*Proceedings of Joint Conference of The Korean Data Analysis Society and Korean Data And Information Science Society, October 17-18, 2003, pp. 93-100*

# A Comparision on CERES & Robust-CERES

*Kwang-Sik Oh*[1], *Soo-Hee Do*[2], *Daehak Kim*[3]

## Abstract

It is necessary to check the curvature of selected covariates in regression diagnostics. There are various graphical methods using residual plots based on least squares fitting. The sensitivity of LS fitting to outliers can distort their residuals, making the identification of the unknown function difficult to impossible. In this paper, we compare combining conditional expectation and residual plots(CERES Plots) between least square fit and robust fits using Huber M-estimator. Robust CERES will be far less distorted than their LS counterparts in the presence of outliers and hence, will be more useful in identifying the unknown function.

*Keywords* : Huber M-estimate, Robust Fit, CERES Plots

## 1. 서언

부분잔차도표는 Ezeckiel에 의해 제안 된 이후, Larsen과 McCleary(1972)가 Partial Residual Plot 이라는 이름으로 회귀변수들과 반응변수 사이의 관계를 조사하는데 사용하기 시작하였다. 그 이후 많은 연구자들이 부분잔차도표의 사용에 관한 연구를 했으며, Mallows(1986)는 Augmented Partial Residual Plot을 제안하였다. Cook(1993)은 이들 부분잔차도표들의 기대되는 유용성을 조사하고 확장된 부분잔차도표로서 조건부기대치잔차도표( CERES Plot)를 제안하였으며 회귀변수들 사이에 상관성이 있고 $g(.)$가 선형이 아닐 경우에는 CERES도표가 더욱 유용함을 실제 예를 통하여 밝힌 바 있다. 또한 Cook(1997)은 부분잔차도표와 Added Variable Plot을 비교 검토한 후 그 역할에 대하여 연구한 바 있으며 회귀변수와 반응변수의 관계를 검토하는데는 부분잔차도표가 더 유용함을 밝혔다.

한편 최소제곱법이 이상치에 민감하게 반응하여 잘못 된 결과를 초래 할 수도 있다는 가정 하에서 로버스트 추정방법을 회귀분석에 적용하기 시작하였으며, Hettmansperger (1990, 1993, 1994)

[1]Prof, Dept of Informational Statistics, Catholic University of Daegu, 712-702, Korea.
E-mail: ohkwang@cu.ac.kr
[2]Lecturer, Catholic University of Daegu
[3]Prof, Dept of Informational Statistics, Catholic University of Daegu, 712-702, Korea.

는 로버스트 잔차에 대한 연구를 하였다. Mckean과 Sheather(2000)은 로버스트 추정 방법에 기초한 부분잔차도표를 제안하고 이상치가 있는 경우에는 유용함을 밝혔다. 그리고 Oh와 Do(2003)는 확장된 부분잔차도표인 CERES도표에 로버스트 방법을 적용 한 로버스트 CERES도표를 제안하였다.

본 연구에서는 CERES도표와 후버의 M-추정치를 이용한 로버스트-CERES도표를 비교해 보고자 한다. 2절에서는 조건부기대치잔차도표(CERES Plots)와 후버의 M-추정치를 이용한 로버스트-조건부기대치잔차도표(Robust-CERES Plots)를 구하는 과정을 간단히 설명하고, 3절에서는 예를 통하여 그 유용성을 비교하고자 한다.

## 2. 조건부기대치 잔차도표

본 절에서는 로버스트 추정방법을 이용한 부분잔차도표를 확장하여 CERES 도표를 제안하고 그 성질을 조사하고자 한다. 다음의 회귀모형 (2.1)을 참 모형으로 간주한다.

$$y = \alpha_0 + \boldsymbol{\alpha}_1^T \boldsymbol{x}_1 + g(x_2) + \varepsilon \tag{2.1}$$

여기에서, $y$ 는 반응변수, $p\times 1$ 회귀변수 $\boldsymbol{x}$ 를 $(p-1)\times 1$ 벡터 $\boldsymbol{x}_1$ 와 첨가하고자 하는 회귀변수 $x_2$로 분할하고, $g$ 는 평균이 0 인 미지함수이고, $E(g)=0$, $E(\varepsilon)=0$, 그리고 $Var(\varepsilon)=\sigma^2$라고 가정한다.

CERES 도표는 모형 (2.2)를 이용하여 조건부기대치 잔차를 구한다.

$$y = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{x}_1 + \boldsymbol{\beta}_2^T \boldsymbol{m}(x_2) + \varepsilon \tag{2.2}$$

여기에서 $\boldsymbol{m}(x_2) = E(\boldsymbol{x}_1 | x_2) - E(\boldsymbol{x}_1)$이다. 모형(2.2)에서 $\boldsymbol{x}_1$과 $\boldsymbol{m}(x_2)$는 선형 독립이 된다. 로버스트 추정에 의해 적합 된 회귀계수와 잔차를 $\widetilde{\boldsymbol{\beta}}_1$, $\widetilde{\boldsymbol{\beta}}_2$ , 그리고 $\tilde{e}$ 라고 하면, 로버스트 CERES 잔차는 다음과 같이 정의할 수 있다.

$$(\ robust CERES)_i = \tilde{e}_i + \widetilde{\boldsymbol{\beta}}_2^T \boldsymbol{m}(x_{i2}) \tag{2.3}$$

여기에서, $\boldsymbol{m}(x_{i2}) = E(\boldsymbol{x}_{i1} | x_{i2}) - E(\boldsymbol{x}_{i1})$이다.

로버스트 CERES도표는 $\{\ \tilde{e} + \widetilde{\boldsymbol{\beta}}_2^T \boldsymbol{m}(x_2),\ x_2\}$가 된다. 특별한 경우로 $\boldsymbol{m}(x_2)$가 $x_2$의 선형이면 부분잔차도표(Partial Residual Plot)와 같고, $\boldsymbol{m}(x_2)$가 $x_2$의 이차형이면 첨가된 부분잔차도표(Augmented Partial Residual Plots)가 된다.

본 연구에서는 로버스트 추정방법으로 후버의 M-추정치를 고려한다. 이 추정법은 매우 효율적인 로버스트추정법일 뿐만 아니라 계산을 빨리 할 수 있는 장점이 있으므로 다이나믹하게 그래프를

사용할 수 있게 된다. 후버의 M-추정치는 목적함수 $\rho$가 볼록함수(Convex Function)이므로 Cook(1993)의 Lemma 4.1에 의해서 다음과 같은 성질을 얻을 수 있다.

**[성질 2.1]**

모형 (2.2)에서 구한 후버의 M-추정치 $\widetilde{\beta}_0$, $\widetilde{\beta}_1$은 참 모형 (2.1)의 $\alpha_0 + E(g(\boldsymbol{x}_2))$ 와 $\boldsymbol{\alpha}_1$의 피셔-일치추정치이다.

한편 식(2.3)의 $\tilde{e}_i$ 는

$$\tilde{\boldsymbol{e}} = \{ \alpha_0 + \boldsymbol{\alpha}_1{}^T \boldsymbol{x}_1 + g(x_2) + \varepsilon \} - \{ \widetilde{\beta}_0 + \widetilde{\boldsymbol{\beta}}_1^T \boldsymbol{x}_1 + \widetilde{\beta}_2 x_2 \}$$

와 같으므로, 식(2.3)이 다음과 같이 표현되고,

$$(\ robustCERES) = (\ \alpha_0 - \widetilde{\beta}_0) + (\ \boldsymbol{\alpha}_1{}^T - \widetilde{\boldsymbol{\beta}}_1^T)\ \boldsymbol{x}_1 + g(x_2) + \boldsymbol{\varepsilon}$$

$\widetilde{\boldsymbol{\beta}}_1^T$ 가 $\boldsymbol{\alpha}_1{}^T$ 에 수렴하므로, 로버스트 CERES도표는 다음과 같이 표현된다.

$$\{ \tilde{\boldsymbol{e}} + \widetilde{\boldsymbol{\beta}}_2^T \boldsymbol{m}(x_2),\ x_2 \} \xrightarrow{p} \{ g(x_2) - E(g(x_2)) + \varepsilon_i,\ x_2 \} \tag{2.4}$$

즉, 로버스트 CERES도표가 함수 $g(x)$의 곡률을 잘 표현 해준다.

로버스트-CERES 도표를 구하는 과정은 크게 다음의 세 단계를 통하여 구한다.

**[단계 1]** 모형(2.2)의 $\boldsymbol{m}(x_2)$를 추정한다.

실제 계산에는 $\boldsymbol{m}(x_2)$의 특징을 잘 표현해주는 유사한 값을 사용하여도 무방하므로, $\boldsymbol{m}(x_2) = E(\boldsymbol{x}_1 | x_2) - E(\boldsymbol{x}_1)$를 비모수적지법을 이용하여 추정한 값을 실제 계산에 사용한다. $E(\boldsymbol{x}_1 | x_2)$의 j-번째 원소의 추정치 $\widehat{E}(x_{1j} | x_2)$는 $\{ x_{1j},\ x_2 \}$도표, 여기에서 j=1...p-1, 에서 평활방법(Smoothing Method)인 LOESS-Curve를 이용하거나 또는 커널함수 비모수추정을 통하여 추정치 $\widehat{\boldsymbol{m}}(x_2) = \widehat{E}(\boldsymbol{x}_1 | x_2) - \widehat{E}(\boldsymbol{x}_1)$를 얻을 수 있다.

**[단계 2]** 후버의 M-추정법을 이용하여 로버스트 회귀계수 $\widetilde{\boldsymbol{\beta}}_1$, $\widetilde{\boldsymbol{\beta}}_2$ ,그리고 잔차 $\tilde{\boldsymbol{e}}$를 구한다. 이 때 모형(2.2)의 $\boldsymbol{m}(x_2)$는 [단계 1]에서 구한 추정치 $\widehat{\boldsymbol{m}}(x_2)$를 사용한다.

**[단계 3]** 로버스트 CERES도표 $\{ \widetilde{e}_i + \widetilde{\boldsymbol{\beta}}_2^T \widehat{\boldsymbol{m}}(x_{i2}),\ x_{i2} \}$를 그린다.

## 3. 예를 통한 비교

예를 통하여 CERES도표와 로버스트 CERES 도표를 비교하고자 한다. 4개의 다른 함수에 대하여 오차항을 고려하지 않은 경우에 대한 비교가 [예3.1]이고, 이상치가 존재하는 경우는 [예3.2]이다.

**[예 3.1]**

100개의 자료를 모형 $y = x_1 + x_2 + g(x_3)$에서 생성하여 조사한다. 적용하는 함수 형태는 다음과 같다.

1. $g_1(x_3) = (x_3-3)^2 + 9$
2. $g_2(x_3) = (x_3-3)^3 + 2$
3. $g_3(x_3) = \dfrac{1}{(x_3+\frac{1}{2})} - 2$
4. $g_4(x_3) = \dfrac{1}{1+e^{-x_3}}$

여기에서 $x_3$는 일양분포 *uniform* (1,26)에서 추출하고, $x_1 = x_3^{-1} + N(0, 0.1^2)$, $x_2 = \log(x_3) + N(0, 0.25^2)$이다.

CERES 도표는 그림(A)이고, 로버스트 CERES 도표는 그림(B)이다. $g(x_3)$를 표현하는 그림은 그림(C)이다. <그림 3.1>과 <그림 3.2>에서 CERES 도표보다 로버스트 CERES 도표가 $g(x_3)$를 잘 표현 해 주고 있음을 알 수 있다.

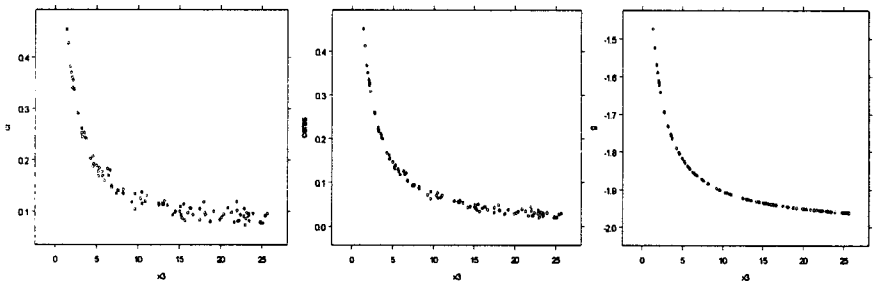
**[예 3.2]**

이상치가 있는 경우를 조사하기 위하여 [예3.1]의 모형에서 처음부터 10개까지의 자료에 $N(0, 0.1^2)$ 그리고 $N(0, 0.01^2)$의 오차항을 더한 후의 결과를 <그림3.3> 에서 <그림3.6>까지 표현하였다. 각각의 그림에서 그림(A)와 그림(B)는 10개의 자료에 오차항 $N(0, 0.1^2)$를 더한 것이고, (C)와 (D)는 오차항 $N(0, 0.01^2)$을 더한 후의 그림이다. (A)와 (C)는 CERES 도표이고, (B)와 (D)는 로버스트 CERES 도표이다.

## 4. 결 론

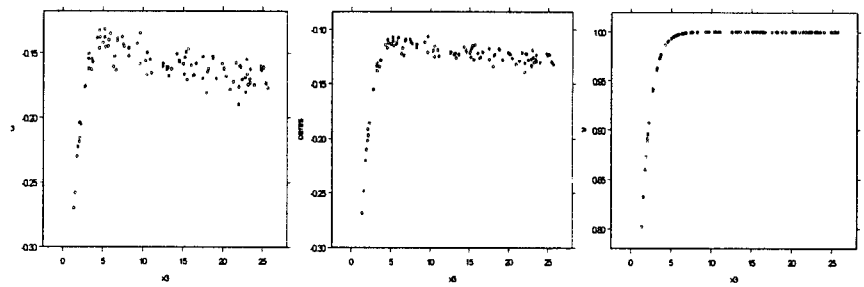
3절에서 설명한 예제 3.1 과 3.2를 통하여 CERES 도표와 로버스트 CERES 도표를 살펴보았다. 예제에서 알 수 있듯이 이상치가 존재하는 경우 CERES 도표나 로버스트 CERES 도표 모두 이상치의 존재를 탐색할 수 있었고 이상치의 존재 유무에 관계없이 함수 $g(\cdot)$의 모습을 잘 나타내고

있음을 알수 있었다. 더욱 로버스트 CERES 도표는 CERES 도표보다 $g(\cdot)$의 모습을 더 잘 나타내고 있음을 발견하였다. 이상치의 탐색에 본 논문에서 제안한 로버스트 CERES 도표를 사용하는 것도 좋은 대안으로 사료된다.



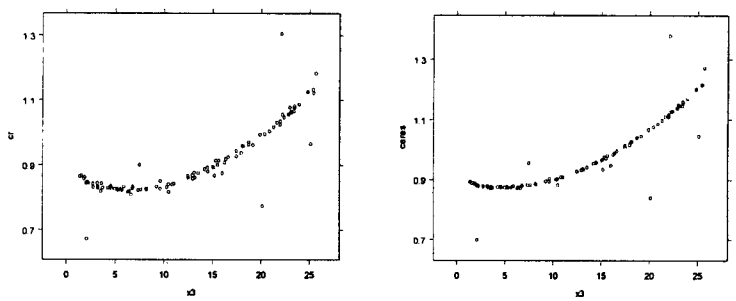
(A) (B) (C)

<그림3.1> [예제3.1]에서 $g_3(x_3)=\frac{1}{(x_3+\frac{1}{2})}-2$ 인 경우의 CERES 도표와 로버스트 CERES 도표



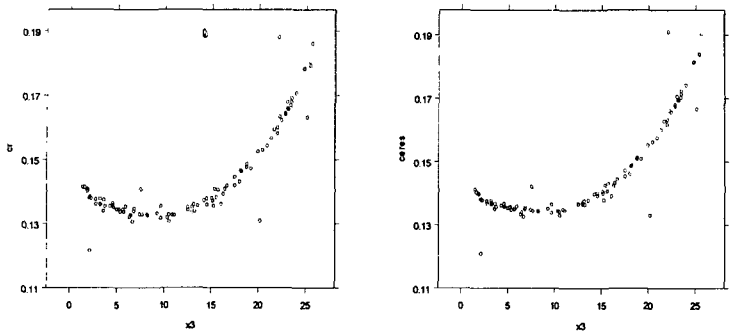
(A) (B) (C)

<그림3.2> [예제3.1]에서 $g_4(x_3)=\frac{1}{1+e^{-x_3}}$ 인 경우의 CERES 도표와 로버스트 CERES 도표



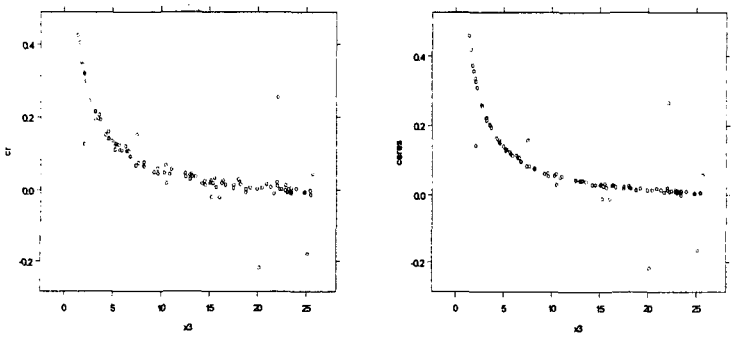
(A) (B)

<그림3.3> [예제3.2]에서 $g_1(x_3)=(x_3-3)^2+9$ 인 경우의 CERES 도표와 로버스트 CERES 도표

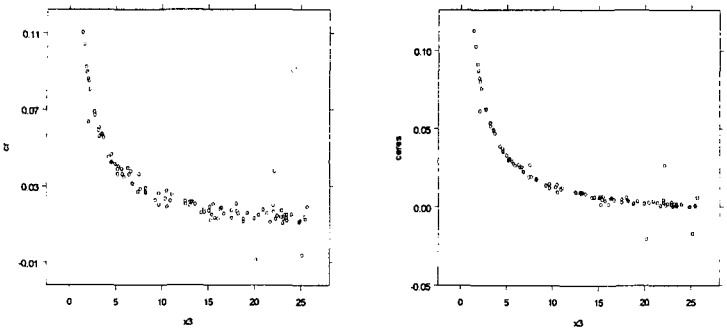



(C) (D)

<그림3.4> [예제3.2] 에서 $g_2(x_3) = (x_3 - 3)^3 + 2$인 경우의 CERES 도표와 로버스트 CERES 도표

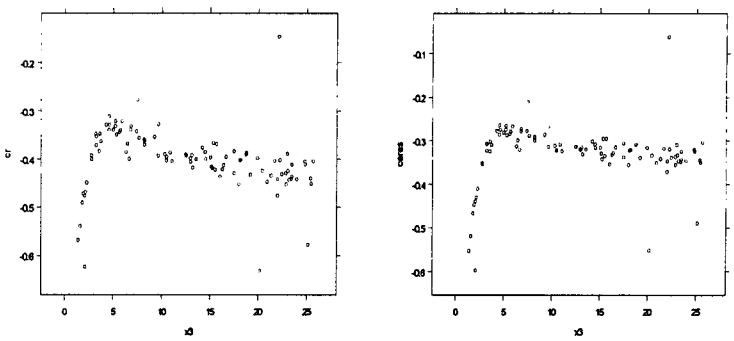



(A) (B)

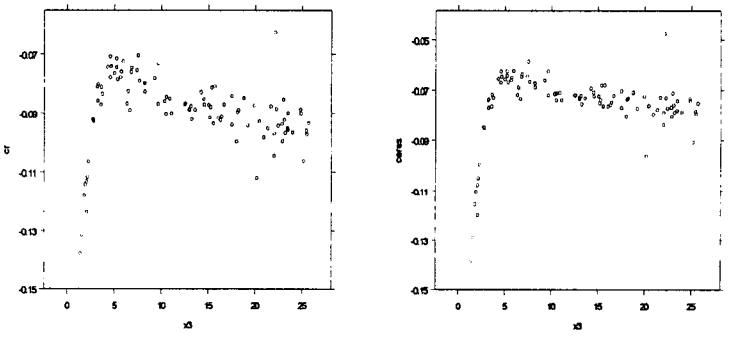



(C) (D)

<그림3.5> [예제3.2]에서 $g_3(x_3) = \dfrac{1}{(x_3 + \frac{1}{2})} - 2$인 경우의 CERES 도표와 로버스트 CERES 도표



(A) (B)



(C) (D)

<그림3.6> [예제3.2]에서 $g_4(x_3)=\frac{1}{1+e^{-x_3}}$ 인 경우의 CERES 도표와 로버스트 CERES 도표

## 참고문헌


[1] Cook, R. D. (1993), Exploring Partial Residual Plots, *Technometrics*, Vol 35, 351-362.

[2] Cook, R. D., and Weisberg, S. (1994). *An Introduction to Regression Graphics*, New York: Wiley.

[3] Berk, K. N., and Booth, D. E. (1995), Seeing a Curve in Multiple Regression, *Technometrics*, Vol 37, 385-398.

[4] Hettmansperger, T. P. (1990), Regression Diagnostics for Rank-Based Methods, *Journal of American Statistical Association*, Vol 85, 1018-1028.

[5] _____(1993), The Use and Interpretation of Residuals Based on Robust Estimation, *Journal of American Statistical Association*, Vol 88, 1254-1263.

[6] _____(1994), Robust and High Breakdown Fits of Polynomial Models, *Technometrics*, Vol

36, 409-413.

[7] Larsen, W. A. and McCleary, S. J. (1972), the Use of Partial Residual Plots in Regression Analysis, *Technometrics,* Vol 14, 781-790.

[8] Mallow, C. L. (1986), Augmented Partial Residual Plots, *Technometrics,* Vol 28, 313-320.

[9] Mckean, J. W. and Sheather, S. J. (2000), Partial Residual Plots Based on Robust Fits, *Technometrics,* Vol 42, 249-261.

[10] Oh, K. S. (2001), Regression Diagnostics Using Residual Plots, *The Korean Communications in Statistics,* vol. 8, 311-317.

[11] Oh, K. S., and Do, S. H. (2003), CERES Plots Based on Robust Fits, *The Journal of Korean Data Analysis Society,* Vol. 5, 233-241.