Journal of Korean
Data & Information Science Society
2006, Vol. 17, No. 1, pp. 205 ~ 211

Asymptotic Properties of Outlier Tests in Nonlinear Regression[1)]

Myung-Wook Kahng[2)]

Abstract

For a linear regression model, the necessary and sufficient condition for the asymptotic consistency of the outlier test statistic is known. An analogous condition for the nonlinear regression model is considered in this paper.

Keywords : Asymptotic distribution, Likelihood ratio test, Mean shift outlier model

1. Introduction

We consider the problem of testing for multiple outliers in nonlinear regression. We shall proceed by first given a specific model for multiple outliers, assuming the suspect set of outliers is known. Given the specific mean shift model, several standard approaches to obtaining test statistics for outliers are discussed by Kahng(1995). These include likelihood ratio tests, Wald tests, and score tests.

Due to nonlinearity, inference of parameters in nonlinear regression models does not enjoy any tactable finite sample optimality property. A general approach to the theoretical study in nonlinear regression models is thus asymptotic. Much of the work was done by first assuming the consistency of nonlinear least square estimators and then proving the asymptotic normality, constructing the confidence region, testing hypothesis, etc. The relatively harder questions of consistency were first rigorously proved by Jennrich(1969) and Wu(1981).

For a linear regression model, the necessary and sufficient condition for the

[1)] This research was supported by the Sookmyung Women's University Research Grants 2004.

[2)] Professor, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea
E-mail : mwkahng@sookmyung.ac.kr

asymptotic consistency of the outlier test statistic is known. An analogous condition for the nonlinear model is considered in this paper.

2. Outlier Models and Test in Nonlinear Regression

The standard nonlinear regression model can be expressed as

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

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 p-dimensional unknown parameter θ, and ϵ_i is error. We assume that f is twice continuously differentiable in θ, and errors ϵ_i are i.i.d normal random variables with mean 0 and variance σ^2. In matrix notation we may write,

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

where $\boldsymbol{y}$ is an n-dimensional vector with elements $y_1, \cdots, y_n$, $\boldsymbol{X}$ is an $n \times q$ matrix with rows $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$, ϵ is an n-dimensional vector with elements $\epsilon_1, \cdots, \epsilon_n$, and $\boldsymbol{f}(\boldsymbol{X}, \theta) = (f(\boldsymbol{x}_1, \theta), \cdots, f(\boldsymbol{x}_n, \theta))^T$.

Suppose we suspect in advance that m cases indexed by an m-dimensional vector $\boldsymbol{I} = (i_1, \cdots, i_m)$ are outliers. It can be helpful to write the model as

$$\begin{cases} y_i = f(\boldsymbol{x}_i, \theta) + \delta_i + \epsilon_i, & i \in \boldsymbol{I} \\ y_i = f(\boldsymbol{x}_i, \theta) + \epsilon_i, & i \notin \boldsymbol{I} \end{cases}$$

which is called the mean shift outlier model. In matrix notation we may write,

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

where $\delta = (\delta_{i_1}, \cdots, \delta_{i_m})^T$, and $\boldsymbol{D} = (\boldsymbol{d}_1, \cdots, \boldsymbol{d}_m)$, and $\boldsymbol{d}_j$ is the i_j-th standard basis vector for $\boldsymbol{R}^n$.

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

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

where $S(\theta, \delta) = (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \theta) - \boldsymbol{D}\delta)^T(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \theta) - \boldsymbol{D}\delta)$. Given σ^2, (2.3) is maximized with respect to $\phi = (\theta, \delta)$ when $S(\theta, \delta)$ is minimized, with minimum at the least squares estimate $\hat{\phi} = (\hat{\theta}_{(I)}, \hat{\delta})$. Furthermore, $\partial L/\partial\sigma^2 = 0$ has solution $\sigma^2 = S(\theta, \delta)/n$, which gives a maximum for given ϕ as the second derivative is negative. This suggests that $\hat{\phi} = (\hat{\theta}_{(I)}, \hat{\delta})$ and $\hat{\sigma}^2_{(I)} = S(\hat{\theta}_{(I)}, \hat{\delta})/n$ are the maximum likelihood estimates. When $\delta = \mathbf{0}$, the maximum likelihood estimates are $\phi_0 = (\hat{\theta}, \mathbf{0})$ and $\hat{\sigma}^2 = S(\hat{\theta}, \mathbf{0})/n$, which are the maximum likelihood estimates of model (2.1)

Let $\boldsymbol{e}$ be the n-dimensional ordinary residual vector, $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \hat{\theta})$. We define $\boldsymbol{y}_I$, ϵ_I, and $\boldsymbol{e}_I$ to be m-dimensional vectors whose j-th elements are y_{i_j}, ϵ_{i_j}, and e_{i_j}, respectively, and $\boldsymbol{X}_I$ to be an $m \times q$ matrix whose j-th row is $\boldsymbol{x}^T_{i_j}$. Also we define $\boldsymbol{y}_{(I)}$, $\epsilon_{(I)}$, and $\boldsymbol{e}_{(I)}$ to be vectors $\boldsymbol{y}$, ϵ, and $\boldsymbol{e}$, respectively, with cases indexed by $\boldsymbol{I}$ deleted and $\boldsymbol{X}_{(I)}$ be matrix with rows indexed by $\boldsymbol{I}$ deleted. Least squares estimation of the parameter δ will give a value of zero for the residuals indexed by $\boldsymbol{I}$ in model (2.2). This means that the observations indexed by $\boldsymbol{I}$ will make no contribution to estimate θ, and thus the least squares estimate of θ in model (2.2) is the same as that in the deletion model,

$$\boldsymbol{y}_{(I)} = \boldsymbol{f}(\boldsymbol{X}_{(I)}, \theta) + \epsilon_{(I)}. \tag{2.4}$$

The resulting estimates of θ from (2.4) or from (2.2) will be called $\hat{\theta}_{(I)}$, from which it is immediate that $\hat{\delta} = \boldsymbol{y}_I - \boldsymbol{f}(\boldsymbol{X}_I, \hat{\theta}_{(I)})$.

The testing of the hypothesis $\delta = \mathbf{0}$ is equivalent to testing whether the set $\boldsymbol{I}$ of m cases are outliers. Thus outlier identification and testing are formally equivalent to solving and testing a subset regression. The likelihood ratio test statistic for testing $Ho: \delta = \mathbf{0}$ against $H_1: \delta \neq \mathbf{0}$ is

$$\begin{aligned} LR &= 2\,[L(\hat{\phi}) - L(\phi_0)] \\ &= n\,[\log S(\hat{\theta}, \mathbf{0}) - \log S(\hat{\theta}_{(I)}, \hat{\delta})]. \end{aligned} \tag{2.5}$$

3. Asymptotic Properties

Under the standard nonlinear regression model (2.1), Seber and Wild(1989, pp. 577-581) prove that the asymptotic distribution of the likelihood ratio statistic for the null hypothesis that satisfies a set of constraints is the chi-square distribution under appropriate regularity conditions (Serfling, 1980, p. 154; Amemiya, 1983. p. 351). They assume that $\hat{\theta}$ and the maximum likelihood estimate of θ under the null hypothesis are asymptotically normal, and that $\boldsymbol{A}$ is positive definite in the neighborhood of θ^*, the true value of θ, where $\boldsymbol{A} = -\,\partial^2 L(\theta)/\partial\theta\partial\theta^T$.

To prove that the asymptotic distribution of LR in (2.5) is chi-square, it is enough to show that the above two conditions hold in the mean shift outlier model (2.2) assuming that these conditions are satisfied in the standard nonlinear regression model (2.1). Specifically, the two conditions required are:

Condition 1 : $\hat{\theta}_{(I)}$ and $\hat{\delta}$ are asymptotically normal.

Condition 2 : $\boldsymbol{M}$ is positive definite in the neighborhood of ϕ^*,

where $\boldsymbol{M} = -\,\partial^2 L/\partial\phi\partial\phi^T$, $\phi = (\theta, \delta)$, and $\phi^* = (\theta^*, \delta)$ is the true ϕ .

Condition 1 is satisfied and can be proved as follows. The maximum likelihood estimate $\hat{\delta}$ is defined to be $\boldsymbol{y}_I - \boldsymbol{f}(\boldsymbol{X}_I, \hat{\theta}_{(I)})$ and can be rewritten as:

$$\begin{aligned}\hat{\delta} &= [\boldsymbol{y}_I - \boldsymbol{f}(\boldsymbol{X}_I, \theta^*)] + [\boldsymbol{f}(\boldsymbol{X}_I, \theta^*) - \boldsymbol{f}(\boldsymbol{X}_I, \hat{\theta}_{(I)})] \\ &= [\epsilon_I + \delta] + [\boldsymbol{f}(\boldsymbol{X}_I, \theta^*) - \boldsymbol{f}(\boldsymbol{X}_I, \hat{\theta}_{(I)})].\end{aligned}$$

Jennrich(1969) and Wu(1981) show that $\hat{\theta}$ is asymptotically normal and converges to θ^* almost surely. For fixed m, this implies that $\hat{\theta}_{(I)}$ is asymptotically normal and converges to θ^* almost surely, and that $\boldsymbol{f}(\boldsymbol{X}_I, \theta^*) - \boldsymbol{f}(\boldsymbol{X}_I, \hat{\theta}_{(I)})$ converges to zero almost surely from continuity of $\boldsymbol{f}$ as long as the regularity conditions for $\hat{\theta}$ estimated with data $(\boldsymbol{X}, \boldsymbol{y})$ apply to estimating from $(\boldsymbol{X}_{(I)}, \boldsymbol{y}_{(I)})$ as well. Thus, $\hat{\delta}$ is the sum of two independent terms $\epsilon_I + \delta$ which is normal, and a second term that is asymptotically normal.

For condition 2, we have

$$-\frac{\partial^2 L}{\partial\phi\partial\phi^T} = \begin{pmatrix} -\dfrac{\partial^2 L(\theta,\delta)}{\partial\theta\partial\theta^T} & -\dfrac{\partial^2 L(\theta,\delta)}{\partial\theta\partial\delta^T} \\ -\dfrac{\partial^2 L(\delta,\delta)}{\partial\delta\partial\theta^T} & -\dfrac{\partial^2 L(\theta,\delta)}{\partial\delta\partial\delta^T} \end{pmatrix},$$

where each of the derivatives is given by

$$\begin{aligned}\frac{\partial^2 L(\theta,\delta)}{\partial\theta\partial\theta^T} &= -\frac{1}{\sigma^2}\left(\left(\frac{\partial\boldsymbol{f}}{\partial\theta^T}\right)^T\left(\frac{\partial\boldsymbol{f}}{\partial\theta^T}\right) - \left(\frac{\partial^2\boldsymbol{f}}{\partial\theta\partial\theta^T}\right)(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X},\theta) - \boldsymbol{D}\delta)\right) \\ \frac{\partial^2 L(\theta,\delta)}{\partial\theta\partial\delta^T} &= -\frac{1}{\sigma^2}\left(\frac{\partial\boldsymbol{f}}{\partial\theta^T}\right)^T\boldsymbol{D} \\ \frac{\partial^2 L(\delta,\delta)}{\partial\theta\partial\delta^T} &= -\frac{1}{\sigma^2}\boldsymbol{D}^T\boldsymbol{D} = -\frac{1}{\sigma^2}\boldsymbol{I}_m.\end{aligned}$$

Now, $\boldsymbol{M}$ can be written as

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22} \end{pmatrix} = \frac{1}{\sigma^2} \begin{pmatrix} \boldsymbol{V}^{*\,T}\boldsymbol{V}^* - \sum_{i=1}^{n} \epsilon_i \boldsymbol{W}_i^* & \boldsymbol{V}^{*\,T}\boldsymbol{D} \\ \boldsymbol{D}^T\boldsymbol{V}^* & \boldsymbol{I}_m \end{pmatrix},$$

where $\boldsymbol{V}^* = \boldsymbol{V}(\theta^*)$ and $\boldsymbol{W}_i^* = \boldsymbol{W}_i(\theta^*)$ are $\partial \boldsymbol{f}/\partial\theta^T$ and $\partial f(\boldsymbol{x}_i, \theta)/\partial\theta\partial\theta^T$ evaluated in the neighborhood of θ^*, respectively. To show that $\boldsymbol{M}$ is positive definite is equivalent to proving that all pivots of $\boldsymbol{M}$ are positive. Since we assume that the upper left sub-matrix, $\boldsymbol{M}_{11}$, is positive definite, the first p pivots are positive. The determinant of $\boldsymbol{M}$ can be written as

$$\begin{aligned} det(\boldsymbol{M}) &= (\sigma^2)^{-(p+m)}\ det(\boldsymbol{M}_{22})\, det(\boldsymbol{M}_{11} - \boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21}) \\ &= (\sigma^2)^{-(p+m)}\ det(\boldsymbol{V}^{*\,T}\boldsymbol{V}^{*} - \sum\epsilon_i \boldsymbol{W}_i^* - \boldsymbol{V}_I^{*\,T}\boldsymbol{V}_I^*) \\ &= (\sigma^2)^{-(p+m)}\ det(\boldsymbol{V}_{(I)}^{*\,T}\boldsymbol{V}_{(I)}^* - \sum\epsilon_i \boldsymbol{W}_i^*). \end{aligned}$$

Let $\boldsymbol{M}_k$ be the $(p+k)\times(p+k)$ upper left sub-matrix of $\boldsymbol{M}$ for $k = 1, \cdots, m$. The determinant of $\boldsymbol{M}_k$ can be written as:

$$det(\boldsymbol{M}_k) = (\sigma^2)^{-(p+k)}\ det(\boldsymbol{V}_{(J)}^{*\,T}\boldsymbol{V}_{(J)}^* - \sum_{i=1}^{n} \epsilon_i \boldsymbol{W}_{\boldsymbol{i}}^*) \quad \text{for } \boldsymbol{J} = (i_1, \cdots, i_k).$$

We can calculate the $(p+k)$-th pivot c_k as a ratio of two determinants:

$$c_k = \frac{det(\boldsymbol{M}_k)}{det(\boldsymbol{M}_{k-1})} \quad \text{for} \quad k = 1, \cdots, m,$$

where $\boldsymbol{M}_0 = \boldsymbol{M}_{11}$. Because $\boldsymbol{M}_{11}$ has a positive determinant, all c_k are positive if all determinants of $\boldsymbol{M}_k$ are positive. Thus condition 2 is satisfied if the following condition holds:

$$det(\boldsymbol{V}_{(J)}^{*\,T}\boldsymbol{V}_{(J)}^* - \sum_{i=1}^{n} \epsilon_i \boldsymbol{W}_i^*) > 0 \quad \text{for} \quad \boldsymbol{J} \in \boldsymbol{I}. \tag{3.1}$$

Therefore, this condition (3.1) should be added to the conditions required to get asymptotic distributions. This leads to the following.

Significant levels of likelihood ratio tests can be found either from the asymptotic distribution of LR, which is the chi-square distribution with m degrees of freedom or by an F-approximation (Gallant, 1987, p. 57; Seber and Wild, 1989, p. 198),

$$F_{LR} = \frac{[S(\hat{\theta}, \mathbf{0}) - S(\hat{\theta}_{(I)}, \hat{\delta})]/m}{S(\hat{\theta}_{(I)}, \hat{\delta})/(n-p-m)} \tag{3.2}$$

which is approximately distributed as the F-distribution with m numerator degrees of freedom and $n-p-m$ denominator degrees of freedom when H_0 is true.

4. Remarks

In practice, we usually do not have a priori knowledge of the suspected cases, thus the procedure based on the first Bonferroni inequality (Miller, 1966) should be used to find the significant levels of outlier tests in (2.5) and (3.2). Under this procedure, we use the following rejection rule; $\max\{LR\} > \chi^2_{\alpha/l}(m, n-p-m)$ or $\max\{F\} > F_{\alpha/l}(m, n-p-m)$, where $l = \binom{m}{n}$, $\chi^2_\alpha(m)$ is the upper α point of the chi-square distribution with m degrees of freedom and $F_\alpha(m, n-p-m)$ is the upper α point of the F-distribution with m and $n-p-m$ degrees of freedom.

In linear the regression model, $\boldsymbol{y} = \boldsymbol{X}\theta + \epsilon$, after deletion of m observations the effect on the residual sum of squares can be written as (see Cook and Weisberg, 1982, p.19; Atkinson, 1985, p. 21)

$$S(\hat{\theta}_{(I)}, \hat{\delta}) - S(\hat{\theta}, \mathbf{0}) = -\boldsymbol{e}_I^T(\boldsymbol{I}_m - \boldsymbol{H}_I)^{-1}\boldsymbol{e}_I \ , \tag{4.1}$$

where $\boldsymbol{H}_I$ is the $m \times m$ minor of $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^T\boldsymbol{X}^T$ with rows and columns indexed by $\boldsymbol{I}$. Using (4.1) we can rewrite (2.5) and (3.2) as

$$LR = n\log\left[\frac{(n-p)s^2}{(n-p)s^2 - \boldsymbol{e}_I^T(\boldsymbol{I}_m - \boldsymbol{H}_i)^{-1}\boldsymbol{e}}\right] \tag{4.2}$$

and

$$F_{LR} = \frac{(n-p-m)[\boldsymbol{e}_I^T(\boldsymbol{I}_m - \boldsymbol{H}_i)^{-1}\boldsymbol{e}]}{m[(n-p)s^2 - \boldsymbol{e}_I^T(\boldsymbol{I}_m - \boldsymbol{H}_i)^{-1}\boldsymbol{e}]} \, . \tag{4.3}$$

Since all quantities in (4.2) and (4.3) can be calculated from the fit for all n observations we can calculate test statistics LR and F_{LR} without refitting the deletion model. However, in nonlinear models, (4.1) does not hold exactly. We need to refit the deletion model to find the test statistics LR in (2.5) and F_{LR} in (3.2).

References

1. Amemiya, T. (1983), Non-linear regression models. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of Economics,* Vol. I, 333-398, North-Holland, Amsterdam.
2. Atkinson, A. C. (1985), *Plots, Transformations, and Regression,* Oxford University Press, Oxford.
3. Cook, R. D. and Weisberg, S. (1982), *Residuals and influence in regression,* Chapman and Hall, New York.
4. Gallant, A. G. (1987), *Nonlinear Statistical Models,* John Wiley and Sons, New York.
5. Jennrich, R. I. (1969), Asymptotic properties of nonlinear least square estimators, *Annals of Mathematical Statistics,* Vol. 40, 633-643.
6. Kahng, M. (1995), Testing outliers in nonlinear regression, *Journal of the Korean Statistical Society,* Vol. 24, No.2.
7. Miller, R. (1966), *Simultaneous Inference,* McGraw Hill, New York.
8. Seber, G. A. F. and Wild, C. J. (1989), *Nonlinear Regression,* John Wiley and Sons, New York.
9. Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics,* John Wiley and Sons, New York.
10. Wu, C. F. (1981), Asymptotic theory of nonlinear least squares estimation, *Annal of Statistics,* Vol. 9, 501-513.

[received date Nov. 205, accepted date : Jan. 2006]