응용통계연구 제 **13**권 1호, *2000*년, *pp. 133-143*

회귀분석에서의 3차원 편잔차그림 *

강명욱[1] 이정아[2]

요 약

비선형성이 존재하는 두 개의 설명변수가 모형에 선형으로 포함되는 경우 두 설명변수가 연관성이 약하면 각각의 변수에 대한 2차원 편잔차그림이 비선형성의 존재와 형태를 잘 나타낸다. 그러나 두 변수가 연관성이 강하면 3차원 편잔차그림이 필요하며 2차원 편잔차그림으로는 알아낼 수 없는 비선형성에 대한 탐지가 가능하다.

주요용어: 편잔차그림, 3차원그림, 회전그림.

1. 서론

회귀분석에서는 관심의 대상이 되는 변수들간의 관계를 잘 설명해 주는 모형을 찾아내는 과정인 모형개발과 더불어 사용되는 모형과 가정의 타당성을 진단하여야 한다. 이러한 진단에서 쉬우면서도 유용하게 쓸 수 있는 방법은 그림을 이용한 방법이다. 그림을 이용한 방법은 Cook과 Weisberg(1982, 1994), Atkinson(1985), Cook(1998) 등에 의해 꾸준한 연구가 진행되면서 통계학의 한 분야로 자리잡고 있다. 또한 컴퓨터 환경의 급속한 발전으로 동적인 그림(dynamic graphics)을 이용한 방법이 Huber(1987)에 의해 처음 시도되었다.

회귀모형의 기본 가정 중 함수의 선형성 가정의 진단을 위해 사용되어지는 가장 일반적인 방법은 2차원 산점도를 이용하는 것이다. 회귀분석에서 그림을 이용한 진단방법 중 가장 활발히 연구되어온 분야로 잔차산점도(residual plot), 추가변수그림(added variable plot), 편잔차그림(partial residual plot), 덧편잔차그림(augmented partial residual plot), CERES그림(combining conditional expectation and residual plot) 등이 있다.

잔차산점도는 함수의 선형성 가정의 검토뿐만 아니라 오차항의 가정들에 대한 검토에도 널리 사용되어지는 그림이다. Cox(1958)에 의해 처음 소개된 추가변수그림은 Belsley, Kuh와 Welsch(1980), Cook과 Weisberg(1982) 등에 의해 회귀진단에 사용되어 왔다. Cook과 Weisberg(1989)는 3차원의 추가변수그림을 제시하였다. 편잔차그림은 Ezekiel(1924)에 의해 소개되었고 Larsen과 McCleary(1972)에 의해 편잔차그림이라 불리게 되었고 Cook과 Weisberg(1982), Atkinson(1985), Chatterjee와 Hadi(1988) 등에 의해 회귀진단의 도구로 사용되면서 그 효용성에 대한 많은 연구가 있어 왔다. 최근에는 Berk와 Booth(1995), Berk(1998) 등에 의해 3차원 편잔차그림에 대한 연구들이 진행 중이다.

* 본 연구는 숙명여자대학교 1998년도 교비연구비 지원에 의해 수행되었음.

1) (140-742) 서울시 용산구 청파동 숙명여자대학교 통계학과 부교수
E-mail: mwkahng@sookmyung.ac.kr
2) (140-742) 서울시 용산구 청파동 숙명여자대학교 통계학과

편잔차그림을 개선한 덧편잔차그림은 Mallows(1986)에 의해 소개되었고 이를 이용하여 Cook(1993)은 선형회귀모형에서 변수의 비선형성을 파악하는 방법을 제시하고 덧편잔차그림을 일반화한 CERES그림이라는 새로운 진단방법도 소개하였다. 동적인 그림을 이용한 방법으로는 회전그림(rotating plot)과 같은 3차원 그림을 회귀진단에 이용하는 방법이 Cook과 Weisberg(1989), Berk와 Booth(1995), Berk(1998) 등에 의해 활발히 연구되어 왔다.

본 논문에서는 비선형성의 검토에 유용하다고 알려져 있는 2차원 편잔차그림의 이론적인 근거를 알아보고 3차원 편잔차그림을 이용함으로써 2차원 편잔차그림보다 완화된 제약조건 하에서 비선형성으로 인한 모형구성의 오류를 찾아낼 수 있는지를 알아보고자 한다.

2. 2차원 편잔차그림

반응변수 y와 $p-1$개의 설명변수 $\boldsymbol{x}^T=(1,x_1,x_2,...,x_{p-1})$의 n개 관찰값에 대해 다음과 같은 선형회귀모형을 생각하자.

$$y_i=\boldsymbol{x}_i^T\boldsymbol{\beta}+\epsilon_i,\quad i=1,\cdots,n, \tag{2.1}$$

여기서 $\boldsymbol{x}_i^T$는 1과 각 설명변수의 i번째 관찰값들로 이루어진 차수 p인 설명변수벡터이고 $\boldsymbol{\beta}$는 차수가 p인 회귀계수벡터이다. 오차항 ϵ_i는 서로 독립적이고 평균 $E(\epsilon_i)=0$, 분산 $Var(\epsilon_i)=\sigma^2$인 정규분포를 따른다고 가정한다. 또한, 설명변수와 오차항간에는 상관관계가 없다고 가정한다.

회귀분석에서는 일반적으로 함수의 선형성과 오차항의 등분산성, 정규성, 독립성을 가정하며 이러한 가정들의 타당성을 알아보기 위해 사용되어지는 가장 일반적인 방법은 모형 (2.1)에서 얻어지는 잔차와 y의 추정값을 각각 세로축과 가로축으로 하는 $\{e_i,\ \hat{y}_i\}$의 산점도를 그려보는 것이다. 이같은 잔차산점도에서 비선형의 형태가 보인다면 선형회귀모형 (2.1)의 일부 혹은 전체에 대한 비선형성의 검토가 필요하다.

Mansfield와 Conerly(1987)에 의하면, 이러한 비선형 형태는 반응변수 y와 선형의 관계를 갖지 않는 어떠한 설명변수가 모형 (2.1)에서와 같이 선형으로 모형에 포함되었음을 의미한다. 따라서 잔차와 각 설명변수 x_j를 각각 세로축과 가로축으로 하는 잔차-설명변수 산점도를 추가적으로 그려보아 비선형의 형태를 보이는 변수를 적절히 변환하여 선형으로 모형에 포함시키는 것을 고려해 보아야 한다. 모형 (2.1)에서 얻어지는 잔차산점도에서 비선형의 형태가 나타나고 $\{e_i,\ x_{ij}\}$의 잔차-설명변수 산점도 중 어느 한 변수에 대해서 비선형의 형태가 보인다고 하자.

이러한 비선형성을 포함하는 모형을 설정하기 위하여 설명변수벡터 $\boldsymbol{x}^T$를 $\boldsymbol{x}_1^T$와 비선형성이 의심되는 설명변수 x_2로 분할하면 모형 (2.1)은 아래와 같이 쓸 수 있다.

$$y_i=\boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1+\beta_2x_{i2}+\epsilon_i,\quad i=1,\cdots,n, \tag{2.2}$$

모형 (2.2)에서 얻어진 잔차를 이용한 잔차산점도와 x_2의 잔차-설명변수 산점도 $\{e_i,\ x_{i2}\}$에서 비선형의 형태가 보인다면 설명변수 x_2의 비선형성을 포함하는 아래와 같은 모형을 생각할 수 있다.

$$y_i=\boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1+g(x_{i2})+\epsilon_i,\quad i=1,\cdots,n, \tag{2.3}$$

여기서 g는 형태가 알려지지 않은 비선형함수이고 모형 (2.3)을 $p-1$개의 설명변수와 반응변수의 관계를 나타내는 참모형(true model)이라고 할 수 있다.

위와 같이 잔차를 이용한 산점도를 통해 비선형성의 존재를 확인할 수는 있지만 산점도상에 나타나는 비선형 형태가 비선형함수 g의 형태를 보여주는 것은 아니다. Mansfield와 Conerly(1987)는 이러한 산점도가 변환의 필요성에 대한 판단의 기준이 되기는 하지만 어떠한 형태로 변환해야 하는지에 대한 추가적인 정보는 줄 수 없다고 하였다. 이러한 문제를 해결하기 위해 편잔차그림이 사용된다.

선형회귀모형에서 비선형성의 탐색에 유용하다고 알려져 있는 편잔차그림은 설명변수들 사이에 연관성이 약할 경우 변환의 필요성과 적절한 변환의 형태에 대한 정보를 제공할 수 있다. 여기서 연관성이라 함은 상관분석에서 상관계수를 통하여 두 변수사이의 직선관계를 알아보는 상관성 문제와 대비하여 두 변수의 선형 및 비선형 모두를 포함하는 관계를 의미한다. 비선형성이 의심되는 설명변수 x_2의 편잔차그림은 모형 (2.2)에 적합시킨 후 얻어지는 잔차 e에 $\hat{\beta}_2 x_2$를 더한 편잔차 $e+\hat{\beta}_2 x_2$를 세로축으로 하고 관심의 대상이 되는 변수 x_2를 가로축으로 하는 $\{e_i+\hat{\beta}_2 x_{i2},\ x_{i2}\}$의 산점도이다.

모형 (2.3)이 참모형이라면 x_2에 대한 편잔차그림의 세로축을 이루게 되는 편잔차는 다음과 같다.

$$e_i+\hat{\beta}_2 x_{i2} = \boldsymbol{x}_{i1}^T(\boldsymbol{\beta}_1-\hat{\boldsymbol{\beta}}_1)+g(x_{i2})+\epsilon_i \tag{2.4}$$

x_2의 편잔차그림이 g의 형태를 잘 나타내기 위해서는 모형 (2.2)에 적합하여 얻은 $\hat{\boldsymbol{\beta}}_1$이 모형 (2.3)에서의 실제 모수 $\boldsymbol{\beta}_1$에 근접해야 한다. $\boldsymbol{x}_1$과 x_2가 연관성이 약할 경우 $\hat{\boldsymbol{\beta}}_1$의 값은 실제 모수 $\boldsymbol{\beta}_1$에 근접하게되고 식 (2.4)의 우변은 $g(x_{i2})+\epsilon_i$가 되어 x_2의 편잔차그림은 g의 형태를 잘 나타내게 된다. 반면에 $\boldsymbol{x}_1$과 x_2가 연관성이 강한 경우는 편잔차그림에 대한 조심스런 접근이 요구된다. Cook(1993)에 의하면, $\boldsymbol{x}_1$과 x_2가 연관성이 강하더라도 조건부 기대값 $E(\boldsymbol{x}_1 \mid x_2)$이 x_2의 선형함수이고 모형 (2.3)가 사실이면 모형 (2.2)에서 얻은 최소제곱추정량 $\hat{\boldsymbol{\beta}}_1$은 $\boldsymbol{\beta}_1$의 일치추정량이 된다. 따라서 $\boldsymbol{x}_1$과 x_2가 연관성이 강하더라도 위의 조건이 만족된다면 편잔차그림은 g의 형태를 잘 나타낸다고 할 수 있다.

그러나 $\boldsymbol{x}_1$과 x_2가 강한 연관성이 있고 조건부 기대값 $E(\boldsymbol{x}_1 \mid x_2)$이 x_2의 선형함수로 표현되지 않는다면 x_2의 편잔차그림은 g의 비선형 형태에 대한 잘못된 정보를 주거나 비선형성의 존재조차 알아내지 못할 수도 있다. 또한 두 개 이상의 설명변수가 비선형의 형태를 보이는 경우에는 2차원 편잔차그림의 효용성이 떨어질 수도 있다. 특히 비선형의 형태를 보이는 변수들간에 강한 연관성이 있는 경우에는 2차원 편잔차그림을 신뢰할 수 없게 된다. 이와 같은 문제점을 해결하기 위해서 3차원 편잔차그림이 필요하다.

3. 3차원 편잔차그림

모형 (2.1)에서 얻어지는 잔차산점도에서 비선형의 형태가 보이고 두 개의 설명변수에 대한 잔차-설명변수 산점도에서 비선형의 형태가 나타나면 3차원 편잔차그림을 사용해야 한다. 앞의 2절과 마찬가지로 $\boldsymbol{x}^T$는 비선형성이 의심되는 부분과 그렇지 않은 나머지 부분으로 분할할 수 있다. 비선형성이 의심되는 설명변수를 x_2와 x_3라고 하면 설명변수벡터 $\boldsymbol{x}^T$

는 1과 $p-3$개의 설명변수로 이루어진 벡터 $\boldsymbol{x}_1^T$와 비선형성이 의심되는 두 개의 설명변수 (x_2, x_3)로 분할할 수 있고 참모형은 다음과 같이 표현할 수 있다.

$$y_i = \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1 + g(x_{i2}, x_{i3}) + \epsilon_i, \quad i = 1, \cdots, n, \tag{3.1}$$

여기서 $g(\cdot,\cdot)$는 형태가 알려지지 않은 비선형 이변량함수(function of two variables)이다. 비선형성이 의심되는 변수 $(x_2, x_3)^T$와 $\boldsymbol{x}_1$간에는 연관성이 약하거나 연관성이 강하더라도 $(x_2, x_3)^T$에 대한 $\boldsymbol{x}_1$의 조건부 기대값이 x_2와 x_3의 선형함수로 표현된다고 가정한다. 모형 (3.1)은 x_2와 x_3간에 연관성이 약한 경우와 강한 경우로 나누어 생각해 볼 수 있다.

3.1. 연관성이 약한 두 변수

비선형 형태가 나타나는 두 변수 x_2와 x_3간에 연관성이 약하다면 x_2, x_3 각각에 대한 비선형 효과를 나타내는 별개의 비선형함수를 포함하는 아래의 모형이 적용될 수 있다.

$$y_i = \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1 + g(x_{i2}) + g(x_{i3}) + \epsilon_i, \quad i = 1, \cdots, n, \tag{3.2}$$

여기서 g_1과 g_2는 형태가 알려지지 않은 비선형함수이다. 만약 비선형 효과를 무시할 수 있다면 g_1과 g_2가 선형함수가 되고 모형 (3.2)는 아래와 같이 된다.

$$y_i = \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad i = 1, \cdots, n, \tag{3.3}$$

3차원 편잔차그림은 모형 (3.3)에 적합시킨 후 얻어지는 잔차에 $\hat{\beta}_2 x_2$와 $\hat{\beta}_3 x_3$를 더한 편잔차 $e + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$를 수직축으로 하고 관심의 대상이 되는 두 설명변수 x_2, x_3를 두 개의 서로 직각인 수평축으로 하는 $\{e_i + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3},\ x_{i2},\ x_{i3}\}$의 산점도이다. 모형 (3.2)가 참모형이라면 3차원 편잔차그림에서 수직축을 이루게 되는 편잔차는 다음과 같다.

$$e_i + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} = \boldsymbol{x}_{i1}^T(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1) + g(x_{i2}) + g(x_{i3}) + \epsilon_i, \tag{3.4}$$

2절과 마찬가지로 3차원 편잔차그림이 x_2와 x_3의 비선형 형태를 잘 나타내기 위해서는 모형 (3.3)에 적합하여 얻은 $\hat{\boldsymbol{\beta}}_1$이 모형 (3.2)에서의 실제 모수 $\boldsymbol{\beta}_1$에 근접해야 한다.

$\boldsymbol{x}_1$과 $(x_2, x_3)^T$사이에 연관성이 약하거나 $(x_2, x_3)^T$에 대한 $\boldsymbol{x}_1$의 조건부 기대값이 x_2와 x_3의 선형함수로 표현되는 경우 Cook(1993)이 제시한 결과를 적용하면 $\hat{\boldsymbol{\beta}}_1$은 $\boldsymbol{\beta}_1$의 일치추정량이 된다. 따라서 식 (3.4)의 우변은 $g(x_{i2}) + g(x_{i3}) + \epsilon_i$가 되어 3차원 편잔차그림은 비선형성이 의심되는 두 설명변수 x_2와 x_3의 비선형 형태를 잘 나타내게 된다.

비선형 형태를 보이는 두 변수간에 연관성이 약하다면 3차원 편잔차그림뿐만이 아니라 x_2, x_3 각각에 대한 2차원 편잔차그림만으로도 g_1과 g_2에 관한 충분한 정보를 얻을 수 있다. 먼저 x_2의 편잔차그림은 모형 (3.3)에 적합시킨 후 얻어지는 잔차에 $\hat{\beta}_2 x_2$를 더한 편잔차를 세로축으로 하고 x_2를 가로축으로 하는 $\{e_i + \hat{\beta}_2 x_{i2},\ x_{i2}\}$의 산점도이며 x_3에 대한 편잔차그림은 $\{e_i + \hat{\beta}_3 x_{i3},\ x_{i3}\}$이다.

모형 (3.2)가 참모형이라면 x_2의 2차원 편잔차그림에서 세로축을 이루는 편잔차는 다음과 같다.

$$e_i + \hat{\beta}_2 x_{i2} = \boldsymbol{x}_{i1}^T(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1) + g_1(x_{i2}) + g_2^*(x_{i3}) + \epsilon_i, \tag{3.5}$$

여기에서 $g_2^*(x_{i3}) = g_2(x_{i3}) - \hat{\beta}_3 x_{i3}$이다.

만약 g_2가 선형이면 식 (3.5)의 우변 세번째 항은 $g_2^*(x_{i3}) = (\beta_3 - \hat{\beta}_3)x_{i3}$이고 이러한 상황은 2절에서와 같은 경우로 $(\boldsymbol{x}_1, x_3)^T$와 x_2 사이에 연관성이 약하거나 $E(\boldsymbol{x}_1 \mid x_2)$와 $E(x_3 \mid x_2)$가 모두 x_2의 선형함수이면 x_2의 2차원 편잔차그림은 g_1의 형태를 잘 나타낸다.

반면에 g_2가 비선형함수일 경우에는 관심의 대상이 되는 두 변수 x_3와 x_2사이에 연관성이 약하고, $\boldsymbol{x}_1$과 x_2사이에 연관성이 약하거나 연관성이 강하더라도 $E(\boldsymbol{x}_1 \mid x_2)$가 x_2의 선형함수이면 $\hat{\boldsymbol{\beta}}_1$은 $\boldsymbol{\beta}_1$의 일치추정량이 된다. 또한 x_3의 비선형함수인 $g_2(x_3)$와 x_2는 특별한 연관성이 없을 것이다. 따라서 x_2의 2차원 편잔차그림은 g_1을 잘 나타낼 수 있다. 물론 g_2의 비선형성으로 인하여 편잔차그림에서 나타나는 비선형의 형태가 g_2가 선형인 경우보다 좀 더 퍼져 보이는 두꺼운 모양으로 나타날 수는 있다.

지금 우리는 $\boldsymbol{x}_1$과 $(x_2, x_3)^T$사이에 연관성이 약하고 관심의 대상이 되는 두 변수 x_2와 x_3사이에도 연관성이 약한 경우를 고려하므로 2차원 편잔차그림으로 x_2의 비선형함수 형태를 찾을 수 있고 x_3에 대해서도 동일한 결과를 얻을 수 있다.

Berk와 Booth(1995), Berk(1998)에 의하면 관심의 대상이 되는 변수 x_2와 x_3 각각에 대한 2차원 편잔차그림은 3차원 편잔차그림을 이용해서도 얻을 수 있다. 3차원 편잔차그림에서 편잔차축을 중심으로 회전하면 연속된 2차원 그림이 나타난다. 3차원 편잔차그림에서 수직축인 편잔차와 수평축 중이 하나인 x_2를 두 축으로 하는 2차원 산점도 자체가 x_2의 편잔차그림이라고 할 수는 없다. 3차원 편잔차그림에서의 편잔차 $e + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$는 x_2의 2차원 편잔차그림에서 세로축을 이루는 편잔차 $e + \hat{\beta}_2 x_2$와 $\hat{\beta}_3 x_3$만큼 차이가 나므로 3차원 편잔차그림을 이용하여 x_2의 편잔차그림을 얻기 위해서는 각 점에서 $\hat{\beta}_3 x_3$를 빼주는 수정이 요구된다. 이러한 수정은 3차원 산점도에 나타난 모든 점을 3차원 편잔차그림에서의 편차를 x_2와 x_3에 회귀시켜 나타나는 회귀평면에 평행인 평면상에서 x_2축에 수직인 방향으로 편잔차-x_2 평면에 이동시키면 이 평면에 나타나는 점들의 2차원 산점도가 x_2의 2차원 편잔차그림이 된다. 이는 3차원 편잔차그림에서 x_2축을 중심으로 하는 회전에서 나타나는 연속되는 2차원 산점도 중 퍼짐의 정도가 가장 좁은 모양의 산점도와 동일하다.

이상에서 알아본 바와 같이 관심의 대상이 되는 두 변수간에 연관성이 약할 경우에는 2차원 편잔차그림을 그려봄으로써 변환을 위한 올바른 정보를 얻을 수 있고 이러한 2차원 편잔차그림은 3차원 편잔차그림을 이용하여 얻을 수도 있다.

3.2. 연관성이 강한 두 변수

비선형성이 의심되는 두 변수 x_2와 x_3가 강한 연관성이 있다면 이들이 서로 영향을 주기 때문에 모형 (3.2)와 같이 두 변수가 각각 별개의 비선형함수 형태로 모형에 포함된다고 볼 수 없다. 이러한 경우에는 두 변수의 비선형함수를 포함하는 모형 (3.1)이 적절하며 각각의 설명변수에 대한 2차원 편잔차그림만으로는 비선형 형태에 대한 충분한 정보를 얻기 어려울 뿐만 아니라 잘못된 정보를 줄 수도 있다. 따라서 두 변수를 함께 고려하는 3차원 편잔차그림을 이용함으로써 2차원 편잔차그림만으로는 알아낼 수 없었던 비선형 형태에 대한 정보를 얻을 수 있다.

모형 (3.1)이 설명변수와 반응변수의 관계를 나타내는 참모형이라면 3차원 편잔차그림

$\{e_i + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3},\ x_{i2},\ x_{i3}\}$에서 수직축을 이루게 되는 편잔차는 다음과 같다.

$$e_i + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} = \boldsymbol{x}_{i1}^T(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1) + g(x_{i2}, x_{i3}) + \epsilon_i, \tag{3.6}$$

$\boldsymbol{x}_1$과 $(x_2,\ x_3)^T$간에 연관성 약하거나 $\boldsymbol{x}_1$의 조건부 기대값 $E(\boldsymbol{x}_1 \mid x_2,\ x_3)$이 x_2와 x_3의 선형함수로 표현된다면 모형 (3.3)에 적합하여 얻은 $\hat{\boldsymbol{\beta}}_1$은 모형 (3.1)에서의 실제 모수 $\boldsymbol{\beta}_1$의 일치추정량이 된다. 따라서 식 (3.6)의 우변은 $g(x_{i2}, x_{i3}) + \epsilon_i$가 되고 3차원 편잔차그림은 g의 형태를 잘 나타냄을 알 수 있다.

편잔차축을 중심으로 3차원 편잔차그림을 회전할 때 보여지는 연속된 2차원의 산점도 중 비선형의 형태가 가장 명확하게 드러나는 산점도를 생각해 보자. 이 산점도의 가로축은 회전각도 θ의 함수인 $z(\theta)$라 표현할 수 있다. 즉, $z(\theta)$란 3차원 편잔차그림을 편잔차축을 중심으로 회전할 때 나타나는 2차원 산점도 중 편잔차를 세로축으로 하고 x_2를 가로축으로 하는 산점도를 시작으로 하여 각도 θ만큼 회전한 후에 나타나는 산점도에서의 가로축으로 x_2와 x_3의 선형결합 $z(\theta) = cos(\theta)\, x_2 + sin(\theta)\, x_3$으로 표현할 수 있다.

만약 연관성이 강한 두 설명변수 x_2와 x_3의 어떠한 비선형함수 $g(x_2,\ x_3)$를 두 변수의 선형결합에 대한 비선형함수 $g^*(z(\theta))$로 근사할 수 있다면 모형 (3.1)의 근사모형으로 다음의 모형을 생각할 수 있다.

$$y_i = \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1 + g^*(z_i(\theta)) + \epsilon_i, \quad i = 1, \cdots, n, \tag{3.7}$$

만약 비선형함수 g^*를 선형함수라 가정하면 모형 (3.7)은 다음과 같이 표현될 수 있다.

$$y_i = \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1 + \beta^* z_i(\theta) + \epsilon_i, \quad i = 1, \cdots, n, \tag{3.8}$$

따라서 우리가 알고자하는 비선형 형태 g^*는 $z(\theta)$에 대한 2차원 편잔차그림을 이용해 얻을 수 있다. $\boldsymbol{x}_1$과 $(x_2,\ x_3)^T$간에 연관성이 약하거나, 강한 연관성이 있더라도 $\boldsymbol{x}_1$의 조건부 기대값 $E(\boldsymbol{x}_1 \mid x_2,\ x_3)$이 x_2와 x_3의 선형함수로 표현된다면 모형 (3.8)에 적합하여 얻은 $\hat{\boldsymbol{\beta}}_1$은 모형 (3.7)에서의 실제 모수 $\boldsymbol{\beta}_1$의 일치추정량이 된다. 따라서 $z(\theta)$의 편잔차그림은 g^*의 형태를 잘 나타내게 되고 이는 앞에서 편잔차축을 중심으로한 3차원 편잔차그림의 회전에서 비선형의 형태가 가장 명확하게 드러나는 2차원 산점도와 동일한 그림이 된다.

4. 예제

두 설명변수가 비선형으로 반응변수에 영향을 주는 경우 3차원 편잔차그림이 변환을 위한 비선형함수 형태를 제시할 수 있는지를 알아본다.

다음과 같은 함수관계를 갖는 모형을 실제 모형이라 하자.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 e^{(x_3 - x_2)} + \epsilon \tag{4.1}$$

설명변수 x_2는 $-.99$부터 $.99$까지 $.01$의 간격으로 얻어진 199개의 값을 관찰값으로 한다. 설명변수 x_1과 오차항 ϵ은 각각 $N(-3,\ 5^2)$와 $N(0,\ 3^2)$에서 생성한다. 설명변수 x_3는 $x_3 = 0.7x_2 + \delta$를 이용하여 x_2와 선형의 관계가 되도록 하고 δ는 $U(0,\ 1)$에서 생성한다. 따라서

x_2와 x_3는 선형의 관계가 있고 x_1과 $(x_2, x_3)^T$는 연관성이 없을 것으로 기대된다. 실제 모형 (4.1)의 회귀계수를 각각 $\beta_0 = 3$, $\beta_1 = 5$, $\beta_2 = 50$으로 하고 주어진 회귀함수를 이용하여 반응변수 y의 199개 값을 결정한다.

잔차산점도에서 비선형의 형태가 보이고 각 설명변수 x_1, x_2, x_3에 대한 잔차-설명변수 산점도 중 x_2와 x_3의 잔차-설명변수 산점도에서 비선형의 형태가 보이므로 x_2와 x_3가 비선형의 형태로 반응변수 y에 영향을 줄 것으로 예상되었다. 또한 비선형의 형태를 보이는 두 변수 x_2와 x_3간의 상관계수가 0.8335로 강한 상관관계를 보이므로 연관성이 있는 두 변수가 함께 어떠한 비선형함수의 형태를 보일 것으로 기대된다.

먼저 2차원 편잔차그림을 보면 비선형성이 의심되는 두 설명변수 x_2와 x_3 각각의 2차원 편잔차그림은 뚜렷한 직선의 형태를 보인다. 따라서 2차원 편잔차그림은 변환을 위한 어떠한 정보도 제공할 수 없음을 알 수 있다.

x_1을 수직축으로 하고 x_2와 x_3를 두 개의 서로 직각인 수평축으로 하는 3차원 산점도를 그려보면 3차원 공간 상에 흩어져 있는 점들이 x_1-축을 중심으로 하여 회전하는 동안 평면의 형태를 유지함이 확인되어 x_1의 조건부 기대값이 x_2와 x_3의 선형함수로 표현 가능하다고 할 수 있다. 그러므로 $(x_2, x_3)^T$의 3차원 편잔차그림은 두 변수의 비선형함수를 제시할 수 있을 것으로 기대된다.

그림 4.1은 3차원 편잔차그림으로 편잔차축을 중심으로 회전시키면 회전한 각도에 따라 강하고 약한 정도의 차이가 있기는 하지만 어떠한 비선형의 함수 형태가 유지되고 있음을 확인할 수 있고 이것이 3차원 편잔차그림으로 알아낼 수 있는 x_2와 x_3의 비선형함수라고 할 수 있다.

그림 4.2는 3차원 편잔차그림의 편잔차축을 중심으로한 회전에서 보여지는 연속된 2차원 산점도 중 가장 명확한 곡선이 보이는 산점도이다. 이것은 그림 4.1에서 편잔차를 세로축으로 하고 x_2를 가로축으로 하는 산점도를 시작으로 하여 132° 회전된 상태이다. 이 회전각도를 이용하면 이 산점도의 가로축 $z(132°) \cong (x_3 - x_2)/\sqrt{2}$는 실제 모형에서 나타난 x_2와 x_3의 관계인 $x_3 - x_2$에 비례함을 알 수 있다. 이 그림에서 나타나는 비선형 곡선은 x_2와 x_3의 선형결합인 $z_0 = x_3 - x_2$의 비선형함수를 나타내며 이것은 지수함수로 볼 수 있다.

만약 3차원 편잔차그림에서 제시한 지수함수가 적절하다면 z_0를 지수변환한 후 얻어지는 e^{z_0}를 추가되는 변수로 하는 모형에서의 2차원 편잔차그림은 직선의 형태를 보일 것이라고 짐작할 수 있다. 실제로 편잔차그림을 그려본 결과 뚜렷한 직선의 형태를 볼 수 있었으며 따라서 더 이상의 변환이 필요하지 않음을 알 수 있고 3차원 편잔차그림이 제시한 비선형함수가 두 변수 x_2와 x_3의 비선형 문제를 해결해 준다고 할 수 있다. 또한 z_0를 e^{z_0}로 지수변환한 모형에 적합하여 얻어진 잔차산점도는 변환 전에 보였던 비선형의 모습이 사라지고 모든 점들이 아무런 유형없이 흩어져 있고 z_0의 잔차-설명변수 산점도에서도 모든 점들이 2차원 평면상에 고루 퍼져 있음을 확인할 수 있다.

다음은 Ryan, Joiner와 Ryan(1985)이 제시한 31개의 흑체리나무를 조사하여 나무의 높이(H), 지표면 4.5피트에서의 나무의 지름(D) 및 부피(V)를 기록한 자료에 대해 생각해 본다. 이 자료는 Cook과 Weisberg(1982), Atkinson(1985)에 의해 분석된 바 있다. 잔차산점도인 그림 4.3에서 뚜렷한 비선형 형태를 볼 수 있다. 따라서 D와 H가 비선형의 형태로 반응

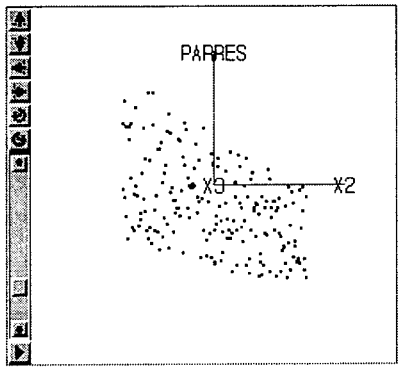

그림 4.1: $(x_2,\ x_3)^T$의 편잔차그림

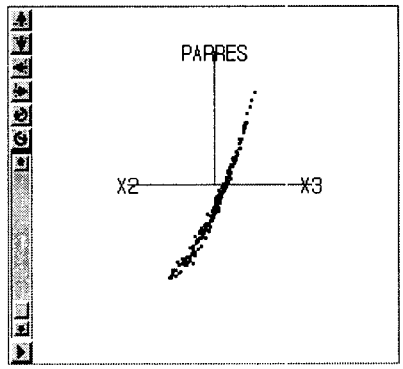

그림 4.2: $(x_2,\ x_3)^T$의 편잔차그림

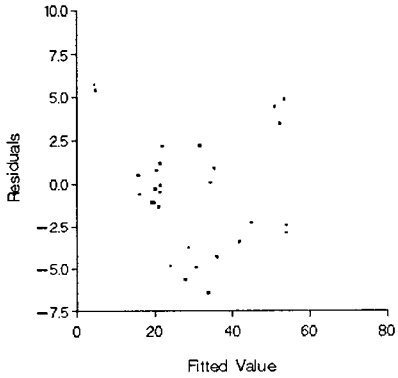

그림 4.3: 잔차산점도

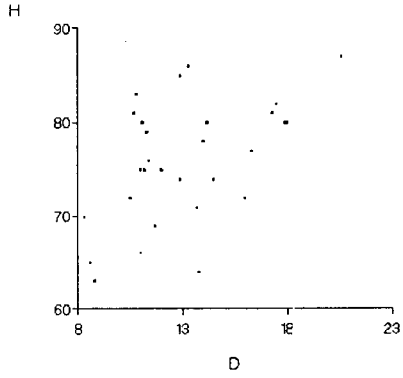

그림 4.4: 산점도 $\{H_i,\ D_i\}$

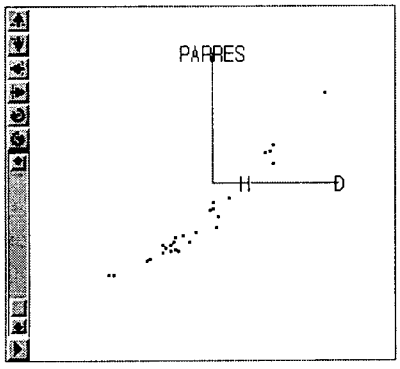

그림 4.5: $(D,\ H)^T$의 편잔차그림

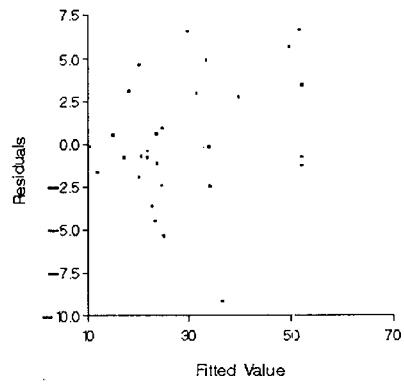

그림 4.6: 잔차산점도

변수 V에 영향을 줄 것이라고 기대할 수 있다. 산점도 $\{H_i,\ D_i\}$인 그림 4.4에서 두 설명변수 D와 H 사이에 연관성이 있음을 알 수 있다. 각각의 설명변수 D와 H에 대한 편잔차그림에서는 비선형의 형태가 거의 나타나지 않았다. 그림 4.5는 3차원 편잔차그림의 편잔차축을 중심으로한 회전에서 보여지는 연속된 2차원 산점도 중 가장 명확한 곡선이 보이는 산점도이다. 이것은 편잔차를 세로축으로 하고 D를 가로축으로 하는 산점도를 시작으로 하여 14°회전된 상태이다. 이 회전 각도를 이용하면 이 산점도의 가로축은 $z(14°) \cong D - H/4$가 된다. 이 그림에서 나타나는 비선형 곡선은 D와 H의 선형결합인 $z_0 = D - H/4$의 비선형 함수를 나타내며 지수함수로 볼 수 있다. 만약 3차원 편잔차그림에서 제시한 지수함수가 적절하다면 z_0를 지수변환한 후 얻어지는 e^{z_0}를 추가되는 변수로 하는 모형에서의 편잔차그림은 직선의 형태를 보일 것이라고 짐작할 수 있다. 실제로 e^{z_0}의 편잔차그림을 그려보면 직선의 형태가 나타남을 확인할 수 있다. 또한 잔차산점도인 그림 4.6은 변환 전에 보였던 비선형의 모습이 사라지고 모든 점들이 아무런 유형 없이 흩어져 있음을 확인할 수 있다.

5. 결론

선형회귀분석에서는 모든 설명변수들이 선형으로 모형에 포함된다고 가정한다. 따라서 최종 모형의 결정에 앞서 선형의 가정이 타당한지에 대한 검토가 필요하다. 본 논문에서는 비선형성의 탐색에 유용하다고 알려져 있는 2차원 편잔차그림에 대해 간단히 소개하였고 비선형성이 의심되는 설명변수가 두 개일때 사용되어질 수 있는 3차원 편잔차그림에 대해 설명하였다. 비선형성이 의심되는 두 변수들 간에 연관성이 약할때는 각 변수에 대한 2차원 편잔차그림만으로도 비선형성의 탐색과 변환을 위한 함수 형태의 파악이 가능함을 확인하였고 이는 3차원 편잔차그림을 통해서도 얻어질 수 있음을 알아보았다. 또한 두 변수간 연관성이 강할 경우 2차원 편잔차그림의 문제점을 지적하고 이를 해결하기 위해 3차원 편잔차그림을 이용함으로써 2차원 편잔차그림에서는 얻을 수 없었던 비선형성의 탐색 및 변환을 위한 비선형함수의 제시가 가능함을 확인하였다.

본 논문에서 다루지는 않았지만 비선형성이 의심되는 두 개의 설명변수와 나머지 변수들간의 관계를 선형으로 볼 수 없을 경우 덧편잔차그림과 CERES그림을 3차원으로 확장하여 적용시킬 수 있다면 비선형성의 탐색 및 변환을 위한 좀 더 일반화된 방법으로의 개선이 가능할 것이다.

참고문헌

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

[2] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics*, John Wiley & Sons, New York.

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

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

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

[6] Cook, R.D. (1993). Exploring partial residualplots, *Technometrics*, Vol. 35, 351-362.

[7] Cook, R.D. (1998). *Regression Graphics*, John Wiley & Sons, New York.

[8] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman & Hall, New York.

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

[10] Cook, R.D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*, John Wiley & Sons, New York.

[11] Cox, D.R. (1958). *Planning of Experiments*, John Wiley & Sons, New York.

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

[13] Huber, P. (1987). Experiences with three-dimentional scatterplots, *Journal of the American Statistical Association*, Vol. 82, 448-453.

[14] Larsen, W.A. and McCleary, S.J. (1972). The use of partial residual plots in regression analysis, *Technometrics*, Vol. 14, 781-790.

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

[16] Mansfield, E.R. and Conerly, M.D. (1987). Diagnostic value of residual and partial residual plots, *Journal of the American Statistical Association*, Vol. 41, 107-116.

[17] Ryan, T., Joiner, B. and Ryan, B. (1985). *Minitab Student Handbook, 2nd Ed.*, Duxbury: Belmont, CA.

[1999년 3월 접수, 1999년 12월 채택]

Three Dimensional Partial Residual Plots in Regression Analysis *

Myung-Wook Kahng [1] Jung-A Lee [2]

ABSTRACT

The structure and usefulness of partial residual plot as basic tools for dealing with curvature as a function of selected predictor variable in multiple regression analysis are explored. If a predictor is included in the model in a linear form when it actually has a nonlinear relationship with the response variable, this plot displays the correct functional form of predictor except in instances when there are nonlinear relationship among the predictors. In situations where nonlinearity exists in two predictors, we extend the idea of a partial residual plot to three dimension. A two-dimensional partial residual plot of each predictor is useful for indicating nonlinear relationships when there is low association between two predictors. In case of high association, a three-dimensional partial residual plot is able to display nonlinearity that is not evident in two-dimensional plots. This is illustrated by the data.

Keywords: Partial residual plot; Three dimensional plot; Rotating plot.

* This Research was supported by the Sookmyung Women's University Research Grants in 1998.

1) Associate Professor, Department of Statistics, Sookmyung Women's University.
E-mail: mwkahng@sookmyung.ac.kr

2) Department of Statistics, Sookmyung Women's University.