Journal of the Korean
Statistical Society
Vol. 24, No. 2, 1995

Testing Outliers in Nonlinear Regression [†]

Myung-Wook Kahng [1]

ABSTRACT

Given the specific mean shift outlier model, several standard approaches to obtaining test statistic for outliers are discussed. Each of these is developed in detail for the nonlinear regression model, and each leads to an equivalent distribution. The geometric interpretations of the statistics and accuracy of linear approximation are also presented.

KEYWORDS: Outlier, Mean shift outlier model, Likelihood ratio test, Wald test, Score test, Curvature.

1. INTRODUCTION

In this article, we consider the problem of testing for multiple outliers in nonlinear regression. We proceed by first specifying a mean shift outlier model, assuming the suspect set of outliers is known. Given this model, several standard approaches to obtaining test statistics for outliers are discussed. These

[†] This paper was supported by NON DIRECT RESEARCH FUND, Korea Research Foundation, 1993.

[1] Department of Statistics, Sookmyung Women's University, Seoul, 140–742, Korea.

include likelihood ratio tests, Wald tests, and score tests. Each of these is developed in detail for the nonlinear regression model.

In the linear regression model, various statistical tests have been proposed for detecting and rejecting outliers by Anscombe (1960), Anscombe and Tukey (1963), Rosner (1975), and others. Examples of these tests and other relative references can be found in Beckman and Cook (1983).

Bates and Watts (1980) propose measures of intrinsic and parameter-effects curvatures for assessing the adequacy of the linear approximation. Relatively small values for both the maximum intrinsic curvature and the maximum parameter-effects curvature indicate that the linear approximation is reasonable. This procedure applies only to full parameters. Cook and Goldberg (1986) extend this idea to develop curvature measures for an arbitrary parameter subset, and is used here to assess the validity of linearization-based test.

2. OUTLIERS IN NONLINEAR REGRESSION

The standard nonlinear regression model can be expressed as

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\theta}) + \epsilon_i,\ i = 1, 2, \ldots, n,$$

in which the i-th response y_i is related to the q-dimensional vector of known explanatory variable $\boldsymbol{x}_i$ through the known model function f, which depends on p-dimensional unknown parameter vector $\boldsymbol{\theta}$, and ϵ_i is error. We assume that f is twice continuously differentiable in $\boldsymbol{\theta}$, and errors ϵ_i are independent, identically distributed 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 $y_1, y_2, \ldots, y_n$, $\boldsymbol{X}$ is an $n \times q$ matrix with rows $\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_n^T$, $\boldsymbol{\epsilon}$ is an n-dimensional vector with elements $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$, and $\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) = (f(\boldsymbol{x}_1, \boldsymbol{\theta}), f(\boldsymbol{x}_2, \boldsymbol{\theta}), \ldots, f(\boldsymbol{x}_n, \boldsymbol{\theta}))^T$.

Suppose we suspect in advance that m cases indexed by an m-dimensional vector $I = (i_1, i_2, \ldots, i_m)$, are outliers. It can be helpful to write the model as

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

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

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

where $\boldsymbol{\delta} = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{im})^T$, and $\boldsymbol{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_m)$, 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 $L(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2)$ and obtain

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

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. When $\boldsymbol{\delta} = \boldsymbol{0}$, the maximum likelihood estimates are $\boldsymbol{\phi}_0 = (\hat{\boldsymbol{\theta}}, \boldsymbol{0})$ and $\hat{\sigma}^2 = S(\hat{\boldsymbol{\theta}}, \boldsymbol{0})/n$, which are the maximum likelihood estimates of model (2.1).

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

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

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

The testing of the hypothesis $\boldsymbol{\delta} = \mathbf{0}$ is equivalent to testing whether m cases in the set I are outliers. In the next section, we consider procedures for testing $H_0 : \boldsymbol{\delta} = \mathbf{0}$ against $H_1 : \boldsymbol{\delta} \neq \mathbf{0}$.

3. OUTLIER TEST

3.1 Likelihood Ratio Test

The likelihood ratio statistic was introduced by Neyman and Pearson (1928). Let $p(Y|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2)$ be the likelihood function for model (2.2). For this particular hypothesis this statistic is given by

$$\Lambda = \frac{\sup\limits_{\boldsymbol{\theta}, \sigma^2} p(Y|\boldsymbol{\theta}, \mathbf{0}, \sigma^2)}{\sup\limits_{\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2} p(Y|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2)}$$

with small value of Λ providing evidence against the null hypothesis. Letting $L(\phi)$ be the log-likelihood evaluated at ϕ, we write

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

Significance level of likelihood ratio tests can be found from the asymptotic distribution of LR, which is the chi-square distribution with m degrees of freedom denoted as $\chi^2(m)$ or by F-approximation (Gallant, 1987, p. 57; Seber and Wild, 1989, p. 198),

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

which is approximately distributed as the F-distribution with m and $n-p-m$ degrees of freedom denoted as $F(m, n-p-m)$ when H_0 is true.

In practice, we usually do not have a prior knowledge of the suspected cases, thus the procedure based on the first Bonferroni inequality (Miller, 1981) should be used to find the significance level of tests. Under this procedure, we use the following rejection rule: max $LR > \chi^2_{\alpha/l}(m)$ or max $F_{LR} > F^2_{\alpha/l}(m, n-p-m)$, where $l = \binom{n}{m}$, $\chi^2_\alpha(m)$ is the upper α point of the chi-square distribution with m degrees of freedom, and $F^2_{\alpha/l}(m, n-p-m)$ is the upper α point of F-distribution with m and $n-p-m$ degrees of freedom. These Bonferroni-type bounds can be useful for small data sets or for situations where only few tests need to be examined, but these critical values are likely to be very conservative. The significance methods developed by Andrews and Pregibon (1978) could also be used, however, they do not provide much help when there are more than 30 observations.

3.2 Wald Test

A second test statistic for the test $\boldsymbol{\delta} = \mathbf{0}$ was proposed by Wald(1943) and is given by,

$$WD = \hat{\boldsymbol{\delta}}^T [\mathrm{Var}(\hat{\boldsymbol{\delta}})]^{-1} \hat{\boldsymbol{\delta}}.$$

Given ϵ_i are i.i.d. $N(0, \sigma^2)$ and under appropriate regularity conditions, we have asymptotically $\hat{\boldsymbol{\phi}} \sim N(\boldsymbol{\phi}, \sigma^2 \boldsymbol{B}^{-1})$ (Seber and Wild, 1989, p. 24), where

$$\boldsymbol{B} = \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\phi}}\bigg|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}} \right)^T \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\phi}}\bigg|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}} \right) = \begin{bmatrix} \widetilde{\boldsymbol{V}}^T \widetilde{\boldsymbol{V}} & \widetilde{\boldsymbol{V}}^T \boldsymbol{D} \\ \boldsymbol{D}^T \widetilde{\boldsymbol{V}} & \boldsymbol{D}^T \boldsymbol{D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix},$$

and $\widetilde{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}}_{(I)})$, which is $\partial \boldsymbol{f}/\partial \boldsymbol{\theta}^T$ evaluated at $\hat{\boldsymbol{\theta}}_{(I)}$. Partitioning $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\delta})$ and $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\theta}}_{(I)}, \hat{\boldsymbol{\delta}})$, we have $\hat{\boldsymbol{\delta}} \sim N(\boldsymbol{\delta}, \sigma^2 (\boldsymbol{B}^{-1})_{22})$, approximately. Using the usual rule to calculate the inverse of a partitioned matrix, we have

$$\begin{aligned} \boldsymbol{B}_{22}^{-1} &= (\boldsymbol{B}_{22} - \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12})^{-1} \\ &= (\boldsymbol{I}_m - \boldsymbol{D}^T\widetilde{\boldsymbol{V}}(\widetilde{\boldsymbol{V}}^T\widetilde{\boldsymbol{V}})^{-1}\widetilde{\boldsymbol{V}}^T\boldsymbol{D})^{-1} \\ &= (\boldsymbol{I}_m - \widetilde{\boldsymbol{H}}_I)^{-1}, \end{aligned}$$

where $\widetilde{\boldsymbol{H}}_I$ is the $m \times m$ minor of $\widetilde{\boldsymbol{H}} = \widetilde{\boldsymbol{V}}(\widetilde{\boldsymbol{V}}^T\widetilde{\boldsymbol{V}})^{-1}\widetilde{\boldsymbol{V}}^T$ with rows and columns indexed by $\boldsymbol{I}$.

The variance of $\hat{\boldsymbol{\delta}}$ is given by

$$\mathrm{Var}(\hat{\boldsymbol{\delta}}) = \sigma^2(\boldsymbol{I} - \widetilde{\boldsymbol{H}}_I)^{-1},$$

and $\mathrm{Var}(\hat{\boldsymbol{\delta}})$ can be estimated by replacing σ^2 by $\hat{\sigma}^2_{(I)}$. Now the Wald test for the hypothesis $\boldsymbol{\delta} = \boldsymbol{0}$ is

$$WD = \frac{1}{\hat{\sigma}^2_{(I)}}\,\hat{\boldsymbol{\delta}}^T(\boldsymbol{I} - \widetilde{\boldsymbol{H}}_I)\hat{\boldsymbol{\delta}}.$$

Like the likelihood ratio test, significance level of the Wald test can be found from the asymptotic distribution of WD, which is $\chi^2(m)$ under appropriate regularity conditions or by F-approximation (Gallant, 1987, p. 48; Seber and Wild, 1989, p. 198),

$$F_{WD} = \frac{1}{m{s^2}_{(I)}}\,\hat{\boldsymbol{\delta}}^T(\boldsymbol{I} - \widetilde{\boldsymbol{H}}_I)\hat{\boldsymbol{\delta}},$$

which is approximately distributed as $F(m, n-p-m)$ when H_0 is true, where ${s^2}_{(I)} = S(\hat{\boldsymbol{\theta}}_{(I)}, \hat{\boldsymbol{\delta}})/(n-p-m)$. When the suspected cases for outliers are unknown, the Bonferroni significance level should be used to carry out the above test. Since $\widetilde{\boldsymbol{V}}$ depends on $\hat{\boldsymbol{\theta}}_{(I)}$, a separate fitting of the deletion model for each set I is necessary to calculate WD.

3.3 Score Test

The score test is a widely applicable method of test construction that provides a convenient alternative to the likelihood ratio test. The score statistic, due originally to Rao(1947) and developed further by Silvey(1959) is

$$S = \boldsymbol{U}(\boldsymbol{\phi}_0)^T \mathcal{I}(\boldsymbol{\phi}_0)^{-1} \boldsymbol{U}(\boldsymbol{\phi}_0),$$

where

$$\boldsymbol{U}(\boldsymbol{\phi}) = \frac{\partial L(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}, \text{ and } \mathcal{I}(\boldsymbol{\phi}) = E(\boldsymbol{I}(\boldsymbol{\phi})) = -E\Big[\frac{\partial^2 L(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}\Big].$$

The score statistic for the test $\boldsymbol{\delta} = \boldsymbol{0}$, considered by Kahng (1993), is given by,

$$S = \frac{1}{\hat{\sigma}^2} \boldsymbol{e}_I{}^T (\boldsymbol{I}_m - \widehat{\boldsymbol{H}}_I)^{-1} \boldsymbol{e}_I,$$

where $\widehat{\boldsymbol{H}}_I$ is the $m \times m$ minor of $\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{V}}(\widehat{\boldsymbol{V}}^T \widehat{\boldsymbol{V}})^{-1} \widehat{\boldsymbol{V}}^T$ with rows and columns indexed by I, $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$ is $\partial \boldsymbol{f} / \partial \boldsymbol{\theta}^T$ evaluated at $\hat{\boldsymbol{\theta}}$. The asymptotic distribution of S is $\chi^2(m)$ under appropriate regularity conditions. Also we have F-approximation (Gallant, 1987, p. 87; Seber and Wild, 1989, p. 198),

$$F_S = \frac{(n-p-m)[\boldsymbol{e}_I{}^T (\boldsymbol{I}_m - \widehat{\boldsymbol{H}}_I)^{-1} \boldsymbol{e}_I]}{m\ [(n-p)s^2 - \boldsymbol{e}_I{}^T (\boldsymbol{I}_m - \widehat{\boldsymbol{H}}_I)^{-1} \boldsymbol{e}_I]},$$

which has $F(m, n-p-m)$ distribution approximately, where $s^2 = S(\hat{\boldsymbol{\theta}}, \boldsymbol{0})/(n-p)$. Again, the Bonferroni significance level should be used to carry out this test if the suspected cases are unknown.

The previous two tests are based on the maximum likelihood estimate, $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\theta}}_{(I)}, \hat{\boldsymbol{\delta}})$ that require refitting of the nonlinear regression model $\binom{n}{m}$ times when the location of outliers is unknown. However, the score test does not require the knowledge of the maximum likelihood estimate $\hat{\boldsymbol{\phi}}$.

3.4 Comparison of Test Statistics

The likelihood ratio statistic compares the height of the likelihood at $\hat{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}_0$. The Wald test statistic compares $\hat{\boldsymbol{\phi}}$ to its standard error. The score test statistic compares the derivatives of the log-likelihood of ϕ_0 to its standard error.

Buse (1982) suggests a simple diagram to compare these three statistics. Suppose that the vector $\boldsymbol{\delta}$ consists of only one element. If we now plot the log-likelihood function, then the value of $\frac{1}{2}LR$ can be read directly from Figure 1. In this figure we note that the distance $\frac{1}{2}LR$ depends on the distance $\hat{\delta}(=\hat{\delta}-\delta_0)$ and the curvature of the log-likelihood function. Instead of considering the difference in log-likelihood, the Wald test takes the squared distance between $\hat{\delta}$ and $\delta_0(=0)$ weighted by the curvature of the log-likelihood function evaluated at $\hat{\delta}$. This curvature is identical to the curvature of the quadratic approximation of the log-likelihood whose first and second derivatives are the same as those of the log-likelihood at $\hat{\delta}$. This is illustrated in Figure 2. The score test takes the squared departure of the slope of the log-likelihood function evaluated at δ_0 from the slope evaluated at $\hat{\delta}$, which is zero, weighted by the inverse of the curvature evaluated at δ_0. This curvature is identical to the curvature of the quadratic approximation of the log-likelihood whose first and second derivatives are the same as those of the log-likelihood at δ_0, not at $\hat{\delta}$. This is illustrated in Figure 3.

If the log-likelihood function is exactly quadratic, which is the linear case, the log-likelihood function and the two quadratic approximations are identical. In this case, the inequality $W \geq LR \geq SC$ holds. This ordering was first established by Berndt and Savin (1977). Although this inequality no longer holds for nonlinear models, Mizon (1977) found that $W \geq LR$ most of the time in his samples.

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 $H_0 : \boldsymbol{\delta} = \mathbf{0}$.

4. ACCURACY OF LINEAR APPROXIMATION

In Section 3 we presented three standard approaches to obtaining test statistics for outliers. One of the tests based on the linear approximation, namely the score test, is easy to calculate, but can be quite different from likelihood ratio tests. In this section, the accuracy of the test, in which the test is based on a linear approximation, is investigated using curvature measures.

4.1 Curvatures

We begin with the standard nonlinear model (2.1). Suppose $\boldsymbol{\theta}$ is close to $\hat{\boldsymbol{\theta}}$, then we have the following quadratic Taylor expansion:

$$\boldsymbol{f}(\boldsymbol{X},\boldsymbol{\theta}) \approx \boldsymbol{f}(\boldsymbol{X},\hat{\boldsymbol{\theta}}) + \widehat{\boldsymbol{V}}\boldsymbol{\kappa} + \frac{1}{2}\boldsymbol{\kappa}^T\widehat{\boldsymbol{W}}\boldsymbol{\kappa}, \tag{4.1}$$

where $\boldsymbol{\kappa} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{W}} = \boldsymbol{W}(\hat{\boldsymbol{\theta}})$ is $\partial^2\boldsymbol{f}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ evaluated at $\hat{\boldsymbol{\theta}}$. If we ignore the quadratic term, we have the linear approximation for $\boldsymbol{\theta}$ in the vicinity of $\hat{\boldsymbol{\theta}}$

$$\boldsymbol{f}(\boldsymbol{X},\boldsymbol{\theta}) \approx \boldsymbol{f}(\boldsymbol{X},\hat{\boldsymbol{\theta}}) + \widehat{\boldsymbol{V}}\boldsymbol{\kappa}. \tag{4.2}$$

This linear approximation amounts to approximating the expectation surface in the neighborhood of $\hat{\boldsymbol{\theta}}$ by the tangent plane at $\hat{\boldsymbol{\theta}}$. An important assumption used in this method is that the expectation surface is flat, so that the tangent plane provides an accurate approximation.

The validity of the linear approximation (4-2) will depend on the magnitude of the quadratic term $\boldsymbol{\kappa}^T\widehat{\boldsymbol{W}}\boldsymbol{\kappa}$ in (4-1) relative to the linear term $\widehat{\boldsymbol{V}}\boldsymbol{\kappa}$. To make this comparison Bates and Watts(1980, 1981) split the quadratic term into two orthogonal components, projections onto the tangent plane and normal to the tangent plane. They define two measures for comparing each quadratic component with the linear term, namely the maximum parameter-effects curvatures and the maximum intrinsic curvatures.

4.2 Subset Curvatures

We next turn to the parameter subsets. Consider the partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_i$ is $p_i \times 1 (i = 1, 2)$ and $p_1 + p_2 = p$. Suppose that $\boldsymbol{\theta}_2$ is the parameter subset of interest.

Let $\boldsymbol{m}_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_2) = \hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$ be the p_1-dimensional vector-valued function that minimizes $S(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$ for $\boldsymbol{\theta}_2$ fixed and let $\dot{\boldsymbol{m}}_{\boldsymbol{\theta}_1}$ and $\ddot{\boldsymbol{m}}_{\boldsymbol{\theta}_1}$ be the first and second partial derivatives, respectively, of $\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$ with respect to $\boldsymbol{\theta}_2$ evaulated at $\hat{\boldsymbol{\theta}}_2$, and define $\boldsymbol{h}(\boldsymbol{\theta}_2) = \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2), \boldsymbol{\theta}_2)$. With these definitions the quadratic approximation of $\boldsymbol{h}(\boldsymbol{\theta}_2)$ about $\hat{\boldsymbol{\theta}}$ can be written as

$$\boldsymbol{h}(\boldsymbol{\theta}_2) \approx \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}) + \widehat{\boldsymbol{V}} \dot{\boldsymbol{m}}_{\boldsymbol{\theta}_1} \boldsymbol{\kappa}_2 + \frac{1}{2} \boldsymbol{\kappa}_2{}^T \dot{\boldsymbol{m}}_{\boldsymbol{\theta}_1}^T \widehat{\boldsymbol{W}} \dot{\boldsymbol{m}}_{\boldsymbol{\theta}_1} \boldsymbol{\kappa}_2 + \frac{1}{2} \widehat{\boldsymbol{V}} (\boldsymbol{\kappa}_2{}^T \ddot{\boldsymbol{m}}_{\boldsymbol{\theta}_1} \boldsymbol{\kappa}_2), \quad (4.3)$$

where $\boldsymbol{\kappa}_2 = (\boldsymbol{\theta}_2 - \hat{\boldsymbol{\theta}}_2)$. Cook and Goldberg (1986) show that the linear part of the above equation describes the plane tangent to $\boldsymbol{h}$ at $\hat{\boldsymbol{\theta}}_2$ and can be reexpressed as

$$\boldsymbol{h}(\boldsymbol{\theta}_2) \approx \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}) + (\boldsymbol{I}_n - \widehat{\boldsymbol{H}}_1) \widehat{\boldsymbol{V}}_2 \boldsymbol{\kappa}_2, \quad (4.4)$$

where $\widehat{\boldsymbol{V}}_1 = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{V}}_2 = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$ are $\partial \boldsymbol{f} / \partial \boldsymbol{\theta}_1^T$ and $\partial \boldsymbol{f} / \partial \boldsymbol{\theta}_2^T$ evaluated at $\hat{\boldsymbol{\theta}}$, respectively, and $\widehat{\boldsymbol{H}}_1 = \widehat{\boldsymbol{V}}_1 (\widehat{\boldsymbol{V}}_1^T \widehat{\boldsymbol{V}}_1)^{-1} \widehat{\boldsymbol{V}}_1^T$.

The validity of (4.4) depends on the magnitude of the quadratic terms in (4.3) relative to the linear term. The global curvature measures may not be relevant in this case, because they measure the worst possible curvatures in any direction from $\hat{\boldsymbol{\theta}}$. To assess the adequacy of the linear approximation (4.4) we need the subset parameter-effects and intrinsic curvatures which were developed by Cook and Goldberg (1986). The following discussion is based on their work.

4.3 Subset Curvatures for $\boldsymbol{\delta}$

We consider the mean shift outlier model (2.2). Since parameter $\boldsymbol{\delta}$ is the parameter subset of this model, the subset curvatures for $\boldsymbol{\delta}$ can be found by the methods described in Section 4.2.

Let $\boldsymbol{h}(\boldsymbol{\delta}) = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{m_\theta}(\boldsymbol{\delta})) + \boldsymbol{D\delta}$, where $\boldsymbol{m_\theta}(\boldsymbol{\delta}) = (\theta_1(\boldsymbol{\delta}), \theta_2(\boldsymbol{\delta}), \ldots, \theta_p(\boldsymbol{\delta}))^T$ denotes the p-dimensional vector-valued function that maximizes $\boldsymbol{L}(\boldsymbol{\theta}, \boldsymbol{\delta})$ over $\boldsymbol{\theta}$ for given $\boldsymbol{\delta}$, that is, $\boldsymbol{m_\theta}(\boldsymbol{\delta})$ represents that value $\boldsymbol{\theta}$ that minimizes $\boldsymbol{S}(\boldsymbol{\theta}, \boldsymbol{\delta})$ for each value of $\boldsymbol{\delta}$. Following Bates and Watts (1980), and Cook and Goldberg (1986), we assume that the intrinsic curvature of $\boldsymbol{h}$ at $\hat{\boldsymbol{\delta}}$ is negligible. If the intrinsic curvature is zero, then the expectation surface and tangent plane are identical and $\boldsymbol{h}(\boldsymbol{\delta})$ is a curve in this plane. The intrinsic curvature should be calculated to check this assumption, however, experiment has shown that they are typically small.

To obtain precise expressions, we need the following definitions. Define $n \times p$ matrix $\widetilde{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}}_{(I)})$ and $n \times p \times p$ array $\widetilde{\boldsymbol{W}} = \boldsymbol{W}(\hat{\boldsymbol{\theta}}_{(I)})$. We consider the QR decomposition of $n \times (p+m)$ matrix $\boldsymbol{V_\phi}$, namely

$$\boldsymbol{V_\phi} = \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\phi}^T}\right|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}} = (\widetilde{\boldsymbol{V}}, \boldsymbol{D}) = \boldsymbol{QR},$$

where $\boldsymbol{Q}$ is an $n \times (p+m)$ matrix with orthogonal columns and $\boldsymbol{R}$ is a $(p+m) \times (p+m)$ upper triangular matrix. Now partition $\boldsymbol{Q}$, $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$ as

$$\boldsymbol{Q} = [\boldsymbol{Q}_1 \quad \boldsymbol{Q}_2],$$

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ 0 & \boldsymbol{R}_{22} \end{bmatrix},$$

$$\boldsymbol{R}^{-1} = \begin{bmatrix} \boldsymbol{R}_{11}{}^{-1} & (\boldsymbol{R}^{-1})_{12} \\ 0 & \boldsymbol{R}_{22}{}^{-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_{11}{}^{-1} & -\boldsymbol{R}_{11}{}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}{}^{-1} \\ 0 & \boldsymbol{R}_{22}{}^{-1} \end{bmatrix},$$

where $\boldsymbol{Q}_1$ is $n \times p$, $\boldsymbol{Q}_2$ is $n \times m$, $\boldsymbol{R}_{11}$ is $p \times p$, $\boldsymbol{R}_{12}$ is $p \times m$, and $\boldsymbol{R}_{22}$ is $m \times m$. Consider the transformation $\check{\boldsymbol{W}}_{\boldsymbol{\phi}} = \boldsymbol{R}^{-T}\widetilde{\boldsymbol{W}}_{\boldsymbol{\phi}}\boldsymbol{R}^{-1}$, where the i-th face of $\widetilde{\boldsymbol{W}}_{\boldsymbol{\phi}}$ is

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

$\widetilde{\boldsymbol{W}}_i$ is the i-th face of $\widetilde{\boldsymbol{W}}$, and h_i is the i-th element of $\boldsymbol{h}(\boldsymbol{\delta})$. Then the i-th face of $\breve{\boldsymbol{W}}_{\phi}$ can be expressed as

$$
\begin{aligned}
(\breve{\boldsymbol{W}}_{\phi})_i &= \begin{bmatrix} \boldsymbol{R}_{11}^{-T} & 0 \\ (\boldsymbol{R}^{-1})_{12}^T & \boldsymbol{R}_{22}^{-T} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{W}}_i & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{11}{}^{-1} & (\boldsymbol{R}^{-1})_{12} \\ 0 & \boldsymbol{R}_{22}{}^{-1} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{R}_{11}^{-T}\widetilde{\boldsymbol{W}}_i\boldsymbol{R}_{11}^{-1} & \boldsymbol{R}_{11}^{-T}\widetilde{\boldsymbol{W}}_i(\boldsymbol{R}^{-1})_{12} \\ (\boldsymbol{R}^{-1})_{12}^T\widetilde{\boldsymbol{W}}_i\boldsymbol{R}_{11}^{-1} & (\boldsymbol{R}^{-1})_{12}^T\widetilde{\boldsymbol{W}}_i(\boldsymbol{R}^{-1})_{12} \end{bmatrix} \\
&= \begin{bmatrix} (\breve{\boldsymbol{W}}_{\phi 11})_i & (\breve{\boldsymbol{W}}_{\phi 12})_i \\ (\breve{\boldsymbol{W}}_{\phi 21})_i & (\breve{\boldsymbol{W}}_{\phi 22})_i \end{bmatrix},
\end{aligned}
$$

where $(\breve{\boldsymbol{W}}_{\phi jk})_i$ is the i-th face of $\breve{\boldsymbol{W}}_{\phi jk}$.

Under the above settings and by applying the result of Cook and Goldberg (1986), we may define the maximum parameter-effects and intrinsic curvatures of $\boldsymbol{h}$ at $\hat{\boldsymbol{\delta}}$ as

$$
\begin{aligned}
\Gamma_s^T(\boldsymbol{\delta}) &= \sqrt{m}\, s_{(I)} \max_{\|\boldsymbol{b}\|=1} \left\| \boldsymbol{b}^T [\boldsymbol{P}_{\boldsymbol{Q}_2}][\breve{\boldsymbol{W}}_{\phi 22}]\boldsymbol{b} \right\| \\
&= \sqrt{m}\, s_{(I)} \max_{\|\boldsymbol{b}\|=1} \left\| \boldsymbol{b}^T \boldsymbol{A}_{22}\boldsymbol{b} \right\| \qquad (4.5)
\end{aligned}
$$

and

$$
\begin{aligned}
\Gamma_s^{\eta}(\boldsymbol{\delta}) &= 2\sqrt{m}\, s_{(I)} \max_{\|\boldsymbol{b}\|=1} \left\| [\boldsymbol{b}^T \boldsymbol{Q}_2^T][\breve{\boldsymbol{W}}_{\phi 12}]\boldsymbol{b} \right\| \\
&= 2\sqrt{m}\, s_{(I)} \max_{\|\boldsymbol{b}\|=1} \left\| [\boldsymbol{b}^T][\boldsymbol{A}_{12}]\boldsymbol{b} \right\|, \qquad (4.6)
\end{aligned}
$$

where $\boldsymbol{A}$ is the $(p+m)\times(p+m)\times(p+m)$ 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 sub-arrays of $\boldsymbol{A}$ with i-th faces $\boldsymbol{A}_{i12}, \boldsymbol{A}_{i22}, i = p+1, \ldots, p+m$.

Combining the two subset curvatures, the total curvature $\boldsymbol{\Gamma}_s(\boldsymbol{\delta})$ of $\boldsymbol{h}$ at $\hat{\boldsymbol{\delta}}$ is

$$\boldsymbol{\Gamma}_s(\boldsymbol{\delta}) = \sqrt{m}\, s_{(I)} \max_{\|\boldsymbol{b}\|=1} \left(\left\| \boldsymbol{b}^T \boldsymbol{A}_{22} \boldsymbol{b} \right\|^2 + 4 \left\| [\boldsymbol{b}^T][\boldsymbol{A}_{12}]\boldsymbol{b} \right\|^2 \right)^{1/2} . \qquad (4.7)$$

Cook and Goldberg (1986) noted that in terms of the geometric interpretation by Bates and Watts (1981) the intrinsic subset curvature depends only on the fanning and torsion components of $\boldsymbol{A}$ and not on the compansion and arcing components. If $\boldsymbol{\Gamma}_s(\boldsymbol{\delta})$ (or both $\boldsymbol{\Gamma}_s^{\boldsymbol{T}}(\boldsymbol{\delta})$ and $\boldsymbol{\Gamma}_s^{\boldsymbol{\eta}}(\boldsymbol{\delta})$) is sufficiently small, the likelihood and linear confidence regions for $\boldsymbol{\delta}$ will be similar, otherwise we can expect these confidence regions to be dissimilar. Following Ratkowsky (1983, p. 18) and Cook and Goldberg (1986), $1/(2\sqrt{F_\alpha(m, n-p-m)})$ 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 $\boldsymbol{\Gamma}_s(\boldsymbol{\delta})$ is greater than the guide, the linearization based test, namely the score test, is quite different from likelihood ratio tests.

5. EXAMPLE

To illustrate the results of Sections 3 and 4, we present a numerical example using the data and model taken from Ratkowsky (1983, p. 88). The data examines the water content of bean root cells as a function of the distance from tip in 15 cases. The proposed model is the Gompertz model,

$$f(x_i, \boldsymbol{\theta}) = \theta_1 \exp(-\exp(\theta_2 - \theta_3 x_i)).$$

First we assume that we have a single outlier ($m = 1$) with location unknown. For each of the three test procedures, we calculate outlier test statistics for each case. Figure 4(a) shows the pairwise plot of three statisitcs, LR, WD, and S. This plot shows that the relationship between LR, WD, and S are close to linear and the order of 15 values are the same for all three procedures, which implies that we can use any procedure to get the location that has the largest test statistic. Next, we assume that there are two outliers ($m = 2$) and calculate 105 test statistics for each pair from 15 locations. The pairwise plot is shown in Figure 4(b). In this plot we have unusual points which represent the subset of cases 4 and 5. This indicates the disagreement between the test statistic based on likelihood and that on linear approximation. In this example, this does not cause serious problems in finding the most likely outlying cases since this subset has small values of LR, WD, and S, however, if the disagreement occurs at large values of statistics, WD and S may give misleading results.

The subset curvatures for δ are listed in Table 1(a) for a single outlier case. The corresponding guide is $1/(2\sqrt{F_{.05}(1,11)}) = .2272$. The subset curvatures in Table 1(a) are all quite small compared to the guide, indicating reasonable agreement between the test statistic based on the likelihood and that on linear approximation. Also, we calculate the subset curvatures for each of the 105 pairs from 15 locations and the 10 largest total subset curvatures for $\boldsymbol{\delta}$ are listed in Table 1(b). The corresponding guide is $1/(2\sqrt{F_{.05}(2,10)}) = .2468$. For the subset with locations 4 and 5, the subset curvature measure exceeds the guide, indicating inadequacy of the linear approximation. These results agree with the findings in the previous paragraph.

6. REMARKS

The quantities in formulas (4.5), (4.6) and (4.7) can be found or estimated without knowing the response, $\boldsymbol{Y}_I$, of the case which is suspected to be an outlier. This implies that the curvature measures, $\boldsymbol{\Gamma}_s^T(\boldsymbol{\delta})$, $\boldsymbol{\Gamma}_s^\eta(\boldsymbol{\delta})$, and $\boldsymbol{\Gamma}_s(\boldsymbol{\delta})$, do not depend on how large $\boldsymbol{\delta}$ is or how severe the outliers are. We may have larger curvature measures for $\boldsymbol{\delta}$ even if the cases that are being tested have small test statistics.

One problem that has not yet been solved is the following. If the curvature measures (4.5), (4.6) and (4.7) can be found from a fit of the full data set, we can assess the accuracy of the linear approximation based outlier test for each subset of size m prior to fitting all deletion models. Then we can use tests based on the linear approximation, such as the score test, for the subsets in which the linear approximation is valid. This substantially reduces the computational cost for detecting outliers. Thus, it is desirable to express curvature measures as a function of the full set of data, as is usual in this kind of an investigation. However, it is not easy to get this expression in nonlinear regression because $\widetilde{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}}_{(I)})$ and $\widetilde{\boldsymbol{W}} = \boldsymbol{W}(\hat{\boldsymbol{\theta}}_{(I)})$ change when cases are deleted.

Table 1. Subset Curvatures for $\boldsymbol{\delta}$

(a) m = 1

I	$\boldsymbol{\Gamma}_s^T(\delta)$	$\boldsymbol{\Gamma}_s^\eta(\delta)$	$\boldsymbol{\Gamma}_s(\delta)$
3	0.0140	0.0699	0.0713
5	0.0158	0.0613	0.0633
15	0.0103	0.0386	0.0400
4	0.0079	0.0348	0.0357
2	0.0023	0.0284	0.0285
7	0.0038	0.0246	0.0249
6	0.0012	0.0244	0.0244
8	0.0028	0.0171	0.0173
9	0.0003	0.0124	0.0124
10	0.0010	0.0121	0.0121
14	0.0021	0.0110	0.0112
11	0.0001	0.0054	0.0054
1	0.0000	0.0027	0.0027
12	0.0000	0.0027	0.0027
13	0.0001	0.0026	0.0026

$1/(2\sqrt{F_{.05}(1,11)}) = .2272$

(b) m = 2

I	$\boldsymbol{\Gamma}_s^T(\boldsymbol{\delta})$	$\boldsymbol{\Gamma}_s^\eta(\boldsymbol{\delta})$	$\boldsymbol{\Gamma}_s(\boldsymbol{\delta})$
4, 5	0.2278	0.2615	0.3468
3, 4	0.0955	0.2254	0.2448
3, 5	0.1025	0.1067	0.1480
2, 3	0.0286	0.1345	0.1375
14,15	0.0529	0.1150	0.1266
5, 9	0.0362	0.1162	0.1217
2, 5	0.0516	0.1093	0.1209
3,10	0.0290	0.1136	0.1172
5, 6	0.0298	0.1133	0.1172
3, 9	0.0308	0.1118	0.1160

$1/(2\sqrt{F_{.05}(2,10)}) = .2468$

Figure 1. Likelihood Ratio Test.

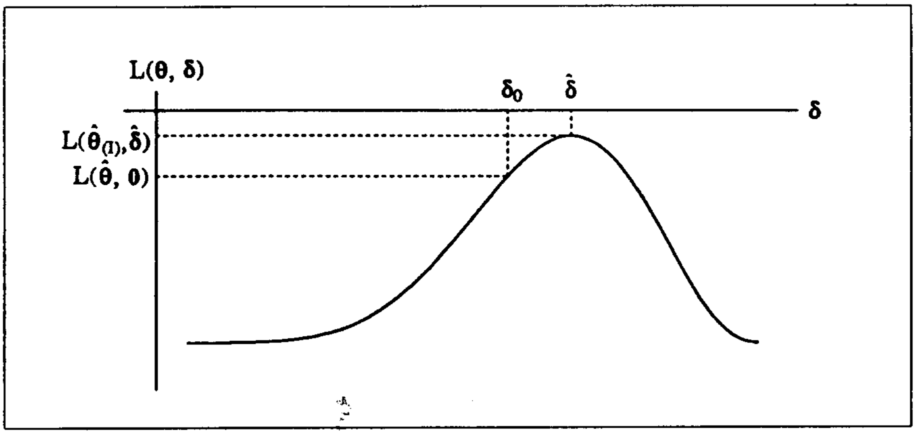


Figure 2. Wald Test.

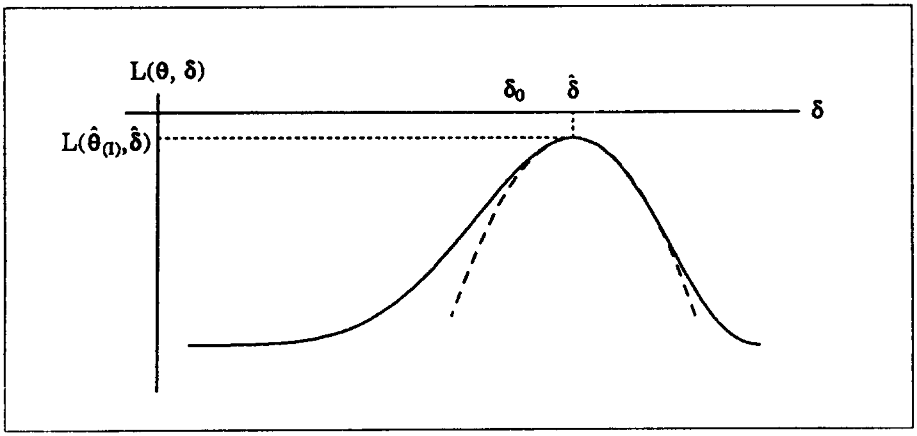


Figure 3. Score Test.

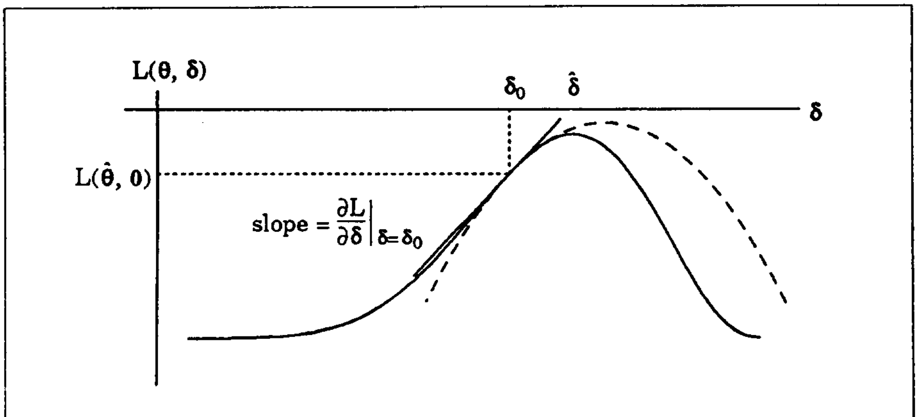
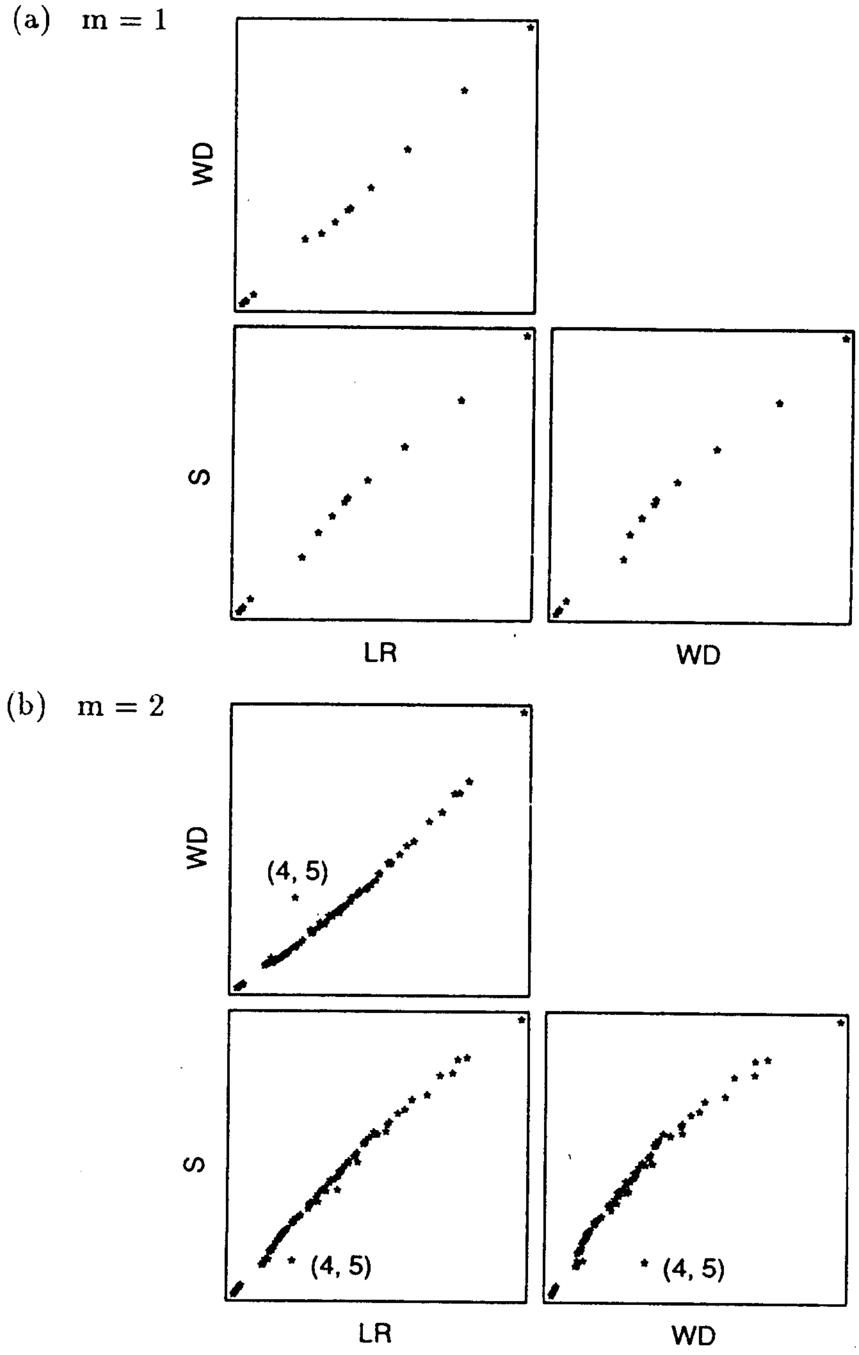


Figure 4. Pairwise Plots of Three Test Statistics.



(a) m = 1

(b) m = 2

REFERENCES

(1) Andrews, D.F. and Pregibon, D. (1978). Finding the Outliers that Matter. *Journal of the Royal Statistical Society Series*, **B**, **40**, 85–93.

(2) Anscombe, F.J. (1960). Rejection of Outliers. *Technometrics*, **2**, 123–167.

(3) Anscombe, F.J. and Tukey, J.W. (1963). The Examination and Analysis of Residuals. *Technometrics*, **5**, 141–160.

(4) 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.

(5) 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.

(6) Beckman, R.J. and Cook, R.D. (1983). Outliers. *Technometrics*, **25**, 119–149.

(7) Berndt, E.R. and Savin, N.E. (1977). Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model. *Econometrica*, **45**, 1263–1278.

(8) Buse, A. (1982). The Likelihood Ratio, Wald, Lagrange Multiplier Tests: An Expository Note. *The American Statistician*, **36**, 153–157.

(9) Cook, R.D. and Goldberg, M.L. (1986). Curvatures for Parameter Subsets in Nonlinear Regression. *The Annals of Statistics*, **14**, 1399–1418.

(10) Gallant, A.R. (1987). *Nonlinear Statistical Models.* John Wiley and Sons, New York.

(11) Kahng, M.-W. (1993). A Score Test for Detection of Outliers in Nonlinear Regression. *Journal of the Korean Statistical Society*, **22**, 201–208.

(12) Miller, R.G., Jr. (1981). *Simultaneous Inference, 2nd ed.* Springer, New York.

(13) Mizon, G.E. (1977). Inferential Procedures in Nonlinear Models: An Application in a UK Industrial Cross Section Study for Factor Substitution and Returns to Scale. *Econometrica*, **45**, 1221–1242.

(14) 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.

(15) 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.

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

(17) Rosner, B. (1975). On the Detection of Many Outliers. *Technometrics*, **17**, 221–227.

(18) Seber, G.A.F. and Wild, C.J. (1989). *Nonlinear Regression*. John Wiley and Sons, New York.

(19) Silvey, S.D. (1959). The Lagrangian Multiplier Test. *The Annals of Mathematical Statistics*, **30**, 389–407.

(20) 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.