(Y_i = y_i|\mu_i, t_i) = \frac{e^{-\mu_i t_i}(\mu_i t_i)^{y_i}}{y_i!}$$

Where

$$\mu_i = t_i\ \mu(X_i'\beta)$$

$$= t_i\ exp(\beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})$$

In other words, the result follows the Poisson distribution for a particular set of the regressor variables' values.

The poisson model has number of issues, It includes Zero-inflation result of model estimate.

Zero-inflated models:

The model class of zero-inflated models (Mullahy 1986; Lambert 1992) is another one that can handle excessive zero counts (Cameron & Trivedi 1998, 2005.). These models consist of two components: a point mass at zero and the count distribution like the geometric, (NB) or Poisson distribution. Due to this, two sources for zeros are there: the count component and the point mass. The unobserved state was modeled by using a binary model (zero vs. count): in the simplest example, it merely has an intercept but could also have regressors.

Formally, the point of the mass at zero and the zero-inflated density is combined. $I_{\{0\}}(y)$ and the count distribution $f_{count}(0; z, y)$. Probability inflates the likelihood of seeing a zero count.$\pi = f_{zero}(0; z, y)$:

$$f_{zeroinfl}(y; x, z, \beta, y) = f_{zero}(0; z, y)\ .\ I_{\{0\}}(y) + \big(1 - f_{zero}(0; z, y)\big)\ .\ f_{count}(y; x, \beta)$$

when $I_{\{0\}}$ is the indicator function and the unobserved probability π of belonging for the point mass component are modelled by a binomial Generalized linear model $\pi = g^{-1}(z'y)$. The corresponding regression equation to the mean given by:

$$\mu_i = \pi_i 0 + (1 - \pi_i).\ exp(x_i'\ \beta)$$

utilizing the link in the canonical log. The zero-inflation model's vector of regressors $z_i$and the regressors in the count component x_i

Wants not to distinction the simple case, $z_i = 1$ is only an intercept. The hypothetical link function $g(\pi)$ in binomial GLMs are the logit link, but other links like the probit is also available. The full set of parameters of β, γ, and Dispersion potential parameter θ (if we used (NB) count model) we can estimate by Maximum likelihood. Inference is usually performed to $\beta \,\&\, \gamma$, when θ is treated as a nuisance parameter even if a (NB) model is use.

hurdle models:

Several experimental count data sets show Excessive dispersion and more zero observations than the Poisson model would predict. The hurdle model, first put forth by Mullahy (1986) in the econometric literature, is one model class capable of capturing both characteristics (Cameron & Trivedi 1998, 2005,). A censored count distribution or a binomial model can be used for the latter. The hurdle model combine count data model more formally. $f_{cont}(y; x, \beta)$ (that was left truncated in $y = 1$) and zero hurdle model $f_{zero}(y; z, \gamma)$ (right-censored in $y = 1$)

$$f_{hurdle}(y;\ x, z\ , \beta, \gamma) = \begin{cases} f_{\text{zero}}(0; z, \gamma) & if\ y = 0 \\ \big(1 - f_{zero}(0; z, \gamma)\big).f_{cont}(y; x, \beta)\big(1 - f_{cont}(0; x, \beta)\big) & if\ y > 0 \end{cases}$$

A model of the parameters β, γ, and likely to be two or one additional scattering parameter θ (whether f_{zero} *or* f_{cont} or both of them (NB) densities) were estimate by MLE , when a benefit of likelihood definition is the ability to maximize the count and hurdle components independently. According to, there is a matching mean regression relationship.

$$log(\mu_i) = x_i'\beta + log\big(1 - f_{zero}(0; z_i, \gamma)\big) - log(1 - f_{count}(0; x_i, \beta))$$

use the canonical log URL once again The most logical specification for using zero model such a barrier is probably the binomial GLM 1. If the same regressors are used, a different insightful interpretation emerges $x_i = z_i$ are use the same count model in both of components $f_{zero} = f_{cont}$.

the test of the hypothesis $\beta = \gamma$ then tests whether the hurdle is required or not.

The Maximum Likelihood (ML) estimator:

We observe data $\{(x_i, y_i)|1 \le i \le n\}$. The number $\boldsymbol{y_i}$ is a realization of the random

variable Y_i. Using independence, the total log-likelihood is given by:

$$log\ L(y_i, \dots, , y_n \backslash \beta, x_i, \dots, x_n = \sum_{i=1}^{n} log\ P(Y_i = y_i \backslash \beta, x_i$$

With, according to:

$$P(Y_i = y_i \backslash \beta, x_i) = \frac{exp(-\mu_i)\mu_i^{y_i}}{y_i!}$$

And $\mu_i = \exp(\beta^t x_i)$. Write now long $L(\beta)$ is a quick way to express the overall likelihood.

It then follows.

$$log\ L(\beta) = \sum_{i=1}^{n} \{-exp(\beta^t x_i) + y_i(\beta^t x_i) - \log(y_i!)\}$$

Therefore (ML) estimator is of course is given by:

$$\hat{\beta}_{ML} = arg_{\beta} max\ log\ L(\beta)$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)x_i = 0$$

With $\hat{y}_i = exp(\beta^t x_i)$ the fitted value of $\boldsymbol{y_i}$ **.** As is customary, the anticipated fitted value has been used as the estimated value of $E[Y_i|x_i]$**.** The vector of the residual is orthogonal to the vectors of the explicative variables, according to this first order requirement.

The advantage of the maximum likelihood framework is that the cov formula is readily available:

$$cov(\hat{\beta}_{ML}) = \left(\sum_{i=1}^{n} x_i x_i^t \hat{y}_i\right)^{-1}$$

Additionally, Wald tests, Lagrange multiplier tests, and Likelihood Ratio tests can now be used to do hypothesis tests.

Genetic Algorithm (GA):

Workability (GA) are based in Darwin's theory of survival of the fittest. It may contains a chromosome, a gene of fitness, set a population, fitness function, selection, breeding and mutation. chromosome's are set Solutions (called population) by their they (GA) is begin with. . depend on we Solutions from one population to create a new population, who is motivated by possibility that a new population will be best from the old population. Furthermore, solutions are select according to their fitness to form new solutions, that is, offsprings. The above process is repeated Until some conditions are met algorithmically, the basic (GA) is outlined As follows:

Stage A (**Start**) Create a chromosomal population at random, or appropriate answers to the issue.

Stage B (**Fitness**) Analyze every chromosome's fitness in the population.

Stage C (**New population**) Repeat the subsequent stages until the new population is finished to create a new population.

1- (**Selection**) Based on their fitness, Select two parents chromosomes from the population. Greater fitness increases the likelihood of being chosen as a parent.

2- (**Crossover**) Cross over the parents in the event of a crossover probability to create additional offspring, or children. If there is no Crossing, the offspring would be a perfect replica of the parents.

3- (**Mutation**) Create new children with a mutation probability at each locus.

4- (**Accepting**) Add fresh progeny to the new population.

Stage D (**Replace**) Use the newly formed population to run the algorithm again.

Stage E (**Test**) Stop and return the most effective solution available to the present population if the final condition is met.

Stage F (Loop) went to **stage B**.

The crossover and mutation operators have a significant impact on the performance of genetic algorithms. In Fig. 1, the block diagram for (GA) is displayed.

Encoding Technique in (GA):

The solution to the problem is transformed into chromosomes using the problem-specific encoding technique is use in (GA). Binary encoding, permutation encoding, value encoding, and tree encoding are some of the different encoding methods utilized in (GA).

Selection Techniques in (GA):

Genetic algorithms (GA) rely on an assessment criterion that provides a Measure the value of any chromosome in the context of the task as the basis for their selection function. In this stage of the genetic algorithm, certain genomes are selected from the collection of chromosomes. The three most popular methods for chromosomal selection are steady state, rank, and roulette wheel.

(GA) Operators:

(GA) can be used to optimize various parameters in any process control application. (GA) uses a variety of operators, including as crossover and mutation, to properly choose the optimal value. The encoding methodology and the requirements of the challenge determine the right crossover and mutation approach to use. **1-Crossover ; 2- Mutation**

Figure no 1. Block Diagram Representation (GA).

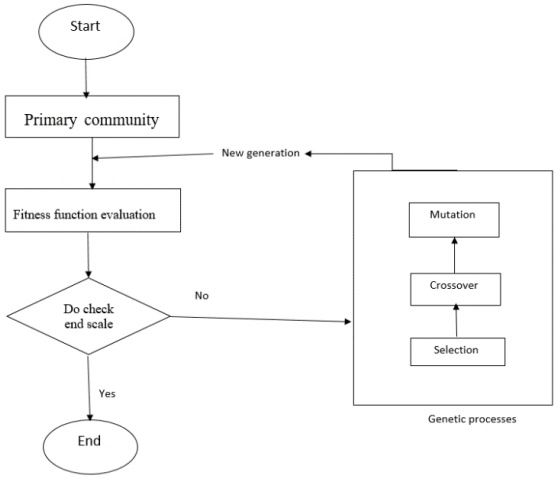

Figure no 1 Displays block diagram of different phases of performance improvement (GA).

Simulation study:

To compare the genetic method with other methods used in the experiment and the extent of its impact on the data, a simulation was conducted consisting of four cases with different sample sizes(50;150; 250), fixed iteration number to (500), different ratio of zeros in data with, (0.1; 0.4; 0.6) the lambda values of the variables (x=2; y=1) are fixed, and the values of the estimated parameters are variable in each case of the simulation, in the first case there are default values and the second is estimated values In(mle) method real data, the third is by (+0.5) the estimated values, and the fourth is(-0.5)

Case 1: in case $\beta = (0.2,0.04,0.1)$

Table 1: MSE for method:

n	p.zero	Mle	nlm	GA
50	0.1	0.2232	0.2255	0.1897
150		0.2108	0.2208	0.1849
250		0.1941	0.2205	0.1731

50		0.2101	0.2263	0.1876
150	0.4	0.1759	0.2214	0.1825
250		0.1089	0.2176	0.1747
50		0.5094	0.2343	0.1785
150	0.6	0.2977	0.2270	0.1745
250		0.2049	0.2254	0.1479

It is clear from the above table that the genetic method (GA) was the best compared to other methods and based on the values of (MSE) as it is clear that when the data's fraction of zeros rises, (MSE) is heading towards a decrease in all methods, but the genetic method was affected less than the rest of the methods and at all studied samples, in addition to that we note that when the sample size increases, the (MSE) is directed to The decline in general

Case 2: in case $\beta = (0.972, -0.002, -0.226)$

Table 2: MSE for method:

N	p.zero	Mle	nlm	GA
50		0.1274	0.0229	0.0433
150	0.1	0.1156	0.0212	0.0351
250		0.0786	0.0196	0.0315
50		0.1433	0.0244	0.0648
150	0.4	0.0879	0.0211	0.0341
250		0.0423	0.0206	0.0321
50		0.1652	0.0267	0.0849
150	0.6	0.1540	0.0250	0.0634
250		0.1255	0.0223	0.0498

It is clear from the above table that the genetic method (GA) was the best compared to other methods and based on the values of (MSE) as it is clear that when the data's fraction of zeros rises, (MSE) is heading towards a decrease in all methods, but the genetic method was affected less than the rest of the methods and at all studied samples, in addition to that we note that when the sample size increases, the (MSE) is directed to The decline in general

Case 3: in case $\beta = (1.472, 0.498, 0.274)$

Table 3: MSE for method:

N	p.zero	mle	Nlm	GA
50	0.1	0.5176	0.2682	0.2075
150		0.4662	0.2665	0.2026
250		0.3945	0.2601	0.1753
50	0.4	0.9486	0.2781	0.1903
150		0.5023	0.2748	0.1800
250		0.4724	0.2720	0.1674
50	0.6	0.6104	0.2776	0.1911
150		0.5408	0.2764	0.1876
250		0.1667	0.2496	0.1599

It is clear from the above table that the genetic method (GA) was the best compared to other methods and based on the values of (MSE) as it is clear that when the data's fraction of zeros rises, (MSE) is heading towards a decrease in all methods, but the genetic method was affected less than the rest of the methods and at all studied samples, in addition to that we note that when the sample size increases, the (MSE) is directed to The decline in general

Case 4: in case $\beta = (0.472, -0.502, -0.726)$

Table 4: MSE for method:

N	p.zero	mle	Nlm	GA
50	0.1	0.3841	0.2676	0.2383
150		0.2734	0.2619	0.2157
250		0.2426	0.2057	0.1934
50	0.4	0.2827	0.2100	0.2915
150		0.2431	0.2040	0.2905
250		0.2363	0.1929	0.2686
50	0.6	0.2540	0.2091	0.3741
150		0.1993	0.2028	0.3300
250		0.1660	0.1979	0.2986

It is clear from the above table that the genetic method (GA) was the best compared to other methods and based on the values of (MSE) as it is clear that when the data's fraction of zeros rises, (MSE) is heading towards a decrease in all methods, but the genetic method was affected less than the rest of the methods and at all studied samples, in addition to that we note that when the sample size increases, the (MSE) is directed to The decline in general

Lung Cancer Data:

It's With every five cases of cancer, there is one case in males, and out of every nine cases of cancer in females, there is one case, Lung cancer ranks second in terms of the speed of its spread compared to the rest of the types of cancer. that originates in the lungs' cellular structure. Many additional cancers, including breast and kidney cancers, has the potential to disseminate (metastasize) to the lungs. There is no referral for the cancer to be lung disease when this occurs. This is so because the location of the original tumor determines what type of cancer it is and how it is treated. It's divided into two basic types: Small Cell Lung Cancer and Non-Small Cell Lung Cancer (NSCLC) for instance, if breast cancer has spread to the lungs, its be treated as metastatic breast cancer rather than lung cancer (SCLC). These varieties develop and disperse differently. They are frequently handled differently.

Lung cancer statistics are used to highlight the performance methodologies. The writers gathered these statistics from an Iraqi medical facility in Al-Naasiria City, Iraq. They reflect the number of lung cancer patients diagnosed each day in Al-Naasiria City between January 1 and December 31, 2021. One response variable (lung cancer) and bacterial water pollutants (T.P.C.) make up these data. Variable (Y) represents the number of people with lung cancer, and variable (X) represents bacterial water pollutants (T.P.C)

Table 5: general statistics.

Variables	Means	SD	Cv
X	5.12	2.1428	41.8512
Y	5.1224	2.1599	42.1658

It is clear from the above table that the value of the arithmetic mean for the variable X amounted to 5.12 with a standard deviation of 2.1428 and a coefficient of variation of 41.8512, and the value of the arithmetic mean for the variable Y amounted to 5.1224 with a standard deviation of 2.1599 and a coefficient of variation of 42.1658.

Table 6: RMSE for the method.

		Par			Expar		
method	RMSE	b_0	b_1	b_2	b_0	b_1	b_2
Mle	0.1838	0.9719	-0.0021	-0.2262	2.6429	0.9979	0.7976
Nlm	0.0904	1.7298	-0.0247	-0.0215	5.6395	0.9756	0.9787
GA	0.0584	2.0653	0.0196	0.0218	7.8880	1.0198	1.0221

The proposed approaches and additional methods are displayed in the table by (RMSE). As the findings of the (GA) technique were the best compared to the other approaches because it produced the lowest estimate, we observe that the regular methods are significantly impacted (RMSE).

Conclusions:

The (GA) was proposed to deal with zero inflation data, improve estimates and obtain better estimator. By using simulation and real data, the results showed that the (GA) is the best compared to other methods (mle and nlm) based on the values of (MSE). The results proved that water pollution (T.P.C) is one of the causes of lung cancer.

Reference

Al-Osh, M., & Alzaid, A. A. (1988). Integer-valued moving average (INMA) process. Statistical Papers, 29(1), 281-300.

Beghin, L., & Macci, C. (2014). Fractional discrete processes: compound and mixed Poisson representations. Journal of Applied Probability, 51(1), 19-36..

Blight, P. A. (1989). Time series formed from the superposition of discrete renewal processes. Journal of Applied Probability, 26(1), 189-195.

Cameron, A. C., & Trivedi, P. K. (2013). Regression analysis of count data (Vol. 53). Cambridge university press.

Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: methods and applications. Cambridge university press.

Giles, D. E. (2007). Modeling inflated count data. In MODSIM 2007 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand, Christchurch, NZ (pp. 919-925).

Giles, D. E. (2010). Hermite regression analysis of multi-modal count data. Economics Bulletin, 30(4), 2936-2945.

Greene, W. H. (2003). Econometric analysis 5th ed Prentice Hall Upper Saddle River.

Box, G. E. (1970). GM Jenkins Time Series Analysis: Forecasting and Control. San Francisco, Holdan-Day.

Gurland, J. (1954). Hypothesis Testing in Time Series Analysis.

Heilbron, D. (1989). Generalized linear models for altered zero probabilities and overdispersion in count data. Unpublished Technical report, University of California, San Francisco, Department of Epidemiology and Biostatistics.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1), 1-14.

McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. Advances in Applied Probability, 20(4), 822-835.

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of econometrics, 33(3), 341-365.

Skvortsov, R., Connell, P., Dawson, P., & Gailis, R. (2007). MODSIM 2007 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand: Citeseer, 657-662.

White, H. (1982). Maximum likelihood estimation of misspecified models. Econometrica: Journal of the econometric society, 1-25.

Winkelmann, R. (2008). Econometric analysis of count data. Springer Science & Business Media.

Wong, W. H. (1986). Theory of partial likelihood. The Annals of statistics, 88-123.

Zeger, S. L. (1988). A regression model for time series of counts. Biometrika, 75(4), 621-629.