The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.3.511
26(3), 511–520

On Rice Estimator in Simple Regression Models with Outliers

Chun Gun Park[a,1]

[a]Department of Mathematics, Kyonggi University

(Received April 6, 2013; Revised May 13, 2013; Accepted May 21, 2013)

Abstract

Detection outliers and robust estimators are crucial in regression models with outliers. In such studies the focus is on detecting outliers and estimating the coefficients using leave-one-out. Our study introduces Rice estimator which is an error variance estimator without estimating the coefficients. In particular, we study a comparison of the statistical properties for Rice estimator with and without outliers in simple regression models.

Keywords: Least squares method, leave-one-out, outliers, Rice estimator, simple regression model.

1. 서론

회귀모형은 오래전부터 다양한 분야에서 이용된 기법 중에 하나이다. 오랜 세월로 인해 회귀모형에 관한 많은 연구가 진행되었지만 아직도 이상치에 관한 기법이 개발되고 있다. 대부분 이상치에 관한 연구는 직접적으로 이상치를 탐색하는 방법과 이상치에 영향을 받지 않고 회귀계수를 추정하는 방법으로 구분할 수 있다. 이러한 기법들은 회귀계수를 추정하는 과정에서 얻어진 정보를 토대로 이상치를 탐색한다 (She와 Owen, 2011). 본 연구에서는 이상치를 탐색하는데 적용되어왔던 전통적인 방법의 비효율성을 개선하고자 일차 차이를 이용한 기법을 이용한 분산 추정량의 통계적 성질을 알아보는 것이다.

일차 차이를 근거로 한 분산 추정량(first order difference based variance estimator)인 Rice 추정량과 유사한 추정량이 여러 연구자들에 의해 개발되었고 그러한 추정량은 비모수 회귀모형에서 오차항이 일정(homogenuous)하고 독립변수에 영향을 받지 않을 경우에 적용될 수 있다고 언급하고 있다 (Gasser 등, 1986; Kay, 1988; Hall과 Marron, 1990; Hall 등, 1991; Thompson 등, 1991). 특히 회귀모형에서 Rice 추정량 및 그와 유사한 추정량에 대한 연구는 전무하다. 그 이유는 관찰치의 차이를 기초로 한 분산 추정량이 불편추정량이 아니고 이상치가 없는 회귀모형에서 최소제곱법(least squares method)에 의해 얻어진 오차 분산 추정량의 통계적 성질이 우월하기 때문이다. 그리고 Rice 추정량 및 이와 유사한 추정량은 회귀모형의 독립변수 및 회귀계수에 영향을 받기 때문에 오차 분산의 추정량으로 적합하지 않다.

따라서 본 논문은 오차 분산을 잘 추정하는 방법을 제시하는 것이 아니라, 이상치 탐색에서 Rice 추정량의 장점을 단순회귀모형으로 적용될 수 있는 간단한 예제를 통하여 알아보는 것이다. 즉 Rice 추정량이

This work was supported by Kyonggi University Research Grant 2012.
[1]Assistant Professor, Department of Mathematics, Kyonggi University, Gyeonggi-do 443-760, Korea.
E-mail: cgpark@kgu.ac.kr

오차 분산을 추정하는데 적합하지는 않지만, 정리 2.1∼정리 3.2로부터 이상치의 유무에 따라 Rice 추정량의 통계적 성질에 차이가 존재하고 이러한 통계적 성질이 *leave-one-out* 기법에서 나타나는 형태가 이상치 탐색에 어떤 정보를 주는지를 실제적으로 살펴보는데 있다.

본 연구의 구성은 다음과 같다. 2장에서는 Rice 추정량의 통계적 성질을 파악하고 3장은 *leave-one-out* 기법을 Rice 추정량에 적용할 경우의 통계적 성질을 유도한다. 4장은 단순회귀모형을 따르는 예제로부터 *leave-one-out* 기법을 Rice 추정량에 적용하여 이상치 탐색에 대한 형태를 관찰한다. 마지막으로 5장에서는 결론과 추후 연구를 제시한다.

2. 통계적 성질

2.1. 단순회귀모형의 일차 차이의 분산 추정량

일차 차이를 이용한 분산 추정량(first order based variance estimator)은 비모수 회귀모형에서 오차분산을 추정하는데 사용되어왔지만 그 평균이 불편추정량이 아닌 관계로 이와 유사한 추정량이 지속적으로 개발되어 왔다 (Tong과 Wang, 2005). 이러한 추정량의 장점은 미지함수를 추정하지 않고 단지 관찰값의 차이를 이용하기 때문에 일정한 조건만 만족하면 미지함수의 형태에 크게 영향을 받지 않는다는 장점이 있다. 이러한 장점에도 불구하고 회귀모형에서는 이러한 연구가 전무하다. 또한 이상치가 존재하는 회귀모형에서 관찰값의 차이를 이용하여 이상치를 탐색하는 연구도 없다.

본 연구에서 다루는 모형은 단순회귀모형으로 국한한다. 그려면 가장 보편적인 단순회귀모형은 다음과 같다.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

여기서 $x_1 < x_2 < \cdots < x_n$, 오차항 ϵ_i은 평균이 0이고 분산이 σ^2인 정규분포를 따르고 독립이라고 가정한다.

일반적으로 식 (2.1)로부터 단순회귀모형을 적합하기 위해서는 회귀계수 $\beta = (\beta_0, \beta_1)^T$을 최소제곱법으로 추정하고 잔차를 이용하여 오차 분산 σ^2를 추정한다. 하지만 이상치가 존재하는 모형인 경우에 최소제곱법을 그대로 적용할 수 없고 이상치의 효과를 제거하거나 이상치에 영향을 받지 않는 로버스트 추정량을 통하여 회귀계수 및 오차분산을 추정할 수 있다. 본 절에서는 일차 차이를 이용한 분산 추정량을 살펴보고 단순회귀모형에서 분산 추정량의 통계적 성질을 알아보는 것이다.

Rice (1984)는 다음과 같은 일차 차이를 이용한 추정량을 제안했다.

$$\widehat{\sigma_R^2} = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2 . \tag{2.2}$$

정리 2.1 식 (2.1)과 식 (2.2)로부터 이상치가 없는 단순회귀모형에 대한 Rice 추정량의 통계적 성질은 다음과 같다. $z = (z_1, \ldots, z_n)^T$, $J = (1, \ldots, 1)^T$, 그리고 $x = (x_1, \ldots, x_n)^T$이면 $z = \beta_0 J + \beta_1 x$이라 가정하자.

(i) $E\left(\widehat{\sigma_R^2}\right) = z^T A z + \sigma^2 = B_1(n, z) + \sigma^2$,

(ii) $\mathrm{Var}\left(\widehat{\sigma_R^2}\right) = \dfrac{4\sigma^2 z^T A^2 z + 2\sigma^4 \mathrm{tr}(A^2)}{\mathrm{tr}(A)^2}$

$$= 2\sigma^2 \frac{\left\{ \sum_{i=2}^{n} (z_i - z_{i-1})^2 - \sum_{i=2}^{n-1} (z_i - z_{i-1})(z_{i+1} - z_i) \right\}}{(n-1)^2} + \sigma^4 \frac{(6n-2)}{2(n-1)^2},$$

(iii) $\mathrm{MSE}\left(\widehat{\sigma_R^2}\right) = \mathrm{Var}\left(\widehat{\sigma_R^2}\right) + \mathrm{Biased}\left(\widehat{\sigma_R^2}\right)^2$

$$= \frac{\left(z^TAz\right)^2 + 4\sigma^2 z^TA^2z + 2\sigma^4\mathrm{tr}(A^2)}{\mathrm{tr}(A)^2},$$

여기서 $B_1(n,z) = \sum_{i=2}^n (z_i - z_{i-1})^2/2(n-1)$0이고 $A = D^tD$0이다.

$$D = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & -1 & 1 \end{pmatrix} \in \mathrm{R}^{(n-1)\times n}.$$

정리 2.1의 증명은 부록을 참조하라.

2.2. 이상치가 있는 단순회귀모형의 일차 차이의 분산 추정량

식 (2.1)로부터 이상치가 존재하는 단순회귀모형은 다음과 같이 정의할 수 있다.

$$y_i = \beta_0 + \beta_1 x_i + d_i I(d_i \neq 0) + \epsilon_i I(d_i = 0), \quad i = 1, \ldots, n, \tag{2.3}$$

여기서 $I(\cdot)$는 지시함수이고 d_i는 i번째 관찰치에 대응되는 이상치이다.

정리 2.2에서 표현된 분산 추정량 $\widehat{\sigma_O^2}$는 $\widehat{\sigma_R^2}$과 동일하고 이상치가 존재하는 단순회귀모형에 적용되었다는 것을 표시하기 위해서 사용된 것이다.

정리 2.2 식 (2.1)와 식 (2.3)으로부터 0상치가 존재하는 Rice 추정량의 통계적 성질은 다음과 같다. $s = z + d$0이고 0상치인 $d = (d_1, \ldots, d_n)^T$은 0상치의 열벡터이고 그 값은 영 또는 상수라고 가정하자.

(i) $E\left(\widehat{\sigma_O^2}\right) = s^TAs + \sigma^2 = B_1(n,x) + B_2(n,d) + B_3(n,x,d) + \sigma^2$,

(ii) $\mathrm{Var}\left(\widehat{\sigma_O^2}\right) = \mathrm{Var}\left(\widehat{\sigma_R^2}\right) + 4\sigma^2 \dfrac{\left(d^TA^2z + z^TA^2d + d^TA^2d\right)}{\mathrm{tr}(A)^2}$,

(iii) $\mathrm{MSE}\left(\widehat{\sigma_O^2}\right) = \mathrm{Var}\left(\widehat{\sigma_O^2}\right) + \mathrm{Biased}\left(\widehat{\sigma_O^2}\right)^2$,

여기서 얻어진 편의(biased)는 다음과 같다.

$$B_1(n,z) = \sum_{i=2}^n \frac{(z_i - z_{i-1})^2}{2(n-1)},$$

$$B_2(n,d) = \sum_{i=2}^n \frac{(d_i - d_{i-1})^2}{2(n-1)},$$

$$B_3(n,z,d) = \sum_{i=2}^n \frac{(z_i - z_{i-1})(d_i - d_{i-1})}{(n-1)}.$$

정리 2.2의 증명은 부록을 참조하라.

정리 2.1과 정리 2.2로부터 이상치의 유무에 따른 단순회귀모형에서 얻어진 Rice 추정량 모두가 편의가 있다. 만약 모든 이상치와 기울기가 양수이면 후자의 추정량의 평균이 전자보다 크다. 하지만 기울기와 이상치의 크기 및 부호에 따라서 추정량들의 편의가 다양하게 변한다. 또한 추정량들의 분산은 기울기와 이상치의 크기 및 부호와 상관없이 항상 후자의 추정량이 크다.

정리 2.1과 정리 2.2로부터 오로지 i-번째 관찰치에 대한 이상치의 유무에 따라 Rice 추정량에 대한 편의의 변화는 이상치의 크기 및 부호, 관찰치의 수, 기울기 그리고 $x_i - x_{i-1}$에 따라 변한다. 또한 이상치가 있는 경우에 대한 추정량의 분산은 이상치의 크기, 부호와 수에 영향을 받는다.

3. Leave-one-out

회귀모형에서 이상치는 회귀계수를 추정하는데 많은 영향을 미치기 때문에 오래전부터 이상치를 탐색하기 위해 많은 진단 방법들이 개발되어왔다. 고전적인 이상치 탐색 방법으로 Cook distance 또는 DFFITS을 들 수 있다. 이 방법들은 *leave-one-out* 기법을 기반으로 개발된 것으로 관련 연구도 많은 편이다 (Hocking, 2003). 이들 방법의 단점은 여러 개의 이상치가 있으면 비록 이상치에 해당되는 하나의 관찰값을 제거하여 회귀계수를 추정하여도 다른 이상치의 영향으로 인해 이상치 탐색이 힘들어질 수 있다는 것이다.

본 연구는 Rice 추정량에 *leave-one-out* 기법을 도입하여 여러 개의 이상치에 대한 Rice 추정량의 통계적 성질을 유도하고 4절에서는 실증예제를 통하여 효과를 관찰하고자 한다.

정리 3.1 정리 2.1로부터 0이상치가 존재하지 않는 단순회귀모형에서 i-번째 관찰값을 제거한 Rice 추정량의 통계적 성질은 다음과 같다. $z_{(i)} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)^T$은 i-번째가 제거된 열벡터라고 가정하자.

(i) $E\left(\widehat{\sigma^2_{R(i)}}\right) = z_{(i)}^T \tilde{A} z_{(i)} + \sigma^2 = \tilde{B}_1(n,x) + \sigma^2,$

(ii) $\mathrm{Var}\left(\widehat{\sigma^2_{R(i)}}\right) = \dfrac{4\sigma^2 z_{(i)}^T \tilde{A}^2 z_{(i)} + 2\sigma^4 \mathrm{tr}(\tilde{A}^2)}{\mathrm{tr}(\tilde{A})^2},$

(iii) $\mathrm{MSE}\left(\widehat{\sigma^2_{R(i)}}\right) = \mathrm{Var}\left(\widehat{\sigma^2_{R(i)}}\right) + \mathrm{Biased}\left(\widehat{\sigma^2_{R(i)}}\right)^2.$

정리 3.2 정리 2.2로부터 0이상치가 존재하는 단순회귀모형에서 i-번째 관찰값을 제거한 Rice 추정량의 통계적 성질은 다음과 같다. $s_{(i)} = z_{(i)} + d_{(i)}$은 i-번째가 제거된 열벡터라고 가정하자.

(i) $E\left(\widehat{\sigma^2_{O(i)}}\right) = s_{(i)}^T \tilde{A} s_{(i)} + \sigma^2 = \tilde{B}_1(n,x) + \tilde{B}_2(n,d) + \tilde{B}_3(n,x,d) + \sigma^2,$

(ii) $\mathrm{Var}\left(\widehat{\sigma^2_{O(i)}}\right) = \mathrm{Var}\left(\widehat{\sigma^2_{R(i)}}\right) + 4\sigma^2 \dfrac{\left(d_{(i)}^T \tilde{A}^2 z_{(i)} + z_{(i)}^T A^2 d_{(i)} + d_{(i)}^T \tilde{A}^2 d_{(i)}\right)}{\mathrm{tr}(\tilde{A})^2},$

(iii) $\mathrm{MSE}\left(\widehat{\sigma^2_{O(i)}}\right) = \mathrm{Var}\left(\widehat{\sigma^2_{O(i)}}\right) + \mathrm{Biased}\left(\widehat{\sigma^2_{O(i)}}\right)^2,$

여기서 $(n-1) \times n$인 행렬 D에서 마지막 행과 열이 제거된 행렬을 $\tilde{D}$는 $(n-2) \times (n-1)$이고 $\tilde{A} = \tilde{D}^T \tilde{D}$이라 하자. 그리고 편의(biased)는 다음과 같다.

$$\tilde{B}_1(n,z) = \sum_{i=2}^{n} \frac{(z_i - z_{i-1})^2}{2(n-2)} + \frac{(z_{i+1} - z_i)(z_i - z_{i-1})}{(n-2)},$$

$$\tilde{B}_2(n,d) = \sum_{i=2}^{n} \frac{(d_i - d_{i-1})^2}{2(n-2)} + \frac{(d_{i+1} - d_i)(d_i - d_{i-1})}{(n-2)},$$

$$\tilde{B}_3(n,z,d) = \sum_{i=2}^{n} \frac{(z_i - z_{i-1})(d_i - d_{i-1})}{(n-2)} + \frac{(z_{i+1} - z_i)(d_i - d_{i-1})}{(n-2)} + \frac{(z_i - z_{i-1})(d_{i+1} - d_i)}{(n-2)}.$$

정리 3.1과 정리 3.2의 증명은 정리 2.1과 정리 2.2의 증명으로부터 쉽게 얻을 수 있다.

4. 이상치 탐색 - 적용예제 중심으로

회귀모형에서 이상치는 회귀계수 추정에 많은 영향을 미치는 요인 중에 하나이다. 따라서 회귀분석을 수행하는데 이상치 탐색이 중요한 절차 중에 하나이다. 본 연구에서 제시한 Rice 추정량이 비록 오차

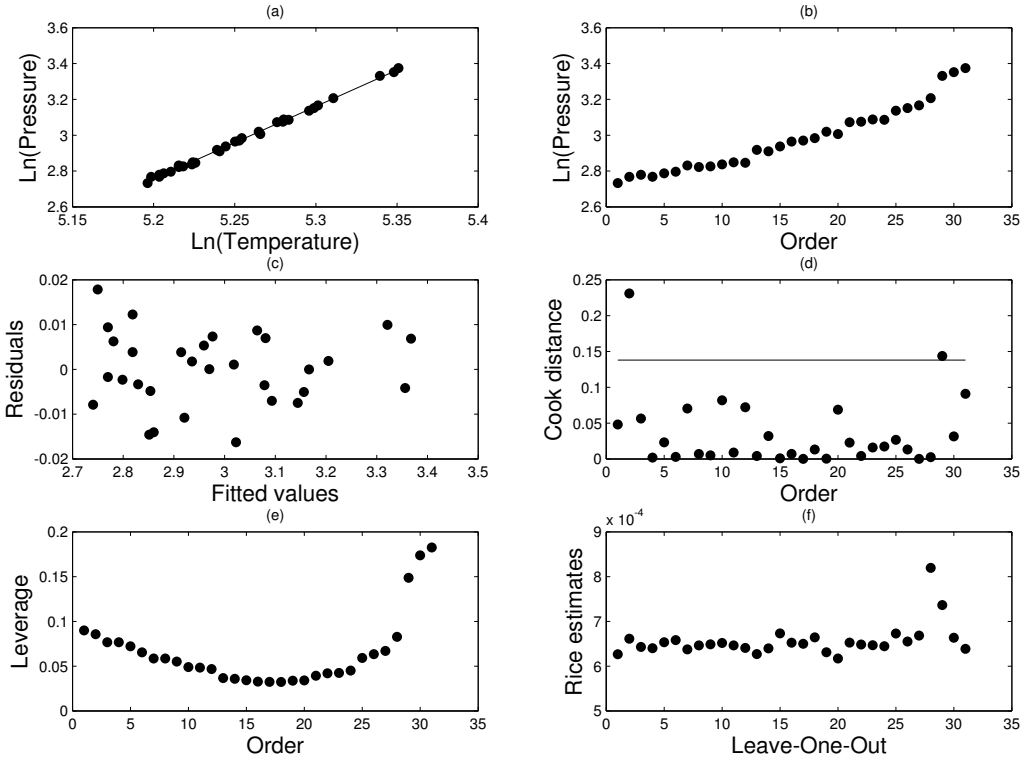

Figure 4.1. Simple regression without an outlier ((a) Plot of Ln(pressure) against Ln(temperature), (b) Plot of Ln(pressure) against order, (c) Residual plot, (d) Cook's distance, (e) Leverage values, (f) Rice estimates with leave-one-out)

분산에 불편추정량이 아니지만 이상치 유무에 따라 독특한 통계적 특성을 지니고 있다. 즉 정리 2.1에서 제시한 통계적 성질에는 오직 독립변수 및 회귀계수에 영향을 받는 반면 정리 2.2로부터 이상치가 존재하는 회귀모형에서 Rice 추정량의 기댓값 및 분산은 이상치의 수 및 크기에도 영향을 받는다. 또한 Rice 추정량이 회귀계수 추정과는 무관하지만 이상치의 문제를 안고 있다. 하지만 직관적으로 정리 2.2의 Rice 추정량에 *leave-one-out*을 적용한 기댓값은 제외될 설명변수 값의 이상치 유무에 따라서 상당한 차이가 있다. 만약 제외될 값이 정상적인 관찰값이면 Rice 추정량은 모든 이상치들이 존재한다. 하지만 제외될 설명변수 값이 이상치를 포함되어 있으면 Rice 추정량은 이상치가 존재하지 않는 경우의 Rice 추정량보다 이상치의 크기에 준한 작은 값이 될 것이다.

이를 실질적으로 확인하기 위해서 Weisberg (2005)가 소개한 압력(설명변수)과 온도(독립변수)의 예제를 통하여 위에서 언급한 내용을 관찰하기로 한다. 적용예제는 이상치가 존재하지 않는 단순선형회귀모형을 따르기 때문에 인위로 몇 개의 이상치들을 설정하여 *leave-one-out*를 통한 Rice 추정량의 특성을 관찰한다.

Figure 4.1은 원래 자료를 회귀분석을 수행한 결과와 *leave-one-out* 방법을 Rice 추정치에 적용한 것이다. 잔차를 분석해 보면 이상치에 대한 특별한 증거는 없지만 Cook distance에는 2개의 이상치가 존재하는 것으로 보인다. 하지만 이러한 증거로 확실히 이상치가 존재할 것이라고 단정하기는 어렵다.

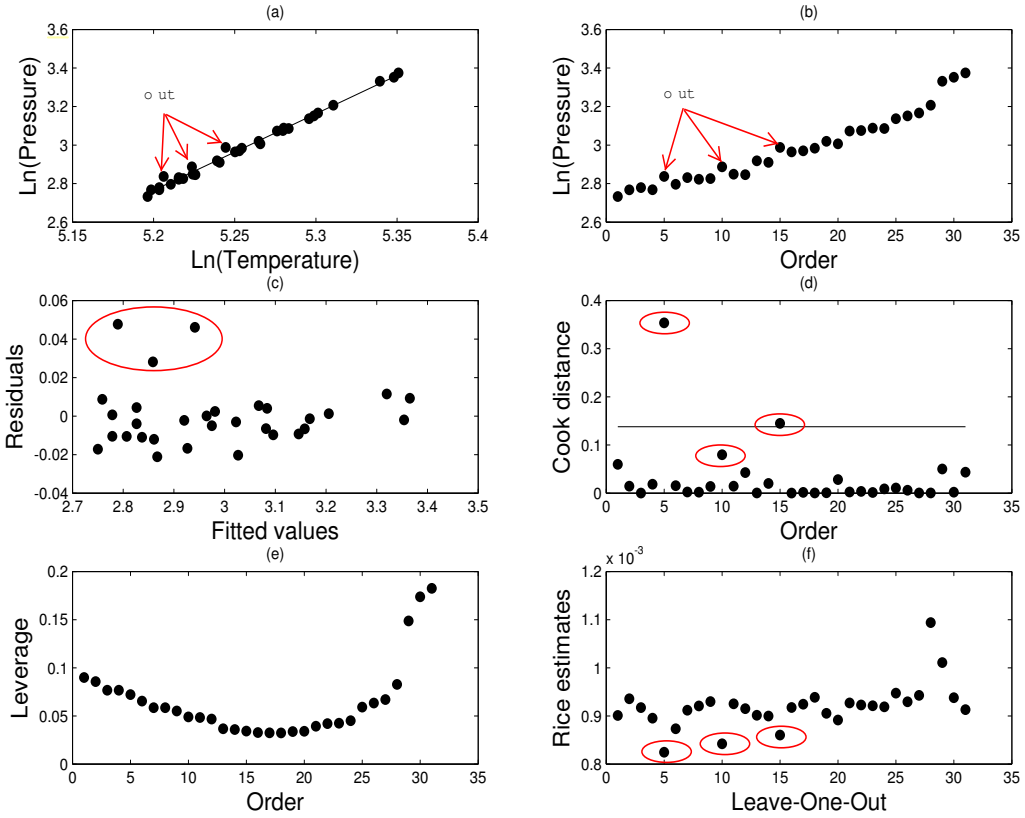

Figure 4.2. Simple regression with several small outliers ((a) Plot of Ln(pressure) against Ln(temperature), (b) Plot of Ln(pressure) against order, (c) Residual plot, (d) Cook's distance, (e) Leverage values, (f) Rice estimates with leave-one-out)

이상치 탐색에서 *leave-one-out* 방법과 Rice 추정치의 특성을 파악하기 위해 원자료에서 특정 위치에 인위적으로 일정한 값을 더하여 이상치 세 개를 생성한다. Figure 4.2는 세 개의 위치(5, 10, 15)에 작은 인위적인 값(= 0.05)을 더하여 이상치를 생성하였다. 잔차 분석으로부터 세 점들이 이상치로 보이지만 Cook distance에서는 두 개만이 이상치로 나타났다.

Figure 4.3은 세 개의 위치(5, 10, 15)에 큰 인위적인 값(= 1)을 더하여 이상치를 생성하였다. 잔차 분석과 Cook distance에서 세 개의이 이상치가 나타났다. 또한 *leave-one-out*를 통한 Rice 추정량에도 동일하게 세 개의 이상치가 발견되었다.

5. 결론 및 추가연구

비록 분산 추정을 하는데 Rice 추정량이 단순회귀모형에서 불편추정량은 아니지만 통계적 성질은 이상치의 유무에 따라 구분되는 점이 있다. 즉 이상치가 없는 경우에 Rice 추정량은 독립변수와 회귀계수에 영향을 받는 반면에 이상치가 존재하는 경우에는 이상치의 수 및 크기에도 영향을 받는다. 본 연구에서 Rice 추정량의 이러한 특성과 *leave-one-out* 기법을 융합하여 이상치 탐색을 실험적으로 수행하였다.

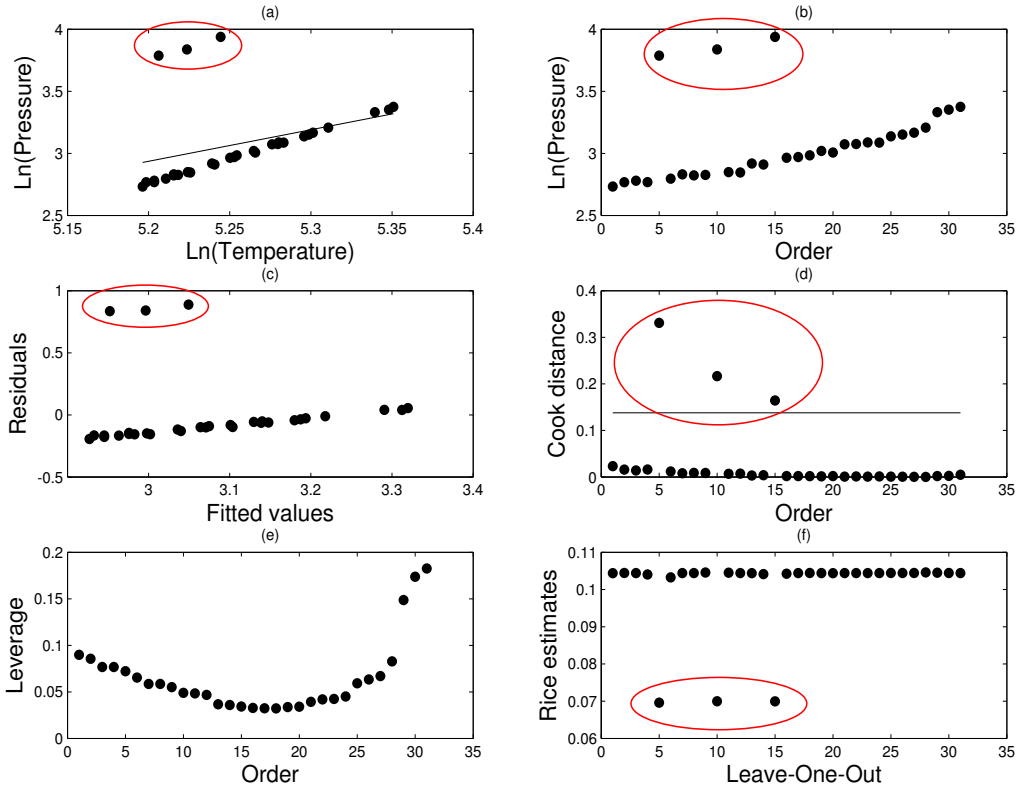

Figure 4.3. Simple regression with several large outliers ((a) Plot of Ln(pressure) against Ln(temperature), (b) Plot of Ln(pressure) against order, (c) Residual plot, (d) Cook's distance, (e) Leverage values, (f) Rice estimates with leave-one-out)

Figure 4.2와 Figure 4.3으로부터 세 개의 이상치에 대한 Cook distance의 결과는 다소 차이는 있지만 잔차분석에서는 이상치라고 판단될 수 있다. 특히 이상치의 크기가 오차분산에 비해서 클 경우에 Cook distance의 크기는 위치에 따라서 변화가 있지만 *leave-one-out* 기법을 통한 Rice 추정량은 거의 일정한 값을 나타냈다. 적용예제로부터 얻은 결과와 정리 4.2에서 제시된 통계적 성질이 거의 일치하는 것으로 보인다. 즉 이상치의 개수에 상관없이 *leave-one-out* 기법을 통한 Rice 추정량은 제거된 i-번째 관찰치가 이상치의 유무에 더 많은 영향을 받지 다른 위치의 이상치에는 큰 영향을 받지 않는다는 것이다.

따라서 본 연구에서는 제한적으로 단순회귀모형에서 이상치 유무에 따른 Rice 추정량을 이용하여 실험적으로 이상치 탐색에 대한 방법을 제시했지만 추후 연구로 *leave-one-out* 기법을 통한 Rice 추정량의 기준값(cut-off)에 대한 연구가 진행되어야 한다.

부록

정리 2.1의 증명

(i) Park (2011, 2012)에 자세히 나와 있어 여기서는 생략한다.

(ii) 대칭행렬 A에 대해서 $A^2 = AA$는 쉽게 구할 수 있다.

$$A = D^t D = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix}_{n\times n}$$

그리고 A와 A^2의 대각원소의 합 그리고 $z^T A^2 z$는 다음과 같이 얻을 수 있다.

$$\begin{aligned} \mathrm{tr}(A) &= 2(n-1), \\ \mathrm{tr}(A^2) &= 6n-2, \\ Az &= [(z_1 - z_2), (z_2 - z_1) - (z_3 - z_2), \ldots, (z_{n-1} - z_{n-2}) - (z_n - z_{n-1}), (z_n - z_{n-1})]^t, \\ z^T A^2 z &= 2\sum_{i=2}^{n}(z_i - z_{i-1})^2 - 2\sum_{i=2}^{n-1}(z_i - z_{i-1})(z_{i+1} - z_i). \end{aligned}$$

따라서

$$\begin{aligned} \mathrm{Var}\left(\widehat{\sigma_R^2}\right) &= \frac{4\sigma^2 z^T A^2 z + 2\sigma^4 \mathrm{tr}(A^2)}{\mathrm{tr}(A)^2} \\ &= 2\sigma^2 \frac{\left\{\sum_{i=2}^{n}(z_i - z_{i-1})^2 - \sum_{i=2}^{n-1}(z_i - z_{i-1})(z_{i+1} - z_i)\right\}}{(n-1)^2} + \sigma^4 \frac{(6n-2)}{2(n-1)^2}. \end{aligned}$$

행렬형태의 증명은 Park(2012)을 참조하라.

(iii) (i)과 (ii)에 의해서 쉽게 증명이 된다.

정리 2.2의 증명

(i) 정리 2.1의 증명과 유사하여 생략한다.

(ii) 정리 2.1과 $s = z + d$에 의해서 다음과 같이 증명된다.

$$\begin{aligned} \mathrm{Var}\left(\widehat{\sigma_O^2}\right) &= \frac{\left\{4\sigma^2 s^t A^2 s + 2\sigma^4 \mathrm{tr}(A^2)\right\}}{\mathrm{tr}(A)^2} \\ &= \frac{\left\{4\sigma^2 z^t A^2 z + 4\sigma^2 d^t A^2 z + 4\sigma^2 z^t A^2 d + 4\sigma^2 d^t A^2 d + 2\sigma^4 \mathrm{tr}(A^2)\right\}}{\mathrm{tr}(A)^2} \\ &= \mathrm{Var}\left(\widehat{\sigma_R^2}\right) + 4\sigma^2 \frac{\left(d^t A^2 z + z^t A^2 d + d^T A^2 d\right)}{\mathrm{tr}(A)^2} \\ &= 2\sigma^2 \frac{\left\{\sum_{i=2}^{n}(z_i - z_{i-1})^2 - \sum_{i=2}^{n-1}(z_i - z_{i-1})(z_{i+1} - z_i)\right\}}{(n-1)^2} + \sigma^4 \frac{(6n-2)}{2(n-1)^2} \\ &\quad + 2\sigma^2 \frac{\sum_{i=2}^{n}(d_i - d_{i-1})(z_i - z_{i-1})}{(n-1)^2} + \sigma^2 \frac{\sum_{i=2}^{n-1} d_{i-1}(z_{i+1} - z_i)}{(n-1)^2} \end{aligned}$$

$$+\sigma^2\frac{\sum_{i=2}^{n-1} d_i\left[(z_i-z_{i-1})-(z_{i+1}-z_i)\right]}{(n-1)^2}-\sigma^2\frac{\sum_{i=2}^{n-1} d_{i+1}(z_i-z_{i-1})}{(n-1)^2}$$
$$+2\sigma^2\frac{\left\{\sum_{i=2}^{n}(d_i-d_{i-1})^2-\sum_{i=2}^{n-1}(d_i-d_{i-1})(d_{i+1}-d_i)\right\}}{(n-1)^2}.$$

여기서

$$Ad=[(d_1-d_2),(d_2-d_1)-(d_3-d_2),\ldots,(d_{n-1}-d_{n-2})-(d_n-d_{n-1}),(d_n-d_{n-1})]^t$$
$$d^tA^2z=2\sum_{i=2}^{n}(d_i-d_{i-1})(z_i-z_{i-1})-\sum_{i=2}^{n-1}\left[(d_i-d_{i-1})(z_{i+1}-z_i)+(d_{i+1}-d_i)(z_i-z_{i-1})\right]$$
$$=2\sum_{i=2}^{n}(d_i-d_{i-1})(z_i-z_{i-1})+\sum_{i=2}^{n-1}d_{i-1}(z_{i+1}-z_i)$$
$$+\sum_{i=2}^{n-1}d_i\left[(z_i-z_{i-1})-(z_{i+1}-z_i)\right]-\sum_{i=2}^{n-1}d_{i+1}(z_i-z_{i+1}).$$

(iii) (i)과 (ii)에 의해서 쉽게 증명이 된다.

References

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression, *Biometrika*, **73**, 625–633.

Hall, P., Kay, J. and Titterington, D. (1991). On estimation of noise variance in two dimensional signal processing, *Advances in Applied Probability*, **23**, 476–495.

Hall, P. and Marron, J. (1990). On variance estimation in nonparametric regression, *Biometrika*, **77**, 415–419.

Hocking, R. R. (2003). *Methods and Applications of Linear Models: Regression and the Analysis of Variance*, Wiley.

Kay, J. (1988). On the choice of the regularisation parameter in image restoration, *Lecture Notes in Computer Science*, **301**, 537–596.

Park, C. G. (2011). Estimation of error variance in nonparametric regression under a finite sample using ridge regression, *Korean Data and Information Science Society*, **22**, 1223–1232.

Park, C. G. (2012). First order difference-based error variance estimator in nonparametric regression with a single outlier, *Communications for the Korea Statistical Society*, **19**, 333–344.

Rice, J. A. (1984). Bandwidth choice for nonparametric regression, *Annals of Statistics*, **12**, 1215–1220.

She, Y. and Owen, A. W. (2011). Outlier detection using Nonconvex Penalized Regression, *Journal of the American Statistical Association*, **106**, 626–639.

Thompson, A., Kay, J. and Titterington, D. (1991). Noise estimation in signal restoration using regularization, *Biometrika*, **78**, 475–488.

Tong, T. and Wang, Y. (2005). Estimating residual variance in nonparametric regression using least squares, *Biometrika*, **92**, 821–830.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd edition, Wiley, Problem 2.2.4, New York.

이상치가 존재하는 단순회귀모형에서 Rice 추정량에 관해서

박천건[a,1]

[a]경기대학교 수학과

(2013년 4월 6일 접수, 2013년 5월 13일 수정, 2013년 5월 21일 채택)

요 약

이상치가 존재하는 회귀모형에서 이상치를 탐색하거나 로버스트 추정량에 대한 연구는 매우 중요하다. 이러한 연구는 *leave-one-out*를 이용하여 회귀계수를 추정하고 잔차를 이용하여 오차 분산을 추정하여 이상치를 탐색하는데 있다. 본 연구는 회귀모형에서 회귀계수를 추정하지 않고 오차 분산을 추정할 수 있는 Rice 추정량의 적용을 소개한 것이다. 특히, 단순회귀모형에서 이상치의 유무에 따라 Rice 추정량의 통계적 성질을 비교하고 이상치 탐색에 있어 어떤 장점이 있는지를 탐색한 연구이다.

주요용어: 최소제곱법, leave-one-out, 이상치, Rice 추정량, 단순회귀모형.

본 연구는 2012학년도 경기대학교 학술연구비(일반연구과제) 지원에 의하여 수행되었음.

[1](443-760) 경기도 수원시 영통구 이의동 산 94-6, 경기대학교 수학과, 조교수. E-mail: cgpark@kgu.ac.kr