QBS
Vol. 36, No. 2, pp. 123-126 (2017)
https://doi.org/10.22283/qbs.2017.36.2.123
pISSN 2288-1344 eISSN 2508-7185 **Original Article**

Quantitative Bio-Science
Official Journal of the Institute of
Natural Science at Keimyung University
Available Online at https://qbs.kmu.ac.kr:442/

Influence and Leverage Measures in Nonlinear Regression

Myung-Wook Kahng[1,*]

[1]Department of Statistics, Sookmyung Women's University, Seoul 04310, Korea
(Received October 25, 2017; Revised November 15, 2017; Accepted November 18, 2017)

ABSTRACT

Assessment of the influence is an important part of regression diagnostics. The measure of influence in linear regression has been extended to nonlinear regression. The connections between measures of influence and leverage are explored. We suggest a modification of the influence measure for assessing the influential observations on the parameter estimates in a nonlinear regression model.

Key words : Distance measure, Influential observation, Intrinsic curvature measure, Jacobian leverage

1. Introduction

In statistical inferences we have implicit assumption that all observations have an equal role in the estimation of model parameters and the subsequent conclusions. An observation is influential if it has a much greater impact on the estimation of parameters and predicted values, compared to the most of the other observations. Detection of influential observation is an important part of the regression diagnostics.

The basic statistics here are the residual and the leverage concerning the influence in linear regression models [1]. The definitions of leverage in nonlinear regression models are considered by Emerson et al. [2]. The main idea is to use the linear approximation of a nonlinear regression model. Bates et al. [3] propose measures of parameter-effects and intrinsic curvature for assessing the validity of the linear approximation. These ideas are refined by Bates et al. [4], and Hamilton et al. [5].

In this paper we consider how the influence measures affect leverage, and we discuss the relationship between influence and leverage measures. A brief review of leverage in nonlinear regression models given in Section 2. We propose a modification of the influence measure in Section 3. In Section 4, we provide example.

* Correspondence should be addressed to Dr. Myung Wook Kahng, Department of Statistics, Sookmyung Women's University, Seoul 04310, Korea. Tel: +82-2-710-9435, Fax: +82-2-710-9283, E-mail: mwkahng@sookmyung.ac.kr

2. Leverages in Nonlinear Regression

Consider the standard nonlinear regression model

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\theta}) + \epsilon_i,\ i = 1, \cdots, n$$

where $\boldsymbol{x}_i$ represents a q-dimensional vector of known explanatory variables associated with the i-th response y_i, $\boldsymbol{\theta}$ is $p \times 1$ vector of unknown parameter, and ϵ_i is error. The response function f is assumed to be known and continuous, one-to-one, and twice continuously differentiable in $\boldsymbol{\theta}$, and ϵ_i is independent random error with zero mean and variance σ^2. In matrix notation the model can be written

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is an $n \times 1$ response vector with elements $y_1, \cdots, y_n$, $\boldsymbol{X}$ is an $n \times q$ matrix of known explanatory variables with rows $\boldsymbol{x}_1^T, \cdots, \boldsymbol{x}_n^T$, $\boldsymbol{\epsilon}$ is an $n \times 1$ vector with elements $\epsilon_1, \cdots, \epsilon_n$, and $\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) = (f(\boldsymbol{x}_1, \boldsymbol{\theta}), \cdots, f(\boldsymbol{x}_n, \boldsymbol{\theta}))^T = (f_1(\boldsymbol{\theta}), \cdots, f_n(\boldsymbol{\theta}))^T = \boldsymbol{f}(\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}$ be the least squares estimate of parameter vector, $\hat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}) = \boldsymbol{f}(\hat{\boldsymbol{\theta}})$ be the predicted response vector.

A tangent plane to the expectation surface at the point $\hat{\boldsymbol{\theta}}$ is used to make inferences about $\boldsymbol{\theta}$ based on the approximated linear model $\boldsymbol{f}(\boldsymbol{\theta})=\boldsymbol{f}(\hat{\boldsymbol{\theta}})+\hat{\boldsymbol{V}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})$. Here, $\boldsymbol{V}=\boldsymbol{V}(\boldsymbol{\theta})=\partial \boldsymbol{f}/\partial \boldsymbol{\theta}^T$ is the $n\times p$ matrix and $\hat{\boldsymbol{V}}=\boldsymbol{V}(\hat{\boldsymbol{\theta}})$. Under this approximation, $\boldsymbol{f}(\boldsymbol{\theta})$ lies in this tangent plane. The tangent plane leverage matrix can be given by $\hat{\boldsymbol{H}}=\hat{\boldsymbol{V}}(\hat{\boldsymbol{V}}^T\hat{\boldsymbol{V}})^{-1}\hat{\boldsymbol{V}}^T$. The diagonal elements $\hat{h}_{ii}$ of $\hat{\boldsymbol{H}}$ are measures of leverage for i-th observation in nonlinear regression model [6].

Emerson et al. [2] discussed the measure of leverages by perturbation schemes. Given the perturbed response vector $\boldsymbol{y}+b\boldsymbol{c}_i$ and the perturbed least square estimate, denoted by $\hat{\boldsymbol{\theta}}(b)$, the perturbed predicted response vector is $\hat{\boldsymbol{y}}(b)=\boldsymbol{f}(\hat{\boldsymbol{\theta}}(b))$, where $\boldsymbol{c}_i$ is an $n\times 1$ vector with i-th element equal to one, and all other elements equal to zero.

Another measure of leverage in nonlinear regression models is referred to by St. Laurent et al. [7] as the Jacobian leverage. The leverages in nonlinear regression model can be obtained as follows [8]:

$$\lim_{b\to 0}\frac{\hat{\boldsymbol{y}}(b)-\hat{\boldsymbol{y}}}{b}=\hat{\boldsymbol{V}}(\hat{\boldsymbol{V}}^T\hat{\boldsymbol{V}}-\sum_{i=1}^{n}e_i\hat{\boldsymbol{W}}_i)^{-1}\hat{\boldsymbol{V}}^T\boldsymbol{c}_i=\hat{\boldsymbol{J}}\boldsymbol{c}_i$$

where

$$\hat{\boldsymbol{J}}=\hat{\boldsymbol{V}}(\hat{\boldsymbol{V}}^T\hat{\boldsymbol{V}}-\sum_{i=1}^{n}e_i\hat{\boldsymbol{W}}_i)^{-1}\hat{\boldsymbol{V}}^T,$$

is the Jacobian leverage matrix. Here $\boldsymbol{W}_i=\boldsymbol{W}_i(\boldsymbol{\theta})=\partial^2 f_i/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T$ is the $p\times p$ Hessian matrix and $\hat{\boldsymbol{W}}_i=\boldsymbol{W}_i(\hat{\boldsymbol{\theta}})$.

3. Influential observations

Assessment of the influence of the observations on the parameter estimate is an important part of influence analysis. Numerous literatures are available for the identification of influential observations in linear regression. The Cook's distance have become very popular among a number of available influence measures.

Consider the linear regression model

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is an $n\times 1$ response vector, $\boldsymbol{X}$ is an $n\times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p\times 1$ vector of unknown parameters, and $\boldsymbol{\epsilon}$ is a vector of independent random variables each with zero mean and variance σ^2. Let $\hat{\beta}$ be the least squares estimate of parameter vector, $\hat{y}=\boldsymbol{X}\hat{\beta}$ be the predicted response vector, and $\boldsymbol{e}=\boldsymbol{y}-\hat{y}$ be the residual vector. Recall that the covariance matrices of the residuals $\boldsymbol{e}=(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$ and predicted values $\hat{y}=\boldsymbol{H}\boldsymbol{y}$ are given by $\sigma^2(\boldsymbol{I}-\boldsymbol{H})$ and $\sigma^2\boldsymbol{H}$, respectively, where $\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. The diagonal elements h_{ii} of the projection matrix $\boldsymbol{H}$ are called leverages and generally indicate the amount of leverage of the response value y_i on the corresponding predicted value $\hat{y}_i$.

Cook [9,10] proposed that the influence of the i-th observation be judged by using the distance measure,

$$D_i=\frac{(\hat{\beta}_{(i)}-\hat{\beta})^T(\boldsymbol{X}^T\boldsymbol{X})(\hat{\beta}_{(i)}-\hat{\beta})}{p\hat{\sigma}^2}$$

where $\hat{\beta}_{(i)}$ is the estimate of β without the i-th data point, and $\hat{\sigma}^2=\boldsymbol{e}^T\boldsymbol{e}/(n-p)$. A large value of D_i indicates that the associated i-th observation has a strong influence on the estimate of β. The magnitude of the distance between $\hat{\beta}$ and $\hat{\beta}_{(i)}$ may be assessed by comparing D_i to the probability points of the central F-distribution with p and $n-p$ degrees of freedom. This is equivalent to finding the level of the confidence ellipsoid centered at $\hat{\beta}$ that passes through $\hat{\beta}_{(i)}$ and entails nothing more than a monotonic transformation of D_i to a familiar scale. The usual computational form of D_i that depends only on the full data set;

$$D_i=\frac{1}{p}r_i^2\frac{h_{ii}}{1-h_{ii}}$$

where r_i is the i-th studentized residual. Clearly, D_i can be larges if either r_i^2 or h_{ii} is large. These two components measure the importance of two characteristics of each data points. The i-th studentized residual r_i, reflecting lack of fit of the model at the i-th case, and potential h_{ii}, reflecting the location of $\boldsymbol{x}_i$. The ratio which is as the Hadi's potential [11] can also be expressed as follows [12]:

$$\frac{h_{ii}}{1-h_{ii}}=\boldsymbol{x}_i^T(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i=\frac{var(\hat{y}_i)}{var(e_i)}$$

where $\boldsymbol{X}_{(i)}$ is $\boldsymbol{X}$ without observation i. It indicates that the leverage of an observation is directly related to its corresponding model sensitivity.

Cook et al. [12] suggested that some statistics developed for linear regression models can be applied to nonlinear regression models if the models are approximately linear in the vicinity of the optimum parameter set. The parameters of nonlinear models can be transformed to yield an approximately linear model if the intrinsic nonlinearity is sufficiently small [3,4]. The matrix $\hat{\boldsymbol{H}}$, residuals $\boldsymbol{e}$, and estimated error variance $\hat{\sigma}^2$ remain invariant under such a transformation [13], suggesting that Cook's distance is a valid measure of influence for models of this type because the measure is based on $\hat{\boldsymbol{H}}$, $\boldsymbol{e}$, and $\hat{\sigma}^2$. The variance-covariance matrix $\sigma^2(\hat{\boldsymbol{V}}^T\hat{\boldsymbol{V}})^{-1}$, however, is not invariant under a transformation of parameters unless the total nonlinearity is small [13]. Therefore if the measure of total nonlinearity indicates that a model is effectively linear, Cook's distance can be accurately computed. Cook's distance is also applicable to models with a high degree of total nonlinearity if the intrinsic nonlinearity is sufficiently low.

A version of Cook's distance for assessing the influence of the observations on the vector of estimated parameters in the nonlinear regression model is proposed by Cook et al. [12]. The explicit expression of this measure is given by

$$D_i=\frac{(\hat{\boldsymbol\theta}_{(i)}-\hat{\boldsymbol\theta})^T(\hat{\boldsymbol V}^T\hat{\boldsymbol V})(\hat{\boldsymbol\theta}_{(i)}-\hat{\boldsymbol\theta})}{p\hat\sigma^2}$$

where $\hat{\boldsymbol\theta}_{(i)}$ is the estimate of $\boldsymbol\theta$ when the i-th observation is excluded from the calculations. The usual computational form of D_i is

$$D_i=\frac{1}{p}r_i^2\frac{\hat h_{ii}}{1-\hat h_{ii}}$$

where r_i is the i-th studentized residual in nonlinear regression

$$r_i=\frac{e_i}{\hat\sigma\sqrt{1-\hat h_{ii}}}$$

and $\hat h_{ii}$ is the i-th diagonal element of $\hat{\boldsymbol H}$, which is referred to as the tangent plane leverage matrix. Another measure of leverage in nonlinear regression models is referred to by St. Laurent et al. [7] as the Jacobian leverage. The Jacobian leverage matrix is given by

$$\hat{\boldsymbol J}=\hat{\boldsymbol V}(\hat{\boldsymbol V}^T\hat{\boldsymbol V}-\sum_{i=1}^n e_i\hat{\boldsymbol W}_i)^{-1}\hat{\boldsymbol V}^T.$$

The discussion about two leverages have led us to suggest a modification of the influence measure $D_{J,i}$

$$D_{J,i}=\frac{1}{p}r_{J,i}^2\frac{\hat j_{ii}}{1-\hat j_{ii}}$$

which is referred to as the i-th Jacobian Cook's distance. Here, $r_{J,i}$ is the i-th Jacobian studentized residual

$$r_{J,i}=\frac{e_i}{\hat\sigma\sqrt{1-\hat j_{ii}}}$$

and $\hat j_{ii}$ is the i-th diagonal element of $\hat{\boldsymbol J}$.

St. Laurent et al. [7] suggested using $\hat{\boldsymbol H}$ whenever possible since it is easier to conduct computations and interpretation of the results is similar to linear regression. For example, the diagonal elements of $\hat{\boldsymbol H}$ have the following properties: $0\leq\hat h_{ii}\leq 1$ and $\sum_{i=1}^n\hat h_{ii}=p$, where p is the number columns of $\hat{\boldsymbol V}$. These properties do generally not hold for the diagonal elements of the Jacobian leverage matrix.

Table 1. Tetracycline data and influence measures

x_i	y_i	$\hat h_{ii}$	$\hat j_{ii}$	r_i	D_i	$r_{J,i}$	$D_{J,i}$
1	0.7	0.978	0.960	1.798	35.747	1.340	10.84
2	1.2	0.617	0.548	−1.502	0.909	−1.383	0.580
3	1.4	0.383	0.375	0.366	0.021	0.363	0.020
4	1.4	0.359	0.353	1.383	0.267	1.376	0.258
6	1.1	0.421	0.375	−0.150	0.004	−0.145	0.003
8	0.8	0.271	0.258	−1.179	0.129	−1.168	0.118
10	0.6	0.258	0.264	−0.866	0.065	−0.869	0.068
12	0.5	0.334	0.334	0.772	0.075	0.772	0.075
16	0.3	0.379	0.360	1.119	0.191	1.102	0.170

4. Example

We consider data on the metabolism of tetracycline from Bates et al. [14], given in Table 1. A proposed model for these data is

$$f(\boldsymbol x,\boldsymbol\theta)=\theta_3[\exp(-\theta_1(\boldsymbol x-\theta_4))-\exp(-\theta_2(\boldsymbol x-\theta_4))].$$

The parameter estimates are $(\hat\theta_1,\hat\theta_2,\hat\theta_3)=(0.149, 0.716, 2.650)$ and $\hat\sigma=0.0448$. We obtain the maximum intrinsic curvature $\Gamma^\eta(\boldsymbol\theta)=.5011$. This curvature measure exceed corresponding guide $1/[\sqrt{F_{.05}(4,5)}]=.4389$, indicating inadequacy of the linear approximation inference. Given in Table 1, we have two versions of leverage, $\hat h_{ii}$, $\hat j_{ii}$ and two versions studentized residuals, r_i, $r_{J,i}$. In addition, the Cook's distance D_i and Jacobian Cook's distance $D_{J,i}$ are calculated for each case. There is a noticeable disagreement between the two measures in some cases.

References

1. Chatterjee S, Hadi AS. Influential observations, high leverage points, and outliers in linear regression. Stat Sci 1986;1:379-416.
2. Emerson JD, Hoaglin DC, Kempthorne PJ. Leverage in least squares additive-plus-multiplicative fits for two-way tables. J Am Stat Assoc 1984;79:329-335.
3. Bates DM, Watts DG. Relative curvature measures of nonlinearity (with discussion). J Roy Stat Soc B Met 1980;42:1-25.
4. Bates DM, Watts DG. Parameter transformations for improved approximate confidence regions in nonlinear least squares. Ann Stat 1981;9:1152-1167.
5. Hamilton DC, Watts DG, Bates DM. Accounting for intrinsic nonlinearity parameter inference regions. Ann Stat 1982;10:

386-393.
6. Ross WH. The geometry of case deletion and the assessments of influence in nonlinear regression. Can J Stat 1987;15:91-103.
7. St Laurent RT, Cook RD. Leverage and superleverage in nonlinear regression. J Am Stat Assoc 1992;87:985-990.
8. Kahng M. Leverage measures in nonlinear regression. J Korean Data Inf Sci Soc 2007;18:229-235.
9. Cook RD. Detection of influential observations in linear regression. Technometrics 1977;19:15-18.
10. Cook RD. Influential observations in linear regression. J Am Stat Assoc 1979;74:169-174.
11. Hadi AS. A new measure of overall potential influence in linear regression. Comput Stat Data An 1992;14:1-27.
12. Cook RD, Weisberg S. Residuals and influence in regression. London: Chapman and Hall; 1982.
13. Seber GAF, Wild CJ. Nonlinear regression. New York: John Wiley and Sons; 1989.
14. Bates DM, Watts DG. Nonlinear regression analysis and its applications. New York: John Wiley and Sons; 1988.