RESIDUALS AND INFLUENCE IN NONLINEAR REGRESSION FOR REPEATED MEASUREMENT DATA

Munsir Ali, Yu Feng, Ali choo, Zamir Ali

Nanjing University of Science and Technology, P.R. China

ABSTRACT

All observations don't have equal significance in regression analysis. Diagnostics of observations is an important aspect of model building. In this paper, we use diagnostics method to detect residuals and influential points in nonlinear regression for repeated measurement data. Cook distance and Gauss newton method have been proposed to identify the outliers in nonlinear regression analysis and parameter estimation. Most of these techniques based on graphical representations of residuals, hat matrix and case deletion measures. The results show us detection of single and multiple outliers cases in repeated measurement data. We use these techniques to explore performance of residuals and influence in nonlinear regression model.

KEY WORDS:

Hat matrix, Cook distance, Residuals, Nonlinear regression models. Mathematics Subject Classification: 62J20,62J02, 62G05,62J05,62J99.

1. INTRODUCTION

Data containing of repeated measurements hold on each of number of individuals appear frequently in biomedical and biological implementations. This kind of modeling data generally implies characterization of the relationship among the measured response of y, measurement factor, or covariate x [11]. In many implementations, the relationship between y and x is nonlinear in unknown parameters of attention.

The expression of repeated measurement on an individual requires definite care in marking the random variation in the data. It is important to recognize random variation among measurements within a given individual and random variation among the individuals. Inferential methods assist these different variance components in the framework of a proper hierarchical statistical model. When the relationship between x and y in the unknown parameters is linear, the framework is that of the classical linear mixed effects model [10]. In this case, Bayesian inferential method is provided satisfactory hierarchical linear model [14]. There is a substantial literature about hierarchical linear model, McCulloch, Casella, and Searle (1992). Linear modeling methods for repeated measurement data are quite advanced and developed, and well recorded in statistical literature, Crowder and Hand (1990), Lindsey(1993), and Diggle, Liang and Zenger(1994).

In this particular work, we aim to indicate residuals data points in nonlinear regression for repeated measurement data and parameter estimation. We use Cook.distacne and Gauss newton method, and we also explore some useful examples for parameter estimation and Outliers detection. The organization of this paper is given as; in section 2, we give some models and parameter estimation; section 3 deals with the diagnostics methods in case of single and multiple Outliers detection by

DOI: 10.5121/oraj.2017.4401

scatterplots and parameter estimation with some applicable examples while section 4 concludes the paper.

2. THE MODEL AND THE PARAMETER ESTIMATION

We introduce hierarchal nonlinear model that forms the fundamental inferential methods and discuss the available techniques for the analysis of repeated measurement data. In the linear case, intra and inter individual variation can assist within the two stages model. The first stage characterizes by a nonlinear regression model with a model for individual covariance structure, and inters individual variability represent in the second stage through individual specific regression parameters.

Let y_{ij} denote the jth response, $j = 1,..,n_i$ for ith individual, $i = 1,..,m,$ taken at a set of conditions sum up by the vector of covariates x_{ij}, so that a sum of $N = \Sigma_{i=1}^{m} ni$ response have been observed. The vector x_{ij} includes variables.

Suppose that, for individual i, the jth response obey the model.

$$y_{ij} = f(x_{ij}, \beta_i) + e_{ij} \qquad (1)$$

Where e_{ij} is a random errorexpression considering unreliability in the response, given the ith individual, with $E(e_{ij} \mid \beta_i) = 0$ Getting the response and errors for the ith individual into the $(n_i \times 1)$ vectors $y_i = [y_{i1},....., y_{ini}]'$, and $e_i = [e_{i1},...., e_{ini}]'$, respectively, and interpreting the $(n_i \times 1)$ vector.

$$y_i = f(x_i, \beta_i) + e_i, \quad (2)$$

where $E(e_i \mid \beta_i) = 0$.

The model given in (1) and (2) describes the organizing and random variation association with measurement on the ith individual.

If for nonlinear regression $\varepsilon_i \sim N(0, \Sigma_i)$, then y on the parameter β of score function $\dot{L}(\beta)$ observation information matrix $-\ddot{L}(\beta)$ and fisher information matrix $I(\beta)$ respectively.

Computational of nonlinear least square estimates need to use the iterative numerical algorithm. $\dot{L}(\hat{\theta}) = 0$, we may use Taylor expansion at point θ_0

$$\dot{L}(\hat{\theta}) = \dot{L}(\theta_0) + \ddot{L}(\hat{\theta})(\hat{\theta} - \theta_0) + o\| \hat{\theta} - \theta_0 \| = 0$$

$$\theta^{i+1} = \theta^i + [-\ddot{L}(\theta^i)]^{-1} \dot{L}(\theta^i),\ i = 1,2,..... \qquad (3)$$

Until $\left|\theta^{i+1} - \theta^i\right| < \delta$, δ is an advance fixed value. Gauss newton method has some important properties.

3. STATISTICAL DIAGNOSTICS FOR NONLINEAR MODELS WITH REPEATED MEASUREMENT DATA

In statistics, Cook's distance is an often used to estimate the influential points of a data [12].Data points with huge residuals (outliers) and/or high leverage may misrepresent the outcome and accuracy of a regression.

$$D_{ij} = D_{ij}(U^T U, p'\hat{\sigma}^2) = \frac{(\hat{\beta}_{(ij)} - \hat{\beta})^T (U^T U)(\hat{\beta}_{(ij)} - \hat{\beta})}{p'\hat{\sigma}^2} \quad (4)$$

Where $U = \dfrac{\partial f(x,\beta)}{\partial \beta}$, Cook distance gives squared distance from $\hat{\beta}$ to $\hat{\beta}_{(i)}$ relative to the fixed geometry of $U^T U$. The values of $D_i(U^T U, p'\hat{\sigma}^2)$ can be converted to a familiar probability scale by comparing calculated values to the $F(p', n-p')$ distribution.

Cook distance in multiple cases:

$$D_i = D_i(U^T U, p'\hat{\sigma}^2) = \frac{(\hat{\beta}_{(i)} - \hat{\beta})(U^T U)(\hat{\beta}_{(i)} - \hat{\beta})}{p'\hat{\sigma}^2} \quad (5)$$

D_i Can be expressed in multidimensional analogues of the r_i , and v_{ii}. The results are obtained by first expressing $\hat{\beta}_{(i)}$ as a function of $\hat{\beta}$:

$$\hat{\beta}_{(i)} = (U_{(i)}^T U_{(i)})^{-1} U_{(i)}^T Y_{(i)} = (U^T U - U_i^T U_i)^{-1}(X^T Y - X_i^T Y_i) \quad (6)$$

The inverse of (6)

$$\hat{\beta}_{(i)} = [(U^T U)^{-1} + (U^T U)^{-1} U_i^T (I - V_i)^{-1} U_i (U^T U)^{-1}][U^T Y - X_i^T Y_i]$$

$$= \hat{\beta} - (U^T U)^{-1} U_i^T [-(I - V_i)^{-1} X_i \hat{\beta} + (I + (I - V_i)^{-1} V_i) Y_i] \quad (7)$$

$$\hat{\beta}_{(i)} = \hat{\beta} - (U^T U)^{-1} U_i^T (I - V_i)^{-1} e_i \quad (8)$$

Substituting into (6) ahead to the form:

$$D_i = \frac{e_i^T (I - V_i)^{-1} V_i (I - V_i)^{-1} e_i}{p'\hat{\sigma}^2} \quad (9)$$

Single case Cook distance:

$$\hat{\beta}^{I}_{(ij)} = \hat{\beta} + [I_{(ij)}(\hat{\beta})]^{-1} \dot{L}_{ij}(\hat{\beta}) \quad (10)$$

In this case, $I(\beta) = U^T \Sigma^{-1} U$, and $\dot{L}(\beta) = U^T \Sigma^{-1} U e$

Replacing into (4), we get the form

$$D_{ij} = [U_{ij}^T \Sigma_{ij}^{-1} U_{ij}]^{-1} U_{ij}^T \Sigma_{ij}^{-1} e_{ij} \quad (11)$$

Multiple cases Cook distance

$$\hat{\beta}_{(i)} = \hat{\beta} + [I_{(i)}(\hat{\beta})]^{-1} \dot{L}_{(i)}(\hat{\beta})$$

Substituting into (6), this form gets

$$D_i = [U_{(i)}^T \Sigma_{(i)}^{-1} U_{(i)}]^{-1} U_{(i)} \Sigma^{-1} U_{(i)} e_{(i)} \quad (12)$$

Example 1:

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

	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 consider two examples to calculate Gauss newton method

$$y = \beta_1 \exp(-\beta_2 x) + \beta_3 \exp(-\beta_4 x),\ \beta_1, ..., \beta_4 > 0, \quad (13)$$

We examine Gauss Newton method

$$\beta^{i+1} = \beta^i + [U^T \Sigma^{-1} U]^{-1} U^T \Sigma^{-1} e$$

Using MATLAB's convention for representing Jacobin matrix U which is equal to $U = \frac{\partial f(\beta)}{\partial \beta}$ where $\Sigma = In$ a known case, and $e = y - f(\beta)$,

We chose initial values of β, $\beta_0 = [0.7, 0.6, 0.54, 0.5]$, after 5 iterations we obtained $\hat{\beta} = [0.75, 0.65, 0.50, 0.45]$ .which is satisfied under condition $\left\| \beta^{i+1} - \beta^i \right\| < 10^{-4}$.

Example 2:

We consider another example to compute Gauss newton method.
The result of estimation of the parameters in based on 11 responses for the fifth subjects are given in Table I. Using Matlab to calculate G-N method and get parameter estimations.
We choose initial values,

$$\beta_0 = [1.0000, 1.2000, -1.1000, -1.2000],$$

then use Gauss newton method to estimate the values of β. After 5 iterations, we obtained $\hat{\beta} = [1.2715,\ 1.0408,\ -1.2327,\ -1.5069]$, and we satisfied under this condition $\left\|\beta^{i+1} - \beta^{i}\right\| < 10^{-4}$.

Example.3.

We consider Table I, we focus on fifth subject to detect single case outlier. Where $U = \dfrac{\partial f(\beta)}{\partial \beta}$, $\sum = In$ and e is unobservable error $y - f(\beta)$.

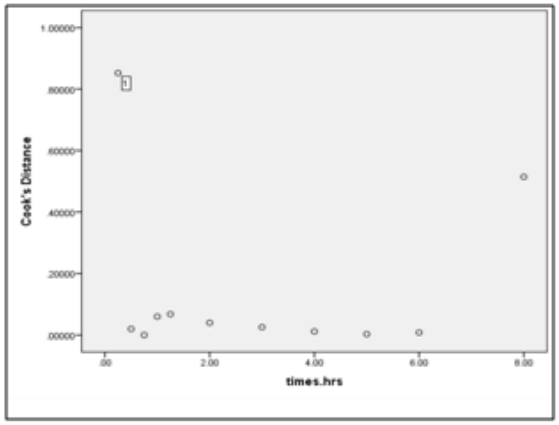

Fig.1. Scatter plot for the table I (fifth individual) under model (11).

In the above scatterplot, we obtained cook's distance and found outlier in a set of predicted values. First observation of our data set is an outlier which is indicated in (figure.1).

Example.4

We consider another example to detect multiple outliers' cases.

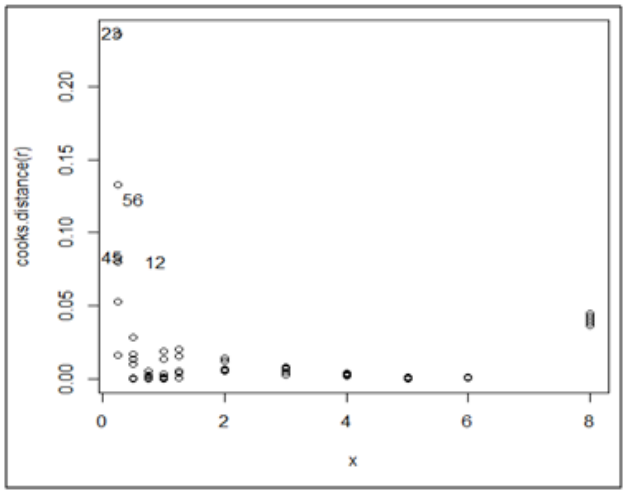

Figure. 2. Scatter plot for the table I under model (12).

We obtained cooks. Distance and found four values that fall far from other data points. So we consider these (23, 56, 45, 12) points outliers in 66 observations data. The outliers are designated in (figure.2) cook's distance plot.

4. CONCLUSION:

It is well understood that all observations of a data set don't play the same role 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 established Gauss newton method for parameter estimation and as well we extended rebut version of Cook. Distance in single and multiple cases to detect outliers data points for repeated measurement data.

REFERENCES:

[1] Ayinde, K., Lukman, A.F. and Arowolo, O. (2015) "Robust Regression Diagnostics of Influential Observations in Linear Regression Model". Open Journal of Statistics,vol.5, pp273-283.

[2] Altman, N. & Krzywinski, M.(2016) "Analyzing outliers influential or nuisance". .Nature methods, vol.13, pp281-282.

[3] 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,vol. 46, pp3159-3171.

[4] Cook, R.D and Tsai, C.L. (1985)"Residual in nonlinear regression", Biometrika, vol. 72, No.1, pp23-29.

[5] Cook R.D. (1979)"Influence observations in linear regression", J.Amer.statist.Assoc, vol.74, pp169-74.

[6] Cook R.D, and presscot. (1981)"Approximation significance levels for detecting outlier in linear regression", Technometrics, vol.23,pp59-64.

[7] Ellenberg, J.H. (1976)"Testing of a single outlier from a general regression model", Biometrics, vol. 32, pp637-45.

[8] Vonesh, E.F. (1992)"Nonlinear models for the analysis of longitudinal data", Statistics in medicine, vol. 11, pp1929-1954.

[9] Solomon P.J. and cox D.R. (1992)"Nonlinear components for variance models", Biometrikka,vol. 79, pp1-11.

[10] Cook R.D. (1979)"Influence observation in liner regression", J.Am.statist.assoc,vol. 74, pp169-174.

[11] Diggle, P. J. (1988)"An approach to the analysis of repeated measurements", Biometrics, vol. 44, pp959-971.

[12] PREGIBON, D. (1981) "Logistic regression diagnostics", Annual of statistics, vol.9, pp705-724.

[13] Anscombe, F.J. (1961) "Examination of residuals, Proc.fouth Berkeley symp" vol. 1, pp1-36.

[14] MARIE DAIDIAN and DAVID M.GILTINAN .march. (1995) "Nonlinear models for repeated measurement data".

AUTHOR

Munsir Ali, school of science, department of statistics Nanjing University of science and technology, P.R china.