American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-7, Issue-3, pp-352-357
www.ajer.org
Research Paper Open Access

Diagnostics of Single and Multiple Outliers on Likelihood Distance

Munsir Ali[1], Zamir Ali[2] Ali Choo[2]and Zeinab Ebrahimpour
School of Science, Nanjing University of Science and Technology,
Nanjing 210094, P.R. China.
Corresponding author: Munsir Ali

ABSTRACT:*A problem often resist in statistical analysis is that there are may exist some extremely small or large observations. Diagnostics of these observations is a crucial tool of statistical analysis. In this paper, we proposed likelihood distance to detect outliers data points for repeated measurement data. The results indicate us single and multiples outliers data cases. The method has been used to explore the presentation of outliers in nonlinear regression models.*
KEYWORD: *Nonlinear least square, influence, diagnostics, log likelihood. Mathematics Subject Classification: 62J20, 62J02, 62G05,62J05,62J99.*

Date of Submission: 17-03-2018 Date of acceptance: 01-04-2018

I. INTRODUCTION

When some or a small group of observations are different in few way from the bulk of the data,the model fitting process may be substantially affected because all observations are forced into the same regression. Observations that make estimates deviate from that structure should be identified and omitted from the model fitting in influential cases. It is well known that all observations don't play equal role in deciding diverse results from a regression analysis. For regression analysis detection of outliers is very important step. The amount of literature and research on influence analysis for nonlinear regression models is not as large as in the linear case.

Statistical analysis with and without these outliers may result totally different results. In practice, the outliers may occur for a variety of causes, including system errors, discrepancy in data transcription and some other cases unrelated to study. The idea has been adopted by many researchers to develop varies techniques for the detection of outliers in different fields under different conditions. For example, the idea is to measure the influence of outlier is based on the difference between two log likelihood function with their maximum likelihood estimates using the whole data set with potential outlier (CHO and Tse)[3]. The result test statistic may be asymptotically compared to a χ^2 distribution(Cook and Weisberg)[10].

A general approach to influence

Influence measures of the $i-th$ case on the ML estimate $\hat{\theta}$ can be based on the sample influence cure $SIC_i \propto \hat{\theta}-\hat{\theta}_i$ represent MLE of θ calculated without the $i-th$ case. While this idea is simple, this maybe computationally costly to apply since $n+1$ MLE are required, where each of that may need iteration. In this situation, it may be useful to consider quadratic approximation of $L_{(i)}$, the log likelihood acquired after removing the $i-th$ case.

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

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

$$\hat{\theta}^1_{(i)}=\hat{\theta}-(\ddot{L}_{(i)}(\hat{\theta}))^{-1}\dot{L}_{(i)}(\hat{\theta}) \quad (2)$$

We mention to $\hat{\theta}^1_{(i)}$ as a one-step approximation to $\hat{\theta}_{(i)}$, it would be obtained same by a single step of Newton's method using $\hat{\theta}$ as beginning values to maximize $L_{(i)}(\theta)$.

If there is not too much difference between $\hat{\theta}_{(i)}$ and $\hat{\theta}$, $L_{(i)}(\theta)$ is locally quadratic, then one-step estimator should be close to the fully iterated value. For that influential cases, $\hat{\theta}-\hat{\theta}_{(i)}$ is "large", the accuracy of the one-step estimator is likely to be lower. But an accurate approximation to $\hat{\theta}_{(i)}$ will not be required as long as $\hat{\theta}-\hat{\theta}_{(i)}$ is sufficiently "large" to draw our notice for further consideration.

In the linear SD problems, elliptical norms of the sample influence cure give a sufficiently rich class metrics for ordering cases of influence. In further general problems, this class can be overly restrictive, particularly if elliptical confidence forms are not approximate.

II. LOG LIKELIHOOD DISTANCE

We define a log likelihood distance LD_i as

$$LD_i=2[L(\hat{\theta})-L(\hat{\theta}_{(i)})] \quad (3)$$

Use the one-step estimator,

$$LD_i=2[L(\hat{\theta})-L(\hat{\theta}^1_{(i)})] \quad (4)$$

It's convenient to be in the general class with $t(\theta)=L(\theta)$, and LD_i is not needed a function of just the sample influence cure (SIC) for θ. The measure LD_i and LD_i^1 may be explained in term of the asymptotic confidence region (Cox and Hinkley, 1974, chapter 9).

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

Where $\chi^2(\alpha;q)$ is the upper, α point of the chi-square distribution with $q\ df$ and q is the dimension of θ. Log likelihood distance maybe calibrate by comparison to the $\chi^2(q)$ distribution.

If the log likelihood form is approximately elliptical, then LD_i can be useful approximated by Taylor expansion of $L(\hat{\theta}_{(i)})$ throughout $\hat{\theta}$,

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

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

$$LD_i\cong(\hat{\theta}_{(i)}-\hat{\theta})^T(-\ddot{L}(\hat{\theta}))(\hat{\theta}_{(i)}-\hat{\theta}) \quad (5)$$

Different approximation can be acquired by replacing observed information $-\ddot{L}(\hat{\theta})$ in (1) by the anticipated information matrix, assessed at $\hat{\theta}$.

III. NONLINEAR LEAST SQUARE

The Model of nonlinear regression is given by

$$yi = f(x_j, \theta) + \sigma\varepsilon_j, \quad j = 1, 2, ..., n \quad (6)$$

Where $f(x_j, \theta)$ is a scalar value function which is nonlinear in the q-vector of unknown parameters θ, and ε_j is independent and identical distributed $N(0,1)$. The maximum likelihood estimate $\hat{\theta}$ of θ can be get by minimizing the residual sum of square

$$G(\theta) = \sum_{j=1}^{n} (y_i - f(x_j, \theta))^2 \quad (7)$$

The problem of resolving $\hat{\theta}$ can be deal as a special case of the general unconstrained maximization problem. Even though special technique that use the fact $G(\theta)$ is quadratic is quite appreciate (Kennedy and Gentle 1980, section 10.3).

The problem of evaluating influence in nonlinear least squares can be approached to using the general method outlined (5). In specially, one-step estimators $\hat{\theta}_i^1$ of the vectors $\hat{\theta}_{(i)}$ that minimize the objective functions

$$G_{(i)}(\theta) = \sum_{j \neq 1} (y_j - f(x_{j,}\theta))^2, \qquad i = 1, 2...., n \quad (8)$$

It can be found by application of the result given by (2).However; particularly interesting result can be obtained if we allow further approximation. We assume that, in a neighborhood about $\hat{\theta}$, $f(x_{j,}\theta)$ is approximately linear,

$$f(x_j, \theta) \cong f(x_j, \hat{\theta}) + z_j^T(\theta - \hat{\theta}) \quad (9)$$

Where $\hat{z}_j^T$ is the $j-th$ row of the $n \times q$ Jacobian matrix $\hat{Z}$,

$$\hat{z}_j^T = \left[\frac{\partial f(x_j, \theta)}{\partial \theta_1}, ..., \frac{\partial f(x_{j,}, \theta)}{\partial \theta_q} \right]_{\theta = \hat{\theta}} \quad (10)$$

If the approximation (9) is substituted into $G_i(\theta)$ defined by (8) the resulting object function is minimized at

$$\hat{\theta}_{(i)}^1 = \hat{\theta} + (\hat{Z}_{(i)}^T \hat{Z}_{(i)})^{-1} \hat{Z}_{(i)}^T e_{(i)}$$

Where e is the n-vector with elements $e_j = y_j f(x_j, \hat{\theta})$. This sequence has close similarity to obtain that by using a single step of the Gauss newton method (Kennedy and Gentle, 1980). This equation simplified that

$\hat{Z}^T e = 0$,explaining $\hat{v}_{ii} = \hat{z}_i^T (\hat{Z}^T \hat{Z})^{-1} \hat{z}_i$, we get

$$\hat{\theta}_i^1 = \hat{\theta} - \frac{(\hat{Z}^T \hat{Z})^{-1} \hat{z}_i e_i}{1 - \hat{v}_{ii}} \quad (11)$$

When this specific algorithm is used to make the one-step estimators, the nonlinear least squares problem is essentially replace by linear least square. Most of the diagnostic and residual analysis for linear least square can be anticipated to apply at least approximately in nonlinear least squares.

In particular, approximate residual (4), where $\hat{\sigma}^2 = G(\hat{\theta})/n$, Elliptical norm of the sample influence cure is

$$D_i(\hat{Z}^T\hat{Z}, q\hat{\sigma}^2) = (\hat{\theta}-\hat{\theta}_i)^T(\hat{Z}^T\hat{Z})(\hat{\theta}-\hat{\theta}_i)/q\hat{\sigma}^2$$

When $\hat{\theta}_i$ is replaced by one-step approximation $\hat{\theta}_i^1$, this norm becomes

$$D_i^1(\hat{Z}^T\hat{Z}, q\hat{\sigma}^2) = \frac{\hat{r}_i^2}{q}\frac{\hat{v}_{ii}}{1-\hat{v}_{ii}} \quad (12)$$

We give extension for log likelihood distance LD_i, to nonlinear regression for repeated measurement data.
One case of likelihood distance

$$LD_{ij} = 2[L(\theta) - L(\theta_{(ij)}] \quad (13)$$

$$LD_{ij} = [\hat{\theta}_{(ij)} - \hat{\theta}]^T[-\ddot{L}(\hat{\theta})(\hat{\theta}_{(ij)} - \hat{\theta})$$

Substituting into (5) leads to the form

$$LD_{ij} = \left(\frac{(U_{ij}^T U_{ij})^{-1}u_{ij}e_{ij}}{1-v_{ii}}\right)^T \left(U^T\Sigma^{-1}U\right)\left(\frac{(U_{ij}^T U_{ij})^{-1}u_{ij}e_{ij}}{1-v_{ii}}\right) \quad (14)$$

Multiple case of likelihood distance

$$LD_i = 2[L(\theta) - L(\theta_{(i)}] \quad (15)$$

$$LD_i = [\hat{\theta}_{(i)} - \hat{\theta}]^T[-\ddot{L}(\hat{\theta})(\hat{\theta}_{(i)} - \hat{\theta})$$

Finally, substituting into (5), this form becomes

$$LD_i = \left(\frac{(U_i^T U_i)^{-1}u_i e_i}{1-v_{ii}}\right)^T \left(U^T\Sigma^{-1}U\right)\left(\frac{(U_i^T U_i)^{-1}u_i e_i}{1-v_{ii}}\right) \quad (16)$$

We observe the data in table I that taken from a study reported by Kwan et al. (1876) of the pharmacokinetics of indomethacin following bolus intravenous injection of the same dose in six human volunteers, for each subject plasma concentrations of indomethacin were measured at 11 times intervals regarding from 15 to 8 hours post-injection[11].

TABLE. I: PLASMA CONCENTRATIONS $(\mu g/ml)$ FOLLOWING INTRAVENOUS INJECTION OF INDOMETHACIN FOR SIX HUMAN

IV. SUBJECTS

Time (hrs.)	1	2	3	4	5	6
0.25	1.50	2.03	2.72	1.85	2.05	2.31
0.50	0.94	1.63	1.49	1.39	1.04	1.44
0.75	0.78	0.71	1.16	1.02	0.81	1.03
1.00	0.48	0.70	0.80	0.89	0.39	0.84
1.25	0.38	0.64	0.80	0.59	0.30	0.64
2.00	0.19	0.36	0.39	0.40	0.23	0.42
3.00	0.12	0.32	0.22	0.16	0.13	0.24
4.00	0.11	0.20	0.12	0.11	0.11	0.17
5.00	0.08	0.25	0.11	0.10	0.08	0.13
6.00	0.07	0.12	0.08	0.07	0.10	0.10
8.00	0.05	0.08	0.08	0.07	0.06	0.09

We propose two examples to calculate likelihood distance

$$y = \beta_1\exp(-\beta_2 x) + \beta_3\exp(-\beta_4 x),\ \beta_1,.....,\beta_4 > 0, \quad (17)$$

We consider Table I, to calculate singular and multiple outlier cases where $U = \frac{\partial f(\beta)}{\partial \beta}$, $\sum = In$, e is unobservable error $y - f(\beta)$ and $v = u_i^T (U^T U)^{-1} u_i$

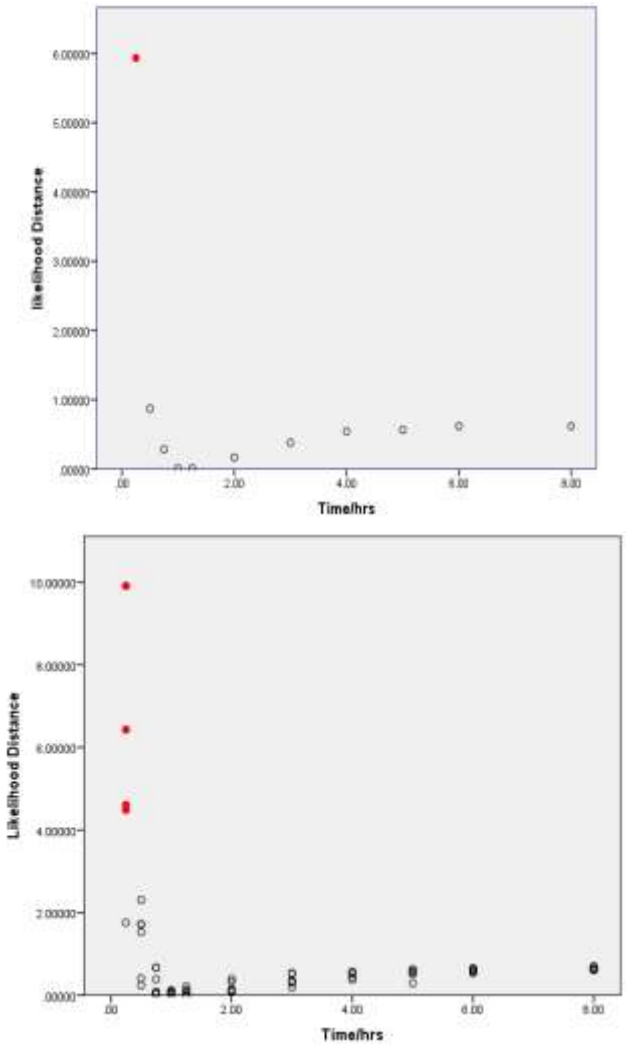

Fig: a **Fig: b**

Fig.a.Scatter plot oflikelihood Distance for the table I (fifth individual) under model (14).which show us one case outlier and fig: b indicates multiple outliers cases under the model (16) for complete data set of table I.

V. CONCLUSION

It is well understood that all observations of a data set don't have equal importance in the result of regression analysis. For example, the character of the regression line maybe determine by only a few observations, while most of the data is somewhat ignored. Such observations that highly influence the results of the analysis are called influential observations. It is important, for many causes, to be able to detect influential observations. In this paper, we calculated likelihood distance for one and multiple cases outliers.

REFERENCES

[1]. Altman, N. &Krzywinski, M.2016. "Analyzing outliers influential or nuisance". .*Nature methods*, 13,281-282.
[2]. Law, M. & Jackson, D. 2017. "Residual plot for linear models with censored outcome data: A refined method for visualizing residual uncertainty". *Communication in statistics simulation and computation*, 46, 3159-3171.
[3]. Chow SC, Tse SK: 1990. "outliers detection in bioavailability / bioequivalence studies." *Statis Med*, 9, 549-558.
[4]. Cook R.D. 1979. "Influence observations in linear regression", *J. Amer. statist.Assoc*, 74,169-74.
[5]. Cook R.D, and presscot. 1981. "Approximation significance levels for detecting outlier in linear regression", *Technometrics*, 23, 59-64.
[6]. Ellenberg, J.H. 1976. "Testing of a single outlier from a general regression model", *Biometrics*, 32, 637-45.

[7]. Vonesh, E.F. 1992. "Nonlinear models for the analysis of longitudinal data",*Statistics in medicine*, 11, 1929-1954.
[8]. Solomon P.J. and cox D.R. 1992."Nonlinear components for variance models", *Biometrikka*, 79, 1-11.
[9]. Cook R.D. 1979. "Influence observation in liner regression", *J. Am. statist.assoc*, 74, 169-174.
[10]. Cook R.D , Weisberg R. 1982. "Residuals and influence and regression", *Chapman and Hall*,
[11]. Anscombe, F.J. 1961. "Examination of residuals, Proc.fouth Berkeley symp" 1, 1-36.
[12]. MARIE DAIDIAN and DAVID M.GILTINAN .march.1995.. *"Nonlinear models for repeated measurement data".*

Munsir Ali " Diagnostics of Single and Multiple Outliers on Likelihood Distance." American Journal of Engineering Research (AJER), vol. 7, no. 3, 2018, pp.352-357.