The Korean Communications
in Statistics Vol. 7, No. 1, 2000
pp. 1-11

CERES Plot in Nonlinear Regression[1)]

Myung-Wook Kahng[2)] and Hye-Wook Jeong[3)]

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 nonlinear regression. 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. In situations where nonlinearity exists in two predictors, we extend the idea of CERES plot to three dimensions. This is illustrated by simulated data.

1. 서 론

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

편잔차그림은 Ezekiel(1924)에 의해 처음 제시되었고 Cook과 Weisberg(1982), Atkinson(1985), Chatterjee와 Hadi(1988) 등에 의해 회귀진단의 도구로 사용되었다. 편잔차그림을 개선한 덧편잔차그림은 Mallows(1986)에 의해 소개되었고 Cook(1993)은 더욱 일반화시킨 CERES그림을 새로운 진단방법으로 제시하였다. 또한 Cook과 Weisberg(1989), Berk와 Booth(1995), Berk(1998) 등에 의해 회전그림(rotating plot), 애니메이션그림(animation plot)과 같은 3차원 그림을 회귀진단에 적용하는 방법이 활발히 연구되어 왔다.

이러한 연구는 주로 선형회귀분석을 대상으로 이루어져 왔는데 비선형회귀분석에서는 추가되는 변수가 선형으로 모형에 추가되기보다는 비선형의 형태로 추가될 가능성이 더욱 높으므로 비선형 회귀분석에서 비선형성의 진단은 선형회귀분석에 비해 더욱 중요한 문제이다. 본 논문에서는 비선형성의 진단에 사용되는 여러 그림들의 이론적인 근거를 알아보고 선형회귀분석에서 뿐만 아니라 비선형회귀분석에도 적용할 수 있는지를 알아보고자 한다.

---

1) This research was supported by the Sookmyung Women's University Research Grants in 1999.
2) Associate Professor, Department of Statistics, Sookmyung Women's University, Seoul, 140-742 KOREA
3) Lecturer, Department of Statistics, Sookmyung Women's University, Seoul, 140-742 KOREA

2. 비선형회귀모형

반응변수 y와 q개의 설명변수 $\boldsymbol{x}^T=(x_1, x_2, \cdots, x_q)$의 n개의 관측값에 대하여 다음과 같은 함수관계가 알려진 비선형회귀모형을 생각하자.

$$y_i = f(\boldsymbol{x}_i; \boldsymbol{\theta}) + \varepsilon_i, \quad i=1,2,\cdots,n \tag{2.1}$$

여기서 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iq})^T$는 설명변수의 i번째 관찰값들로 이루어진 설명변수벡터이고 $\boldsymbol{\theta}$는 $p\times 1$인 회귀계수벡터이다. 이 때 ε_i는 오차항으로 서로 독립적이고 평균 $E(\varepsilon_i)=0$, 분산 $Var(\varepsilon_i)=\sigma^2$인 정규분포를 따른다고 가정한다.

모형 (2.1)에 포함시키려는 새로운 설명변수 z가 $g(z)$의 형태로 추가된다고 생각하면 모형은 다음과 같고 이것을 설명변수와 반응변수의 관계를 나타내는 참모형(true model)이라고 하자.

$$y_i = f(\boldsymbol{x}_i; \boldsymbol{\theta}) + g(z_i) + \varepsilon_i, \quad i=1,2,\cdots,n \tag{2.2}$$

여기서 $g(\cdot)$은 선형 또는 형태가 알려지지 않은 비선형함수이다.

3. CERES그림

3.1 편잔차그림과 덧편잔차그림

비선형회귀분석에서도 설명변수 z에 대한 변환의 필요성을 확인하기 위하여 잔차산점도와 잔차-설명변수산점도를 그려보고 만약 비선형의 형태가 나타나면 z가 어떠한 형태의 비선형함수 $g(z)$로 변환되어야 할 지를 알아보기 위해 편잔차그림을 사용할 수 있다.

만약 모형 (2.2)에서 $g(z)$의 함수형태가 z에 대한 선형함수라면 이 모형은 다음과 같은 선형모형이 된다.

$$y_i = f(\boldsymbol{x}_i; \boldsymbol{\beta}) + az_i + \delta_i, \quad i=1,2,\cdots,n \tag{3.1}$$

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

모형 (2.2)가 참모형이라면 편잔차그림의 세로축을 이루게 되는 편잔차는 다음과 같다.

$$e_i+\hat{a}z_i = f(\boldsymbol{x}_i; \boldsymbol{\theta}) - f(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}) + g(z_i) + \varepsilon_i \tag{3.2}$$

모형 (3.1)에 적합하여 얻은 $\hat{\boldsymbol{\beta}}$이 모형 (2.2)에서의 실제모수 $\boldsymbol{\theta}$에 근접하면 $f(\boldsymbol{x}_i; \boldsymbol{\theta}) - f(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}})$

이 0에 가까워지고 식 (3.2)에 의해 편잔차그림은 $\{g(z_i)+\varepsilon_i, z_i\}$의 산점도가 되어 $g(z)$의 형태를 잘 묘사한다.

만약 $g(z)$가 z에 대한 선형함수라면 모형 (3.1)은 참모형 (2.2)에 근사적인 모형이 되고 $\hat{\boldsymbol{\beta}}$이 $\boldsymbol{\theta}$에 근접하고 편잔차그림은 $g(z)$의 형태를 잘 나타내게 된다. $g(z)$가 z에 대한 비선형인 경우에도 Kahng과 Kim(1998)에 의하면 모형 (2.2)가 참모형이고 조건부기대값 $E(\boldsymbol{v}\mid z)$이 z와 연관성이 약하거나 z에 대해 선형이면 $\hat{\boldsymbol{\beta}}$은 $\boldsymbol{\theta}$의 일치추정량이 되므로 편의가 없어져서 편잔차그림은 $g(z)$의 형태를 나타낼 수 있다. 여기서 $\boldsymbol{v}=\{v_j\}=\{\partial f/\partial\theta_j\}=\partial f/\partial\boldsymbol{\theta}$ 이다.

그러나 $E(\boldsymbol{v}\mid z)$가 z에 대해 선형함수가 아니고 특히 설명변수들간의 연관성이 강한 경우 편잔차그림은 $g(z)$의 형태에 대한 올바른 정보를 주지 못하며 비선형성을 감지할 수 없다. 이러한 경우 모형 (3.1)에 z의 2차함수를 추가시킨 다음과 같은 모형을 생각하자.

$$y_i=f(\boldsymbol{x}_i;\boldsymbol{\beta})+\alpha_1 z_i+\alpha_2 z_i^2+\delta_i,\quad i=1,2,\cdots,n \tag{3.3}$$

덧편잔차그림은 모형 (3.3)에 적합시킨 후 얻어지는 잔차 e에 $\hat{\alpha}_1 z$와 $\hat{\alpha}_2 z^2$를 더한 덧편잔차를 세로축으로 하고 관심의 대상이 되는 변수 z를 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 z과 z^2 각각의 회귀계수추정값이다.

모형 (2.2)가 참모형이라면 덧편잔차그림의 세로축을 이루는 덧편잔차는 다음과 같다.

$$e_i+\hat{\alpha}_1 z_i+\hat{\alpha}_2 z_i^2=f(\boldsymbol{x}_i;\boldsymbol{\theta})-f(\boldsymbol{x}_i;\breve{\boldsymbol{\beta}})+g(z_i)+\varepsilon_i \tag{3.4}$$

만약 $g(z)$가 근사적으로 z의 2차함수이면 모형 (3.3)은 참모형 (2.2)에 근사적인 모형이 되고 모형 (3.3)에 적합하여 얻은 $\breve{\boldsymbol{\beta}}$은 편잔차그림에서 사용된 모형 (3.1)에서 구한 $\hat{\boldsymbol{\beta}}$에 비해 $\boldsymbol{\theta}$에 보다 근접하게 된다. 따라서 식 (3.4)에 의해 덧편잔차그림은 $\{g(z_i)+\varepsilon_i, z_i\}$의 산점도가 되어 $g(z)$의 형태를 잘 나타낸다. $g(z)$가 z의 선형이나 2차함수가 아닌 경우에도 Kahng과 Kim(1998)에 의하면 모형 (2.2)가 참모형이고 조건부기대값 $E(\boldsymbol{v}\mid z)$이 z의 2차함수이면 $\breve{\boldsymbol{\beta}}$는 $\boldsymbol{\theta}$의 일치추정량이 되므로 편의가 없어져서 덧편잔차그림은 $g(z)$의 형태를 나타낼 수 있다.

3.2 CERES 그림

모형에 추가되는 변수 z의 형태가 선형이나 2차함수가 아닌 비선형함수이거나 조건부기대값 $E(\boldsymbol{v}\mid z)$가 z의 선형 또는 2차함수가 아닌 경우 다음과 같은 모형을 생각하자.

$$y_i=f(\boldsymbol{x}_i;\boldsymbol{\phi})+\boldsymbol{\alpha}^T\boldsymbol{m}(z_i)+\xi_i,\quad i=1,2,\cdots,n \tag{3.5}$$

여기서 $\boldsymbol{m}(z)=\{m_j(z)\}$는 $p\times 1$인 벡터이고 조건부기대값 $E(\boldsymbol{v}\mid z)$로 정의하며 $\boldsymbol{m}(z)$의 값

은 사전에 알려져 있지 않기 때문에 추정해야 한다. $m_j(z), j=1,2,\cdots,p$는 각각 다음과 같이 모형 (2.1)에서 구한 θ의 최소제곱추정량인 $\tilde{\theta}$에서 계산된 $\tilde{v}_j$를 반응변수로 하고 z에 의해 비모수회귀(nonparametric regression)시켜 구한 적합값으로 추정한다. 여기서

$$\tilde{v} = v(\tilde{\theta}) = \{\tilde{v}_j\} = \{\partial f(x;\theta)/\partial\theta_j|_{\theta=\tilde{\theta}}\},$$
$$\hat{m}(z) = \{\hat{m}_j(z)\} = \{\hat{E}(\tilde{v}_j \mid z)\}, \quad j=1,2,\cdots,p \tag{3.6}$$

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

$$y_i = f(x_i;\phi) + a^T\hat{m}(z_i) + \xi_i, \quad i=1,2,\cdots,n \tag{3.7}$$

추가변수 z의 CERES그림은 모형 (3.7)에 적합시킨 후 얻어지는 잔차 e에 $\hat{a}^T\hat{m}(z)$를 더한 CERES를 세로축으로 하고 추가변수 z를 가로축으로 하는 $\{e_i+\hat{a}^T\hat{m}(z_i), z_i\}$의 산점도이다. 이 때 $\hat{a}^T$는 모형 (3.7)에서 구한 $\hat{m}(z)$의 회귀계수추정값벡터이다.

모형 (2.2)가 참모형이라면 CERES그림의 세로축을 이루게 되는 CERES는 다음과 같다.

$$e_i+\hat{a}^T\hat{m}(z_i) = f(x_i;\theta) - f(x_i;\hat{\phi}) + g(z_i) + \varepsilon_i \tag{3.8}$$

모형 (3.7)에 적합하여 얻은 $\hat{\phi}$이 참모형 (2.2)에서의 실제모수 θ에 근접하면 식 (3.8)에 의해 CERES그림은 $\{g(z_i)+\varepsilon_i, z_i\}$의 산점도가 되어 $g(z)$의 형태를 잘 묘사한다.

만약 $g(z)$가 $m_j(z), j=1,2,\cdots,p$의 선형함수라면 $g(z)$는 $a^T\hat{m}(z)$로 근사가 가능하며 이 때 모형 (3.7)은 참모형인 (2.2)의 근사모형이 되고 $\hat{\phi}$은 θ에 근접하게 된다. 따라서 CERES그림은 $g(z)$의 형태를 잘 나타내게 된다.

편잔차그림과 덧편잔차그림에서 $E(v \mid z)$가 각각 z의 선형 또는 이차함수일 때 모형 (3.5)의 $a^Tm(z)$를 z의 선형과 이차함수로 하는 모형 (3.1)과 (3.3)에서 β의 최소제곱추정량이 θ에 대한 일치추정량이 되고 이러한 경우 편의가 없어져서 편잔차그림과 덧편잔차그림이 $g(z)$의 형태를 잘 나타낸다고 하였다. 따라서 $g(z)$가 $m_j(z)$의 선형함수가 아닌 경우에도 모형 (3.7)에서 $m(z)$를 $E(v \mid z)$로 정의하고 $\hat{E}(\tilde{v} \mid z)$로 추정하였으므로 $E(v \mid z)$는 $m_j(z)$의 선형함수로 볼 수 있으며 이 모형에서 구한 $\hat{\phi}$은 θ의 일치추정량이라 할 수 있고 식 (3.8)에서 편의가 없어지므로 CERES그림이 $g(z)$의 형태를 잘 묘사하게 된다. 이는 다음을 통하여 확인할 수 있다.

$f(x;\theta)$를 $\hat{\phi}$ 근방의 θ에 대하여 Taylor급수전개에 의해 전개하고 2차이상 미분한 부분은

무시할 수 있다고 가정하면 $f(\boldsymbol{x};\boldsymbol{\theta})$는 다음과 같이 선형화 된다.

$$\begin{aligned} f(\boldsymbol{x};\boldsymbol{\theta}) &\cong f(\boldsymbol{x};\hat{\boldsymbol{\phi}})+(\boldsymbol{\theta}-\hat{\boldsymbol{\phi}})^{T}\cdot\left(\frac{\partial f(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\phi}}}\right) \\ &= f(\boldsymbol{x};\hat{\boldsymbol{\phi}})+(\boldsymbol{\theta}-\hat{\boldsymbol{\phi}})^{T}\cdot\hat{\boldsymbol{v}} \end{aligned} \tag{3.9}$$

여기서 $\hat{\boldsymbol{v}}=\boldsymbol{v}(\hat{\boldsymbol{\phi}})$이다. 물론 곡률(curvature)을 검토하여 선형근사의 정확성을 확인하는 과정이 필요하지만 여기서는 곡률이 크지 않고 선형근사가 적절하다고 가정한다.

식 (3.9)을 이용하면 식 (3.8)의 CERES는 다음과 같다.

$$e_i+\hat{\boldsymbol{a}}^{T}\hat{\boldsymbol{m}}(z_i)\cong(\boldsymbol{\theta}-\hat{\boldsymbol{\phi}})^{T}\cdot\hat{\boldsymbol{v}}_i+g(z_i)+\varepsilon_i \tag{3.10}$$

여기서 $\hat{\boldsymbol{v}}_i$는 i번째 관측값에서 계산된 $\hat{\boldsymbol{v}}$이다.

정리 3.1: 모형 (2.2)가 참모형이라면 모형 (3.7)에서 얻은 최소제곱추정량 $\hat{\boldsymbol{\phi}}$은 모형 (2.2)에서의 실제모수 $\boldsymbol{\theta}$의 일치추정량이 된다.

이 정리는 Cook(1993)의 Lemma 4.1에서 선형회귀모형의 설명변수벡터 $\boldsymbol{x}$를 $\hat{\boldsymbol{v}}$로 바꾸어 도출할 수 있고 Lemma 4.1의 증명과 같은 방법으로 정리 3.1을 증명할 수 있다. 따라서 식 (3.8)에서의 편의가 없어지므로 CERES그림은 추가하는 변수 z의 형태를 잘 묘사할 수 있다.

3.3 3차원 CERES 그림

추가하려는 설명변수가 두 개일 경우 설명변수 각각의 잔차산점도를 통하여 설명변수의 비선형성을 확인한 후 어떠한 형태로 모형에 포함되어야 할 것인지를 알아보기 위해 3차원 편잔차그림이나 3차원 CERES그림을 사용할 수 있다.

모형 (2.1)에 추가시키려는 설명변수 z_1과 z_2가 $g(z_1,z_2)$의 형태로 추가되면 추가변수의 효과를 나타내는 다음의 모형으로 생각할 수 있고 이 모형을 참모형이라고 하자.

$$y_i=f(\boldsymbol{x}_i;\boldsymbol{\theta})+g(z_{i1},z_{i2})+\varepsilon_i,\quad i=1,2,\cdots,n \tag{3.11}$$

여기서 $g(\cdot,\cdot)$는 선형 또는 형태가 알려지지 않은 비선형함수이다.

조건부기대값 $E(\boldsymbol{v}\mid z_1,z_2)$가 z_1과 z_2의 선형함수로 표현되지 않을 경우 3차원 편잔차그림은 $g(\cdot,\cdot)$에 대한 정보를 제대로 제공하지 못하게 된다. 이러한 문제점을 해결하기 위해서 보다 완화된 조건에서도 적절한 정보를 제공해 주는 3차원 CERES그림을 적용해 볼 수 있다.

설명변수 z_1과 z_2가 서로 연관성이 강한 경우 $g(z_1,z_2)$가 선형함수의 형태를 나타내지 않거나 $E(\boldsymbol{v}\mid z_1,z_2)$가 z_1과 z_2의 선형함수가 아니면 다음과 같은 모형을 생각해 볼 수 있다.

$$y_i=f(\boldsymbol{x}_i;\boldsymbol{\phi})+\boldsymbol{a}_1^{T}\boldsymbol{m}_1(z_{i1})+\boldsymbol{a}_2^{T}\boldsymbol{m}_2(z_{i2})+\xi_i,\quad i=1,2,\cdots,n \tag{3.12}$$

여기서 $\boldsymbol{m}_1(z_1)=\{m_{1j}(z_1)\}$과 $\boldsymbol{m}_2(z_2)=\{m_{2j}(z_2)\}$는 각각 $E(\boldsymbol{v}\mid z_1)$과 $E(\boldsymbol{v}\mid z_2)$로 정의한다. $m_{1j}(z_1)$, $j=1,2,\cdots,p$는 (2.1)에서 구한 $\tilde{\boldsymbol{\theta}}$에서 계산된 $\tilde{v}_j$를 반응변수로 하고 z_1에 대해 비모수회귀시켜 구한 적합값으로 추정한다. 또한 $m_{2j}(z_2)$도 같은 방법으로 추정한다. 즉,

$$\begin{aligned}\widehat{\boldsymbol{m}}_1(z_1)&=\{\widehat{m}_{1j}(z_1)\}=\{\widehat{E}(\tilde{v}_j\mid z_1)\},\quad j=1,2,\cdots,p\\ \widehat{\boldsymbol{m}}_2(z_2)&=\{\widehat{m}_{2j}(z_2)\}=\{\widehat{E}(\tilde{v}_j\mid z_2)\},\quad j=1,2,\cdots,p\end{aligned}\tag{3.13}$$

이다. 모형 (3.12)에서 $\boldsymbol{m}_1(z_{i1})$과 $\boldsymbol{m}(z_{i2})$를 식 (3.13)의 $\widehat{\boldsymbol{m}}_1(z_{i1})$과 $\widehat{\boldsymbol{m}}_2(z_{i2})$로 바꾸면 다음과 같은 모형이 된다.

$$y_i=f(\boldsymbol{x}_i;\boldsymbol{\phi})+\boldsymbol{\alpha}_1^T\widehat{\boldsymbol{m}}_1(z_{i1})+\boldsymbol{\alpha}_2^T\widehat{\boldsymbol{m}}_2(z_{i2})+\xi_i,\quad i=1,2,\cdots,n\tag{3.14}$$

CERES그림은 모형 (3.14)에 적합하여 얻어진 잔차 e에 $\widehat{\boldsymbol{\alpha}}_1^T\widehat{\boldsymbol{m}}_1(z_1)$과 $\widehat{\boldsymbol{\alpha}}_2^T\widehat{\boldsymbol{m}}_2(z_2)$를 더한 CERES를 수직축으로 하고 z_1과 z_2를 서로 직각을 이루는 두 개의 수평축으로 하는 $\{e_i+\widehat{\boldsymbol{\alpha}}_1^T\widehat{\boldsymbol{m}}_1(z_{i1})+\widehat{\boldsymbol{\alpha}}_2^T\widehat{\boldsymbol{m}}_2(z_{i2}),z_{i1},z_{i2}\}$의 3차원 산점도이다. 이 때 $\widehat{\boldsymbol{\alpha}}_1^T$, $\widehat{\boldsymbol{\alpha}}_2^T$는 모형 (3.14)에서 구한 $\widehat{\boldsymbol{m}}_1(z_{i1})$과 $\widehat{\boldsymbol{m}}_2(z_{i2})$에 대한 각각의 회귀계수추정값벡터이다.

모형 (3.11)이 참모형이라면 3차원 CERES그림의 수직축을 이루는 CERES는 다음과 같다.

$$e_i+\widehat{\boldsymbol{\alpha}}_1^T\widehat{\boldsymbol{m}}_1(z_{i1})+\widehat{\boldsymbol{\alpha}}_2^T\widehat{\boldsymbol{m}}_2(z_{i2})=f(\boldsymbol{x}_i;\boldsymbol{\theta})-f(\boldsymbol{x}_i;\widehat{\boldsymbol{\phi}})+g(z_{i1},z_{i2})+\varepsilon_i\tag{3.15}$$

3차원 CERES그림이 $g(z_1,z_2)$의 형태를 잘 나타내기 위하여 모형 (3.14)에 적합하여 얻은 $\widehat{\boldsymbol{\phi}}$은 모형 (3.11)에서의 실제모수 $\boldsymbol{\theta}$에 근접해야 한다. $g(z_1,z_2)$이 $m_{1j}(z_1)$, $m_{2j}(z_2)$, $j=1,2,\cdots,p$의 선형함수로 나타나는 경우 $g(z_1,z_2)$는 $\boldsymbol{\alpha}_1^T\widehat{\boldsymbol{m}}_1(z_1)+\boldsymbol{\alpha}_2^T\widehat{\boldsymbol{m}}_2(z_2)$으로 근사가 가능하며 이 때 모형 (3.14)는 모형 (3.11)의 근사모형이 되고 $\widehat{\boldsymbol{\phi}}$는 $\boldsymbol{\theta}$에 근접하게 된다. 따라서 3차원 CERES그림은 $g(z_1,z_2)$의 형태를 잘 묘사한다. $g(z_1,z_2)$이 $m_{1j}(z_1)$, $m_{2j}(z_2)$의 선형함수로 나타나는 않는 경우에도 모형 (3.14)에서 $\boldsymbol{m}_1(z_1)$과 $\boldsymbol{m}(z_2)$는 각각 $\widehat{E}(\tilde{\boldsymbol{v}}\mid z_1)$과 $\widehat{E}(\tilde{\boldsymbol{v}}\mid z_2)$로 추정하였으므로 $E(\boldsymbol{v}\mid z_1)$과 $E(\boldsymbol{v}\mid z_2)$는 $m_{1j}(z_1)$과 $m_{2j}(z_2)$의 선형함수로 볼 수 있다. 따라서 $\widehat{\boldsymbol{\phi}}$는 $\boldsymbol{\theta}$의 일치추정량이라 할 수 있고 3차원 CERES그림이 $g(z_1,z_2)$의 형태를 잘 나타내게 된다.

3차원 CERES그림을 수직축을 중심으로 회전할 때 보여지는 연속된 2차원의 산점도 중 비선형의 형태를 가장 잘 나타내는 산점도를 생각해 보자. 이 산점도의 가로축은 회전각도 ω의 함수인 $z(\omega)$라 표현할 수 있다. 즉 $z(\omega)$는 3차원 CERES그림을 수직축으로 회전할 때 나타나는 2차원 산점도 중 CERES를 세로축으로 하고 z_1을 가로축으로 하는 산점도를 시작으로 하여 ω만큼 회

전한 후에 나타나는 산점도의 가로축으로 다음과 같이 z_1과 z_2의 선형결합으로 표현할 수 있다.

$$z(\omega) = z_1\cos(\omega) + z_2\sin(\omega) \tag{3.16}$$

연관성이 강한 두 설명변수 z_1과 z_2의 비선형함수 $g(z_1, z_2)$를 두 변수의 선형결합에 대한 비선형함수 $g^*(z(\omega))$로 근사할 수 있다면 다음의 모형을 참모형 (3.11)의 근사모형이라 할 수 있다.

$$y_i = f(\boldsymbol{x}_i;\ \boldsymbol{\theta}) + g^*(z(\omega)) + \boldsymbol{\varepsilon}_i, \quad i = 1,2,\cdots,n \tag{3.17}$$

따라서 $g^*(\cdot)$의 형태는 $z(\omega)$의 2차원 CERES그림을 통해 얻을 수 있고 이는 3차원 CERES그림의 수직축에 대한 회전에서 비선형형태를 가장 잘 나타내는 산점도와 일치한다.

두 설명변수 z_1과 z_2간에 연관성이 약한 경우 모형 (3.11)에서 함수 $g(z_1, z_2)$를 z_1과 z_2 각각의 함수인 $g_1(z_1)$과 $g_2(z_2)$로 분리하여 포함하는 다음의 모형을 생각할 수 있다.

$$y_i = f(\boldsymbol{x}_i;\ \boldsymbol{\theta}) + g_1(z_{i1}) + g_2(z_{i2}) + \varepsilon_i, \quad i = 1,2,\cdots,n \tag{3.18}$$

따라서 3차원 CERES그림은 $g_1(z_1) + g_2(z_2)$의 형태를 잘 묘사하게 된다. 이런 경우에는 z_1, z_2 각각의 2차원 CERES그림이 $g_1(z_1)$과 $g_2(z_2)$의 형태를 잘 나타내게 된다.

4. 예 제

다음의 모형을 참모형이라 하자.

$$y = \theta_1\theta_2 x_1/(100 + \theta_1 x_1) + \theta_3 e^z/30 + \varepsilon \tag{4.1}$$

설명변수 x_1과 ε은 각각 $U(1,6)$와 $N(0, .01^2)$에서 50개의 값을 생성하고 회귀계수를 각각 $\theta_1 = 0.1$, $\theta_2 = 100$, $\theta_3 = 1$ 으로 하였다. 추가되는 설명변수 z는 $E(\boldsymbol{v} \mid z)$가 z의 선형이나 2차함수가 아닌 비선형관계가 되도록 하기 위하여 다음과 같이 생성하였다. 주어진 $\boldsymbol{\theta}$와 x_1의 값에서 계산된 50개의 $\boldsymbol{v}$벡터를 행으로 하는 행렬 $\boldsymbol{V}$의 첫 번째 열벡터인 $\mathbf{v}_1 = \partial f/\partial\theta_1$을 구하고 $z = \log \mathrm{v}_1 + \delta$의 관계를 갖도록 하였다. 이 때 δ는 $N(0, .01^2)$에서 생성하였고 x_1, z, $\boldsymbol{\theta}$를 이용하여 모형 (4.1)에 적용시켜 50개의 y값을 결정하였다.

우선 잔차산점도를 이용하여 추가하려는 설명변수 z의 비선형성을 확인한 결과 비선형성이 탐지되었다. 추가변수가 어떠한 형태로 변환되어 모형에 추가되어야 하는지 알아보기 위하여 편잔차그림을 적용해 볼 수 있다. 편잔차그림을 그리기 위해 다음의 모형을 생각해 보자.

$$y = \beta_1\beta_2 x_1/(100 + \beta_1 x_1) + \alpha z + \delta \tag{4.2}$$

모형 (4.2)에 적합하여 구한 최소제곱추정값을 이용하여 행렬 $\widehat{\boldsymbol{V}}$을 구한 후 첫 번째와 두 번째

열벡터를 $\hat{\mathbf{v}}_1$, $\hat{\mathbf{v}}_2$라 한다. 이 때 $E(\boldsymbol{v} \mid z)$와 z의 관계는 $\{\hat{v}_{i1}, z_i\}$와 $\{\hat{v}_{i2}, z_i\}$의 산점도로 확인할 수 있다. 산점도 $\{\hat{v}_{i1}, z_i\}$인 <그림 1>에서 뚜렷한 비선형곡선이 보이고 $\{\hat{v}_{i2}, z_i\}$의 산점도 역시 <그림 1>과 유사하므로 $E(\boldsymbol{v} \mid z)$은 z에 대해 선형이 아니며 z의 편잔차그림이 추가되는 z의 함수형태를 올바르게 나타낼 것으로 기대할 수 없다. 실제로 편잔차그림인 <그림 2>는 어떤 특정한 함수의 형태를 취하지 않고 실제의 함수형태인 지수함수를 나타내주지 못한다.

이제는 좀더 조건이 완화된 덧편잔차그림을 얻기 위해 다음의 모형을 생각해 보자.

$$y = \beta_1\beta_2 x_1/(100 + \beta_1 x_1) + \alpha_1 z + \alpha_2 z^2 + \delta \tag{4.3}$$

모형 (4.3)에 적합하여 얻은 최소제곱추정값들을 이용하여 행렬 $\breve{\boldsymbol{V}}$를 구한 후 첫 번째 열벡터와 두 번째 열벡터를 각각 $\breve{\mathbf{v}}_1$, $\breve{\mathbf{v}}_2$라 하자. 덧편잔차그림이 추가되는 설명변수 z의 함수의 형태를 잘 나타내기 위한 조건이 만족되는지 알아보기 위하여 $\{\breve{v}_{i1}, z_i\}$와 $\{\breve{v}_{i2}, z_i\}$의 산점도를 그려본 결과 이 산점도들도 <그림 1>과 유사한 형태를 나타낸다. 따라서 $E(\boldsymbol{v} \mid z)$이 z에 대해 선형 혹은 2차함수의 형태를 갖지 않음을 확인할 수 있고 이 자료는 덧편잔차그림의 조건을 만족시키지 못하므로 덧편잔차그림도 추가되는 설명변수 z의 변환의 형태에 대한 올바른 정보를 제공해 주지 못할 것으로 예상된다. 실제로 덧편잔차그림을 그려보면 <그림 2>와 거의 유사하고 어떤 특별한 형태를 나타내지 않으므로 실제의 함수인 지수함수의 형태를 나타내지 못한다.

CERES그림을 구하기 위해 다음의 모형을 생각해 보자.

$$y = \phi_1\phi_2 x_1/(100 + \phi_1 x_1) + \boldsymbol{\alpha}^T \boldsymbol{m}(z) + \xi \tag{4.4}$$

모형 (4.4)의 $\boldsymbol{m}(z)$는 비모수회귀적합에 의한 적합값 $\widehat{E}(\tilde{\boldsymbol{v}} \mid z)$을 이용하여 추정할 수 있다. 여기서 $\tilde{\boldsymbol{v}} = \boldsymbol{v}(\tilde{\boldsymbol{\theta}})$이고 이 때 추정된 $\hat{\boldsymbol{m}}(z)$을 대입하여 다음의 모형을 설정한다.

$$y = \phi_1\phi_2 x_1/(100 + \phi_1 x_1) + \boldsymbol{\alpha}^T \hat{\boldsymbol{m}}(z) + \xi \tag{4.5}$$

모형 (4.5)를 적합하여 구한 최소제곱추정값 $\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\boldsymbol{\alpha}}^T$를 이용하여 CERES그림을 그려보면 <그림 3>와 같고 <그림 2>의 편잔차그림이나 덧편잔차그림에서 나타나지 않았던 z의 실제함수인 지수함수형태를 나타냄을 알 수 있다. 조건부기대값 $E(\boldsymbol{v} \mid z)$이 z에 대해 선형이나 2차함수의 형태가 아닌 비선형함수이더라도 CERES그림은 z의 변환의 형태를 잘 나타내 주는 것을 확인할 수 있다. 따라서 CERES그림을 통해 추가되는 설명변수 z가 지수함수 형태로 변환되어 모형에 추가되어야 한다는 것을 알 수 있다.

이제 지수함수로 변환된 변수를 추가되는 변수로 생각하자. 이 때 변수의 변환형태가 적절했다면 변환된 변수에 대한 CERES그림이 선형의 형태를 나타내야 할 것이다. 설명변수 z를 지수함수 형태로 변환한 $z' = e^z$를 추가변수로 생각하고 z'에 대한 CERES그림을 그려보면 <그림 4>와 같이 직선의 형태를 나타내므로 지수함수를 취한 변환이 적절하였음을 알 수 있다.

비선형회귀모형에서 추가하려는 설명변수가 두 개일 경우에 설명변수간에 연관성이 강한 경우

와 약한 경우로 나누어 생각할 수 있다. 먼저 2개의 설명변수가 서로 연관성이 약한 경우는 앞에서 설명한 바와 같이 두 설명변수 각각에 대한 2차원 CERES그림을 통해 비선형성의 진단과 비선형함수의 형태를 알아볼 수 있으므로 여기서는 두 설명변수가 서로 연관성이 강한 경우에 대해서만 적용해 보고자 한다.

다음의 모형을 참모형이라 하자.

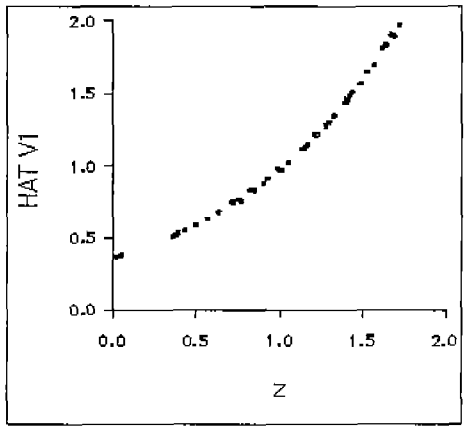
$$y = \theta_1\theta_2 x_1/(100+\theta_1 x_1) + \theta_3 e^{\theta_4(z_1-z_2)} + \varepsilon \tag{4.6}$$

설명변수 x_1과 ε은 각각 $U(1,6)$와 $N(0,1)$에서 100개의 값을 생성하고 $\theta_1=1$, $\theta_2=100$, $\theta_3=2$, $\theta_4=1$으로 하였다. 추가되는 설명변수 z_1과 z_2는 $E(\boldsymbol{v}\mid z_1)$가 z_1과 z_2에 대해 비선형관계가 되도록 하기 위하여 다음과 같이 생성하였다. 주어진 $\boldsymbol{\theta}$와 x_1의 값에서 계산된 100개의 $\boldsymbol{v}$벡터를 행으로 하는 행렬 $\boldsymbol{V}$의 첫 번째 열벡터인 $\mathbf{v}_1$을 구하고 $z_1 = 2\log \mathrm{v}_1 + \delta$와 $z_2 = \sqrt{z_1} + \tau$의 관계를 갖도록 하였다. 이 때 δ와 τ는 각각 $N(0,.3^2)$와 $N(0,.5^2)$에서 생성하였고 x_1, z_1, $\boldsymbol{\theta}$를 이용하여 y의 100개의 값을 결정하였다.

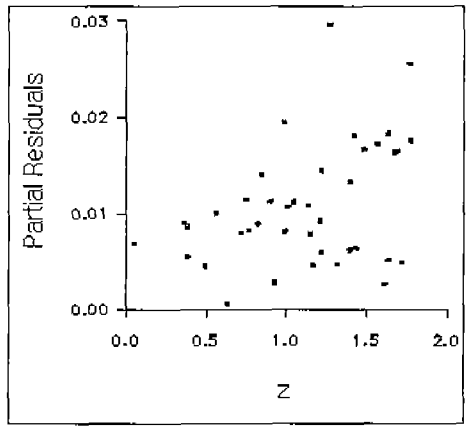
먼저 v_1을 수직축으로 하고 z_1과 z_2를 두 개의 서로 직각인 수평축으로 하는 3차원 산점도 $\{v_{i1}, z_{i1}, z_{i2}\}$에서 수직축을 중심으로 회전시키면 3차원 공간상에 흩어져 있는 점들이 비선형곡면의 형태를 나타내므로 v_1과 $(z_1,z_2)^T$간에는 비선형의 관계가 있음을 알 수 있다. 즉 $E(\boldsymbol{v}\mid z_1,z_2)$는 z_1과 z_2에 대해 비선형함수 관계가 있으므로 3차원 편잔차그림은 우리가 찾고자 하는 z_1과 z_2의 비선형함수를 올바르게 나타낼 것으로 기대할 수 없다. 따라서 이 자료의 경우 CERES그림이 적절한 방법이라 할 수 있다. z_1과 z_2의 3차원 CERES그림을 그려보면 <그림 5>과 같고 수직축을 중심으로 회전시키면 회전하는 각도에 따라 차이가 있지만 어떤 비선형 곡면의 형태를 나타내고 있는 것을 확인할 수 있고 이것이 비선형함수 $g(z_1,z_2)$의 형태이다.

<그림 6>은 3차원 CERES그림의 수직축을 중심으로 회전하여 나타나는 연속적인 2차원 산점도 중 가장 명확한 곡선이 보이는 산점도이다. 이 산점도는 CERES를 세로축으로 하고 z_1를 가로축으로 하는 산점도를 시작으로 하여 315°만큼 회전된 상태이다. 회전각도 315°를 이용하면 이 산점도의 가로축은 z_1-z_2가 됨을 확인할 수 있다. 즉 $z(315^\circ) = (z_1-z_2)/\sqrt{2}$이 된다. 이 산점도에서 나타나는 비선형형태는 z_1과 z_2의 선형결합인 $z_0 = z_1 - z_2$의 비선형함수를 나타내며 그림의 형태로 보아 지수함수에 근사하다는 것을 알 수 있다.

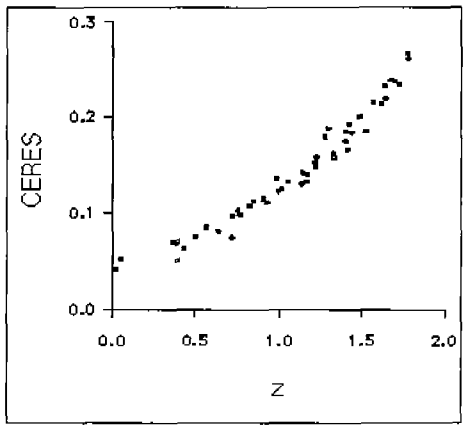
이제 지수함수로 변환된 변수를 추가되는 변수로 생각하자. 이 때 지수함수 형태의 변환이 적절했다면 변환된 추가변수 e^{z_0}의 2차원 CERES그림은 선형의 형태를 나타내야 할 것이다. 새로운 추가변수 $z' = e^{z_0} = e^{z_1-z_2}$의 CERES그림을 그려보면 <그림 4>와 유사하게 직선의 형태를 보이므로 지수함수의 형태로 변환하는 것이 적절하다고 할 수 있다.



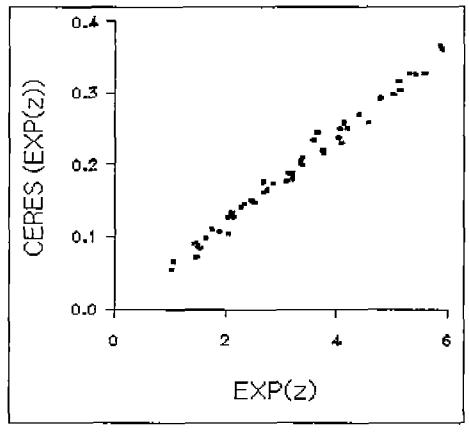
<그림 1> $\{\hat{v}_{t1}, z\}$의 산점도



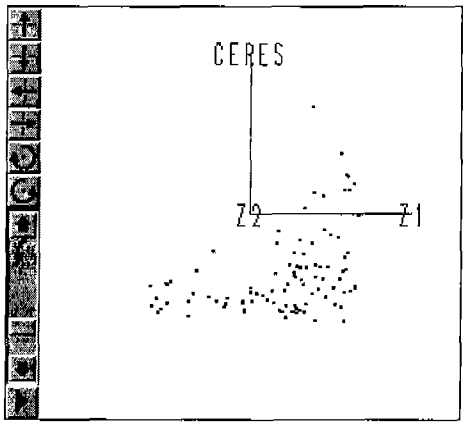
<그림 2> z에 대한 편잔차그림



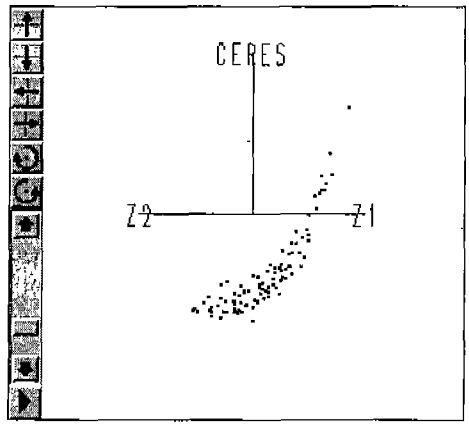
<그림 3> z에 대한 CERES그림



<그림 4> $z' = e^z$에 대한 CERES그림



<그림 5> 3차원 CERES그림



<그림 6> 3차원 CERES그림

5. 결 론

비선형회귀분석에서 새로운 설명변수를 모형에 추가하고자 할 경우 추가변수의 비선형성에 대한 진단은 선형회귀분석에 비해 더욱 중요한 문제라고 할 수 있다. 이러한 진단에는 설명변수의 비선형성 존재에 대한 탐색과 비선형 형태에 대한 제시가 포함된다.

모형에 포함되어 있던 변수들과 새로 추가되는 변수 사이에 연관성이 약한 경우 또는 연관성이 강하더라도 일정한 조건이 충족되면 편잔차그림과 덧편잔차그림이 추가되는 변수의 유의성과 변환의 필요성, 비선형 형태까지도 제시할 수 있다. 본 논문에서는 이러한 조건이 충족되지 않는 경우에도 CERES그림은 같은 기능을 수행할 수 있음을 확인하였다. 또한 추가되는 변수가 두 개인 경우 3차원의 CERES그림을 통하여 2차원의 그림에서는 얻을 수 없었던 비선형성의 탐색과 비선형함수의 제시가 가능함을 확인하였다.

비선형회귀모형에서 기존의 모형에 새로운 설명변수가 추가되는 경우 새로운 형태의 비선형함수의 모형이 필요할 수 있다. 그러나 본 논문에서는 추가되는 설명변수의 비선형함수가 가법적(additive)으로 모형에 추가되는 경우로 제한하여 생각해 보았다. 일반적인 경우는 좀 더 연구가 필요하다고 생각된다.

참 고 문 헌

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

[2] Berk, K. N. (1998), Regression diagnostic plots in 3-D, *Technometrics,* Vol. 40, 39-47.

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

[4] Chatterjee, S. and Hadi, A. S. (1988), *Sensitivity Analysis in Linear Regression,* John Wiley & Sons: New York.

[5] Cook, R. D. (1993), Exploring partial residual plots, *Technometrics,* Vol. 35, 351-362.

[6] Cook, R. D. and Weisberg, S. (1982), *Residuals and influence in regression,* Chapman & Hall: New York.

[7] Cook, R. D. and Weisberg, S. (1989), Regression diagnostics with dynamic graphics, *Technometrics,* Vol. 31, 277-311.

[8] Ezekiel, M. (1924), A method for handling curvilinear correlation for any number of variables, *Journal of the American Statistical Association,* Vol. 19, 431-453.

[9] ahng, M. and Kim, J. (1998), Partial residual plot in nonlinear regression, *The Korean Communications in Statistics,* Vol. 3, 571-580.

[10] allows, C. L. (1986), Augmented partial residuals, *Technometrics,* Vol.28, 313-319.