Communications of the Korean Statistical Society
2009, Vol. 16, No. 1, 163–168

Assessing the Accuracy of Outlier Tests in Nonlinear Regression

Myung Wook Kahng[1,a], Bu-Yong Kim[a]

[a]Dept.of Statistics, Sookmyung Women's Univ.

Abstract

Given the specific mean shift outlier model, the standard approaches to obtaining test statistics for outliers are discussed. Accuracy of outlier tests is investigated using subset curvatures. These subset curvatures appear to be reliable indicators of the adequacy of the linearization based test. Also, we consider obtaining graphical summaries of uncertainty in estimating parameters through confidence curves. The results are applied to the problem of assessing the accuracy of outlier tests.

Keywords: Mean shift outlier model, outlier test, curvature measure, confidence curves.

1. Introduction

A standard paradigm for the analysis of nonlinear models is to assume that results for the linear model hold at least approximately for large enough sample sizes. Beginning with Beale (1960) and more importantly with the papers of Bates and Watts (1980, 1981), it has become increasingly clear that this approximation may not be at all adequate in some problems. Researchers have developed methods both to determine when linear approximations are inadequate and to give alternative procedures when the standard methods fail.

Section 2 considers the problem of testing outliers in nonlinear regression. Given the specific mean shift model, standard approaches to obtaining test statistics for outliers are discussed. The test based on linear approximation, namely the score test, is easy to calculate, but quite different from likelihood ratio test. Accuracy of the linear approximation is investigated using subset curvature measures in Section 3.

We consider obtaining graphical summaries of uncertainty in estimation in nonlinear regression models using confidence curves in Section 4. The results of confidence curves are applied not only to the problem of finding significance levels for outlier tests, but also to assessing the accuracy of the test based on linear approximation. We provide an example in Section 5.

2. Outlier Tests in Nonlinear Regression

The standard nonlinear regression model can be expressed as

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

in which the i^{th} response y_i is related to the q-dimensional vector of known explanatory variables $\boldsymbol{x}_i$ through the known model function f, which depends on the p-dimensional unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

This research was supported by the Sookmyung Women's University Research Grants 2007.

[1] Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea. E-mail: mwkahng@sm.ac.kr

and ϵ_i is error. We assume that f is twice continuously differentiable in $\boldsymbol{\theta}$ and errors ϵ_i are *i.i.d.* normal random variables with mean 0 and variance σ^2. In matrix notation we may write,

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{y}$ is an n-dimensional vector with elements $y_1, \ldots, y_n$, $\boldsymbol{X}$ is an $n \times q$ matrix with rows $\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T$, $\boldsymbol{\epsilon}$ is an n-dimensional vector with elements $\epsilon_1, \ldots, \epsilon_n$ and $\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) = (f(\boldsymbol{x}_1, \boldsymbol{\theta}), \ldots, f(\boldsymbol{x}_n, \boldsymbol{\theta}))^T = (f_1(\boldsymbol{\theta}), \ldots, f_n(\boldsymbol{\theta}))^T = \boldsymbol{f}(\boldsymbol{\theta})$. Given the response vector $\boldsymbol{y}$, the least squares estimate of $\boldsymbol{\theta}$ is denoted $\hat{\boldsymbol{\theta}}$, the predicted response vector is $\hat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}) = \boldsymbol{f}(\hat{\boldsymbol{\theta}})$ and the residual vector is $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}) = \{e_i\}$. A tangent plane approximation to the expectation surface at $\hat{\boldsymbol{\theta}}$ is used to make inferences about $\boldsymbol{\theta}$ through the derived linear model $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{f}(\hat{\boldsymbol{\theta}}) + \hat{\boldsymbol{V}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta}) = \partial \boldsymbol{f} / \partial \boldsymbol{\theta}^T$ is the $n \times p$ matrix and $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$ is $\partial \boldsymbol{f} / \partial \boldsymbol{\theta}^T$ evaluated at $\hat{\boldsymbol{\theta}}$.

Suppose we suspect in advance that m cases indexed by an m-dimensional vector $\mathbf{I} = (i_1, \ldots, i_m)$ are outliers. A useful framework used to study outliers is the mean shift outlier model,

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) + \boldsymbol{D}\boldsymbol{\delta} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{\delta} = (\delta_{i_1}, \ldots, \delta_{i_m})^T$, $\boldsymbol{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_m)$ and $\boldsymbol{d}_j$ is the i_j^{th} standard basis vector for $\boldsymbol{R}^n$. The testing of the hypothesis $\boldsymbol{\delta} = \mathbf{0}$ is equivalent to testing whether cases in the set $\mathbf{I}$ are outliers.

We denote the log-likelihood for model (2.2) by $L(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2)$ and obtain

$$\begin{aligned} L\left(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2\right) &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{\delta})^T (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{\delta}) \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} S(\boldsymbol{\theta}, \boldsymbol{\delta}), \end{aligned} \tag{2.3}$$

where $S(\boldsymbol{\theta}, \boldsymbol{\delta}) = (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{\delta})^T (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{\delta})$. Given σ^2, (2.3) is maximized with respect to $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\delta})$ when $S(\boldsymbol{\theta}, \boldsymbol{\delta})$ is minimized at the least squares estimates $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\theta}}_{(I)}, \hat{\boldsymbol{\delta}})$. Furthermore, $\partial L / \partial \sigma^2 = 0$ has solution $\sigma^2 = S(\boldsymbol{\theta}, \boldsymbol{\delta})/n$, which gives a maximum for given $\boldsymbol{\phi}$ as the second derivative is negative. This suggests that $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\theta}}_{(I)}, \hat{\boldsymbol{\delta}})$ and $\hat{\sigma}^2_{(I)} = S(\hat{\boldsymbol{\theta}}_{(I)}, \hat{\boldsymbol{\delta}})/n$ are the maximum likelihood estimates. Under the null hypothesis $\boldsymbol{\delta} = \mathbf{0}$, the maximum likelihood estimates are $\boldsymbol{\phi}_0 = (\hat{\boldsymbol{\theta}}, \mathbf{0})$ and $\hat{\sigma}^2 = S(\hat{\boldsymbol{\theta}}, \mathbf{0})/n$, which are the maximum likelihood estimates of model (2.1).

Now, we consider procedures for testing $H_0 : \boldsymbol{\delta} = \mathbf{0}$ against $H_1 : \boldsymbol{\delta} \neq \mathbf{0}$. Three asymptotically equivalent test methods are discussed by Kahng (1995), namely the likelihood ratio (LR) test introduced by Neyman and Pearson (1928), Wald (WD) test proposed by Wald (1943) and the score (S) test due originally to Rao (1947) and developed further by Silvey (1959). The three test statistics are defined as follows:

$$\begin{aligned} \mathrm{LD} &= 2\left\{L(\hat{\boldsymbol{\phi}}) - L(\boldsymbol{\phi}_0)\right\} \\ &= n\left\{\log S\left(\hat{\boldsymbol{\theta}}, \mathbf{0}\right) - \log S\left(\hat{\boldsymbol{\theta}}_{(I)}, \hat{\boldsymbol{\delta}}\right)\right\}, \\ \mathrm{WD} &= \hat{\boldsymbol{\delta}}^T \left\{\mathrm{Var}(\hat{\boldsymbol{\delta}})\right\}^{-1} \hat{\boldsymbol{\delta}} \\ &= \frac{1}{\hat{\sigma}^2_{(I)}} \hat{\boldsymbol{\delta}}^T (\boldsymbol{I} - \tilde{\boldsymbol{H}}_I) \hat{\boldsymbol{\delta}}, \\ S &= \boldsymbol{U}(\boldsymbol{\phi}_0)^T \boldsymbol{I}(\boldsymbol{\phi}_0)^{-1} \boldsymbol{U}(\boldsymbol{\phi}_0) \\ &= \frac{1}{\hat{\sigma}^2} \boldsymbol{e}_I^T \left(\boldsymbol{I} - \hat{\boldsymbol{H}}_I\right)^{-1} \boldsymbol{e}_I, \end{aligned}$$

where $\tilde{\boldsymbol{H}}_I$ is $m \times m$ minor of $\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{V}}(\tilde{\boldsymbol{V}}^T\tilde{\boldsymbol{V}})^{-1}\tilde{\boldsymbol{V}}^T$ with rows and columns indexed by $\mathbf{I}$, $\tilde{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}}_{(I)})$ is $\partial \boldsymbol{f}/\partial \boldsymbol{\theta}^T$ evaluated at $\hat{\boldsymbol{\theta}}_{(I)}$, $\boldsymbol{U}(\boldsymbol{\phi}) = \partial L(\boldsymbol{\phi})/\partial \boldsymbol{\phi}$, $\boldsymbol{I}(\boldsymbol{\phi}_0) = \partial^2 L(\boldsymbol{\phi})/\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^T$, $\hat{\boldsymbol{H}}_I$ is the $m \times m$ minor of $\hat{\boldsymbol{H}} = \hat{\boldsymbol{V}}(\hat{\boldsymbol{V}}^T\hat{\boldsymbol{V}})^{-1}\hat{\boldsymbol{V}}^T$ with rows and columns indexed by $\mathbf{I}$ and $\boldsymbol{e}_I$ is m-dimensional vector whose j^{th} element is e_{i_j}.

Suppose that only the i^{th} case is suspected as being an outlier. In this single outlier case, the outlier model (2.2) becomes

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) + \delta \boldsymbol{d}_i + \boldsymbol{\epsilon}, \tag{2.4}$$

where $\boldsymbol{d}_i$ is the i^{th} standard basis vector for $\boldsymbol{R}^n$. The three statistics take the form

$$\mathrm{LD} = n\left\{\log S\left(\hat{\boldsymbol{\theta}}, 0\right) - \log S\left(\hat{\boldsymbol{\theta}}_{(i)}, \hat{\delta}\right)\right\},$$

$$\mathrm{WD} = \frac{\hat{\delta}^2\left(1 - \tilde{h}_{ii}\right)}{\hat{\sigma}^2_{(i)}},$$

$$S = \frac{e_i^2}{\hat{\sigma}^2\left(1 - \hat{h}_{ii}\right)},$$

where $\tilde{h}_{ii}$ and $\hat{h}_{ii}$ are the i^{th} diagonal elements of $\tilde{\boldsymbol{H}}$ and $\hat{\boldsymbol{H}}$, respectively. The three statistics differ in computational features. Unlike the other two tests, the score test requires only quantities calculated under the null hypothesis. Nevertheless, all three statistics are invariant under the reparametrization of $\boldsymbol{\theta}$ and have the same asymptotic distribution under the null hypothesis $H_0 : \boldsymbol{\delta} = \mathbf{0}$.

3. Accuracy of Linear Approximation

Let's consider the mean shift outlier model (2.4). Since parameter δ is the parameter subset of this model, the subset curvatures for δ can be found by the methods described in Cook and Goldberg (1986).

Let $\boldsymbol{h}(\delta) = \boldsymbol{f}(\boldsymbol{X}, \tilde{\boldsymbol{\theta}}(\delta)) + \delta \boldsymbol{d}_i$, where $\tilde{\boldsymbol{\theta}}(\delta)$ denotes the p-dimensional vector-valued function that maximizes $L(\boldsymbol{\theta}, \delta, \sigma^2)$ over $\boldsymbol{\theta}$ given δ. To obtain precise expressions, we need the following definitions. Define $n \times p \times p$ array $\tilde{\boldsymbol{W}} = \boldsymbol{W}(\hat{\boldsymbol{\theta}}_{(i)})$ which is $\partial^2 \boldsymbol{f}/\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T$ evaluated at $\hat{\boldsymbol{\theta}}_{(i)}$. We consider the QR decomposition of $n \times (p+1)$ matrix $\boldsymbol{V}_\phi$, namely

$$\boldsymbol{V}_{\boldsymbol{\phi}} = \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\phi}^T}\right|_{\phi=\hat{\phi}} = \left(\tilde{\boldsymbol{V}}, \boldsymbol{d}_i\right) = \boldsymbol{QR},$$

where $\boldsymbol{Q}$ is an $n \times (p+1)$ matrix with orthogonal columns and $\boldsymbol{R}$ is a $(p+1) \times (p+1)$ upper triangular matrix. Now consider the transformation $\tilde{\boldsymbol{W}}_\phi = \boldsymbol{R}^{-T}\tilde{\boldsymbol{W}}_\phi \boldsymbol{R}^{-1}$, where the i^{th} face of $\tilde{\boldsymbol{W}}_\phi$ is

$$\left(\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}\right)_i = \left.\frac{\partial^2 h_i}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^T}\right|_{\phi=\hat{\phi}} = \begin{bmatrix} \tilde{\boldsymbol{W}}_i & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix},$$

$\tilde{\boldsymbol{W}}_i$ is the i^{th} face of $\tilde{\boldsymbol{W}}$ and h_i is the i^{th} element of $\boldsymbol{h}(\delta)$.

Under the above settings and by applying the result of Kahng (1995), we may define the maximum parameter-effects and intrinsic curvatures of $\boldsymbol{h}$ at $\hat{\delta}$ as

$$\Gamma^{\tau}_s(\delta) = \hat{\sigma}_{(i)} \max_{\|\boldsymbol{b}\|=1} \left\| \boldsymbol{b}^T \boldsymbol{A}_{22} \boldsymbol{b} \right\|,$$

$$\Gamma^{\eta}_s(\delta) = 2\hat{\sigma}_{(i)} \max_{\|\boldsymbol{b}\|=1} \left\| [\boldsymbol{b}^T][\boldsymbol{A}_{12}]\boldsymbol{b} \right\|,$$

where $\boldsymbol{A}$ is the $(p+1)\times(p+1)\times(p+1)$ parameter-effects curvature array $\boldsymbol{A} = [\boldsymbol{Q}^T][\breve{\boldsymbol{W}}_\phi]$ (Bates and Watts, 1981), $\boldsymbol{A}_{12}$ and $\boldsymbol{A}_{22}$ are the subarrays of $\boldsymbol{A}$.

If $\Gamma_s^\tau(\delta)$ and $\Gamma_s^\eta(\delta)$ are sufficiently small, the likelihood and linear confidence regions for δ will be similar, otherwise we can expect these confidence regions to be dissimilar. Following Ratkowsky (1983, p.18) and Cook and Goldberg (1986), $1/[2\sqrt{F_\alpha(1,\, n-p-1)}]$ may be used as a rough guide for judging the size of these curvatures. This method can be used to judge the adequacy of the test procedures which are based on the linear approximation. When $\Gamma_s^\tau(\delta)$ and $\Gamma_s^\eta(\delta)$ are greater than the guide, the linearization based test, namely the score test, is quite different from the likelihood ratio test.

4. Confidence Curves for Parameter δ

Cook and Weisberg (1990) give the graphical alternative to likelihood and Wald confidence intervals for a component of the parameter vector. The reason for using confidence curves is that likelihood intervals can have a different shape for each significance level $1-\alpha$. Wald intervals are always symmetric, so if we have a 95% interval, we can always obtain the 90% and 99% intervals in a simple way. With the likelihood intervals, this is not so. Consequently, a graphical summary that does not depend on level is desirable.

The confidence curves include two curves - a likelihood confidence curve and Wald confidence curve. Applying the result of Kahng (2003), the likelihood confidence curves are the set of points defined by

$$\begin{cases} \sqrt{\left\{S\left(\tilde{\boldsymbol{\theta}}(\delta),\, \delta\right) - S\left(\hat{\boldsymbol{\theta}}_{(i)},\, \hat{\delta}\right)\right\}/s_{(i)}^2}, & \text{on the horizontal axis,} \\ \delta, & \text{on the vertical axis,} \end{cases} \tag{4.1}$$

where $s_{(i)}^2 = S(\hat{\boldsymbol{\theta}}_{(i)}, \hat{\delta})/(n-p-1)$. This is a modification of the graphical summary of the standard profile log-likelihood which has δ on the horizontal axis and $S(\tilde{\boldsymbol{\theta}}(\delta), \delta)$ on the vertical axis. The plot (4.1) will be curves, with the amount of curvature giving information about the nonlinearity of the model. To this plot, two straight lines passing through $(0, \hat{\delta})$ with slope $\pm\mathrm{se}(\hat{\delta})$ are added. These two straight lines represent the Wald interval. At any point on the horizontal axis of the confidence curve plot, the interval between the upper and lower curves provides a confidence interval for δ for some level of $1-\alpha$. The confidence level can be determined from the calibrating distribution for either the Wald or likelihood procedure, which, in the scale of the plot, is $t(\nu)$, where ν is the corresponding degrees of freedom. If $t_\nu(u)^{-1}$ is the inverse of the t cumulative distribution function with ν degrees of freedom evaluated at u, then the confidence level at a point along the horizontal axis is $1 - 2t_\nu^{-1}\left[\sqrt{\{S(\tilde{\boldsymbol{\theta}}(\delta), \delta) - S(\hat{\boldsymbol{\theta}}_{(i)}, \hat{\delta})\}/s_{(i)}^2}\right]$.

The Wald and the likelihood regions are tangent at the maximum likelihood estimate $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\theta}}_{(i)}, \hat{\delta})$. If the likelihood is exactly quadratic, that is, if the log-likelihood is exactly normal, then the likelihood curves are the same as the Wald curves(two straight lines). Non-normality is reflected in the likelihood confidence curves failing to be straight.

5. Example

The data on the metabolism of tetracycline were presented in Bates and Watts (1988) and are reproduced in Table 1. In this experiment, a tetracycline compound was administered orally to a subject

Table 1: Tetracycline data

i	x_i	y_i
1	1	0.7
2	2	1.2
3	3	1.4
4	4	1.4
5	6	1.1
6	8	0.8
7	10	0.6
8	12	0.5
9	16	0.3

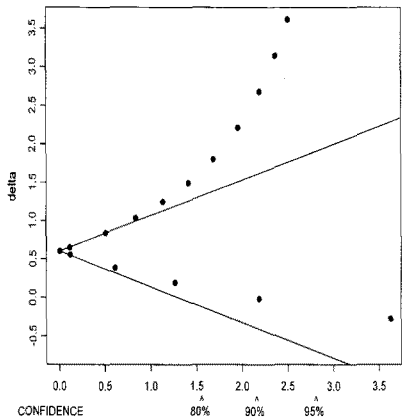

Figure 1: *Confidence curves for* δ

and the concentration(y_i) of tetracycline hydrochloride in the serum in micrograms per milliliter was measured over a period(x_i) of 16 hours.

We consider the outlier problem. Suppose that case 1 is suspected to be an outlier. The proposed model is the following:

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

The subset curvatures for δ are $\Gamma_s^{\tau}(\delta) = .5470$, $\Gamma_s^{\eta}(\delta) = .3653$. The corresponding guide is $1/[2\sqrt{F_{.05}(1, 4)}] = .180$, indicating inadequacy of the linear approximation.

Now we consider the confidence curves discussed in Section 4. The confidence curves for δ is given in Figure 1. In this figure, the likelihood confidence curves are severely curved and dissimilar to the Wald confidence curves. This indicates that estimations, inferences and diagnostics based on the linear approximation about δ may give misleading results.

References

Bates, D. M. and Watts, D. G. (1980). Relative curvature measures of nonlinearity (with discussion), *Journal of the Royal Statistical Society, Series B,* **42**, 1–25.

Bates, D. M. and Watts, D. G. (1981). Parameter transformations for improved approximate confidence regions in nonlinear least squares, *The Annals of Statistics,* **9**, 1152–1167.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications,* John Wiley & Sons, New York.

Beale, E. M. L. (1960). Confidence regions in nonlinear estimation (with discussion), *Journal of the Royal Statistical Society, Series, B,* **22**, 41–76.

Cook, R. D. and Goldberg, M. L. (1986). Curvatures for parameter subsets in nonlinear regression, *The Annals of Statistics,* **14**, 1399–1418.

Cook, R. D. and Weisberg, S. (1990). Confidence curves in nonlinear regression, *Journal of the American Statistical Association,* **85**, 544–551.

Kahng, M. W. (1995). Testing outliers in nonlinear regression, *Journal of the Korean Statistical Society,* **24**, 419–437.

Kahng, M. W. (2003). Confidence curves for a function of parameters in nonlinear regression, *Journal of the Korean Statistical Society,* **32**, 1–10.

Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, *Biometrika,* **20A**, 175–240 and 263–294.

Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation, *Proceedings of the Cambridge Philosophical Society,* **44**, 50–57.

Ratkowsky, D. A. (1983). *Nonlinear Regression Modeling: A Unified Practical Approach,* Marcel Dekker, New York.

Silvey, S. D. (1959). The Lagrangian multiplier test, *The Annals of Mathematical Statistics,* **30**, 389–407.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Transactions of the American Mathematical Society,* **54**, 426–482.

Received October 2008; Accepted November 2008