SEBHA UNIVERSITY JOURNAL OF PURE & APPLIED SCIENCES VOL.21 NO. 4 2022
DOI: 10.51984/JOPAS.V21I4.2273

مجلة جامعة سبها للعلوم البحتة والتطبيقية
Sebha University Journal of Pure & Applied Sciences

Journal homepage: www.sebhau.edu.ly/journal/index.php/jopas

Application of outliers detection techniques in nonlinear regression

Abdelgadir Khalifa Alsalem[a], Alsaidi M. Altaher[b]

[a]Department of Mathematical, Faculty of Education Ubari / University of Sebha, Libya
[b]Department of Statistics, Faculty of Science/University of Sebha, Libya

Keywords:
Outliers
nonlinear regression
linear approximation
Design matrix

ABSTRACT

outlier's detection is very essential issue due to their responsibility for producing interpretative problem in linear as well as in nonlinear regression analysis. Much work has been accomplished on the identification of outlier in linear regression, but not as in nonlinear regression. This paper aims to evaluate several outlier detection techniques for nonlinear regression based on Studentized Residuals, Hadi Potential, Cook Distance, Difference in Fits and Atkinson's Distance). The main idea is to use the linear approximation of a nonlinear model and consider the gradient as the design matrix. Subsequently, the detection techniques are formulated. A real life data showed that among the five measures, only Difference in Fits and Cook Distance consistently capable of identifying the correct outlier.

تطبيق تقنيات الكشف عن القيم المتطرفة في الانحدار اللاخطي

*عبدالقادر خليفة السالم[1] و السعيدي ا لمهدي ا لطاهر[2]

[1]قسم الرياضيات, كلية التربية أوباري, جامعة سبها, ليبيا
[2]قسم الإحصاء, كلية العلوم, جامعة سبها, ليبيا

الكلمات المفتاحية:
القيم المتطرفة
الانحدار اللاخطي
التقريب الخطي
مصفوفة التصميم

الملخص

يعد اكتشاف القيم المتطرفة ضرورياً جداً نظراً لمسؤوليتها عن إنتاج مشكلة تفسيرية في تحليل الانحدار الخطي وكذلك في تحليل الانحدار غير الخطي. تم إنجاز الكثير من العمل على تحديد القيم المتطرفة في الانحدار الخطي، ولكن ليس كما في الانحدار غير الخطي. تهدف هذه الورقة إلى عرض العديد من التقنيات للكشف عن القيم المتطرفة للانحدار اللاخطي اعتمادا على عدة مقاييس كالأخطاء المعيارية. الفكرة الرئيسية هي استخدام التقريب الخطي لنموذج غير خطي واعتبار التدرج كمصفوفة تصميم ثم صياغة تقنيات الكشف. تظهر الدراسة أنه من بين المقاييس الخمسة، فقط مقياس Difference in Fits ومقياسAtkinson's Distance القادر باستمرار على تحديد القيم المتطرفة الصحيحة.

1. Introduction

Nonlinear regression is one of the most popular and widely used models in analyzing the effect of explanatory variables on a response variable when the underling regression function is nonlinear. It has many applications in scientific research such as in dose response studies conducted in agricultural sciences, toxicology and other biological sciences; see [1] and [2]. With the presence of outliers in the data, the ordinary least squares LS method provides misleading values for the parameters of the nonlinear regression, and predictions may no longer be reliable, see [3]. Outliers are those observations that deviate markedly from other members of the observations or data points which are unusually large or small from the majority of the observations. They are also called the abnormal data. Outliers can arise due to measurement or recording error, natural variation of the underlying distribution, or a sudden alteration in the operating system. An exact definition of an outlier often depends on hidden assumptions regarding the data structure and the applied detection method. Yet, some definitions are regarded general enough to cope with various types of data and methods. [4] defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that is was generated by a different mechanism. [5] defines an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data. [6] indicate that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. [7] stated that an outlier is a value or an observation which is positioned outside the general mold of a distribution. [8] designate outliers to be the values or observations that deviate from the sketch that is laid down by the best part of the data. An outlier is a value or observation

Corresponding author:
E-mail addresses: ab.alsalem1@sebhau.edu.ly, (A. M. Altaher) als.altaher@sebhau.edu.ly
Article History: Received 15 June 2022 - Received in revised form 09 August 2022 - Accepted 03 October 2022

which is located at anomalous distance from the rest of the values or observations in a random sample taken from a population. [9] defined the outlier to be a value or observation which is far away from the bulk of the data. Many statistics practitioners have been using residuals for the identification of outliers. The use of residuals resulting from the ordinary least squares (OLS) estimates will give a misleading conclusion because the residuals are functions of leverages and true errors. There are many measure to identification of outliers in linear regression [see for example, Hadi [10], Habshah et al. [11], Cook and Weisberg [12], Belsley et al. [13], Anscombe and Tukey [14] and the discussion on the properties of Atkinson's distances in [15] and [16]). However, not much work has been explored in the formulation of the outlier's identification method in nonlinear regression. Cook and Weisberg [12] and Fox et al.[17] introduced a measure for the identification of outliers in nonlinear model, which is based on the OLS method.

The remainder of this paper is organized as follows: Section 2 gives a brief review to the Studentized Residuals, Hadi Potential, Elliptic Norm (Cook Distance), Difference in Fits and Atkinson's Distance. In section 3 real data applications are present. Conclusions are drawn in section 4.

1. Methods and Models

This section is devoted to introduce the theoretical descriptions of Studentized Residuals, Hadi Potential, Elliptic Norm (Cook Distance), Difference in Fits and Atkinson's Distance. We shall first introduce the concept of the hat matrix in nonlinear Regression. In nonlinear regression, the linear approximation of function model is used, and replaces the explanatory matrix in linear regression, by the gradient of the function model. The linear approximation form can be derived by expanding the function model

$$y \equiv \eta(\theta) + \varepsilon \quad (1)$$

Where $y = [y_1, y_2, \dots, y_n]^T$ is $n \times 1$ response vector, $\eta(\theta) = [f(x_1;\theta), \dots, f(x_n;\theta)]$ is $n \times 1$ vector of function models $f(x_i;\theta)'s$, $x_i = [x_{i1}, x_{i2}, \dots, x_{ik}]^T$ is k dimensional predictor (design) vector, $\varepsilon = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n]^T$ is $n \times 1$ vector of iid residuals, around the true value θ^*

$$\eta(\theta) \cong \eta(\theta^*) + \dot{V}(\theta - \theta^*) \quad (2)$$

Where $\dot{V} = \frac{\partial f(x;\theta)}{\partial \theta}$ is n×p gradient matrix computed at estimated point. Based on this approximation, an equivalent measure for equation of the Hat matrix which is called as tangent plane leverage matrix is given by

$$H = \dot{V}(\dot{V}^T\dot{V})^{-1}\dot{V}^T \quad (3)$$

This leverage matrix in nonlinear plays a similar role as the Hat Matrix in linear form. Linear regression uses the Hat matrix as a beginning idea of influence detection tool, and creates several statistical measures for outlier detection. Next, we shall briefly discuss some outlier's detection measures.

1.1 Studentized Residuals

This measure (hereafter refer as t_i) is used for identifying outliers. Suppose h_{ii} is the diagonal of leverage matrix H based on gradient in equation (3), the studentized residuals are defined by

$$t_i = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \quad (4)$$

where $\hat{\sigma}_{(-i)}$ is the estimated standard deviation in the absence of the *i'th* observation. The residual, denoted as: $r_i = y_i - f(x_i;\hat{\theta})$ is obtained from the NLS, M and MM estimates. The *i'th* observation is considered as an outlier if $|t_i| > 2.5$ or 3 [14] and [18].

1.2 Hadi Potential

Hadi [10] proposed Hadi's potential denoted as p_{ii} to detect high leverage points or large residuals :

$$p_{ii} = \frac{h_{ii}}{1-h_{ii}} \quad (5)$$

Where h_{ii} is the i'th diagonal element of H. Hadi [10] proposed a cut-off point for p_{ii} as :

$$Median(p_{ii}) + c.MAD(p_{ii}) \quad (6)$$

Where MAD represents the Mean Absolute Deviance defined by:

$$MAD(p_{ii}) = Median|p_{ii} - Median(p_{ii})|/0.6745 \quad (7)$$

c is an appropriately chosen constant such as 2 or 3.

1.3 Elliptic Norm (Cook Distance)

The Cook Distance (hereafter is refereed as CD) which is defined by Cook and Weisberg [12], is used to assess the influential observations. An observation is influence if the value of CD is greater than one. They defined CD as

$$CD_i(\dot{V}^T\dot{V}, p\hat{\sigma}^2) = \frac{(\theta-\hat{\theta}_{(-1)})^T(\dot{V}^T\dot{V})(\theta-\hat{\theta}_{(-1)})}{p\hat{\sigma}^2} \quad (8)$$

where $\hat{\theta}_{(-1)}$ is the parameter estimates when the i'th observation is removed. When $\hat{\theta}_{(-1)}$ is replaced by the linear approximation, this norm changes to

$$CD_i(\dot{V}^T\dot{V}, p\hat{\sigma}^2) = \frac{t_i^2}{p}\frac{h_{ii}}{1-h_{ii}} \quad (9)$$

where t_i and p is the studentized residual and the number of parameters in the model, respectively. With the cut of point equal to 1, that is the expectation of 50% confidence ellipsoid of parameter estimates.

1.4 Difference in Fits

Difference in Fits, denoted by DFFITS, is another diagnostics measure used in measuring the influence, defined by Belsley et al.[13]. For the i'th observation, DFFITS is defined as:

$$DFFITS_i = \left(\sqrt{\frac{h_{ii}}{1-h_{ii}}}\right)|d_i| \quad (10)$$

where d_i is the deleted studentized residual. They considered observation is an outlier when DFFITS exceeds the cut of point equals to $2\sqrt{p/n}$.

1.5 Atkinson's Distance

Atkinson distance (hereafter refer as C_i) for observation i was developed by Atkinson [15] and it is used to detect the influential observation. Atkinson defined the Atkinson's distance as follows:

$$C_i = \left(\sqrt{\frac{n-p}{p}\frac{h_{ii}}{1-h_{ii}}}\right)|d_i|, \qquad i = 1, \dots, n. \quad (11)$$

where d_i is the deleted studentized residuals. He suggested a cut-off value equals to 2.

2. *Results and Discussion*

In this section we summarize and discus the numerical results from real life data, a set of real data which is referred to as the Drug concentration data and Tumor metastasis data. Kenakin [19] used a set of responses to the concentration of an agonist in a functional assay. In drug concentration data, observation 5 has an outlier in the response direction. The model associated with this data is Michaelis-Menten model [20] and [21], expresses the reaction velocity as a function of concentration of substrate as

$$y_i = \frac{\beta_0 x_i}{\beta_1 + x_i} + \varepsilon_i$$

Where response variable y_i is velocity and predictor variable x_i is substrate; the parameter is β_0 the maximum reaction velocity and β_1 denotes concentration of substrate. However, different true parameters are possible as long as convergence occurs in optimization process. In Tumor metastasis data, it is taken from Tabatabai et al.[22] which was primarily collected by Tubiana & Koscielny [23]. The given data is clean as there is no outlier present. Then outliers are inserted in X direction, both XY direction and in Y direction. In X space shifting the value in observation 12 from 90 to 2. In the Y space changing the Y value in observation 6 from 0.55 to 1. And in both XY space transforming Y value in observation 7 from 0.56 to 3 and X value in observation 12 from 90 to 2. The model associated with this data is Exponential Model, it is a two parameter model given by the following relationship

$$y_i = \beta_0\, \beta_1^{\,x_i} + \varepsilon_i$$

Where β_0, β_1 are parameters, x is independent predictor, y response predictor, ε is random variable.

Having examined Tables 1 to 4 carefully, we have noted many important features: In Table 1 suggest that most of the diagnostic measures that are based on OLS fail to identify the observation 5 as outlier. The DFFITS and C_i identify correctly observation 5 as outlier. The results of tables 2,3 and 4 can be observed the best performance has been achieved by the DFFITS and C_i. An important feature to note is that the Cook Distance $\boldsymbol{CD_i}$ and Hadi Potential $\boldsymbol{P_{ii}}$ is the worst one and fail to identify any outlier.

Table 1: Five outlier measures based on OLS for drug concentration data.

Cut of point Index	t_i 3.0	CD_i 1.0	P_{ii} (0.896-1.118)	$DFFITS_i$ 1.07	C_i 2.000
1	-0.303	0.030	0.020	-2.146	0.480
2	-0.587	0.147	0.126	-1.651	0.928
3	-0.666	0.341	0.523	-0.921	1.053
4	0.190	0.090	0.451	0.283	0.300
5	2.072	0.804	0.301	3.775	3.276
6	-0.785	0.394	0.503	-1.107	1.241
7	-1.000	0.605	0.730	-1.171	1.582

Table 2: Five outlier measures based on OLS for tumor metastasis data with outlier in the X direction

Cut of point Index	t_i 3.0	CD_i 1.0	P_{ii} (0.249-0.303)	$DFFITS_i$ 0.816	C_i 2.000
1	-0.670	0.191	0.162	-1.663	1.498
2	-0.601	0.161	0.143	-1.589	1.344
3	-0.504	0.135	0.143	-1.333	1.128
4	-0.309	0.075	0.117	-0.902	0.690
5	-0.385	0.088	0.105	-1.188	0.860
6	-0.276	0.060	0.094	-0.902	0.618
7	3.030	0.655	0.094	9.902	6.774
8	-0.221	0.052	0.109	-0.670	0.494
9	-0.153	0.045	0.170	-0.371	0.342
10	-0.196	0.079	0.322	-0.346	0.439
11	-0.390	0.232	0.710	-0.463	0.872
12	0.505	0.162	0.205	1.116	1.129

Table 3: Five outlier measures based on OLS for tumor metastasis data with outlier in the Y direction.

Cut of point Index	t_i 3.0	CD_i 1.0	P_{ii} (0.288-0.351)	$DFFITS_i$ 0.816	C_i 2.000
1	-1.304	0.387	0.176	-3.106	2.917
2	-1.068	0.302	0.160	-2.669	2.389
3	-0.726	0.206	0.160	-1.814	1.623
4	-0.044	0.011	0.135	-0.121	0.099
5	-0.320	0.078	0.120	-0.924	0.716
6	0.047	0.011	0.102	0.146	0.104
7	0.064	0.014	0.101	0.202	0.144
8	0.221	0.051	0.108	0.673	0.494
9	0.445	0.130	0.171	1.077	0.996
10	0.297	0.128	0.375	0.485	0.664
11	-0.394	0.295	1.123	-0.372	0.882
12	2.835	0.910	0.206	6.247	6.340

Table 4: Five outlier measures based on OLS for tumor metastasis data with outlier in the XY direction.

Cut of point Index	t_i 3.0	CD_i 1.0	P_{ii} (0.209-0.246)	$DFFITS_i$ 0.816	C_i 2.000
1	-1.326	0.358	0.146	-3.473	2.965
2	-1.072	0.284	0.141	-2.858	2.396
3	-0.683	0.181	0.141	-1.821	1.527
4	0.077	0.020	0.130	0.214	0.173
5	-0.250	0.062	0.123	-0.714	0.559
6	2.613	0.611	0.109	7.898	5.842
7	0.164	0.038	0.108	0.498	0.366
8	0.306	0.069	0.100	0.967	0.685
9	0.505	0.120	0.112	1.507	1.130
10	0.261	0.076	0.170	0.633	0.584
11	-0.498	0.206	0.343	-0.850	1.114
12	-0.699	0.475	0.925	-0.727	1.563

3. *Conclusion*

In this paper, a linear approximation of a nonlinear model is formulated and subsequently leverage matrix based on the gradient is formed. The outlier measures for nonlinear regression are then formulated by incorporating the leverage matrix and the commonly used detection measures, namely Studentized Residuals t_i, Hadi Potential p_{ii}, Cook Distance CD_i, Difference in Fits *DFFITS* and Atkinson's Distance c_i. The results of the study clearly reveal that the proposed measures Difference in Fits *DFFITS* and Atkinson's Distance c_i are the best outlier measures in nonlinear regression because they consistently can identify outliers correctly in different outliers scenarios.

References.

[1]- D. M. Bates and D. G. Watts, *Nonlinear Regression Analysis and Its Applications*. Wiley, 2007.

[2]- G. A. F. Seber and C. J. Wild, *Nonlinear Regression*. Wiley, 2003.

[3]- P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in statistics*, Springer, 1992, pp. 492–518.

[4]- D. M. Hawkins, *Identification of outliers*, vol. 11. Springer, 1980.

[5]- R. A. Johnson and D. W. Wichern, "Applied multivariate statistical analysis. Prentice Hall, Englewood Cliffs, NJ.," *Appl. Multivar. Stat. Anal. Prentice-Hall, Englewood Cliffs, NJ.*, 1992.

[6]- V. Barnett and T. Lewis, *Outliers in Statistical Data*. Wiley, 1994.

[7]- D. S. Moore and G. P. McCabe, *Introduction to the Practice of Statistics*. WH Freeman, 1999.

[8]- F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics: the approach based on influence functions*, vol. 196. John Wiley & Sons, 2011.

[9]- R. G. Staudte and S. J. Sheather, *Robust estimation and testing*, vol. 918. John Wiley & Sons, 2011.

[10]- A. S. Hadi, "A new measure of overall potential influence in linear regression," *Comput. Stat. Data Anal.*, vol. 14, no. 1, pp. 1–27, 1992.

[11]- M. Habshah, M. R. Norazan, and A. H. M. Rahmatullah Imon, "The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression," *J. Appl. Stat.*, vol. 36, no. 5, pp. 507–520, 2009.

[12]- R. D. Cook and S. Weisberg, *Residuals and influence in regression*. New York: Chapman and Hall, 1982.

[13]- D. A. Belsley, E. Kuh, and R. E. Welsch, "Regression Diagnostics John Wiley & Sons," *New York*, 1980.

[14]- F. J. Anscombe and J. W. Tukey, "The examination and analysis of residuals," *Technometrics*, vol. 5, no. 2, pp. 141–160, 1963.

[15]- A. C. Atkinson, "Regression diagnostics, transformations and constructed variables," *J. R. Stat. Soc. Ser. B*, pp. 1–36, 1982.

[16]- A. C. Atkinson, "Masking unmasked," *Biometrika*, vol. 73, no. 3, pp. 533–541, 1986.

[17]- T. Fox, D. Hinkley, and K. Larntz, "Jackknifing in nonlinear regression," *Technometrics*, vol. 22, no. 1, pp. 29–33, 1980.

[18]- K. S. Srikantan, "Testing for the single outlier in a regression model," *Sankhyā Indian J. Stat. Ser. A*, pp. 251–260, 1961.

[19]- T. Kenakin, "A pharmacology primer: theory, application and methods," 2009.

[20]- A. J. Stromberg and D. Ruppert, "Breakdown in nonlinear regression," *J. Am. Stat. Assoc.*, vol. 87, no. 420, pp. 991–997, 1992.

[21]- D. E. Herwindiati, M. A. Djauhari, and M. Mashuri, "Robust multivariate outlier labeling," *Commun. Stat. Comput.*, vol. 36, no. 6, pp. 1287–1294, 2007.

[22]- M. A. Tabatabai, J. J. Kengwoung-Keumo, W. M. Eby, S. Bae, U. Manne, M. Fouad, and K. P. Singh, "A New Robust Method for Nonlinear Regression," *J. Biom. Biostat.*, vol. 5, no. 5, p. 211, 2014.

[23]- M. Tubiana and S. Koscielny, “The natural history of breast cancer: implications for a screening strategy,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 19, no. 5, pp. 1117–1120, 1990.