The Korean Communications
in Statistics Vol. 5, No. 3, 1998
pp. 571-580

비선형회귀분석에서 편잔차그림에 대한 연구[1)]

강 명 욱[2)], 김 정 혜[3)]

요 약

선형회귀분석에서 새로운 변수가 모형에 추가될 때 변수변환의 필요성과 적절한 변환의 형태를 진단하는 기능이 있다고 알려져 있는 편잔차그림과 덧편잔차그림을 비선형회귀모형에 적용하고 이 그림들이 기능을 제대로 수행할 수 있는 조건을 알아보았다.

1. 서 론

회귀모형의 개발에서 모형에 포함되는 적절한 변수의 선택은 매우 중요하고도 어려운 문제이다. 새로운 변수가 기존의 모형에 추가되는 경우 이 변수가 모형에 미치는 영향력과 기여도는 이미 모형에 포함되어있는 변수의 영향을 받는다. 따라서 모형개발의 과정에서 추가되는 변수의 영향력을 알아보고 또한 선형의 형태로 추가할 것인지 아니면 비선형의 형태로 변환하여 추가할 것인지에 대한 진단을 하여야 한다. 이러한 진단은 그림을 통한 방법으로 가능하며 그 중 대표적이고 지금까지 연구가 활발히 진행되어온 것으로 추가변수그림(added variable plot)과 편잔차그림(partial residual plot), 편잔차그림을 확장시킨 덧편잔차그림(augmented partial residual plot)과 CERES그림(combining conditional expectation and residual plot)이 있다.

추가변수그림은 Cox(1958)가 처음 제안하였고 1980년대에 Belsley, Kuh와 Welsch(1980), Henderson과 Velleman(1981), Cook과 Weisberg(1982)에 의해 회귀진단에 이용되었으며 Cook과 Weisberg(1989)는 3차원의 추가변수그림을 제안하였고 Cook(1996)은 추가변수그림을 통한 선형회귀모형에서 비선형성의 탐색에 관해 발표하였다.

편잔차그림은 Ezekiel(1924)이 처음 도입한 이후로 Larsen과 McCleary(1972)가 편잔차그림이라 명명하였고 Wood(1973)는 이것을 성분잔차그림(component-plus-residual plot)이라 불렀다. 이 방법도 1980년대에 들어 Cook과 Weisberg(1982), Atkinson(1985), Chatterjee와 Hadi(1988)에 의해 회귀진단의 도구로 사용되었다. 이 그림은 주목적이라고 할 수 있는 비선형성의 탐색뿐만이 아니라 이상점 및 영향력이 큰 관측값을 찾는데도 이용되었다. Mallows(1986)는 편잔차그림을 개선한 덧편잔차그림을 소개하였고 Cook(1993)은 선형회귀모형에서 추가되는 변수의 비선형성을 파악하는 방법을 제시하고 CERES그림이라는 새로운 진단방법도 소개하였고 Berk(1998)은 이것을 3차원으로 확장하였다.

본 논문에서는 지금까지 선형모형에 국한되어 연구되어 온 그림들을 비선형모형에도 적용할 수 있는지 알아보려고 한다

1) 본 연구는 숙명여자대학교 1996년도 교비연구비 지원에 의해 수행되었음.
2) (140-742) 서울특별시 용산구 청파동 2가 53-12 숙명여자대학교 통계학과 부교수.
3) (140-742) 서울특별시 용산구 청파동 2가 53-12 숙명여자대학교 통계학과.

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.

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

여기서 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iq})^T$이고 $f(\boldsymbol{X}; \boldsymbol{\theta}) = \left[f(\boldsymbol{x}_1; \boldsymbol{\theta}), f(\boldsymbol{x}_2; \boldsymbol{\theta}), \cdots, f(\boldsymbol{x}_n; \boldsymbol{\theta})\right]^T$라고 하면 위의 비선형회귀모형은 다음과 같이 표현된다.

$$\boldsymbol{y} = f(\boldsymbol{X}; \boldsymbol{\theta}) + \boldsymbol{\varepsilon} \tag{2.1}$$

모수 $\boldsymbol{\theta}$는 차수가 p인 회귀계수벡터이며, $\boldsymbol{\varepsilon}$은 차수가 n인 오차항벡터이고 평균 $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, 분산 $Var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$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.

$$\boldsymbol{y} = f(\boldsymbol{X}; \boldsymbol{\theta}) + \boldsymbol{g}(\boldsymbol{z}) + \boldsymbol{\varepsilon} \tag{2.2}$$

여기서 $\boldsymbol{g}(\boldsymbol{z}) = [g(z_1), \cdots, g(z_n)]^T$이다. $g(z)$가 변수 z의 선형함수 또는 선형에 가깝다고 하면 모형 (2.2)는

$$\boldsymbol{y} = f(\boldsymbol{X}; \boldsymbol{\beta}) + \alpha \boldsymbol{z} + \boldsymbol{\delta} \tag{2.3}$$

이 된다. 우선 모형 (2.3)에서 $\boldsymbol{\beta}$와 α의 최소제곱추정값을 각각 $\hat{\boldsymbol{\beta}}$과 $\hat{\alpha}$이라고 하자.

3. 그림을 통한 진단방법

3.1 편잔차그림

추가변수그림은 설명변수들 사이에 연관성이 클 경우에는 추가되는 변수의 함수 $g(z)$를 왜곡되게 표현한다. 따라서 추가변수그림의 형태로 추가여부를 결정할 수는 있으나 이것으로 추가되는 변수의 형태까지 파악하기에는 곤란하다. 이러한 형태를 파악하고 적절한 변환을 찾기 위해서는 함수 $g(z)$를 좀 더 정확히 묘사하는 방법이 필요하며 이를 위해 편잔차그림이 이용된다.

편잔차그림은 모형 (2.3)에 적합시켜 얻어지는 잔차 $e_i = y_i - \hat{y}_i = y_i - f(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}) - \hat{\alpha} z_i$에 $\hat{\alpha} z_i$를 더해 얻어지는 편잔차(partial residual) $e_i + \hat{\alpha} z_i$를 세로축으로 하고 추가되는 변수의 값 z_i를 가로축으로 하는 $\{e_i + \hat{\alpha} z_i,\ z_i\}$의 산점도이다. 이 때 편잔차벡터 $\boldsymbol{e} + \hat{\alpha} \boldsymbol{z}$는 다음과 같다.

$$\begin{aligned} e+\hat{\alpha}z &= [f(X;\theta)+g(z)+\varepsilon]-[f(X;\hat{\beta})+\hat{\alpha}z]+\hat{\alpha}z \\ &= f(X;\theta)-f(X;\hat{\beta})+g(z)+\varepsilon \end{aligned} \tag{3.1}$$

최소제곱추정값 $\hat{\beta}$이 실제모수 θ에 근접하면 $f(X;\theta)\cong f(X;\hat{\beta})$이고 식 (3.1)에 의해 편잔차그림은 $\{g(z_i)+\varepsilon_i,\ z_i\}$가 되어 $g(z)$를 잘 묘사할 수 있음을 알 수 있다. 따라서 편잔차그림은 $\hat{\beta}$에 의해 결정된다고 할 수 있다.

만약 $g(z)$가 z에 대해 선형 또는 선형에 가깝다면 모형 (2.3)은 참모형인 (2.2)에 대한 근사적인 모형이고 $\hat{\beta}$은 실제모수인 θ에 근접하고 $f(X;\theta)-f(X;\hat{\beta})$이 0에 가깝게 된다. 따라서 편잔차그림은 강한 선형을 보이고 추가되는 변수인 z는 변환 없이 추가될 수 있다. 이에 반해 $g(z)$가 z에 대해 강한 비선형이라면 모형 (2.3)은 z의 변환에 의해 개선되어야 한다. 이 경우 $\hat{\beta}$이 θ에 근접한다고 기대하기가 어렵고 $f(X;\theta)-f(X;\hat{\beta})$이라는 편의(bias)가 발생하므로 편잔차그림이 $g(z)$의 형태를 나타낸다고 말할 수는 없다. 그러나 편잔차그림이 $g(z)$가 선형인 경우에만 효과가 있고 비선형인 경우는 함수의 형태를 감지하는 기능이 전혀 없는 것은 아니다. 그것은 $g(z)$의 형태와 관계없이 $\hat{\beta}$이 실제모수 θ에 근접하는 경우가 있기 때문이다.

비선형함수인 $f(X;\theta)$를 Taylor 급수전개에 의해 전개하고 $\hat{\beta}$근방의 θ에 대하여 2차이상 미분한 부분을 무시할 수 있다고 가정하면 다음과 같이 선형화 할 수 있다.

$$f(X;\theta)\cong f(X;\hat{\beta})+\hat{V}(\theta-\hat{\beta}) \tag{3.2}$$

여기서 $\hat{V}$은 $\theta=\hat{\beta}$일 때 $\partial f/\partial\theta$의 값이며 식 (2.2)의 양변에서 $f(X;\hat{\beta})$을 빼주면 식 (3.2)에 의해 다음과 같이 되며

$$y-f(X;\hat{\beta})\cong\hat{V}(\theta-\hat{\beta})+g(z)+\varepsilon \tag{3.3}$$

행렬 $\hat{V}$이 선형모형에서의 X행렬의 역할을 한다는 것을 알 수 있다. 또한 식 (3.3)은 선형모형의 특성을 가지므로 Cook(1993)이 선형모형에 적용한 방법을 그대로 사용할 수 있다. 따라서 $\hat{V}$과 z가 서로 연관성이 없으면 $\hat{\beta}$이 실제모수 θ에 근접하고 편잔차그림은 $g(z)$를 잘 표현할 수 있음을 알 수 있다. Cook(1996)은 이 때 z의 편잔차그림과 추가변수그림은 일치한다고 하였다. 반면 $\hat{V}$과 z가 연관성이 있을 경우 편잔차그림에 대한 조심스런 접근이 요구된다.

식 (3.2)를 이용하면 식 (3.1)의 편잔차벡터는

$$e+\hat{\alpha}z\cong\hat{V}(\theta-\hat{\beta})+g(z)+\varepsilon \tag{3.4}$$

이다. 만약 $\hat{\beta}$과 실제모수 θ가 상당히 차이가 있어 식 (3.4)의 우변의 $\hat{V}(\theta-\hat{\beta})$이 소멸되지

않는 경우 편잔차그림은 $g(z)$의 형태를 나타내는 기능에 편의를 갖는다고 할 수 있다. 식 (3.4)의 양변에 조건부 기대값을 취하면

$$E(e+\hat{\alpha}z \mid z) \cong E(\hat{V} \mid z)(\theta-\hat{\beta}) + g(z)$$

이고 이 식은 편잔차벡터와 z의 회귀함수이다. 만약 $\hat{\beta}$과 실제모수 θ가 상당히 차이가 나는 경우 $\hat{V}$과 z의 종속성이 편잔차그림에 상당한 편의를 제공해 준다. 그러나 $E(\hat{V} \mid z)$이 z에 무관하거나 z에 선형이라면 편의는 없어진다고 보아도 된다. 이러한 편잔차의 조건부 기대값벡터의 각 원소를 첨자를 생략하고 표현하면 다음과 같다.

$$E(e+\hat{\alpha}z \mid z) \cong (\theta-\hat{\beta})^T E(\hat{v} \mid z) + g(z) \tag{3.5}$$

여기서 $\hat{v}$은 $\theta=\hat{\beta}$일 때 $\partial f/\partial\theta$의 값이고 위에서의 설명은 다음의 정리에서 확인된다.

정리 3.1 : 조건부 기대값 $E(\hat{v} \mid z)$이 a_0+a_1z이고 모형 (2.2)가 사실이면 모형 (2.3)에서 얻은 $\hat{\beta}$은 θ의 일치(consistent)추정량이다. 이때 a_0와 a_1은 $p\times 1$ 벡터이다.

이 정리는 다음에 나오는 정리 3.2의 특별한 경우이며 Cook(1993)이 제시한 Lemma 2.1, 3.1, 4.1에서 설명변수벡터 x_1을 $\hat{v}$으로 바꾸어 이러한 정리를 도출할 수 있고 Lemma 4.1의 증명과 같은 방법으로 증명할 수 있다.

정리 3.1에 의하면 $E(\hat{v} \mid z)$이 z의 선형으로 표현되면 식 (3.5)에서 편의는 소멸되고 편잔차그림은 $g(z)$의 형태를 나타낸다. $E(\hat{v} \mid z)$의 선형성은 p개의 $\{\hat{v}_{ij}, z_i\}$, $j=1,\cdots,p$의 산점도를 통하여 확인할 수 있다. 이러한 산점도에서 강한 비선형성이 나타나지 않는다면 편잔차그림에서 나타난 형태가 바로 $g(z)$의 형태라고 할 수 있다. 그러나 이러한 산점도에서 비선형성이 보이고 따라서 정리 3.1에서의 충분조건이 만족되지 않는다면 편잔차그림에서 나타난 결과가 $g(z)$의 형태라고 볼 수 없다. 다음 절에서는 이러한 문제의 해결방법으로 덧편잔차그림을 생각한다.

3.2 덧편잔차그림

덧편잔차그림은 조건부 기대값이 선형이어야 한다는 편잔차그림에서의 제약을 보다 완화시킨 그림이라 할 수 있다. 덧편잔차그림은 변수들 사이에 강한 연관성이 있을 경우 편잔차그림보다 추가되는 변수의 형태에 더 근접하게 접근된다.

모형 (2.3)에 추가되는 변수 z의 2차항을 더한 다음과 같은 모형을 고려한다.

$$y = f(X;\phi) + \alpha_1 z + \alpha_2 z\cdot z + \xi \tag{3.6}$$

여기서 $z\cdot z$는 추가되는 설명변수 z_i들의 제곱을 원소로 하는 벡터이다.

이 모형에 대한 최소제곱추정값을 $\hat{\phi}$, $\hat{\alpha}_1$, $\hat{\alpha}_2$이라 하면 덧편잔차그림은 위 모형 (3.6)에서 얻은 잔차 e_i에 $\hat{\alpha}_1 z_i + \hat{\alpha}_2 z_i^2$를 더한 덧편잔차(augmented partial residual) $e_i + \hat{\alpha}_1 z_i + \hat{\alpha}_2 z_i^2$를 세로축으로 하고 추가되는 변수의 값 z_i를 가로축으로 하는 $\{e_i + \hat{\alpha}_1 z_i + \hat{\alpha}_2 z_i^2,\ z_i\}$의 산점도이다. 식 (3.2)에서와 같이 비선형함수 $f(X;\theta)$는 모형 (3.6)에서 얻은 최소제곱추정값 $\hat{\phi}$에 대해 선형화 할 수 있고 이를 이용하면 덧편잔차벡터는 다음과 같다.

$$e + \hat{\alpha}_1 z + \hat{\alpha}_2 z \cdot z \cong \tilde{V}(\theta - \hat{\phi}) + g(z) + \varepsilon$$

여기서 $\tilde{V}$는 $\theta = \hat{\phi}$일 때 $\partial f / \partial \theta$ 값이며 회귀함수를 얻기 위해 양변에 조건부 기대값을 취하면 다음과 같다.

$$E(e + \hat{\alpha}_1 z + \hat{\alpha}_2 z \cdot z \mid z) \cong E(\tilde{V} \mid z)(\theta - \hat{\phi}) + g(z)$$

이 회귀함수가 $g(z)$가 되기 위해서는 모형 (3.6)에서 얻은 $\hat{\phi}$이 θ에 근접해야 한다. 이러한 덧편잔차의 조건부 기대값벡터의 각 원소를 첨자를 생략하고 표현하면 다음과 같다.

$$E(e + \hat{\alpha}_1 z + \hat{\alpha}_2 z^2 \mid z) \cong (\theta - \hat{\phi})^T E(\tilde{v} \mid z) + g(z)$$

여기서 $\tilde{v}$는 $\theta = \hat{\phi}$일 때 $\partial f / \partial \theta$의 값이고 편잔차그림에서의 정리 3.1과 유사한 다음의 정리가 도출된다.

정리 3.2 : 조건부 기대값 $E(\tilde{v} \mid z)$이 $a_0 + a_1 z + a_2 z^2$이고 모형 (2.2)가 사실이면 모형 (3.6)에서 얻은 $\hat{\phi}$은 θ의 일치추정량이다. 이때 a_0, a_1과 a_2는 $p \times 1$ 벡터이다.

따라서 $E(\tilde{v} \mid z)$이 z의 2차식이라면 덧편잔차그림은 $g(z)$의 형태를 잘 나타낸다고 할 수 있다. 이 조건은 편잔차그림에서와 같이 $\{\hat{v}_{ij},\ z_i\}$의 산점도로 검토할 수 있고 정리 3.1에서의 충분조건보다 완화된 조건이라고 볼 수 있을 뿐만 아니라 $g(z)$를 근사하는데 1차함수인 αz보다는 2차함수인 $\alpha_1 z + \alpha_2 z^2$가 좀 더 정확하므로 $\hat{\phi}$이 $\hat{\beta}$보다 θ에 더 근접한다고 볼 수 있기 때문에 덧편잔차그림이 편잔차그림보다 나은 효과가 있다고 할 수 있다.

4. 예 제

지금까지 알아본 비선형모형에서의 편잔차그림과 덧편잔차그림으로 추가되는 변수의 유의성의 확인과 변환형태의 제시가 가능한가를 생성한 자료를 이용하여 알아보려고 한다. 실제모형은 다음과 같다.

$$y = \theta_1\theta_2 x_1/(1+\theta_1 x_1) + e^{\theta_3 x_2} + \varepsilon \tag{4.1}$$

x_1과 x_2는 각각 정규분포 $N(150, 4^2)$와 균일분포 $U(0, 1)$에서 생성하고 y는 $\theta_1=3$, $\theta_2=10$, $\theta_3=5$라 하고 회귀함수의 값을 구하여 여기에 정규분포 $N(0, 7^2)$에서 생성된 오차값을 더하여 44개의 자료를 생성한다. 위의 비선형모형 (4.1)에서 변수 x_2는 비선형형태로 반응변수 y에 영향을 준다. 그렇다면 x_1만으로 구성된 모형 $y = \theta_1\theta_2 x_1/(1+\theta_1 x_1) + \varepsilon$에 x_2를 추가할 때 적절한 변환형태인 지수함수가 편잔차그림으로 제시되는지 확인해 본다.

만약 $g(x_2)$가 x_2의 선형에 가깝다면 모형 $y = \beta_1\beta_2 x_1/(1+\beta_1 x_1) + \alpha x_2 + \delta$를 생각할 수 있고 편잔차그림이 추가되는 변수 x_2의 비선형성을 탐지할 수 있으려면 $\hat{\boldsymbol{v}}$과 x_2가 연관성이 없거나 연관성이 있다면 조건부 기대값 $E(\hat{\boldsymbol{v}} \mid x_2)$이 x_2의 선형이어야 한다. $E(\hat{\boldsymbol{v}} \mid x_2)$과 x_2와의 관계는 [그림 1]의 산점도 $\{\hat{v}_{i1}, x_{i2}\}$와 $\{\hat{v}_{i2}, x_{i2}\}$로 확인할 수 있다. 이때 $\hat{v}_{i1}$과 $\hat{v}_{i2}$은 각각 $f(x_1; \boldsymbol{\theta}) = \theta_1\theta_2 x_1/(1+\theta_1 x_1)$이라 하고 $\boldsymbol{\theta} = \hat{\boldsymbol{\beta}}$일 때의 $\partial f/\partial\theta_1$과 $\partial f/\partial\theta_2$의 44개의 값으로 구성된 벡터 $\hat{\mathbf{v}}_1$과 $\hat{\mathbf{v}}_2$의 i번째 원소이다. 즉, $\hat{\mathbf{v}}_1$과 $\hat{\mathbf{v}}_2$은 $\hat{\boldsymbol{V}}$행렬의 첫 번째와 두 번째 열 벡터이다. 두 그림에서의 모든 점들이 아무런 유형 없이 흩어져 있음이 보인다. 따라서 $\hat{\boldsymbol{v}}$은 x_2와 연관성이 없으므로 편잔차그림은 추가되는 x_2의 함수형태를 잘 나타낼 수 있으리라고 기대할 수 있다. 편잔차그림인 [그림 2]에서 기대한 대로 뚜렷한 비선형곡선을 볼 수 있고, x_2가 지수함수로 추가되어야 하는 바를 보이고 있다.

이 경우 $\hat{\boldsymbol{v}}$이 x_2와 연관성이 없으므로 덧편잔차그림은 편잔차그림과 거의 같은 그림이 되고 이는 또한 y의 x_2에 대한 주변그림과도 같을 것이라고 기대할 수 있다. 따라서 $\hat{\boldsymbol{v}}$과 x_2가 연관성이 없을 때에는 단순한 주변그림을 그려서 변수의 변환형태를 탐지할 수도 있다. 본 논문에 보이지는 않았으나 이 두 그림은 편잔차그림과 흡사하다는 것을 확인할 수 있었다.

이상의 그림들에서 추가되는 변수 x_2를 지수함수의 형태로 변환하여야 한다는 것을 알 수 있다. 그렇다면 실제로 추가되는 변수 x_2를 지수함수로 변환하고 이것을 추가되는 변수로 생각하여 이에 대한 편잔차그림을 그린다면 직선 형태를 보일 것이라는 예측을 해 볼 수 있다. 즉, $z = e^{x_2}$를 추가되는 변수로 하여 편잔차그림을 그려보면 [그림 3]이며 뚜렷한 직선을 볼 수 있다. 이는 변수 x_2가 지수함수의 형태로 추가되어야 한다는 강력한 증거이다. 이상에서 $\hat{\boldsymbol{v}}$과 x_2가 연관성이 없을 때에 변수 x_2의 추가 필요성과 추가되는 변수의 함수형태를 편잔차그림으로 알아내어 정확한 모형을 설정할 수 있음을 알 수 있다.

만약 $\hat{\boldsymbol{v}}$과 x_2가 연관성이 있는 경우에는 조건부 기대값 $E(\hat{\boldsymbol{v}} \mid x_2)$이 x_2의 선형이라는 조건을 만족해야만 편잔차그림이 변수의 비선형성을 탐지할 수 있다. 그러나 이러한 조건이 만족되지 않는다면 조건부 기대값에 대한 제약이 보다 완화된 덧편잔차그림을 그려보는 것이 바람직하다. 함수관계가 알려진 다음의 모형을 참모형이라고 생각하자.

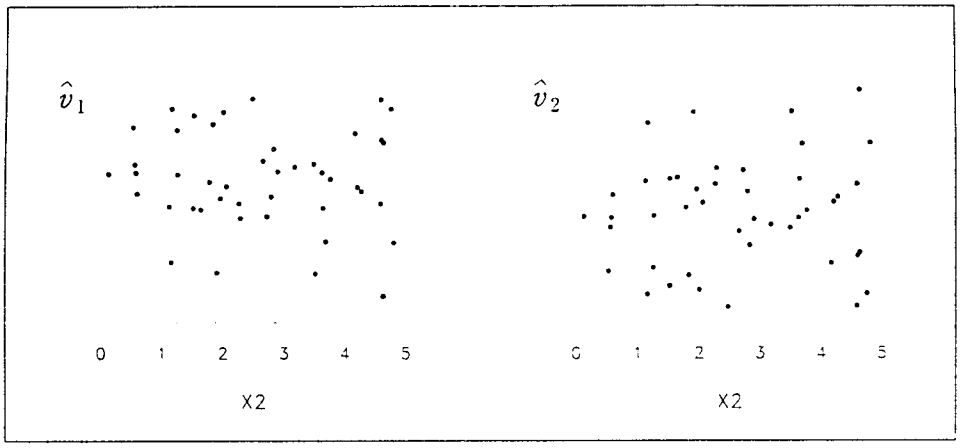
$$y = \theta_1 \theta_2 x_1 / (100 + \theta_1 x_1) + e^{\theta_3 x_2}/10 + \varepsilon \qquad (4.2)$$

여기서 $\theta_1 = .03$, $\theta_2 = 1000$, $\theta_3 = 1$이라 하고 $\boldsymbol{x}_1$은 $N(10, 4^2)$, $\boldsymbol{\varepsilon}$은 $N(0, .5^2)$에서 각각 44개의 난수를 생성하여 구성한다. $\boldsymbol{x}_2$는 조건부 기대값 $E(\hat{\boldsymbol{v}} \mid x_2)$이 x_2에 대해 비선형이 되도록 다음과 같이 생성한다. 주어진 $\theta_1, \theta_2, \theta_3$의 값과 벡터 $\boldsymbol{x}_1$을 이용하여 $\boldsymbol{V}$행렬의 첫 번째 열벡터 $\mathbf{v}_1 = \{v_{i1}\}$을 구하고 벡터 $\mathbf{v}_1$의 각 원소의 제곱근을 원소로 하는 벡터 $\boldsymbol{x}_2$를 생성한다. 이렇게 구한 $\mathbf{v}_1$과 $\boldsymbol{x}_2$를 이용하여 산점도 $\{\hat{v}_{i1}, x_{i2}\}$를 그려보면 비선형이 될 것이다. 이제 $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{\varepsilon}$을 이용하고 모형 (4.2)를 적용하여 $\boldsymbol{y}$를 생성한다.

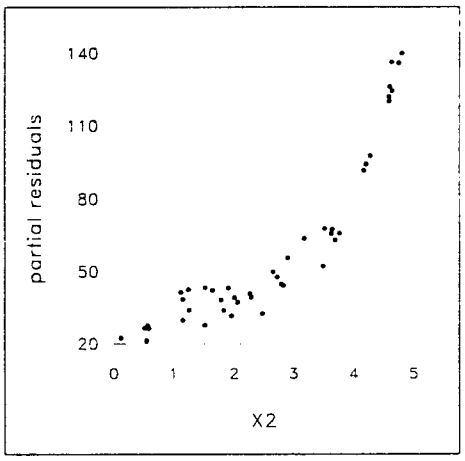
우선 편잔차그림을 그리기 위하여 모형 $y = \beta_1 \beta_2 x_1 / (100 + \beta_1 x_1) + \alpha x_2 + \delta$를 생각한다. 이 모형에서 회귀계수의 최소제곱추정값을 구하고 이를 이용하여 $\hat{\boldsymbol{V}}$을 찾아 첫 번째와 두 번째 열벡터를 $\hat{\mathbf{v}}_1$, $\hat{\mathbf{v}}_2$이라 한다. $E(\hat{\boldsymbol{v}} \mid x_2)$의 관계를 알아보기 위하여 $\{\hat{v}_{i1}, x_{i2}\}$와 $\{\hat{v}_{i2}, x_{i2}\}$의 산점도를 그려보면 [그림 4]와 같다. [그림 4-a]의 산점도 $\{\hat{v}_{i2}, x_{i2}\}$에서 뚜렷한 비선형의 곡선이 보이며 따라서 $\hat{\boldsymbol{v}}$과 x_2가 연관성이 있고 조건부 기대값이 x_2의 선형이 아니므로 이 자료에서 편잔차그림은 추가되는 변수 x_2의 함수형태를 잘 나타낼 수 있다고 기대할 수 없다. 실제로 편잔차그림인 [그림 5]는 지수함수 형태가 아닌 선형의 그림으로 $g(x_2)$의 형태를 나타내지 못하는 예상된 결과를 보인다.

편잔차그림의 제약조건보다 완화된 조건에서 함수의 형태를 보이는 덧편잔차그림을 얻기 위해 모형 $y = \phi_1 \phi_2 x_1 / (100 + \phi_1 x_1) + \alpha_1 x_2 + \alpha_2 x_2^2 + \xi$를 생각한다. 이 모형에서 적합한 후 구해지는 회귀계수의 최소제곱추정값을 $\hat{\boldsymbol{\phi}}$이라 하고 $\tilde{\mathbf{v}}_1$와 $\tilde{\mathbf{v}}_2$는 각각 $\theta = \hat{\boldsymbol{\phi}}$일 때의 $\tilde{\boldsymbol{V}}$행렬의 첫 번째와 두 번째 열벡터라 하자. [그림 7]은 $\{\tilde{v}_{i1}, x_{i2}\}$와 $\{\tilde{v}_{i2}, x_{i2}\}$의 산점도이며 이들은 특정한 형태가 나타나지 않거나 상수함수의 형태이다. 따라서 덧편잔차그림의 충분조건을 만족한다고 할 수 있고 이 경우 덧편잔차그림은 $g(x_2)$의 형태를 잘 묘사할 것으로 예상된다.

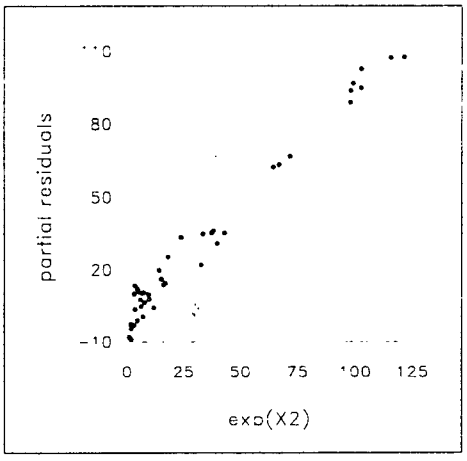
덧편잔차그림을 그려보면 [그림 6]과 같다. 이 덧편잔차그림은 편잔차그림에서는 나타나지 않았던 지수함수 형태의 곡선이 나타난다. 따라서 추가되는 변수 x_2를 지수함수의 형태로 변환하여야 한다는 것을 덧편잔차그림을 통하여 알 수 있다. 그렇다면 지수함수로 변환한 변수 $z = e^{x_2}$를 추가되는 변수로 생각하자. 이 변수가 더 이상의 변환이 필요 없이 추가되어야 한다는 결론을 내릴 수 있으려면 이에 대한 편잔차그림이 직선 형태가 되어야 할 것이다. 편잔차그림을 그려보면 앞의 예에서와 마찬가지로 직선의 형태를 볼 수 있다. 따라서 추가되는 변수 x_2는 더 이상의 변환이 필요하지 않고 지수함수의 형태로 모형에 추가되어야 한다는 것을 알 수 있다.



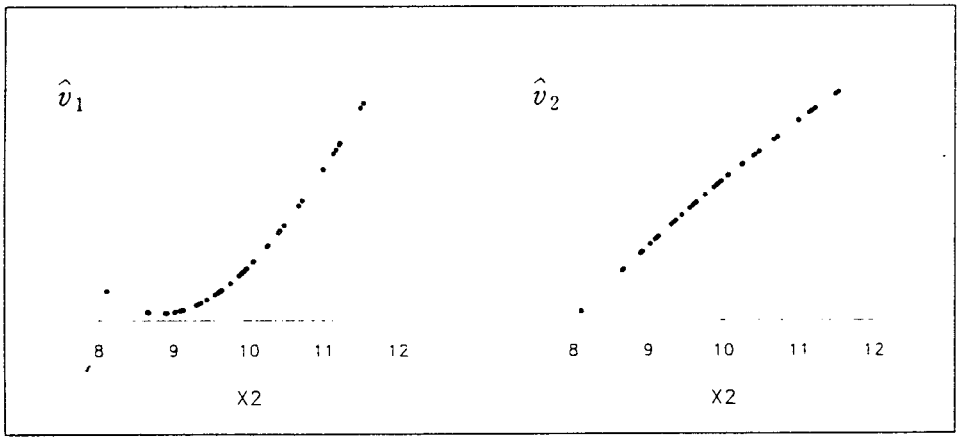
[그림 1-a] 산점도 : $\{\hat{v}_{i1}, x_{i2}\}$ [그림 1-b] 산점도 : $\{\hat{v}_{i2}, x_{i2}\}$



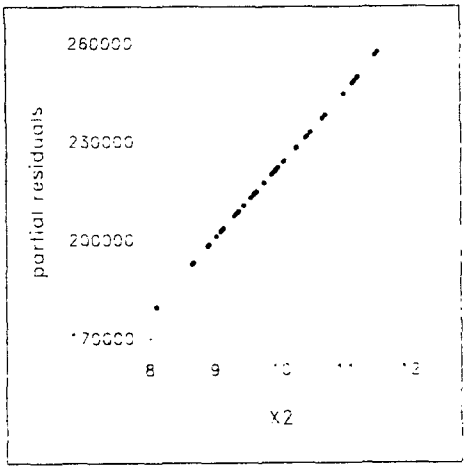
[그림 2] x_2의 편잔차그림



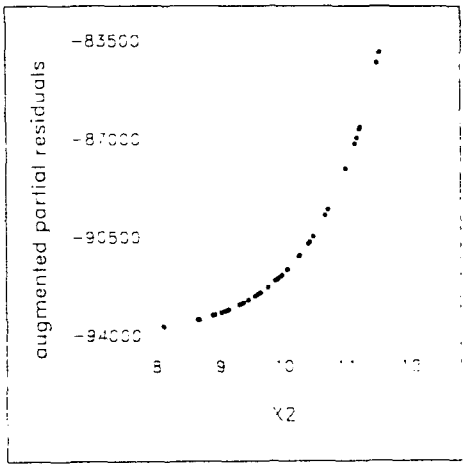
[그림 3] $z = e^{x_2}$의 편잔차그림



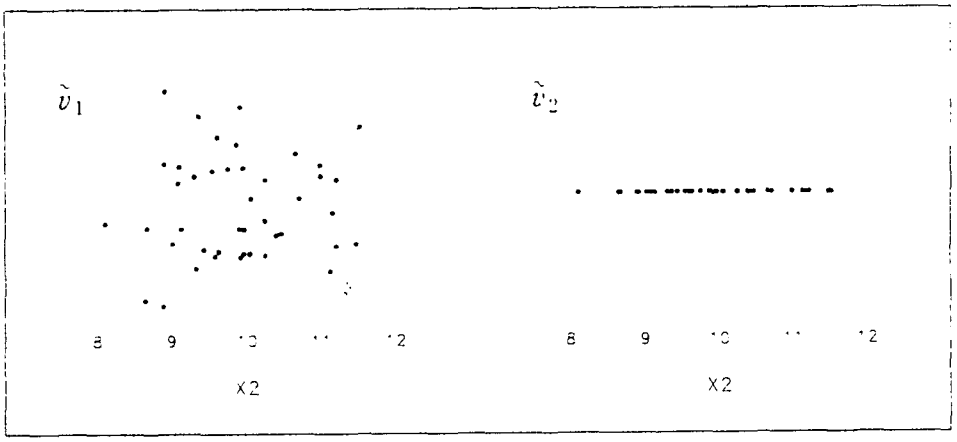
[그림 4-a] 산점도 : $\{\hat{v}_{i1}, x_{i2}\}$ [그림 4-b] 산점도 : $\{\hat{v}_{i2}, x_{i2}\}$



[그림 5] x_2의 편잔차그림



[그림 6] x_2의 덧편잔차그림



[그림 7-a] 산점도 : $\{\tilde{v}_1, x_{i2}\}$

[그림 7-b] 산점도 : $\{\tilde{v}_2, x_{i2}\}$

5. 결 론

기존의 모형에 새로운 변수를 추가하고자 할 때 그 변수의 유의성을 확인해야 하고 추가되는 변수의 적절한 함수형태를 찾아야 한다. 특히 비선형모형에서는 추가되는 변수가 선형으로 추가되기보다는 비선형으로 추가되는 경우가 많으므로 이 문제는 더욱 중요하다. 본 논문에서는 선형모형에서 변수를 추가하고자 할 때 적절한 변수의 선택과 진단에 사용되어온 편잔차그림과 덧편잔차그림을 이론적으로 간략히 설명하였고 이 방법을 비선형모형에 응용하여 적용시켜 보았다. 모형에 포함되어 있던 변수들과 새로 추가되는 변수 사이에 연관성이 없을 경우에는 편잔차그림과 덧편잔차그림이 추가되는 변수의 유의성과 변환의 필요성, 변환의 함수 형태까지도 제시할 수 있었다. 또한 연관성이 있을 경우에도 일정한 조건이 충족되면 연관성이 없을 때와 같은 기능을 수행할 수 있음을 확인하였다.

본 논문에서는 다루지 않았으나 편잔차그림과 덧편잔차그림의 이론을 확장하여 선형모형에서

Cook(1993)이 소개한 새로운 진단방법인 CERES그림을 비선형모형에 적용시킬 수 있다면 더욱 완화된 조건에서도 추가되는 변수의 비선형성 탐지를 위한 좀 더 일반적이고 개선된 방법의 개발이 가능할 것이다.

참 고 문 헌

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

[2] Belsley, D., 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] 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. (1996). Added variable plots and curvature in linear regression, *Technometrics,* Vol. 38, 275-278.

[7] Cook, R. D., and Weisberg, S. (1982). *Residuals and Influence in Regression,* Chapman and Hall: London.

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

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

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

[11] Henderson, H. V., and Velleman, P. F. (1981). Building multiple regression models interactively, *Biometrics,* Vol. 37, 391-412.

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

[13] Mallows, C. L. (1986). Augmented partial residual plots, *Technometrics,* Vol. 28, 313-320.

[14] Wood, F. S. (1973). The use of individual effects and residuals in fitting equations to data. *Technometrics.* Vol. 15, 677-695.