Statistical Diagnostics of Models for Count Data

Zamir Ali[#1], Yu Feng[*2], Ali Choo[#3], Munsir Ali[#4],

[1,2,3,4] *Department of Probability and Mathematical Statistics, School of Science, Nanjing University of Science and Technology, Nanjing 210094, P.R. China*

Abstract

While doing statistical analysis, a problem often resists that there may exist some extremely small or large observation (outliers). To deal with such problem diagnostics of the observations is the best option for the model building process. Mostly analysts use ordinary least square (OLS) method which is utterly failing in the identification of outliers. In this paper, we use the diagnostics method to detect outliers and influential points in models for count data. Gauss-Newton and Likelihood Distance method approach has been treated to detect the outliers in parameter estimation in non-linear regression analysis. We used these techniques to analyze the performance of residual and influence in the non-linear regression model. The results show us detection of single and multiple outlier cases in count data.

Key Words: Poisson *Regression, Outliers, Residuals, Non-linear regression model, Likelihood distance.*

I. INTRODUCTION

Count data is a statistical data type, in which the observation can take only the non-negative integer values i.e. {0, 1, 2……} where these integers coming from counting instead of ranking. Count data is dissimilar from the binary data and ordinal data and their statistical approach are also different. In binary data the observation can take only two values, usually represented by 0 and 1. And ordinal data may also consists of integers where the individual values fall on an unpredictable scale and only the relative ranking is important. While count data include simple counts such as the number of thunderstorm in a calendar year, earthquakes and accidents. The statistical analysis of counts within the framework of discrete parametric distributions for univariate independently identically distributed random variables has a long and rich history (Johnson, Kemp, and Kotz, 2005).

Significant early developments in count models took place in actuarial science, biostatistics, and demography [1]. In recent years these models have also been used extensively in economics, political science, and sociology. The special features of data in their respective fields of application have fueled developments that have enlarged the scope of these models. An important milestone in the development of count data models for regression was the emergence of the "generalized linear models," of which the Poisson regression is a special case [2] [3], first described by Nelder and Wedderburn (1972) and detailed in McCullagh and Nelder (1983, 1989). Building on these contributions, the papers by Gourieroux, Monfort, and Trognon (1984a, b) and the work on longitudinal or panel count data models by Hausman, Hall, and Griliches (1984) have also been very influential in stimulating applied work in the econometric literature [4].

For such type of data Poisson distribution or some modification should be the first choice [6]. The Poisson distribution was derived as a limiting case of the binomial by Poisson (1837). The apparent simplicity of Poisson comes with two restrictive assumptions (Sturman, 1999). First, the variance and mean of the count variable are assumed to be equal. In reality, however, the variance is usually much greater than the mean (i.e., overdispersion) [5] and therefore Poisson models—though widely used to handle count data—may not be well suited to handle some types of count outcomes. Another restrictive assumption of Poisson models is that occurrences of the specified behavior are assumed to be independent of each other [6] [7]. This assumption is also frequently violated. For example, in the case of children's injuries, past injurious experiences are known to be related to future injury risk (Jacques & Finney, 1994).

In this particular work, we aim to represent the residual facts points in models for count data and parameter estimation. In addition, we propose a graphical display and diagnosis for determining the impact of estimation techniques on parameter estimation. Using Gauss Newton and log-likelihood distance technique, some useful examples of parameter estimation and single and multiple outlier's detection are given. The structure of this paper is given, some models and parameter estimation are given in section 2, and the diagnostics method of single and multiple outlier's detection by scatter plot and parameter estimation with some applicable examples is discussed in section 3, and section 4 summarizes the conclusion of this paper.

II. THE MODELS AND THE PARAMETER ESTIMATION

Regression analysis with generalized linear models is based on likelihoods. This section contains the basic inferential tools for parameter estimation, hypothesis testing, and goodness-of-fit tests, whereas more detailed material on model choice and model checking is deferred to Chapter 3.

Given the sample $y_1, y_2, \dots, y_i$ ..., together with the covariates $x_1, x_2, \dots, x_i$ ..., or design vectors $z_1, z_2, \dots, z_i$ ..., a maximum likelihood estimator (MLE) of the unknown parameter vector β in the model $E9y_i|x_i = \mu_i = h(z'\theta)$ is obtained by maximizing the likelihood. To treat the cases of individual data $(i = 1,2, \dots, n)$ and of grouped data $(i = 1,2, \dots, g)$ simultaneously, we omit n or g as the upper limit in summation signs. Thus, sums may run over i from 1 to n or from 1 to g, and weights ω_i have to be set equal to 1 for individual data and equal to n_i for grouped data.

We first assume that the scale parameter ϕ is known. Since ϕ appears as a factor in the likelihood, we may set $\phi = 1$ in this case without loss of generality if we are only interested in a point estimate of β. Note, however, that ϕ (or a consistent estimate) is needed for computing variances of the MLE. Consistent estimation of an unknown ϕ by a method of moments, which is carried out in a subsequent step, is described at the end of this subsection. The parameter ϕ may also be considered as an overdispersion parameter, and it may be treated formally in the same way as a scale parameter. (Note, however, that only the mean $\mu_i = h(z_i'\theta)$ and the variance function are then properly defined, so that one has to start with the expression for the score function $s(\theta)$ instead of the log-likelihood $l(\theta)$ in (1).

We are using l as a generic symbol for log-likelihood. In the case of Poisson responses $(\mu_i = \lambda_i)$ we have

$l_i(\lambda_i) = y_i \log \lambda_i - \lambda_i$

By inserting the mean structure $\mu_i = h(z_i'\theta)$ finally we get

$l_i(\theta) = l_i(h(z_i'\theta))$

According to the log-linear poison model

$$l(\theta) = \sum_i [y_i z_i'\theta - e^{z_i'\theta}] \tag{1}$$

The score function shows how sensitive a likelihood function $[l(\theta; X)]$ is to its parameter θ.

$$s(\theta) = \frac{\partial l}{\partial \theta} = \sum_i s_i(\theta).$$

$$s(\theta) = \frac{\partial l(\theta)}{\partial \theta}$$

$$s(\theta) = \sum_i (y_i - e^{z_i'\theta}) z_i' \tag{2}$$

The observed information matrix is given by,

$$F_{obs}(\theta) = -\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'} = -\frac{\partial s(\theta)}{\partial\theta'}$$

$$= \sum_i \left(z_i'^2 e^{z_i'\theta} \right) \tag{3}$$

The above equation (1) shows the likelihood function and the equation (2) shows the score function and equation shows the observed information matrix with response on the ith individuals.

In case of nonlinear regression $\varepsilon_i \sim N(0, \Sigma_i)$ then the y on the parmeter θ of the observation information matrix $-\ddot{L}(\theta)$, score function $\dot{L}(\theta)$ and the fisher information matrix $I(\theta)$ respectively. Certainly, other iterative approach may be applied to solve the likelihood equations. The Newton-Raphson approach is obtained from Fisher scoring if, the expected information $F(\theta)$ is changed by observed information$F_{obs}(\theta)$. Computational, of nonlinear least square estimate need to apply the iterative numerical set of rules.

$\dot{L}(\hat{\theta}) = 0,$ We may additionally use the Taylor expansion at the point θ_0.

$$\theta^{i+1} = L(\theta^i) + [-\ddot{L}(\theta^i)]^{-1}\dot{L}(\theta^i), \qquad i = 1, 2, \dots \qquad (4)$$

Until$\left|\theta^{i+1} - \theta^i\right| < \delta$, δ is an advance constant value θ^{i+1}Will converge to $\hat{\beta}$ under some regular condition and, the pace of convergence will depend on the choose value of θ_o.

III. STATISTICAL DIAGNOSTICS OF MODELS FOR COUNT DATA

For assessing the fit of, a classical linear regression models diagnostics techniques are in ordinary use [8]. They are designed, to discrepancies among the data and the fitted values as well as discrepancies, among a few data and the rest. Nearly all these tools are based on graphical, presentations of residuals, hat matrix, and case deletion measures [9].

A. A General Approach to Influence

There are several cases of local influence methods [10]. It appeals because it allows the calculation of the effects of individual observations as well as the assessment of the effects of multiple observations. Influential observations are closely related to high leverage observations and outliers [7]. When analyzing high-leverage observations and outliers, you can gain a deeper understanding of the diagnostic measures used to detect influential observations [11].

Measures of the influence of the i^{th} case on the maximum likelihood estimate $\hat{\theta}$ can be based on the sample influence curve $SIC \propto \hat{\theta} - \hat{\theta}_{(i)}$, where $\hat{\theta}_{(i)}$ denotes the ML estimates of θ computed without the i-th case. It may be computationally expensive to implement since $n + 1$ ML estimates are needed, each of which may requires iteration. In such situation, it may be useful to consider quadratic approximation of $L_{(i)}$, the log of likelihood can be obtained after removing the i-th case:

$$L_{(i)}(\theta) \cong L_{(i)}(\hat{\theta}) + (\theta - \hat{\theta})^T \dot{L}_{(i)}(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T \ddot{L}_{(i)}(\hat{\theta})(\theta - \hat{\theta}) \qquad (5)$$

where $L_{(i)}(\hat{\theta})$ is the gradient vector with j-th element $\partial L_{(i)}(\theta)/\partial\theta_j$ evaluated at $\beta = \hat{\theta}$ and $\ddot{L}_{(i)}(\hat{\theta})$ has the (j,k)-th element $\partial^2 L_{(i)}(\theta)/\partial\theta_j\partial\theta_k$, evaluated at $\theta = \hat{\theta}$. If $-\ddot{L}_{(i)}(\hat{\theta})$ is positive definite, the quadratic approximation is maximized at

$$\hat{\theta}^1_{(i)} = \hat{\theta} - (\ddot{\boldsymbol{L}}_{(i)}(\hat{\theta}))^{-1}\dot{\boldsymbol{L}}_{(i)}(\hat{\theta}) \qquad (6)$$

We define a likelihood distance as

$$LD_i = 2[L(\hat{\theta}) - L(\hat{\theta}_{(i)})] \qquad (7)$$

While using one step estimator,

$$LD_i^1 = 2[L(\hat{\theta}) - L(\hat{\theta}^1_{(i)})] \qquad (8)$$

This can be seen easily to be the general class with $t(\theta) = L(\theta)$. The measures LD_i and LD_i^1 can also be explained in terms of the asymptotic confidence region (Cox and Hinkley, 1974)

$\{\theta{:}2[L(\hat{\theta}) - L(\theta)] \le \chi^2(\alpha; q)\}$

Where $\chi^2(\alpha; q)$ is the upper α point of chi-squared distribution with q df, and q is the dimension of θ. Log-likelihood distance can accordingly be calibrated by comparison to the $\boldsymbol{\chi^2}$ $(\boldsymbol{q})$ distribution. If the log-likelihood contours are approximately elliptical then $\boldsymbol{LD_i}$ can be usefully approximated by Taylor expansion of $\boldsymbol{L}(\hat{\theta}_{(i)})$ around $\hat{\theta}$.

$$L(\hat{\theta}_{(i)}) \cong L(\hat{\theta}) + (\hat{\theta}_{(i)} - \hat{\theta})^T \dot{L}(\hat{\theta}) + \frac{1}{2}(\hat{\theta}_{(i)} - \hat{\theta})^T (\ddot{L}(\hat{\theta}))(\hat{\theta}_{(i)} - \hat{\theta})$$

since $\dot{L}(\hat{\theta}) = 0$,

$$LD_i \cong (\hat{\theta}_{(i)} - \hat{\theta})^T (-\ddot{L}(\hat{\theta}))(\hat{\theta}_{(i)} - \hat{\theta}) \tag{9}$$

A different approximation can be obtained by replacing the observed information $-\ddot{L}(\hat{\theta})$ in the above equation be the expected information matrix evaluated at $\hat{\theta}$.

B. Poison Regression and Generalized Linear Model

The standard model for count data is the Poisson regression model, which is a nonlinear regression model. A standard application of Poisson regression is to cross-section data. Typical cross-section data for applied work consist of n independent observations, the i^{th} of which is (y_i, x_i) [12]. The scalar dependent variable y_i is the number of occurrences of the event of interest, and x_i is the vector of linearly independent regressors that are thought to determine y_i. A regression model based on this distribution follows by conditioning the distribution of y_i on a k-dimensional vector of covariates, $x_i' = [x_{1i}, \dots, x_{ki}]$, and parameters β, through a continuous function $\mu(x_i, \theta)$, such that $E[y_i|x_i] = \mu(x_i, \theta)$.

That is, y_i given x_i is Poisson distributed with density

$$f(y_i|x_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \qquad y_i = 0,1,2,\dots \tag{10}$$

In the log-linear version of the model the mean parameter is parameterized as

$\mu_i = \exp(x_i'\theta)$.

The log-likelihood for $\eta = X\theta$ is,

$$L(\eta) = L(X\theta) = \sum_{j=1}^{n}[y_i X_j^T\theta - a_j(X_j^T\theta) + b_j(y_j)] \tag{11}$$

Where,$a_j(z) = n_j \log[1 + \exp(z)]$ and$b_j(z) = \log\binom{n_j}{z}$. The maximum likelihood estimates $\hat{\theta}$ of θ is usually found applying Newton's method. Using different notation defines $\hat{p}_j = \exp(X_j^T\hat{\theta})/[1 + \exp(X_j^T\hat{\theta})]$ and let $\widehat{W}$ be a $n \times n$ diagonal matrix with $j - th$ diagonal$n_j\hat{p}_j(1 - \hat{p}_j)$. Also suppose $\hat{s}$ be a $n - vector$ with $j - th$ element$\hat{s}_j = y_j - n_j\hat{p}_j$. One can show that,

$$\dot{L}_{(i)}(\hat{\eta}) = X_{(i)}^T \hat{s}_{(i)}; \quad \ddot{L}_{(i)}(\hat{\eta}) = -(X_{(i)}^T \widehat{W}_{(i)} X_{(i)}) \tag{12}$$

By using (11), finally we get this equation,

$$\hat{\theta}_{(i)}^1 = \hat{\theta} - \frac{(X^T\widehat{W}X)^{-1}x_i\hat{S}_i}{1 - \hat{v}_{ii}} \tag{13}$$

Where, $\hat{v}_{ii}$ is the $i - th$ diagonal element of $\hat{V} = \widehat{W}^{\frac{1}{2}}X(X^T\widehat{W}X)^{-1}X^T\widehat{W}^{\frac{1}{2}}$. Pregibon [1981] discusses, the accuracy of this one-step approximation and concludes that component wise, the approximation tends to underestimate totally iterated value but that this may be unimportant for identifying, influential cases.

Measures, for the differences $\hat{\theta} - \hat{\theta}_{(i)}\ or\ \hat{\theta} - \hat{\theta}_{(i)}$ can be derived applying elliptical approximation likelihood displacement or alteration in fitted value vectors as discussed in below. Following Pregibon [1981] we will allow for these to characterize, influence for

$$LD_{(ij)}^1(\theta) = (\hat{\theta} - \hat{\theta}_{(ij)}^1)^T (-\ddot{L}(\hat{\theta}) . (\hat{\theta} - \hat{\theta}_{(ij)}^1) \tag{14}$$

We give extension for log likelihood distance LD_i for binomial response data.

One case of likelihood distance

$$LD_{ij} = 2[L(\hat{\theta}) - (\hat{\theta}_{(ij)})] \tag{15}$$

$$LD^1_{(ij)}(\theta) = (\hat{\hat{\theta}} - \hat{\hat{\theta}}^1_{(ij)})^T\ (-\ddot{L}(\hat{\hat{\theta}}).(\hat{\hat{\theta}} - \hat{\hat{\theta}}^1_{(ij)})$$

Putting the value of $\hat{\hat{\theta}} - \hat{\hat{\theta}}^1_{(i)}$ in equation (14) we get,

$$LD^1_{(ij)}(\theta) = \left(\frac{\left[\sum_{ij} {z'_{ij}}^2 e^{z'_{ij}\theta}\right]^{-1} \sum_{ij}(y_{ij} - e^{z'_{ij}\theta}) z'_{ij}}{1-\hat{v}_{ii}} \right)^T \sum_i {z'_i}^2 e^{z'_i\theta} \left(\frac{\left[\sum_{ij} {z'_{ij}}^2 e^{z'_{ij}\theta}\right]^{-1} \sum_{ij}(y_{ij} - e^{z'_{ij}\theta}) z'_{ij}}{1-\hat{v}_{ii}} \right) \quad (16)$$

Multiple case of likelihood distance,

$$LD_i = 2[L(\hat{\theta}) - (\hat{\theta}_{(i)})] \quad (17)$$

$$LD^1_{(i)}(\theta) = (\hat{\hat{\theta}} - \hat{\hat{\theta}}^1_{(i)})^T\ (-\ddot{L}(\hat{\hat{\theta}}).(\hat{\hat{\theta}} - \hat{\hat{\theta}}^1_{(i)})$$

Finally, substituting into (14), this form becomes

$$LD^1_{(i)}(\theta) = \left(\frac{\left[\sum_{ij} {z'_i}^2 e^{z'_i\theta}\right]^{-1} \sum_{ij}(y_i - e^{z'_i\theta}) z'_i}{1-\hat{v}_{ii}} \right)^T \sum_i {z'_i}^2 e^{z'_i\theta} \left(\frac{\left[\sum_i {z'_i}^2 e^{z'_i\theta}\right]^{-1} \sum_{ij}(y_i - e^{z'_i\theta}) z'_i}{1-\hat{v}_{ii}} \right) \quad (18)$$

Example 1:

A company receives shipments of parts from three different suppliers. Each supplier sends the parts in the same sized batch. We need to determine whether one supplier produces fewer defects per batch than the other suppliers. To perform this analysis, we’ll randomly sample batches of parts from all suppliers. The inspectors examine all parts in each batch and record the count of defective parts. We’ll randomly sample 30 batches from each supplier.

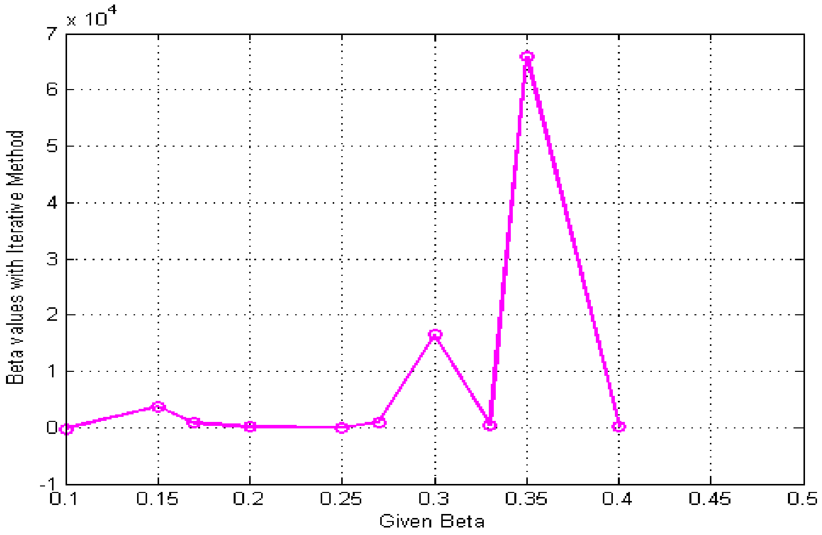
Supplier 1	1, 7, 3, 4, 4, 6, 3, 5, 2, 4, 4, 3, 3, 1, 2, 3, 5, 5, 3, 3, 4, 1, 3, 1, 5 , 6, 4, 6, 2, 4
Supplier 2	1, 10, 4, 2, 0, 4, 8, 5, 8, 4, 4, 7, 6, 9, 4, 4, 5, 5, 4, 6, 12, 8, 2, 6, 4, 8, 7, 5, 5, 4
Supplier 3	3, 5, 1, 7, 6, 10, 9, 4, 3, 7, 5, 2, 2, 6, 4, 4, 3, 9, 1, 5, 3, 8, 4, 3, 6, 1, 7, 4, 7, 11,

Here we consider a bio-exponential characteristic to calculate Gauss newton method,

$$y = \theta_1 \exp(-\theta_2 x) + \theta_3 \exp(-\theta_4 x), \theta_1, \ldots. \theta_4 > 0, \quad (19)$$

We observe the Guass newton method.

$$\hat{\theta}^{(k+1)} = \sum_i [y_i z'_i \theta - e^{z'_i\theta}] + \left[\sum_i {z'_i}^2 e^{z'_i\theta}\right]^{-1} . \sum_i (y_i - e^{z'_i\theta}) z'_i \quad (20)$$



To solve this problem, we use MATLAB. Here, we chose the initial values of $\beta, \beta_o = [0.1,\ 0.15,\ 0.17,\ 0.2,\ 0.25,\ 0.27,\ 0.3,\ 0.33,\ 0.35,\ 0.4]$. After the iteration we get $\hat{\beta} = [-0.019,\ 0.369,\ 0.084,\ 0.009,\ 0,\ 0.096,\ 1.653,\ 0.050,\ 6.597,\ 0.025\ ]$. Which is satisfied under circumstance $\|\beta^{i+1} - \beta^{i}\| < 10^{-4}$ biexponential regression feature to compute Gauss newton.

The result of this estimation of the parameter is primarily based on 10 responses for the third subject are given in the above table.

Example 3:
I took into account table 1, I targeted on the second supplier to detect single case Outlier. Where v_{ii} is the$i-th$ diagonal element $\hat{V} = \widehat{W}^{1/2}X(X^T\widehat{W}X)^{-1}X^T\widehat{W}^{1/2}$

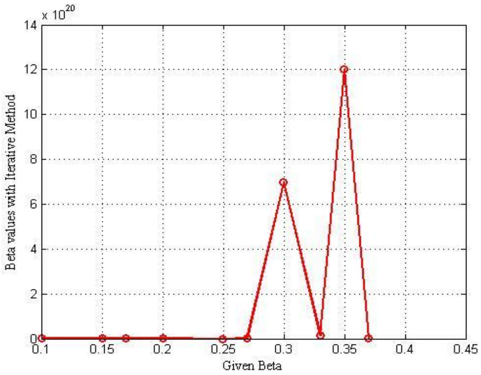


Fig.2 (a): Plot for table 1 for single case outlier under the model (15)

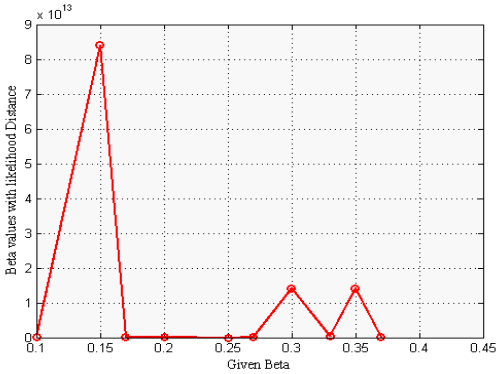


Fig.2 (b): Plot for table 1 for single case outlier under the model (15)

Figure 2 (a) and Figure 2 (b) represent the single and multiple outliers for likelihood distance under model (16) and model (18) respectively.

IV. CONCLUSION

Although a few or a small portion of observations differ from most facts in many respects, the model fitting system may be significantly affected because all observations are forced into the same regression. It is necessary to identify and ignore the observations of the deviation from the shape and the model fit under the affected conditions. As everyone knows, each observation does not play the same role in the regression analysis. As an example, an individual of the regression line can be determined only by some observation, and the maximum value of the information is mostly omitted. Observations of these incredible impact analysis are called influential comments. For regression analysis, detecting outliers can be a critical step. The literature and the number of studies on the impact analysis of nonlinear regression models are not as extensive as linear cases. In this paper, we propose a Gauss-Newton method for parameter estimation and correctly study the rebuttal version of the likelihood distance in single and multiple cases to detect the outlier data points of the count data.

REFERENCES

[1] C.R.D and Presscot, "Approximation significance levels for detecting outlier in linear regression," Technometrics, vol. 23, pp. 59-64, 1981.

[2] Hausman, J.A., B. Hall and Z. Griliches, "Econometric Models for Count Data with an Application to the Patents-R and D Relationship," Econometrica , vol. 52, pp. 909-938, 1984.

[3] Nelder, J.A and R. Wedderburn, "Generalized Linear Models," Journal of the Royal Statistical Society A, vol. 135, pp. 370-384, 1972.

[4] Mullahy and J, "Specification and Testing of some modified count data models," Journal of Econometrics, vol. 33, pp. 341-365, 1986.

[5] Cameron, A.C and P. Trivedi, "Count Data Models for Financial Data" in G.S Maddala and C.R. Rao, eds., Handbook of Statistics, vol. 14, North-Holland: Statistical Methods in Finance, Amsterdam,, 1996.

[6] M. Ali, Z. Ali and A. Choo, "Diagnostics of Single and Multiple outliers on likelihood distance," AJER, vol. 07, pp. 352-357, 2018.

[7] McCullagh, p. and J. Nelder, Generalized Linear Models, edition 1 and 2, Lodon: Chapman and Hall, 1983, 1989.

[8] Cameron, A.C and P. Trivedi, "Economics Models Based on Count Data; Comparison and Application of Some Estimators," Journal of Applied Econometrics, vol. 01, pp. 29-53, 1986.

[9] Ronning, G, R. Jung and i. L. e. al., "Estimation of a First Order Autoregressive Process with Poisson Marginals for Count Data," in Advances in GLIM and Statistical Modelling, New York, Springer Verlag, 1972, pp. 188-194.

[10] C.R.D and J.Amer, "Influence observations in linear regression," Statist Associates, vol. 74, pp. 169-174, 1979.

[11] V. Dujin, M.A.J. and U.B., "Mixture Models for the Analysis of Repeated Count Data," Journal of the Royal Statistical Society C, vol. 44, pp. 473-485, 1995.

[12] Williams and D.A., "Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions," Applied Statistics, vol. 36, pp. 181-191, 1987.

[13] L. Vanegas, L. Rondin and F. Cysneiros, "Diagnostic procedures in Birnbaum Saunders nonliear regression models," Computational Statistics and Data Analysis, vol. 56, pp. 1662-1680, 2012.