한국통계학회논문집 제3권 2호
1996년 8월 pp. 1-11

Computational Methods for Detection of Multiple Outliers in Nonlinear Regression[1)]

Myung-Wook Kahng[2)]

Abstract

The detection of multiple outliers in nonlinear regression models can be computationally not feasible. As a compromise approach, we consider the use of simulated annealing algorithm, an approximate approach to combinatorial optimization. We show that this method ensures convergence and works well in locating multiple outliers while reducing computational time.

1. Introduction

In this article we consider the computational methods for detection of multiple outliers for the nonlinear regression model. We use the likelihood ratio test statistic as an indication of the prospect of the corresponding observations being outliers. Given m outliers from n observations, we consider all $\binom{n}{m}$ partitions of the data set obtained by specifying subsets of size m. If we were to calculate the likelihood ratio test statistic for all partitions of the data, we would examine their sizes, and the largest test statistic is used to detect the m most likely outlying cases. Thus, identifying the set I of m most likely outlying cases implies finding the set I that maximizes the test statistics over all possible subsets of size m, which requires $\binom{n}{m}$ fittings. Even for modest n, if m is bigger than 2 or 3, this can be very expensive.

The procedure that generates the optimal subset of size m using an algorithm which reduces computational time is developed. We also describe a method that does not require refitting models for every subset and examine its accuracy.

---

1) This work was supported by Sookmyung Women's University Research Fund in 1996.
2) Associate Professor, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea.

2. Outliers in Nonlinear Regression

The standard nonlinear regression model can be expressed as

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

in which the i-th response y_i is related to the q-dimensional vector of known explanatory variable x_i through the known model function f, which depends on the p-dimensional unknown parameter vector $\boldsymbol{\theta}$, and ε_i is error. We assume that f is 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,

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

where $\boldsymbol{y}$ is an n-dimensional vector with elements $y_1, y_2, \cdots, y_n$, $\boldsymbol{X}$ is an $n \times q$ matrix with rows $\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \cdots, \boldsymbol{x}_n^T$, $\boldsymbol{\varepsilon}$ is an n-dimensional vector with elements $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$, and $\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) = (f(\boldsymbol{x}_1, \boldsymbol{\theta}), f(\boldsymbol{x}_2, \boldsymbol{\theta}), \cdots, f(\boldsymbol{x}_n, \boldsymbol{\theta}))^T$. Suppose we suspect in advance that m cases indexed by an m-vector $\boldsymbol{I} = (i_1, i_2, \cdots, i_m)$ are outliers. It can be helpful to write the model as

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

In matrix notation we may write,

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

where $\boldsymbol{\delta} = (\delta_{i_1}, \delta_{i_2}, \cdots, \delta_{i_m})^T$, $\boldsymbol{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_m)$ and $\boldsymbol{d}_j$ is the i_j-th standard basis vector for R^n.

Let $\boldsymbol{e}$ be the n-dimensional ordinary residual vector, where $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\theta}}$ is the least squares estimate of $\boldsymbol{\theta}$. We define $\boldsymbol{y}_I$, $\boldsymbol{\varepsilon}_I$, and $\boldsymbol{e}_I$ to be m-vectors whose j-th elements are y_{i_j}, ε_{i_j}, and e_{i_j}, respectively. Also we define $\boldsymbol{y}_{(I)}$, $\boldsymbol{\varepsilon}_{(I)}$, and $\boldsymbol{e}_{(I)}$ to be

vectors $\boldsymbol{y}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{e}$, respectively, with cases indexed by $\boldsymbol{I}$ deleted. Least squares estimation of the parameter $\boldsymbol{\delta}$ will give a value of zero for the residuals indexed by $\boldsymbol{I}$ in model (2.1). This means that the observations indexed by $\boldsymbol{I}$ will make no contribution to estimate $\boldsymbol{\theta}$, and thus the least squares estimate of $\boldsymbol{\theta}$ in model (2.1) is the same as that in the deletion model,

$$y_i = f(x_i, \theta) + \varepsilon_i, \quad \text{for } i \notin I \qquad \text{or} \qquad y_{(I)} = f(X_{(I)}, \theta) + \varepsilon_{(I)}. \tag{2.2}$$

The resulting estimates of $\boldsymbol{\theta}$ from (2.2) or from (2.1) will be called $\hat{\boldsymbol{\theta}}_{(I)}$, from which it is immediate that $\hat{\boldsymbol{\delta}} = \boldsymbol{y}_I - f(\boldsymbol{X}_I, \hat{\boldsymbol{\theta}}_{(I)})$.

The testing of the hypothesis $\boldsymbol{\delta} = \boldsymbol{0}$ is equivalent to testing whether the set $\boldsymbol{I}$ of m cases are outliers. Thus the outlier identification and testing are formally equivalent to solving and testing a subset regression. The likelihood ratio statistic for this particular hypothesis is given by

$$LR = n[\ \log S(\hat{\theta}, 0) - \log S(\hat{\theta}_{(I)}, \hat{\delta})]\ , \tag{2.3}$$

where $S(\theta, \delta) = (y - f(X, \theta) - D\delta)^T (y - f(X, \theta) - D\delta)$. Significance 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 when H_0 is true.

3. Simulated Annealing

The simulated annealing approach to combinatorial optimization was developed by Kirkpatrick, Gelatt, and Vecchi (1983). This algorithm is based on the algorithm by Metropolis who attempted to simulate the behavior of an ensemble of atoms in equilibrium at a given temperature. In the statistical context, Bonomi and Lutton (1984) applied it to solve the traveling salesman problem, Lundy (1985) used it to the construction of evolutionary trees, and Bohachevsky, Johnson, and Stein (1986) and Haines (1987) applied the algorithm to the calculation of exact optimum experimental design. Recently, Atkinson and Weisberg (1991) applied this algorithm to the multiple outlier detection procedure in the linear regression model. In this section we consider the application of the annealing algorithm to identifying the subset of m most outlying cases in the nonlinear regression model.

3.1 Generalized Simulated Annealing Method

In physics, annealing is a thermal process of heating up a solid until it melts, followed by cooling it down until it obtains low energy state in a heat bath. Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) introduced an algorithm for simulating the behavior of the solid in a heat bath to thermal equilibrium. Suppose that a current state, say i, has energy E_i, then a subsequent state j is generated. The energy of the next stage is E_j. If the energy difference, $E_j - E_i$, is less than or equal to 0, state j is accepted as a current state. If the energy difference is greater than 0, state j is accepted with a certain probability which is given by $\exp[\ (E_i - E_j)/(\ c_b T)]$ where T denotes the temperature of the heat bath and c_b is a constant known as the Boltzman constant. This acceptance rule described above is the Metropolis criterion and the algorithm that goes with it is known as the Metropolis algorithm.

By applying the Metropolis algorithm, we can generate a solution for a combinatorial optimization problem. Suppose f is the object function to be minimized or maximized. Simulated annealing is a convenient way of finding the global extremum of a function that has many local extrema. The method is a biased random walk that samples the object function in the space of independent variables. Unlike most optimization methods, the algorithm moves not only in beneficial directions but also in detrimental directions with some probability to allow for escape from local extrema. A typical feature of the simulated annealing algorithm is that, initially, large deteriorations are accepted; as algorithm proceeds, only small deteriorations are accepted and finally, no deteriorations are accepted. The conditional acceptance probability depends on the increment or decrement of the object function, and it becomes smaller as the algorithm proceeds. We say that the system is cooler if this probability is smaller. Formally, we choose the initial point and calculate the initial value of object function f_0. Next, by using an appropriate method a new point is chosen and the new value of the object function f_1 is calculated. A new step is accepted with probability p given by

$$p = \begin{cases} 1 & \text{if } \Delta f \text{ is in the beneficial direction} \\ \exp(-\beta \Delta f) & \text{otherwise} \end{cases}$$

where $\Delta f = f_1 - f_0$ and β is a control parameter. Thus the beneficial steps are accepted unconditionally but the detrimental steps are accepted according to the probability p. The control parameter β should be chosen to satisfy the properties of the system. Execution of the algorithm is terminated if the step remains unchanged for a number of consecutive

searches.

3.2 An Outlier Problem

Atkinson and Weisberg (1991) applied the simulated annealing algorithm to the multiple outliers detection procedure in linear regression models. This algorithm in linear regression may not be very useful because exact computations for moderate number of outliers are not difficult, and methods like least median of squares (Rousseeuw, 1984) are very effective at finding multiple outliers. These factors do not carry over to the nonlinear model. Computations are too expensive to do exactly, and there are no well established robust methods for finding outliers in the literature.

In this subsection, the algorithm of Atkinson and Weisberg (1991) is adopted to work efficiently in the nonlinear model case. The changes made in the basic algorithm include: modification of stopping criterion; use of an approximate, easily computed objective function; and some modification of the basic computational method which requires exact computations much less frequently. These will be discussed later in this section.

Suppose that the value of m is chosen as a number of outliers. To arrive at the initial step, we randomly divide n cases into two subsets of m bad (outliers) and $n-m$ good (inliers). We consider an interchange of a single case currently in the m cases identified as outliers with a single case in the $n-m$ cases identified as inliers. Our goal is to find a subset I of size m which maximizes the likelihood ratio test statistic (2.3), that is, which minimizes $\log S(\hat{\theta}_{(I)}, \hat{\delta})$. We will set $f = \log S(\hat{\theta}_{(I)}, \hat{\delta}) = \log(RSS_{(I)})$ to be the object function.

Suppose we have fitted the model deleting the current m bad cases indexed by I_O and have calculated residual sum of squares RSS_O. If this step is accepted we need to find the new subset I_N by exchanging a randomly chosen case from the bad group with a randomly chosen case from the good group and find the new residual sum of squares RSS_N. The change of the object function is

$$\Delta f = f_{new} - f_{old} = \log\left(\frac{RSS_N}{RSS_O}\right),$$

and the acceptance probability is

$$p = \begin{cases} 1 & \text{if } \Delta f \le 0 \\ \exp(-\beta \Delta f) & \text{if } \Delta f > 0 \end{cases}.$$

As the algorithm proceeds, the value of β increases so that the acceptance probability decreases. The control parameter β increases according to a step function using blocks of n_s searches. For the k-th block of n_s searches, β was fixed at

$$\beta = -2^{k-1}\frac{\log(0.5)}{\log(c)}, \quad k = 1, 2, \cdots$$

where c is a user defined constant. With this control parameter, the probability of acceptance is 50% in the first block if $RSS_N = cRSS_O$. In their algorithm, Atkinson and Weisberg (1991) set a fixed number of blocks prior to the search. In practice, the algorithm may terminate while the value of the object function is still fluctuating, thus leading to premature termination and producing erroneous results. This problem can be overcome by introducing the following rule. The algorithm is terminated if the subset accepted remains unchanged for n_c consecutive searches. We refer to this rule as the stop criterion and n_c as the stop parameter.

The behavior of the algorithm depends on the control and stop parameters. If β is too large or c is too small, the rate of cooling is too fast and the conditional acceptance probability is too small so that local minima may not be avoided. If the parameters n_s or n_c are too small, the minimum may not be found to a sufficient degree of accuracy. In our application, we have found that setting $c = 1.5 \sim 2.0$, $n_s = 50$ and $n_c = 25 \sim 50$ works well and requires about 200~300 searches. Using these parameters, this algorithm converges to the global minimum about 30~90% of the time with most being more than 60%. These percentages are obtained from 12 different data sets each with a different model, each run over 100 times.

The simulated annealing algorithm does not always get the global minimum of the object function; however, if we run this algorithm more than once and keep record of the final RSS's and the subset I's, then we can select the subset I which has the smallest final RSS. This method has a smaller risk in getting a local minimum and still saves the computational time compared to calculating all possible subsets. With a 60% convergence rate, if we run this algorithm 5 times the error rate is only about 1%.

Example : The data for this example are taken from Carr (1960) on the reaction rate of the catalytic isomerization of n-pentane to isopentane and are reproduced in Table 1. A proposed model function for these data is

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\theta_1\theta_3(x_2 - x_3/1.632)}{1+\theta_2 x_1+\theta_3 x_2+\theta_4 x_3}.$$

Box and Hill (1974) and Carroll and Ruppert (1984) also analyze these data. Suppose that three outlying cases exist, then we have the global minimum of the object function or the global maximum of the likelihood ratio test statistic from subset $I=(7, 9, 24)$ with $RSS_{(I)} = 1.19072$. Figure 1 shows the evolution of the residual sum of squares for 242 exchanges with $c=1.5$, $n_s=50$ and $n_c=50$. The algorithm converges to the global minimum of residual sum of squares in 242 searches with 242 fittings. Using this data and model under the above settings, the algorithm converged to the global maximum in 82 out of 100 trials.

3.3 Speeding Up the Simulated Annealing Algorithm

In this section we consider a modified procedure that speeds up the simulated annealing algorithm. Suppose that at the current stage we accept the subset I_O with the residual sum of squares RSS_O. Then, we need to choose the new subset I_N by interchanging a case between the bad and good groups and calculate the residual sum of squares RSS_N. Let N-th case among n cases which is in the outlier group and O-th case among n in the inlier group be interchanged, that is, the N-th case is added to the good group and the O-th case is deleted from the good group. Using the linear approximations (A.5) and (A.6), and with the aid of (A.4), we have the following formula for the distance for the two residual sum of squares before and after interchanging cases.

$$RSS_{AN} - RSS_O = \frac{e_N^2(1-h_{OO}) - e_O^2(1+h_{NN}) + 2h_{NO}\, e_N\, e_O}{(1+h_{NN})(1-h_{OO}) + h_{NO}^2}, \tag{3.1}$$

where

$$e_O = y_O - f(\boldsymbol{x}_O, \hat{\boldsymbol{\theta}}_{(I_o)})$$
$$e_N = y_N - f(\boldsymbol{x}_N, \hat{\boldsymbol{\theta}}_{(I_o)})$$
$$\boldsymbol{H} = \{h_{ij}\} = \hat{\boldsymbol{V}}_o(\hat{\boldsymbol{V}}_o^T\hat{\boldsymbol{V}}_o)^{-1}\hat{\boldsymbol{V}}_o^T, \text{ with } \hat{\boldsymbol{V}}_o = V(\hat{\boldsymbol{\theta}}_{(I_o)}) = \left.\frac{\partial f}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{(I_o)}}.$$

To avoid refitting the deletion model with new subset I_N, we use (3.1) to get approximated residual sum of squares RSS_{AN} for the new subset. This approximation,

RSS_{AN}, is generally more accurate than the approximated RSS given in (A.6) when the number of outliers is three or more, since RSS_{AN} is obtained by exchanging one case while approximated RSS given in (A.6) is obtained by deleting m cases from the full data set. Thus the test statistic calculated using RSS_{AN} has higher degree of accuracy than the one obtained using the other approximated residual sum of squares or the score test statistic.

The modified procedures are as follows. If the RSS_{AN} is larger than RSS_O, we reject the new choice with probability $1-p$ without refitting the deletion model. If we do not reject the new choice we refit the deletion model and find the RSS_N. For the residual sum of squares obtained by refitting, we make a decision using the general procedure described in Section 3.2. In our application, we can save about 30~60% of the computational time with this procedure while keeping the accuracy as high as that of the general procedure.

Example (continued) : The performance of the modified simulated annealing method is illustrated in this example. Figure 2 shows the evolution of the residual sum of squares for a search with same parameters $c=1.5$, $n_s=50$ and $n_c=50$. The algorithm converges to the global minimum of residual sum of squares in a similar manner but it requires 129 fittings of the deletion models in 245 searches. For this example, the convergence rate using the modified procedure is 86% (129 out of 150), which is better than that obtained using the original algorithm.

4. Comments

We discussed the method for finding the subset I of m most likely outlying cases using the simulated annealing algorithm. Once we find these cases we need to test whether they are outliers using the procedures discussed in Section 2. If test turns out to be significant, m cases indexed by I are outliers. Otherwise, we may consider the following procedures. We reduce the number of outliers from m to $m-1$ and find $m-1$ most outlying cases in a similar fashion. We continue these steps until the test results are significant. In this case, m is the maximum number of outliers to be tested. The number of outliers or the maximum number of outliers to check for would depend on the context of the problem and is an area that needs further research.

The most crucial factors in implementing the simulated annealing algorithm are the choice of a suitable scheme that exchanges the cases between two groups and an appropriate conditional acceptance probability that determines the annealing schedule. When the subset

size is large, single case interchanges are not reliable. There may exist a certain conditional acceptance probability that is more appropriate for this algorithm. The perturbation schemes which interchange more than one case and other annealing schedules still need to be explored.

Table 1. Reaction rate for isomerization of n-pentane to ispentane

X_1 : partial pressure of hydrogen, X_2 : partial pressure of n-pentane,
X_3 : partial pressure of isopentane, Y : reaction time

case	X_1	X_2	X_3	Y	case	X_1	X_2	X_3	Y
1	205.8	90.9	37.1	3.541	13	297.3	142.2	10.5	5.686
2	404.8	92.9	36.3	2.397	14	314.0	146.7	157.1	1.193
3	209.7	174.9	49.4	6.694	15	305.7	142.0	86.0	2.648
4	401.6	187.2	44.9	4.722	16	300.1	143.7	90.2	3.303
5	224.9	92.7	116.3	0.593	17	305.4	141.1	87.4	3.054
6	402.6	102.2	128.9	0.268	18	305.2	141.5	87.0	3.302
7	212.7	186.9	134.4	2.797	19	300.1	83.0	66.4	1.271
8	406.2	192.6	134.9	2.451	20	106.6	209.6	33.0	11.648
9	133.3	140.8	87.6	3.196	21	417.2	83.9	32.9	2.002
10	470.9	144.2	86.9	2.021	22	251.0	294.4	41.5	9.604
11	300.0	68.3	81.7	0.896	23	250.3	148.0	14.7	7.754
12	301.6	214.6	101.7	5.084	24	145.1	291.0	50.2	11.590

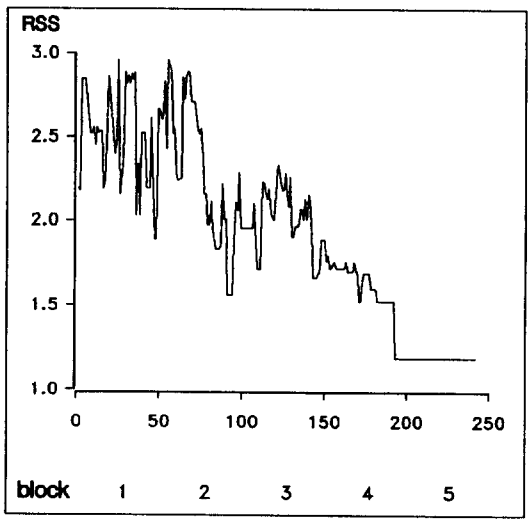

Figure 1. Simulated annealing plot

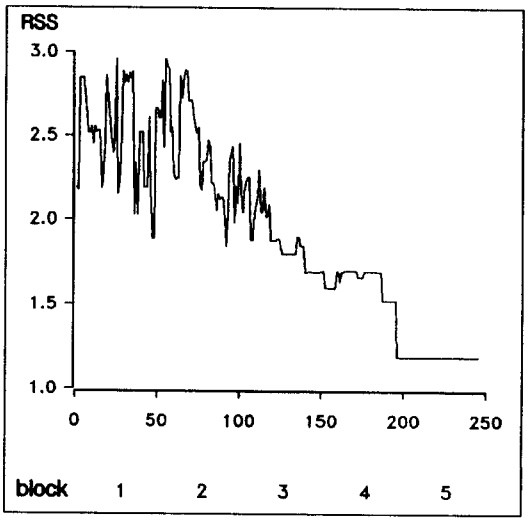

Figure 2. Simulated annealing plot with approximation

References

[1] Atkinson, A. C. and Weisberg, S. (1991). Simulated annealing for detection of multiple outliers using least squares and least median of squares fitting, In Stahel, W. and Weisberg, S. (Eds.), *Directions in Robust Statistics and Diagnostics, Part I,* 7-20, Springer-Verlag: New York.

[2] Bohachevsky, I. O., Johnson, M. E., and Stein, M. L. (1986). Generalized simulated annealing for function optimization, *Technometrics,* Vol. 28, 209-217.

[3] Bonomi, E. and Lutton, J.-L. (1984). The N-city travelling salesman problem: Statistical mechanics and the Metropolis algorithm, *SIAM Review,* Vol. 26, 551-568.

[4] Box, G. E. P. and Hill, W. J. (1974). Correcting inhomogeneity of variance with power transformation weighting, *Technometrics,* Vol. 16, 385-389.

[5] Carr, N. L. (1960). Kinetics of catalytic isomerization of n-pentane, *Industrial Engineering Chemistry,* Vol. 52, 391-396.

[6] Carroll, R. J. and Ruppert, D. (1984). Power transformations when fitting theoretical models to data, *Journal of the American Statistical Association,* Vol. 79, 321-328.

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

[8] Haines, L. M. (1987). The application of the annealing algorithm to the construction of exact optimal designs for linear-regression models, *Technometrics,* Vol. 29, 439-447.

[9] Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices, *SIAM Review,* Vol. 23, 53-60.

[10] Kirkpatrick, S., Gelatt, C. D., Jr., and Vecchi, M. P. (1983). Optimization by simulated annealing, *Science,* Vol. 220, 671-680.

[11] Lundy, M. (1985). Applications of the annealing algorithm to combinatorial problems in statistics, *Biometrika,* Vol. 72, 191-198.

[12] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations using fast computing machines, *The Journal of Chemical Physics,* Vol. 21, 1087-1092.

[13] Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association,* Vol. 79, 871-880.

Appendix

Suppose that $f(X, \theta)$ is approximately linear in a neighborhood about $\hat{\theta}$. Then, we have the following linear Taylor expansion for the cases not included in the subset I,

$$f(X_{(I)}, \theta_{(I)}) \cong f(X_{(I)}, \hat{\theta}) + \hat{V}_{(I)}(\theta_{(I)} - \hat{\theta}), \tag{A.1}$$

where $\hat{V}_{(I)}$ is obtained by deleting the m rows from $\hat{V} = V(\hat{\theta})$ indexed by I. If we use this approximation (A.1), the residual sum of squares for deletion model can be written as

$$\begin{aligned} S(\theta_{(I)}) &= (y_{(I)} - f(X_{(I)}, \theta_{(I)}))^T (y_{(I)} - f(X_{(I)}, \theta_{(I)})) \\ &\cong (y_{(I)} - f(X_{(I)}, \hat{\theta}) - \hat{V}_{(I)}(\theta_{(I)} - \hat{\theta}))^T (y_{(I)} - f(X_{(I)}, \hat{\theta}) - \hat{V}_{(I)}(\theta_{(I)} - \hat{\theta})) \\ &= (e_{(I)} - \hat{V}_{(I)}(\theta_{(I)} - \hat{\theta}))^T (e_{(I)} - \hat{V}_{(I)}(\theta_{(I)} - \hat{\theta})) \end{aligned} \tag{A.2}$$

and is minimized at

$$\hat{\theta}_{(I)} = \hat{\theta} + (\hat{V}_{(I)}{}^T \hat{V}_{(I)})^{-1} \hat{V}_{(I)}{}^T e_{(I)}. \tag{A.3}$$

Let A be a $p \times p$ square matrix and let B and C be the matrices of dimension $p \times m$. Assuming that the inverses exist, Henderson and Searle(1981) verified the following

$$(A - BC^T)^{-1} = A^{-1} + A^{-1}B(I_m - C^T A^{-1} B)^{-1} C^T A^{-1}. \tag{A.4}$$

With formula (A.4), equation (A.3) is simplified to give a more usual form

$$\hat{\theta}_{(I)} = \hat{\theta} + (\hat{V}^T \hat{V})^{-1} \hat{V}_I{}^T (I_m - \hat{H}_I)^{-1} e_I. \tag{A.5}$$

where $\hat{H}_I$ is the $m \times m$ minor of $\hat{H} = \hat{V}(\hat{V}^T \hat{V})^{-1} \hat{V}^T$ with rows and columns indexed by I. By substituting (A.5) into (A.2) and with the fact $e^T \hat{V} = 0$, we have the following equation after simplification:

$$S(\hat{\theta}_{(I)}, \hat{\delta}) - S(\hat{\theta}, 0) = -e_I{}^T (I_m - H)^{-1} e_I. \tag{A.6}$$