Journal of the Korean
Statistical Society
Vol. 22, No. 2, 1993

A Score Test for Detection of Outliers in Nonlinear Regression

Myung-Wook Kahng[1]

ABSTRACT

Given the specific mean shift outlier model, the score test for multiple outliers in nonlinear regression is discussed as an alternative to the likelihood ratio test. The geometric interpretation of the score statistic is also presented.

KEYWORDS: Outlier, Efficient score, Information matrix, Mean shift outlier model, Score test.

1. INTRODUCTION

For routine regression diagnostic work, we prefer hypothesis testing which can be easily constructed using standard regression software. Methods that are based on the maximum likelihood estimates often require special and complicated programs, and are not well suited for this purpose. The score statistic provides a suitable diagnostic test.

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, we discuss standard approaches to obtaining score statistic for outliers and provide its geometric interpretation.

In other applications, diagnostics based on the score test have been proposed by Atkinson(1981, 1982), Cook and Weisberg(1983), Lawrance(1987), Tsai(1988), St. Laurent(1990), and others.

[1] Department of Statistics, Sookmyung Women's University, Yongsan-ku, Seoul, 140-742, Korea

2. OUTLIERS IN NONLINEAR REGRESSION

The standard nonlinear regression model can be expressed as

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

in which the i-th response y_i is related to the q-dimensional vector of known explanatory variable $\mathbf{x}_i$ 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 will write,

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

where $\mathbf{Y}$ is an n-dimensional vector with elements $y_1, y_2, \cdots, y_n$, $\mathbf{X}$ is an $n \times q$ matrix with rows $\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_n^T$, $\boldsymbol{\epsilon}$ is an n-dimensional vector with elements $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$, and $\mathbf{f}(\mathbf{X}, \boldsymbol{\theta}) = (f(\mathbf{x}_1, \boldsymbol{\theta}), f(\mathbf{x}_2, \boldsymbol{\theta}), \cdots, f(\mathbf{x}_n, \boldsymbol{\theta}))^T$.

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

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

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

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

where $\boldsymbol{\delta} = (\delta_{i1}, \delta_{i2}, \cdots, \delta_{im})^T$, and $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_m)$, and $\mathbf{d}_j$ is the i_j-th standard basis vector for $\mathbf{R}^n$.

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

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

where $S(\boldsymbol{\theta}, \boldsymbol{\delta}) = (\mathbf{Y} - \mathbf{f}(\mathbf{X}, \boldsymbol{\theta}) - \boldsymbol{D\delta})^T (\mathbf{Y} - \mathbf{f}(\mathbf{X}, \boldsymbol{\theta}) - \boldsymbol{D\delta})$. Given σ^2, (2.3) is maximized with respect to $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\delta})$ when $\mathrm{S}(\boldsymbol{\theta}, \boldsymbol{\delta})$ 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 $\boldsymbol{\phi} = (\hat{\boldsymbol{\theta}}_{(I)}, \hat{\boldsymbol{\delta}})$ 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).

The testing of the hypothesis $\boldsymbol{\delta} = \mathbf{0}$ is equivalent to testing whether m cases in the set $\mathbf{I}$ 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. SCORE TEST

The score test or Lagrange multiplier 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 given by,

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

where,

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

We have the vector of efficient score,

$$\mathbf{U}(\boldsymbol{\phi}) = \mathbf{U}(\boldsymbol{\theta}, \boldsymbol{\delta}) = (\frac{\partial L(\boldsymbol{\delta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad \frac{\partial L(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial \boldsymbol{\delta}})^T,$$

and the observed information matrix,

$$\mathbf{I}(\boldsymbol{\phi}) = \mathbf{I}(\boldsymbol{\theta}, \boldsymbol{\delta}) = \begin{bmatrix} \mathbf{I}_{11}(\boldsymbol{\theta}, \boldsymbol{\delta}) & \mathbf{I}_{12}(\boldsymbol{\theta}, \boldsymbol{\delta}) \\ \mathbf{I}_{21}(\boldsymbol{\theta}, \boldsymbol{\delta}) & \mathbf{I}_{22}(\boldsymbol{\theta}, \boldsymbol{\delta}) \end{bmatrix} = \begin{bmatrix} -\frac{\partial^2 L(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} & -\frac{\partial^2 L(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\delta}^T} \\ -\frac{\partial^2 L(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\theta}^T} & -\frac{\partial^2 L(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} \end{bmatrix},$$

where each of the first derivatives is given by

$$\frac{\partial L(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial \boldsymbol{\theta}} = \frac{1}{\sigma^2}(\mathbf{Y} - \mathbf{f}(\mathbf{X}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{\delta})^T (\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}^T}),$$
$$\frac{\partial L(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \frac{1}{\sigma^2}(\mathbf{Y} - \mathbf{f}(\mathbf{X}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{\delta})^T \mathbf{D},$$

and the second derivatives are given by

$$\frac{\partial^2 L(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = -\frac{1}{\sigma^2}((\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}^T})^T (\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}^T}) - (\frac{\partial^2 \mathbf{f}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T})(\mathbf{Y} - \mathbf{f}(\mathbf{X}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{\delta})),$$
$$\frac{\partial^2 L(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\delta}^T} = -\frac{1}{\sigma^2}(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}^T})^T \mathbf{D},$$

$$\frac{\partial^2 L(\boldsymbol{\theta},\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^T} = -\frac{1}{\sigma^2}\mathbf{D}^T\mathbf{D} = -\frac{1}{\sigma^2}\mathbf{I}_m.$$

Let $\mathbf{e}$ be the n-dimensional ordinary residual vector, where $\mathbf{e} = \mathbf{Y} - \mathbf{f}(\mathbf{X},\hat{\boldsymbol{\theta}})$. We define $\mathbf{e}_I$ to be m-vectors whose j-th element is e_{ij}. Under the null hypothesis $\boldsymbol{\delta} = \mathbf{0}$, the efficient score $\mathbf{U}(\boldsymbol{\phi}_0)$ is given by

$$\mathbf{U}(\boldsymbol{\phi}_0) = \mathbf{U}(\hat{\theta},\mathbf{0}) = \frac{1}{\hat{\sigma}^2}\begin{bmatrix}\mathbf{0}\\ \\ \mathbf{D}^T(\mathbf{Y}-\mathbf{f}(\mathbf{X},\hat{\boldsymbol{\theta}}))\end{bmatrix} = \frac{1}{\hat{\sigma}^2}\begin{bmatrix}\mathbf{0}\\ \\ \mathbf{e}_I\end{bmatrix},$$

and the observed information $\mathbf{I}(\boldsymbol{\phi}_0)$ is

$$\mathbf{I}(\boldsymbol{\phi}_0) = \mathbf{I}(\hat{\boldsymbol{\theta}},\mathbf{0}) = \begin{bmatrix}\mathbf{I}_{11}(\hat{\boldsymbol{\theta}},\mathbf{0}) & \mathbf{I}_{12}(\hat{\theta},\mathbf{0})\\ \mathbf{I}_{21}(\hat{\boldsymbol{\theta}},\mathbf{0}) & \mathbf{I}_{22}(\hat{\boldsymbol{\theta}},\mathbf{0})\end{bmatrix} = \frac{1}{\hat{\sigma}^2}\begin{bmatrix}\hat{\mathbf{V}}^T\hat{\mathbf{V}} - \sum_{i=1}^{n} e_i\hat{\mathbf{W}}_i & \hat{\mathbf{V}}^T\mathbf{D}\\ \mathbf{D}^T\hat{\mathbf{V}} & \mathbf{I}_m\end{bmatrix},$$

where $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{W}} = \mathbf{W}(\hat{\boldsymbol{\theta}})$ are $\partial\mathbf{f}/\partial\boldsymbol{\theta}^T$ and $\partial^2\mathbf{f}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ evaluated at $\hat{\theta}$, respectively. The expected information matrix is $\mathcal{I}(\boldsymbol{\phi}_0)$

$$\mathcal{I}(\boldsymbol{\phi}_0) = \mathcal{I}(\hat{\boldsymbol{\theta}},\mathbf{0}) = E(\mathbf{I}(\hat{\boldsymbol{\theta}},\mathbf{0})) = \begin{bmatrix}\mathcal{I}_{11}(\hat{\boldsymbol{\theta}},\mathbf{0}) & \mathcal{I}_{12}(\hat{\boldsymbol{\theta}},\mathbf{0})\\ \mathcal{I}_{21}(\hat{\boldsymbol{\theta}},\mathbf{0}) & \mathcal{I}_{22}(\hat{\boldsymbol{\theta}},\mathbf{0})\end{bmatrix} = \frac{1}{\hat{\sigma}^2}\begin{bmatrix}\hat{\mathbf{V}}^T\hat{\mathbf{V}} & \hat{\mathbf{V}}^T\mathbf{D}\\ \\ \mathbf{D}^T\hat{\mathbf{V}} & \mathbf{I}_m\end{bmatrix}.$$

Using $\mathcal{I}(\boldsymbol{\phi}_0)^{-1}$ to estimate variance, the score statistic for the test $\boldsymbol{\delta} = \mathbf{0}$ is

$$\begin{aligned}
S &= \mathbf{U}(\boldsymbol{\phi}_0)^T\mathcal{I}(\boldsymbol{\phi}_0)^{-1}\mathbf{U}(\boldsymbol{\phi}_0)\\
&= \frac{1}{\hat{\sigma}^2}\begin{bmatrix}\mathbf{0} & \mathbf{e}_I^T\end{bmatrix}\begin{bmatrix}\hat{\mathbf{V}}^T\hat{\mathbf{V}} & \hat{\mathbf{V}}^T\mathbf{D}\\ \\ \mathbf{D}^T\hat{\mathbf{V}} & \mathbf{I}_m\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{0}\\ \\ \mathbf{e}_I\end{bmatrix}\\
&= \frac{1}{\hat{\sigma}^2}\mathbf{e}_I{}^T(\mathbf{I}_m - \mathbf{D}^T\hat{\mathbf{V}}(\hat{\mathbf{V}}^T\hat{\mathbf{V}})^{-1}\hat{\mathbf{V}}^T\mathbf{D})^{-1}\mathbf{e}_I\\
&= \frac{1}{\hat{\sigma}^2}\mathbf{e}_I{}^T(\mathbf{I}_m - \hat{\mathbf{H}}_I)^{-1}\mathbf{e}_I,
\end{aligned}$$

where $\hat{\mathbf{H}}_I$ is the $m\times m$ minor of $\hat{\mathbf{H}} = \hat{\mathbf{V}}(\hat{\mathbf{V}}^T\hat{\mathbf{V}})^{-1}\hat{\mathbf{V}}^T$ with rows and columns indexed by $\mathbf{I}$. The asymptotic distribution of S is a chi-square distribution with m degrees of freedom under appropriate regularity conditions. (Gallant, 1987, p. 87; Seber and Wild, 1989, p. 230)

When the candidate cases for outliers are unknown, the test is usually based on the maximum value of S for all subsets of size m. A multiple testing procedure, such as one based on the Bonferroni inequality must be used to find significant levels.

If the observed information $\mathbf{I}(\boldsymbol{\phi}_0)$ is substituted for the expected information $\mathcal{I}(\boldsymbol{\phi}_0)$, the statistic is

$$\begin{aligned} S_o &= \mathbf{U}(\boldsymbol{\phi}_0)^T \mathbf{I}(\boldsymbol{\phi}_0)^{-1} \mathbf{U}(\boldsymbol{\phi}_0) \\ &= \frac{1}{\hat{\sigma}^2} \begin{bmatrix} \mathbf{0} & \mathbf{e}_I^T \end{bmatrix} \begin{bmatrix} \hat{\mathbf{V}}^T\hat{\mathbf{V}} - \sum_{i=1}^{n} e_i \hat{\mathbf{W}}_i & \hat{\mathbf{V}}^T\mathbf{D} \\ \mathbf{D}^T\hat{\mathbf{V}} & \mathbf{I}_m \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{e}_I \end{bmatrix} \\ &= \frac{1}{\hat{\sigma}^2} \mathbf{e}_I{}^T (\mathbf{I}_m - \mathbf{D}^T\hat{\mathbf{V}}(\hat{\mathbf{V}}^T\hat{\mathbf{V}} - \sum_{i=1}^{n} e_i \hat{\mathbf{W}}_i)^{-1} \hat{\mathbf{V}}^T\mathbf{D})^{-1} \mathbf{e}_I. \end{aligned}$$

The asymptotic properties of this score test are not changed (Cox and Hinkley, 1974, p. 302; Atkinson, 1985, p. 93).

4. EXAMPLE

To illustrate the results of the previous section we present a numerical example using the data and the model taken from Clarke (1987). The data examines the weight of cut grass as a function of the weeks after commencement of grazing in a pasture for 13 cases. The proposed model is the Mitcherlitz equation,

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

First we assume that we have a single outlier (m = 1) with location unknown. The score statistics for each case are calculated and listed in Table 1(a). At nominal level 0.05, the score test for a single outlier will reject if maximum S is larger than $\chi^2{}_{0.05/13}(1) = 8.355$, where $\chi^2{}_\alpha(\mathrm{m})$ is the upper α point of the chi-square distribution with m degrees of freedom. Since the maximum score statistic, S = 5.52599 for case 6, is less than this value, no evidence is provided that this case is an outlier. However, if we suspect in advance that case 6 is an outlier, the critical value $\chi^2{}_{0.05}(1) = 3.841$, would suggest that this case is an outlier. Next, we assume that there are two outliers (m = 2) and the 10 largest statistics among 78 test statistics for each pair of 13 cases are listed in Table 1(b). Since the critical value for the multiple outlier test at level 0.05 based on Bonferroni bound is $\chi^2{}_{0.05/78}(2) = 14.705$, none of the pairs would be declared as outliers by this test.

Table 1. Score Statistics

(a) m = 1

I	S
6	5.52599
13	2.22498
7	1.75233
1	1.73959
5	1.58503
12	1.51284
3	1.01087
9	0.95109
2	0.70646
10	0.47200
4	0.02434
8	0.01529
11	0.00451

(b) m = 2

I		S
6,	7	8.93116
6,	12	7.22563
6,	13	7.04883
1,	6	6.55839
5,	6	6.12461
2,	6	6.07855
3,	6	6.02944
6,	9	5.96888
4,	6	5.88413
6,	10	5.76124

5. REMARKS

The likelihood ratio test is based on the maximum likelihood estimate, $\hat{\phi} = (\hat{\theta}_{(I)}, \hat{\delta})$ that require refitting of the nonlinear regression model $\binom{n}{m}$ times when the location of outliers is unknown. The score test does not require the knowledge of the maximum likelihood estimate $\hat{\phi}$; fitting of model (2.1) is all that is needed.

The score statistic compares the derivatives of the log-likelihood at ϕ_0 to its standard error. Buse(1982) suggests a simple diagram which represents the score statistics. Suppose that the vector δ consists of only one element. If we now plot the log-likelihood function, the score statistic 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. This is illustrated in Figure 1. If the log-likelihood function is exactly quadratic, which is the linear model case, the log-likelihood function and the quadratic approximation are identical.

The score test can be quite different from the likelihood ratio test since the former is based on the linear approximation. An important assumption used in these methods is that the expectation surface in the neighborhood of $\hat{\phi}$ is flat, so

that the tangent plane at $\hat{\phi}$ provides an accurate approximation. The accuracy of the score test can be investigated using curvature measures, and needs further study.

REFERENCES

(1) Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika,* 68,13-20.

(2) Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables (with discussion). *Journal of the Royal Statistical Society, Ser. B,* 44, 1-36.

(3) Atkinson, A. C. (1985). *Plots, Transformations and Regression.* Oxford University Press: Oxford.

(4) Buse, A. (1982). The likelihood, Wald, Lagrange multiplier tests: An expository note. *The American Statistician,* 36, 153-157.

(5) Clarke, G. P. Y. (1987). Approximate confidence limits for a parameter function in nonlinear regression. *Journal of the American Statistical Association,* 82, 221-230.

(6) Cook, R. D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika,* 70,1-10.

(7) Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall: London.

(8) Gallant, A. R. (1987). *Nonlinear Statistical Models.* John Wiley & Sons: New York.

(9) Lawrance, A. J. (1987). The score statistics for regression transformation. *Biometrika,* 74,275-279.

(10) Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Comb. Phil. Soc.,* 44, 50-57.

(11) St. Laurent, R. T. (1990). The equivalence of the Milliken-Graybill procedure and the score test. *The American Statistician,* 44, 36-37.

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

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

(14) Tsai, C. L. (1988). Power transformations and reparameterizations in nonlinear regression models. *Technometrics,* 30, 441-448.

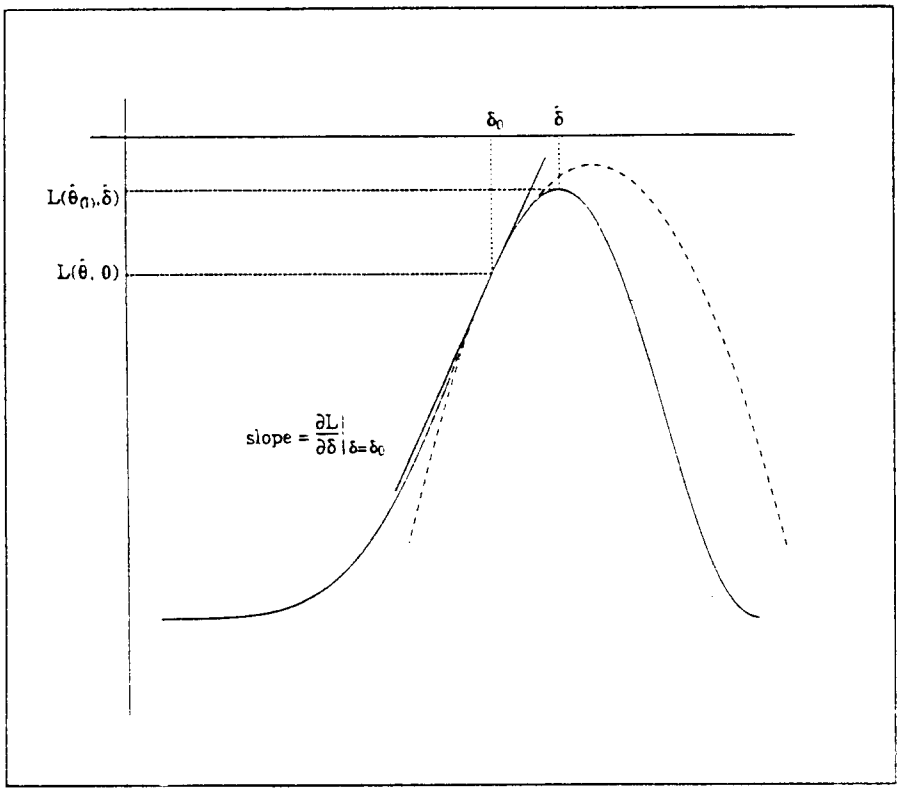

Figure 1. Score Statistic for Testing $H_0 : \delta = \delta_0$