

CERES Plot in Generalized Linear Models¹⁾

Myung-Wook Kahng²⁾ and Eun Jeong Lee³⁾

Abstract

We explore the structure and usefulness of CERES plot as a basic tool for dealing with curvature as a function of the new predictor in generalized linear models. If a predictor has a nonlinear effect and there are nonlinear relationships among the predictors, the partial residual plot and augmented partial residual plot are not able to display the correct functional form of the predictor. Unlike these plots, the CERES plot can show the correct form. This is illustrated by simulated data.

Keywords : generalized linear model, CERES plot, LOWESS

1. 서 론

정규이론에 근거한 선형모형은 설명변수의 선형결합(linear predictor)이 직접 반응변수를 설명한다. 이러한 선형모형은 분산분석, 선형회귀분석 등에서 매우 다양한 모형의 틀을 제공하는 것이 사실이지만 모든 상황에서 충분한 것은 아니다. 반응변수가 이항반응자료처럼 이산형인 경우에는 원칙적으로 일반 선형모형은 적절하지 않고 정규이론에 의한 고전적 선형모형을 확장한 일반화선형모형(generalized linear models)으로 해결이 가능하다.

회귀분석에서 기존에 사용되고 있는 모형에 새로운 설명변수를 추가하고자 할 때 추가변수가 회귀모형에 어떤 영향을 미치는지를 알아보아야 하고 추가될 때의 형태에 대한 판단이 필요하며 선형이 아닌 형태로 추가되어야 한다면 추가변수가 어떤 형태로 변환되어야 하는지를 진단하여야 한다. 이러한 진단은 그림을 통한 방법으로 가능하며 지금까지 활발히 연구가 진행되고 있는 방법으로 추가변수그림(added variable plot)과 편잔차그림(partial residual plot), 그리고 편잔차그림을 확장시킨 덧편잔차그림(augmented partial residual plot)과 CERES그림(combining conditional expectation and residual plot)이 있다.

편잔차그림은 Ezekiel(1924)에 의해 처음 제시되었고 Cook과 Weisberg(1982) 등에 의해 회귀진단의 도구로 사용되었다. 편잔차그림을 개선한 덧편잔차그림은 Mallows(1986)에 의해 소개되었고 Cook(1993)은 더욱 일반화시킨 CERES그림을 새로운 진단방법으로 제시하였다. 이와 같이 그림을 이용한 비선형성의 탐색은 Atkinson(1985), Chatterjee and Hadi(1988), Cook and Weisberg(1999)

1) This research was supported by the Sookmyung Women's University Research Grants 2003.

2) Professor, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea
E-mail : mwkahng@sookmyung.ac.kr

3) Graduate Student, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea

등에 의해 회귀진단의 도구로 많은 연구가 있어 왔다. 최근에는 Kahng and Kim(1998)이 비선형 모형에서의 편잔차그림에 대한 연구로 발전시켰고, Kahng and Jeong(2000)은 CERES그림을 통해 비선형모형에서 추가변수의 유의성 및 비선형성에 대한 진단으로 확장하였다.

본 연구에서는 일반화선형모형에서 CERES그림을 통한 추가변수의 필요성과 함수의 형태에 대해 연구하고자 한다. 2절에서는 일반화선형모형을 소개하고 3절에서는 CERES그림을 일반화선형모형에 적용시켜본다. 4절에서는 일반화선형모형의 하나인 이항회귀모형(binomial regression model)에서 CERES그림을 통하여 추가변수의 형태를 알 수 있는 지를 확인해본다.

2. 일반화선형모형

Nelder and Wedderburn(1972)에 의해 체계화된 일반화선형모형은 지수족(exponential family) 분포와 연결함수(link function)를 이용하여 정규이론에 의한 선형모형을 다음과 같은 두 가지 과정으로 일반화한 것이다. 첫째, 오차의 분포는 정규분포를 포함하는 지수족의 여러 가지 분포를 사용한다. 둘째, 반응변수의 기대값과 설명변수의 선형결합을 연결시키는 연결함수를 설정한다. 고전적 선형모형은 반응변수가 서로 독립적이며 정규분포를 따르고 연결함수가 항등함수(identity function)인 일반화선형모형의 특수한 형태라고 할 수 있다.

먼저 반응변수가 다양한 형태의 확률분포를 가질 수 있도록 하기 위해서는 공통된 특성을 갖는 확률분포의 집합을 정의할 필요가 있는데 확률변수 Y 의 확률밀도함수가

$$f_Y(y; \theta, \phi) = \exp \{ (y\theta - b(\theta)) / a(\phi) + c(y, \phi) \} \quad (2.1)$$

의 형태로 표현되고 ϕ 가 알려져 있다면 Y 의 분포는 지수족에 속한다고 한다. 여기서 $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ 는 특정 실수함수이다. 지수족에 대한 좀 더 자세한 내용은 Barndorff and Nilsen(1978)에 설명되어 있다. 실제로 우리가 사용하고 있는 많은 분포들이 지수족에 속하는데 분산이 알려져 있는 경우의 정규분포, 이항분포, 포아송분포, 감마분포 등이 여기에 해당된다.

식 (2.1)로 부터 Y 에 대한 로그우도함수(log-likelihood function)는 다음과 같이 나타내어진다.

$$l(\theta, \phi; y) = \log f_Y(y; \theta, \phi) = (y\theta - b(\theta)) / a(\phi) + c(y, \phi)$$

Y 의 기대값과 분산은 다음의 두 관계식 $E(\partial l / \partial \theta) = 0$ 과 $E(\partial^2 l / \partial \theta^2) + E(\partial l / \partial \theta)^2 = 0$ 으로부터 $E(Y) = \mu = b(\theta)$, $Var(Y) = b''(\theta)a(\phi)$ 가 된다.

일반화선형모형에서 반응변수 Y 는 설명변수 $\mathbf{x}^T = (1, x_1, x_2, \dots, x_{p-1})$ 의 선형결합 η 를 통해서 결정된다. 이때 선형결합 η 는 연결함수 $g(\cdot)$ 에 의하여 $\mu = E(Y)$ 와 연결된다. 설명변수 \mathbf{x}^T 의 n 개의 관측값에 대하여 선형결합은

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i, p-1}, \quad i = 1, \dots, n \quad (2.2)$$

이다. 선형결합 (2.2)에 추가시키려고 하는 새로운 설명변수 z 가 선형 또는 형태가 알려지지 않은 비선형함수 $h(z)$ 의 형태로 추가된다고 하면 선형결합을 다음과 같이 나타낼 수 있고

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + h(z_i), \quad i = 1, \dots, n \quad (2.3)$$

이 모형 (2.3)을 설명변수와 반응변수의 관계를 나타내는 참모형이라 하자. 여기서 $h(\cdot)$ 는 선형 또는 형태가 알려지지 않은 비선형함수이다.

3. CERES그림

3.1 편잔차그림과 덧편잔차그림

선형모형에서와 마찬가지로 일반화선형모형에서도 추가변수 z 에 대한 변환의 필요성을 확인하기 위하여 잔차산점도와 잔차-설명변수산점도를 그려보고 만약 비선형의 형태가 나타나면 z 가 어떠한 형태의 비선형함수 $h(z)$ 로 변환되어야 할지를 알아보기 위해 편잔차그림을 사용할 수 있다.

만약 모형 (2.3)에서 $h(z)$ 의 함수형태가 z 에 대한 선형함수라면 이 모형은 다음과 같은 모형이 된다.

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} + \alpha z_i, \quad i = 1, \dots, n \quad (3.1)$$

추가변수 z 의 편잔차그림은 모형 (3.1)에 적합 시킨 후 얻어지는 잔차 e 에 $\hat{\alpha}z$ 를 더한 편잔차를 세로축으로 하고 추가변수 z 를 가로축으로 하는 $\{e_i + \hat{\alpha}z_i, z_i\}$ 의 산점도이다. 이 때 $\hat{\alpha}$ 은 모형 (3.1)에서 구한 z 의 회귀계수추정값이다.

모형 (2.3)이 참모형이라면 편잔차그림의 세로축을 이루게 되는 편잔차의 기대값은 다음과 같다.

$$E(e_i + \hat{\alpha} z_i | z_i) = E(\mathbf{x}_i | z_i)^T (\boldsymbol{\beta}^* - \boldsymbol{\beta}) + h(z_i) \quad (3.2)$$

모형 (3.1)의 $\boldsymbol{\beta}$ 가 모형 (2.3)에서의 실제모수 $\boldsymbol{\beta}^*$ 에 근접하면 $\boldsymbol{\beta}^* - \boldsymbol{\beta}$ 가 0에 가까워지고 식 (3.2)에 의해 편잔차그림은 $h(z)$ 의 형태를 잘 묘사한다.

만약 $h(z)$ 가 z 에 대한 선형함수라면 모형 (3.1)은 참모형 (2.3)의 근사적인 모형이 되고 $\boldsymbol{\beta}$ 가 $\boldsymbol{\beta}^*$ 에 근접하고 편잔차그림은 $h(z)$ 의 형태를 잘 나타내게 된다. $h(z)$ 가 z 에 대한 비선형인 경우에도 Cook(1993)에 의하면 모형 (2.3)이 참모형이고 조건부기대값 $E(\mathbf{x} | z)$ 이 z 와 연관성이 약하거나 z 에 대해 선형이면 모형 (3.1)에서의 추정량은 $\boldsymbol{\beta}^*$ 의 일치추정량이 되므로 편의가 없어져서 편잔차그림은 $h(z)$ 의 형태를 나타낼 수 있다.

그러나 $E(\mathbf{x} | z)$ 가 z 에 대해 선형함수가 아니고 특히 설명변수들간의 연관성이 강한 경우 편잔차그림은 $h(z)$ 의 형태에 대한 올바른 정보를 주지 못하며 비선형성을 감지할 수 없다. 이러한 경우 모형 (3.1)에 z 의 이차항을 추가시킨 다음과 같은 모형을 생각하자.

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} + \alpha_1 z_i + \alpha_2 z_i^2, \quad i = 1, \dots, n \quad (3.3)$$

덧편잔차그림은 모형 (3.3)에 적합 시킨 후 얻어지는 잔차 e 에 $\hat{\alpha}_1 z$ 와 $\hat{\alpha}_2 z^2$ 를 더한 덧편잔차를 세로축으로 하고 관심의 대상이 되는 변수 z 를 가로축으로 하는 $\{e_i + \hat{\alpha}_1 z_i + \hat{\alpha}_2 z_i^2, z_i\}$ 의 산점도이다. 이때 $\hat{\alpha}_1$ 과 $\hat{\alpha}_2$ 은 각각 모형 (3.3)에서 구한 z 과 z^2 각각의 회귀계수추정값이다.

모형 (2.3)이 참모형이라면 덧편잔차그림의 세로축을 이루는 덧편잔차의 기대값은 다음과 같다.

$$E(e_i + \hat{\alpha}_1 z_i + \hat{\alpha}_2 z_i^2 | z_i) = E(x_i | z_i)^T (\beta^* - \beta) + h(z_i) \quad (3.4)$$

만약 $h(z)$ 가 근사적으로 z 의 이차함수이면 모형 (3.3)은 참모형 (2.3)의 근사적인 모형이 되고 모형 (3.3)의 β 는 편잔차그림에서 사용된 모형 (3.1)에 비해 β^* 에 보다 근접하게 된다. 따라서 식 (3.4)에 의해 덧편잔차그림은 $h(z)$ 의 형태를 잘 나타낸다. $h(z)$ 가 z 의 선형이나 이차함수가 아닌 경우에도 Cook(1993)과 Kahng and Kim(1998)에 의하면 모형 (2.3)이 참모형이고 조건부기대값 $E(x|z)$ 이 z 의 이차함수이면 모형 (3.3)에서 추정량은 β^* 의 일치추정량이 되므로 편의가 없어져서 덧편잔차그림은 $h(z)$ 의 형태를 나타낼 수 있다.

3.2 CERES그림

일반화선형모형에서도 추가변수 z 의 형태가 선형이나 이차함수일 경우 편잔차그림이나 덧편잔차그림으로 z 의 함수형태를 파악할 수 있다. 그러나 추가변수 z 의 형태가 선형이나 이차함수가 아닌 비선형함수이거나 조건부기대값 $E(x|z)$ 이 z 의 선형 또는 이차함수가 아닌 보다 복잡한 형태의 함수일 경우에는 편잔차그림이나 덧편잔차그림으로는 z 의 함수형태에 대한 정확한 파악이 어렵다. 다음과 같은 선형결합을 생각해보자.

$$\eta_i = \mathbf{x}_i^T \beta + \alpha^T \mathbf{m}(z_i), \quad i = 1, \dots, n \quad (3.5)$$

여기서 $\mathbf{m}(z) = \{m_j(z)\}$ 는 $(p-1) \times 1$ 벡터이고 조건부기대값 $E(x|z)$ 로 정의하며 $\mathbf{m}(z)$ 의 값이 사전에 알려지지 않았기 때문에 $m_j(z), j = 1, \dots, p-1$ 는 각각 x_j 를 반응변수로하고 z 에 의해 비모수회귀시켜 구한 적합값으로 다음과 같이 추정한다.

$$\hat{\mathbf{m}}(z) = \{\hat{m}_j(z)\} = \{\hat{E}(x_j|z), j = 1, \dots, p-1\} \quad (3.6)$$

비모수회귀적합은 이미 알려져 있는 여러 가지 방법 중 어느 것을 사용하여도 그 결과에 거의 차이가 없다(Cook, 1993). 여기서는 그 중 LOWESS방법을 사용한다. 모형 (3.5)에서 $\mathbf{m}(z_i)$ 을 식 (3.6)의 $\hat{\mathbf{m}}(z_i)$ 로 바꾸면 다음과 같은 모형이 된다.

$$\eta_i = \mathbf{x}_i^T \beta + \alpha^T \hat{\mathbf{m}}(z_i), \quad i = 1, \dots, n \quad (3.7)$$

추가변수 z 의 CERES그림은 모형 (3.7)에 적합시킨 후 얻어지는 잔차 e 에 $\hat{\alpha}^T \hat{m}(z)$ 를 더한 CERES를 세로축으로 하고 추가변수 z 를 가로축으로 하는 $\{e_i + \hat{\alpha}^T \hat{m}(z_i), z_i\}$ 의 산점도이다. 이때 $\hat{\alpha}^T$ 은 모형 (3.7)에서 구한 $\hat{m}(z)$ 의 회귀계수 추정값벡터이다.

모형 (2.3)이 참모형이라면 CERES그림의 세로축을 이루는 CERES의 기대값은 다음과 같다.

$$E(e_i + \hat{\alpha}^T \hat{m}(z_i) | z_i) = E(x_i | z_i)^T (\beta^* - \beta) + h(z_i) \quad (3.8)$$

모형 (3.7)의 β 가 참모형 (2.3)에서의 실제모수 β^* 에 근접하면 식 (3.8)에 의해 CERES그림은 $h(z)$ 의 형태를 잘 묘사한다.

만약 $h(z)$ 가 $m_j(z), j=1, \dots, p-1$ 의 선형함수라면 $h(z)$ 는 $\hat{\alpha}^T \hat{m}(z)$ 로 근사가 가능하며 이때 모형 (3.7)은 참모형인 (2.3)의 근사모형이 되고 β 는 실제모수 β^* 에 근접하게 된다. 따라서 CERES그림은 $h(z)$ 의 형태를 잘 나타내게 된다. 하지만 $h(z)$ 가 $m_j(z)$ 의 선형함수가 아니면 β 와 β^* 가 근접한다고 하기 어렵고 편의가 발생하여 CERES그림은 $h(z)$ 을 제대로 나타내지 못한다고 말할 수도 있을 것이다.

편잔차그림과 뒤편잔차그림에서 $E(x|z)$ 가 각각 z 의 선형 또는 이차함수일 때 모형 (3.5)에서의 $\alpha^T m(z)$ 를 z 의 선형과 이차함수로 하는 모형 (3.1)과 (3.3)에서 β 의 추정량이 일치추정량이 되고 이러한 경우 편의가 없어져서 편잔차그림과 뒤편잔차그림이 $h(z)$ 의 형태를 잘 나타낸다고 하였다. 모형 (3.7)에서 $m(z)$ 를 $E(x|z)$ 로 정의하고 $\hat{E}(x|z)$ 로 추정하였으므로 $E(x|z)$ 는 당연히 $m_j(z)$ 의 선형함수로 볼 수 있으며 이 모형에서 β 의 추정량은 실제모수 β^* 의 일치추정량이라 할 수 있고 식 (3.8)에서 편의가 없어지므로 CERES그림이 $h(z)$ 의 형태를 잘 묘사한다. 일치추정량이 된다는 사실은 Cook(1993)에 증명되어 있다.

4. 예 제

일반화선형모형에서 $E(x|z)$ 이 z 에 대해 선형 혹은 이차함수가 아닌 경우 CERES그림이 추가변수의 비선형성의 진단과 비선형함수의 형태를 잘 나타내 주는지를 알아보려고 한다. 다음의 함수 관계를 실제 선형결합이라 하자.

$$\eta = \beta_0 + \beta_1 x + 0.5e^{-z} \quad (4.1)$$

추가변수 z 는 $U(0,5)$ 에서 30개의 값을 생성시키고 설명변수 x 는 $x = \log(1/z) + \epsilon$ 을 이용하여 $E(x|z)$ 가 z 와 비선형관계가 되도록 하였다. 여기서 ϵ 은 $N(0, 0.2^2)$ 에서 생성한다. 또한 회귀계수는 각각 $\beta_0 = 1, \beta_1 = 0.5$ 로 결정하였다. 이와 같이 결정한 x, z, β 를 이용하여 모형 (4.1)에 적용시켜 30개의 η 를 계산하였다. 두 설명변수의 산점도인 $\{x_i, z_i\}$ 를 그려보면 $E(x|z)$ 는 z 에 선형이나 이차함수의 관계가 아님을 알 수 있다.

이때 $\eta = \log[\pi(x)/(1-\pi(x))]$ 라고 하면 $\pi(x)$ 는 $\pi(x) = \exp(\eta)/(1+\exp(\eta))$ 로 30개의 비율 형태의 값이 된다. $\pi(x)$ 가 $m=50$ 번의 시행에서 특정사건이 발생할 확률이라고 생각하고 y 는 m 번의 시행에서 특정사건이 발생한 도수로 보고 $B(50, \pi(x))$ 인 이항분포에 의해 생성하였다. 따라서 반응변수 y 는 이항회귀모형의 반응변수로 볼 수 있다. 만약 선형결합 $\eta = \beta_0 + \beta_1 x$ 에 추가변수 z 가 $h(z)$ 의 형태로 추가된다고 생각하면 모형 (4.1)은 $\eta = \beta_0 + \beta_1 x + h(z)$ 가 된다. 함수 $h(z)$ 의 형태는 <그림 1>과 같다.

이때, 잔차산점도를 이용하여 추가변수 z 의 비선형성을 확인한 결과 비선형성이 탐지되었다. 추가변수가 어떠한 형태로 변환되어 모형에 추가되어야 하는지 알아보기 위하여 편잔차그림이나 덧편잔차그림을 적용해 볼 수 있는데 $E(x|z)$ 는 z 에 선형이나 이차함수가 아니므로 z 의 편잔차그림이나 덧편잔차그림이 추가변수 z 의 함수 형태를 올바르게 나타낼 것으로 기대할 수 없다. 실제 편잔차그림인 <그림 2>는 특정함수의 형태를 취하지 않고 실제함수 형태인 <그림 1>과 같은 지수함수의 형태를 나타내지 못한다. 덧편잔차그림에서도 특정함수의 형태를 보이지 않았다. 따라서 추가변수 z 에 대한 올바른 정보를 제공해 주지 못한다.

CERES그림에 적용하기 위해 다음의 선형결합을 생각해 보자.

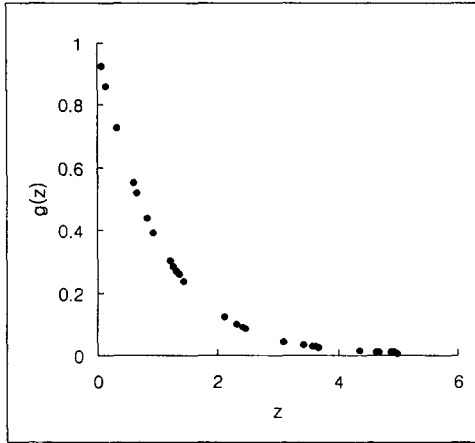
$$\eta = \beta_0 + \beta_1 x + \alpha m(z) \quad (4.2)$$

$m(z)$ 는 사전에 알려지지 않은 값이기 때문에 비모수회귀적합에 의해 $\hat{m}(z) = \{\hat{E}(x|z)\}$ 를 추정할 수 있다. 비모수회귀적합은 이미 알려져 있는 여러 가지 방법 중에 LOWESS방법을 사용하고 모형 (4.2)에서 $m(z)$ 를 추정된 $\hat{m}(z)$ 로 바꾸면 다음과 같이 변형된다.

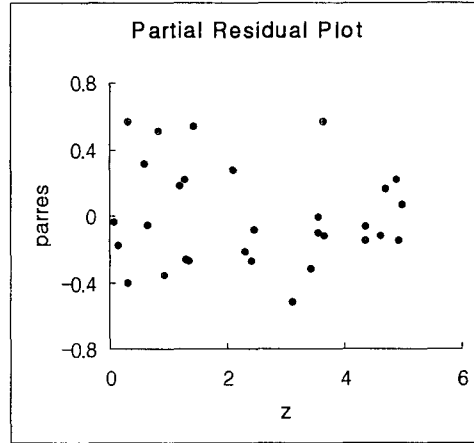
$$\eta = \beta_0 + \beta_1 x + \alpha \hat{m}(z) \quad (4.3)$$

모형 (4.3)을 적합하여 구한 최소제곱추정값 $\hat{\beta}_0, \hat{\beta}_1, \hat{\alpha}$ 를 이용하여 CERES그림을 그려보면 <그림 3>과 같고 편잔차그림이나 덧편잔차그림에서 나타나지 않았던 z 의 실제함수의 형태가 <그림 1>과 같이 지수함수형태로 나타남을 알 수 있다. 따라서 $E(x|z)$ 이 z 에 대해 선형이나 이차함수의 형태가 아닌 비선형함수일 경우에도 CERES그림은 z 의 변환의 형태를 잘 나타내 주는 것을 확인할 수 있다.

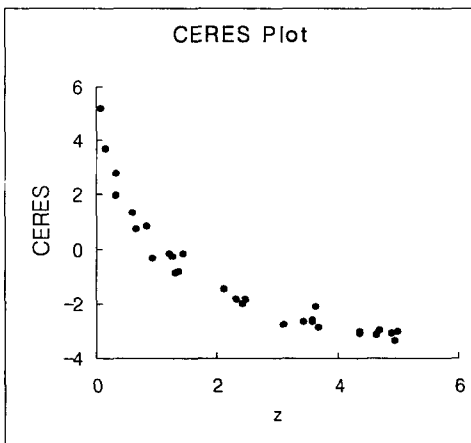
그러나 <그림 3>에서 보이는 비선형함수는 z 의 역함수로 판단할 수도 있다. 변환이 적절하다면 변환된 추가변수의 CERES그림은 선형의 형태를 나타내야 할 것이다. 이를 확인하기 위하여 역함수로 변환된 추가변수 $z' = 1/z$ 의 CERES그림을 그려본 결과 선형의 형태가 아니었으므로 역함수로의 변환은 적절하지 않으며 지수함수로 변환된 추가변수 $z'' = e^{-z}$ 의 CERES그림은 직선형태를 보였다. 따라서 지수함수의 형태로 변환하는 것이 바람직하다고 할 수 있다.



<그림 1> z에 대한 $h(z)$ 의 형태



<그림 2> z에 대한 편잔차그림



<그림 3> z에 대한 CERES그림

6. 결론

본 논문에서는 일반화선형모형에서 새로운 변수를 추가하고자 할 때 그림을 이용하여 그 변수의 유의성을 확인하고 추가변수의 적절한 함수형태를 찾는 방법을 알아보았다. 선형결합에 이미 포함되어 있던 변수와 추가변수 사이에 연관성이 있고 그 형태가 선형이나 이차함수가 아닌 경우에도 CERES그림은 추가변수의 유의성과 변환의 필요성, 비선형 형태까지도 제시할 수 있다. 이러한 일반화선형모형에서의 CERES그림의 기능을 이항회귀모형의 예를 통하여 확인하였다.

참 고 문 헌

- [1] Atkinson, A. C. (1985), *Plots, Transformations, and Regression*, Oxford University Press; Oxford.
- [2] Bandorff and Neilsen, D. E. (1978), *Information and Exponential Families in Statistical Theory*, Wiley; Chichester.
- [3] Chatterjee, S. and Hadi, A. S. (1988), *Sensitivity Analysis in Linear Regression*, John Wiley & Sons; New York.
- [4] Cook, R. D. (1993), Exploring partial residual plots, *Technometrics*, Vol. 35, 351-362.
- [5] Cook, R. D. and Weisberg, S. (1982), *Residuals and influence in regression*, Chapman & Hall; New York.
- [6] Cook, R. D. and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics*, John Wiley and Sons, New York.
- [7] Ezekiel, M. (1924), A method for handling curvilinear correlation for any number of variables, *Journal of the American Statistical Association*, Vol. 19, 431-453.
- [8] Kahng, M. and Jeong, H. (2000), CERES plot in nonlinear regression, *The Korean Communications in Statistics*, Vol. 7, 1-11.
- [9] Kahng, M. and Kim, J. (1998), Partial residual plot in nonlinear regression, *The Korean Communications in Statistics*, Vol. 3, 571-580.
- [10] Mallows, C. L. (1986), Augmented partial residuals, *Technometrics*, Vol. 28, 313-319.
- [11] Nelder, J. A. and Wedderburn, R. W. M. (1972), Generalized linear models, *Journal of Royal Statistical Society, A*, Vol. 135, 370-384.

[2004년 7월 접수, 2004년 10월 채택]