Journal of the Korean Data & Information Science Society
2014, **25**(1), 177–185

http://dx.doi.org/10.7465/jkdi.2014.25.1.177
한국데이터정보과학회지

3차원 잔차산점도를 이용한 로지스틱회귀모형에서 교호작용의 탐색[†]

강명욱[1]

[1]숙명여자대학교 통계학과

접수 2013년 12월 20일, 수정 2014년 1월 6일, 게재확정 2014년 1월 12일

요 약

로지스틱회귀모형에서 설명변수만으로는 충분히 설명이 되지 못하고 설명변수의 변환된 형태인 이차항 또는 교호작용항이 필요한 경우가 있다. 설명변수가 두 개이고 조건부 분포가 이변량 정규분포를 따르는 경우 로지스틱회귀모형에서는 기본적으로 이차항과 교호작용항이 모형에 포함되어야 한다. 하지만 조건부 분포의 분산과 상관계수에 따라 이차항과 교호작용항이 필요하지 않게 되는 경우도 있다. 분산이나 상관계수에 대한 정보는 산점도를 보고 대체적인 판단이 가능하지만 교호작용항의 필요성을 판단하기가 쉽지 않다. 본 논문에서는 3차원 잔차산점도를 이용한 교호작용의 탐색방법을 제시하고 이 방법을 실제 자료에 적용시켜본다.

주요용어: 교호작용, 로그-밀도비, 로지스틱회귀모형, 역회귀, 이항회귀, 3차원 잔차산점도.

1. 서론

일반적인 정규선형모형 (normal linear model)은 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. 이항반응자료처럼 반응변수가 이산형인 경우에는 정규이론에 근거한 선형모형은 적절하지 않고 Nelder와 Wedderburn (1972)에 의해 체계화 된 일반화선형모형 (generalized linear models)으로 해결이 가능하다. 일반화선형모형은 지수족 (exponential family) 분포와 연결함수 (link function)를 이용하여 다음과 같은 두 가지 과정으로 정규이론에 의한 선형모형을 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.

선형회귀모형에서 반응변수의 기댓값은 설명변수들의 선형결합 $\boldsymbol{x}^T\boldsymbol{\beta}$이라고 가정한다. 로지스틱회귀모형에서도 반응변수는 기본적으로 설명변수의 선형결합에 의해 결정된다고 가정한다. 그러나 이러한 모형은 설명변수 간의 교호작용을 감안하지 않은 것이다. 선형모형에서 교호작용의 효과에 대한 검토는 교호작용을 포함하는 모형에서 교호작용의 추가적인 설명력에 대한 유의성 검정을 통해서 확인할 수 있다. Cook과 Weisberg (1989)가 제시한 3차원 잔차산점도는 선형회귀모형에 대한 가정의 검토에 널리 사용되어지는 그림이며 이 그림은 설명변수들 간의 교호작용진단에도 사용될 수 있다. 본 논문에서

[†] 본 연구는 숙명여자대학교 2012년도 교내연구비 지원에 의해 수행되었음.
[1] (140-742) 서울 용산구 청파로 47길 100, 숙명여자대학교 통계학과, 교수. E-mail: mwkahng@sm.ac.kr

는 선형회귀모형에 적용되었던 방법을 확장하여 로지스틱회귀모형에서 설명변수 간의 교호작용을 탐색해 볼 수 있는 방법에 대해 연구하고자 한다. 로지스틱회귀모형에서 그래프를 이용한 연구로는 Kahng 등 (2010)이 있다.

2절에서는 로지스틱회귀모형에서 설명변수가 두 개일 때 이변량 정규분포에 근거한 로그-밀도비를 알아보고 이를 이용하여 두 설명변수에 추가하여 이차항과 교호작용항이 필요한 조건을 알아본다. 3절에서는 선형회귀모형에 적용되었던 방법의 확장하여 일반화선형모형에서 설명변수 간의 교호작용을 탐색해 볼 수 있는 그래픽적인 방법에 대해 연구하고자 한다. 4절에서는 이러한 방법을 구체적으로 적용시켜본다.

2. 로지스틱 회귀모형에서 로그-밀도비

성공여부를 나타내는 확률변수 y가 성공확률이 θ인 베르누이분포를 따른다고 하자. 반응변수를 y로 설명변수를 $\boldsymbol{x} = (x_1, x_2, ..., x_p)^T$로 하는 이항회귀 (binomial regression)의 모형은 $E(y|\boldsymbol{x}) = \theta(\boldsymbol{x}) = m(\boldsymbol{x}^T\boldsymbol{\beta})$로 표현된다. 이항회귀모형은 일반화선형모형의 한 형태로 $m()$는 커널평균함수 (kernel mean function)이고 연결함수 $g()$의 역함수이다. 커널평균함수로 로지스틱함수 (logistic function)를 사용하는 로지스틱회귀모형은 다음과 같다.

$$E(y|\boldsymbol{x}) = \theta(\boldsymbol{x}) = \frac{exp(\boldsymbol{x}^T\boldsymbol{\beta})}{1 + exp(\boldsymbol{x}^T\boldsymbol{\beta})} = m(\boldsymbol{x}^T\boldsymbol{\beta}) \tag{2.1}$$

모형 (2.1)은 로짓 연결함수를 통하여 다음과 같이 선형모형의 형태가 된다.

$$g(\theta(\boldsymbol{x})) = log\left(\frac{\theta(\boldsymbol{x})}{1 - \theta(\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta}$$

모형 (2.1)에서는 $\boldsymbol{x}$의 선형결합인 $\boldsymbol{x}^T\boldsymbol{\beta}$의 함수로 모형을 구성하고 있으나 Cook과 Weisberg (1999)은 $\boldsymbol{u}$의 선형결합인 $\boldsymbol{\eta}^T\boldsymbol{u}$의 함수를 사용하였다. 여기서 $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$는 p개의 설명변수 $\boldsymbol{x}$로부터 구한 벡터이다. 일반적으로 $\boldsymbol{u}$는 $\boldsymbol{x}$의 함수들로 구성된다. 모형 (2.1)에서와 같이 로지스틱함수를 커널평균함수로 하면 다음과 같이 로지스틱회귀모형이 된다.

$$E(y|\boldsymbol{x}) = \theta(\boldsymbol{x}) = \frac{exp(\boldsymbol{\eta}^T\boldsymbol{u})}{1 + exp(\boldsymbol{\eta}^T\boldsymbol{u})} = m(\boldsymbol{\eta}^T\boldsymbol{u}) \tag{2.2}$$

여기서 $\boldsymbol{\eta}^T\boldsymbol{u}$에 대한 방정식을 풀면, 다음과 같이 쓸 수 있고

$$logit(\theta(\boldsymbol{x})) = log\left(\frac{\theta(\boldsymbol{x})}{1 - \theta(\boldsymbol{x})}\right) = \boldsymbol{\eta}^T\boldsymbol{u}$$

로짓 연결함수를 구성하는 $\theta(\boldsymbol{x})/(1 - \theta(\boldsymbol{x}))$를 성공-오즈라 부른다.

Kay와 Little (1987)은 설명변수의 조건부분포, 즉 $\boldsymbol{x}|y$의 분포에 따라 $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$를 적절하게 선택하는 과정을 제시하였다. Cook과 Weisberg (1999), Scrucca와 Weisberg (2004)에 따르면, 설명변수가 하나이고 그 조건부분포가 정규분포라 하면 $\boldsymbol{u}^T = (1, x, x^2)$를 사용하고 분산이 같을 때에는 $\boldsymbol{u}^T = (1, x)$를 사용한다. 또한 Kahng과 Shin (2012)에서는 조건부분포가 좌우대칭이 아니면 감마분포로 보고 $\boldsymbol{u}^T = (1, x, log(x))$를 사용한다.

Kay와 Little (1987)에서와 같이 회귀 $y|\boldsymbol{x}$와 역회귀 (inverse regression) $\boldsymbol{x}|y$사이의 관계를 알아보자. $f(\boldsymbol{x}|y = j)$를 $y = j$가 주어졌을 때, $\boldsymbol{x}$에 대한 확률밀도함수라 하자. 그리고 $f(\boldsymbol{x})$를 주변확률밀도

함수라 하자. 반응변수가 이항변수이므로 로지스틱회귀에서의 평균함수 $E(y|\boldsymbol{x})$는 베이즈공식을 이용하면 다음과 같이 쓸 수 있다.

$$E(y|\boldsymbol{x}) = \theta(\boldsymbol{x}) = P(y=1|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|y=1)P(y=1)}{f(\boldsymbol{x})} \tag{2.3}$$

식 (2.3)에서 $\boldsymbol{x}$가 주어졌을 때 $y=1$에 대한 확률 $\theta(\boldsymbol{x})$를 평균함수라 말할 수 있다. 또한 $\boldsymbol{x}$가 주어졌을 때 $y=0$에 대한 확률 $1-\theta(\boldsymbol{x})$는 다음과 같이 쓸 수 있다.

$$1-\theta(\boldsymbol{x}) = P(y=0|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|y=0)P(y=0)}{f(\boldsymbol{x})} \tag{2.4}$$

식 (2.3)과 식 (2.4)의 두 값의 로그비를 취하면 다음과 같이 로그-오즈를 얻을 수 있다.

$$\begin{aligned} log\left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) &= log\left(\frac{P(y=1)}{P(y=0)}\right) + log\left(\frac{f(\boldsymbol{x}|y=1)}{f(\boldsymbol{x}|y=0)}\right) \\ &= log\left(\frac{P(y=1)}{P(y=0)}\right) + h(\boldsymbol{x}) \end{aligned}$$

따라서 로그-오즈는 두 항의 합이다. 첫 번째 항은 $\boldsymbol{x}$에 의존하지 않는 주변로그-오즈 (marginal log-odds)이고 두 번째 항 $h(\boldsymbol{x})$는 로그-밀도비 (log-density ratio)라고 한다.

만약 $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$의 몇 가지 변환집합에 대해 $h(\boldsymbol{x}) = \boldsymbol{\eta}^T\boldsymbol{u}$과 같이 쓸 수 있다면, 연결함수가 로짓이 된다. 또한, 커널평균함수가 로지스틱이 되고, 예측변수가 $\boldsymbol{\eta}^T\boldsymbol{u}$가 된다. 그러므로 상대적인 통계 정보는 역회귀의 연구에 의해 추출될 수 있다.

조건부 확률밀도함수 $f(\boldsymbol{x}|y=j)$, $j=0,1$가 평균 $\boldsymbol{\mu}_j$와 분산 $\boldsymbol{\Sigma}_j$를 가지는 정규밀도함수, 즉 $(\boldsymbol{x}|y=j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$라면 로그-밀도비를 다음과 같다.

$$h(\boldsymbol{x}) = log\left(\frac{f(\boldsymbol{x}|y=1)}{f(\boldsymbol{x}|y=0)}\right) = c_0 + \boldsymbol{x}^T\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) + \frac{1}{2}\boldsymbol{x}^T\left(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}\right)\boldsymbol{x} \tag{2.5}$$

여기서 $c_0 = [log(|\boldsymbol{\Sigma}_0|/|\boldsymbol{\Sigma}_1|) + (\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1)]/2$이다.

설명변수가 두 개이고 $\boldsymbol{x} = (x_1, x_2)^T$가 이변량 정규분포를 따른다고 하자. 각각의 이변량 정규분포의 기댓값벡터는 $\boldsymbol{\mu}_j = (\mu_{j1}, \mu_{j2})^T$이고 분산-공분산행렬 $\boldsymbol{\Sigma}_j$은 다음과 같이 표현할 수 있다.

$$\boldsymbol{\Sigma}_j = \begin{pmatrix} \sigma_{j1}^2 & \rho_j\sigma_{j1}\sigma_{j2} \\ \rho_j\sigma_{j1}\sigma_{j2} & \sigma_{j2}^2 \end{pmatrix}$$

식 (2.5)에서, $h(\boldsymbol{x})$는 세 개의 항의 합으로 구성되어 있는데 그 중 두 항만 설명변수 $\boldsymbol{x}$의 영향을 받는다. 따라서 $\boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)$에서 일차항인 x_1과 x_2의 포함여부를 파악 할 수 있고 $\boldsymbol{x}^T(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1})\boldsymbol{x}/2$에서 이차항과 교호작용항인 x_1^2, x_2^2, x_1x_2의 포함여부를 파악 할 수 있다.

식 (2.5)의 우변의 마지막 항인 $\boldsymbol{x}^T(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1})\boldsymbol{x}/2$를 보면 이차항인 x_1^2, x_2^2, 교호작용항인 x_1x_2의 포함 여부를 알 수 있다. $\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1} = \{a_{ij}\}, (i,j=1,2)$라 하면 이차항과 교호작용항의 식을 다음과 같이 표현할 수 있고 (Kahng과 Yoon, 2013),

$$\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1})\boldsymbol{x} = \frac{1}{2}(x_1x_2)\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{2}a_{11}x_1^2 + \frac{1}{2}a_{22}x_2^2 + a_{12}x_1x_2$$

이 식을 정리하면 a_{11}, a_{22}, a_{12}는 다음과 같다.

$$\begin{aligned} a_{11} &= \left\{(1-\rho_1^2)\sigma_{11}^2 - (1-\rho_0^2)\sigma_{01}^2\right\}/(1-\rho_0^2)(1-\rho_1^2)\sigma_{01}^2\sigma_{11}^2 \\ a_{22} &= \left\{(1-\rho_1^2)\sigma_{12}^2 - (1-\rho_0^2)\sigma_{02}^2\right\}/(1-\rho_0^2)(1-\rho_1^2)\sigma_{02}^2\sigma_{12}^2 \\ a_{12} &= \left\{(1-\rho_0^2)\rho_1\sigma_{01}\sigma_{02} - (1-\rho_1^2)\rho_0\sigma_{11}\sigma_{12}\right\}/(1-\rho_0^2)(1-\rho_1^2)\sigma_{01}\sigma_{11}\sigma_{02}\sigma_{12} \end{aligned} \tag{2.6}$$

만약 $y=0$과 $y=1$인 경우 x_1의 두 분산이 같고 x_2의 두 분산도 같으며 x_1과 x_2의 두 상관계수까지 동일하면 두 개의 분산-공분산행렬이 같다. 즉 $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$이 된다. 이 경우 이차항 x_1^2, x_2^2과 교호작용항인 x_1x_2는 밀도비에 포함되지 않는다. 따라서 x_1과 x_2만으로 로지스틱회귀모형을 구성하면 된다.

하지만 $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$이 아닌 경우에도 이차항과 교호작용항의 일부나 모두가 필요하지 않을 수도 있다. 우선 두 상관계수의 제곱이 같은 경우를 생각해 보자. $\rho_0^2 = \rho_1^2 = 0$이면 식 (2.6)에서 분자가 모두 0이 되고 이차항과 교호작용항은 밀도비에 포함되지 않는다. 또한 $\sigma_{11}^2 = \sigma_{01}^2$이면 $a_{11} = 0$이고 이차항 x_1^2은 밀도비에 포함되지 않는다. $\sigma_{12}^2 = \sigma_{02}^2$이면 $a_{22} = 0$이고 x_2^2은 밀도비에 포함되지 않는다. 만약 $\rho_0^2 = \rho_1^2$이고 그 값이 0과 1 사이에 있으면 $\sigma_{01}^2 = \sigma_{11}^2$이면 $a_{11} = 0$이고 $\sigma_{02}^2 = \sigma_{12}^2$이면 $a_{22} = 0$이 된다. 또한 $\sigma_{01}\sigma_{02} = \sigma_{11}\sigma_{12}$인 경우에 $a_{12} = 0$이 된다. Kahng과 Yoon (2013)에 의하면 이러한 조건은 $y=0$과 $y=1$에서 x_1와 x_2의 산점도 두 개의 비교를 통하여 파악이 가능하다. 하지만 이러한 산점도를 통한 방법은 이차항에 대한 판단은 용이하지만 교호작용항의 필요성 여부를 파악하기는 수월하지 않다. 다음절에서는 교호작용을 판단할 수 있는 대안을 알아보고자 한다.

3. 3차원 잔차산점도를 이용한 교호작용의 검토

로지스틱모형을 포함하는 일반화선형모형에서 선형결합 $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$는 반응변수 y의 기댓값 $\theta = E(y)$과 연결함수에 의하여 $g(\theta) = \eta$와 같이 연결된다. 설명변수가 2개인 경우 교호작용의 효과를 알아보기 위하여 설명변수벡터 $\boldsymbol{x}^T = (1, x_1, x_2)$를 $\boldsymbol{x}_1^T = (1, x_1)$와 x_2로 분할하면 일반화선형모형은 다음과 같이 표현된다.

$$g(\theta) = \boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_1^T\boldsymbol{\beta}_1 + \beta_2 x_2 \tag{3.1}$$

모형 (3.1)에서 $g(\theta) = \theta$인 선형모형의 경우 Cook과 Weisberg (1989)는 교호작용이 존재하는 경우 잔차 e_i를 수직축으로 하고 $\widehat{y}_{i2.1}$와 $\widehat{y}_{i1}$를 서로 직각을 이루는 수평축으로 하는 3차원 잔차산점도 $\{e_i, \widehat{y}_{i2.1}, \widehat{y}_{i1}\}$의 수직축을 회전시키면 안장형 (saddle-shaped)이나 U자형 (U-shaped)으로 나타난다고 하였다. 여기서 $\widehat{y}_i$와 $\widehat{y}_{i1}$는 각각 $\boldsymbol{x}^T$와 $\boldsymbol{x}_1^T$를 설명변수로 할 때의 적합값이고 $\widehat{y}_{i2.1} = \widehat{y}_i - \widehat{y}_{i1}$이다.

Kahng (2005)에서와 같이 모형 (3.1)의 대안으로 다음의 모형을 생각하고 3차원 잔차산점도가 교호작용을 나타내는 과정을 보자.

$$g(\theta) = \boldsymbol{x}_1^T\boldsymbol{\beta}_1 + \beta_2(\boldsymbol{x}_1)\widetilde{x}_2 \tag{3.2}$$

여기서 $\boldsymbol{x}_i^T$, $\boldsymbol{x}_{i1}^T$, x_{i2}를 각각 $\boldsymbol{x}^T$, $\boldsymbol{x}_1^T$, x_2의 i번째 관측값이라 하고 $\boldsymbol{X}$와 $\boldsymbol{X}_1$는 각각 $\boldsymbol{x}_i^T$와 $\boldsymbol{x}_{i1}^T$를 행으로 하는 행렬이고 $\boldsymbol{x}_2$는 x_{i2}를 원소로 하는 벡터라 하자. $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T$라 하면 $\widetilde{\boldsymbol{x}}_2 = \{\widetilde{x}_{i2}\} = (\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{x}_2$는 $\boldsymbol{x}_2$에서 $\boldsymbol{X}_1$의 영향을 제거한 순수한 $\boldsymbol{x}_2$의 영향력이라고 할 수 있다. 모형 (3.2)는 $\beta_2(\boldsymbol{x}_1)$이 $\boldsymbol{x}_1$에 의존한다는 점에서 모형 (3.1)과는 차이가 있다.

모형 (3.2)의 $\beta_2(\boldsymbol{x}_1)$를 $\boldsymbol{x}_1$에 대하여 Taylor 급수전개를 하고 $\boldsymbol{x}_1 = \boldsymbol{0}$ 근방에서 2차이상 미분 부분을 무시할 수 있다면 $\beta_2(\boldsymbol{x}_1) \approx \beta_2 + \boldsymbol{x}_1^T\dot{\beta}_2$와 같이 선형화 된다. 여기서 $\beta_2 = \beta_2(\boldsymbol{0})$이고 $\dot{\beta}_2$는 $\boldsymbol{x}_1$에 대한

$\beta_2(\boldsymbol{x}_1)$의 일차미분벡터를 $\boldsymbol{x}_1 = \boldsymbol{0}$에서 계산한 값이다. 이러한 선형화를 모형 (3.2)에 적용하면 다음과 같다.

$$g(\theta) \approx \boldsymbol{x}_1^T\boldsymbol{\beta}_1 + \beta_2\widetilde{x}_2 + \widetilde{x}_2\boldsymbol{x}_1^T\dot{\beta}_2 \tag{3.3}$$

모형 (3.3)는 $\boldsymbol{x}_1$과 x_2만 아니라 교호작용항이 포함된 것으로 볼 수 있다. 이 모형의 양변에 $(\boldsymbol{I}-\boldsymbol{P})$를 곱한 후 기대값을 취하면 다음과 같다.

$$\begin{aligned} E(\boldsymbol{e}) &= (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{D}(\widetilde{\boldsymbol{x}}_2)\boldsymbol{X}_1\dot{\beta}_2 \\ &= \boldsymbol{D}(\widetilde{\boldsymbol{x}}_2)\boldsymbol{X}_1\dot{\beta}_2 - \boldsymbol{P}\boldsymbol{D}(\widetilde{\boldsymbol{x}}_2)\boldsymbol{X}_1\dot{\beta}_2 \end{aligned} \tag{3.4}$$

여기서 $\boldsymbol{D}(\widetilde{\boldsymbol{x}}_2) = diag\{\ \widetilde{x}_{i2}\}$이고 $\boldsymbol{e}$는 모형 (3.1)에 적합하여 얻어진 잔차벡터이다. 식 (3.4)에서 두 번째 항은 매우 작아지며 따라서 $E(\boldsymbol{e}) \approx \boldsymbol{D}(\widetilde{\boldsymbol{x}}_2)\boldsymbol{X}_1\dot{\beta}_2$이 된다.

모형 (3.2)의 특별한 경우로 $\beta_2(\boldsymbol{x}_1) = \beta_2(\boldsymbol{x}_1^T\boldsymbol{\beta}_1)$를 생각해 보자. 즉 교호작용은 선형결합 $\boldsymbol{x}_1^T\boldsymbol{\beta}_1$에 의존한다고 하자. 이때 모형 (3.2)는 다음과 같다.

$$g(\theta) = \boldsymbol{x}_1^T\boldsymbol{\beta}_1 + \beta_2(\boldsymbol{x}_1^T\boldsymbol{\beta}_1)\widetilde{\boldsymbol{x}}_2 \tag{3.5}$$

이제 $\eta_1 = \boldsymbol{x}_1^T\boldsymbol{\beta}_1$라 하고 $\beta_2(\boldsymbol{x}_1^T\boldsymbol{\beta}_1)$를 Taylor 급수전개에 의하여 전개하고 $\eta_1 = 0$근방의 η_1에 대하여 2차 이상 미분한 부분을 무시할 수 있다면 $\beta_2(\boldsymbol{x}_1^T\boldsymbol{\beta}_1) = \beta_2(\eta_1) \approx \beta_2 + \widetilde{\beta}_2\eta_1$과 같이 선형화 할 수 있다. 여기서 $\tilde{\beta}_2$는 $\eta_1 = 0$에서 계산된 $\tilde{\beta}_2 = \partial\beta_2(\eta_1)/\partial\eta_1$의 값이다. 이러한 선형화를 모형 (3.5)에 적용하면 다음과 같다.

$$g(\theta) = \boldsymbol{x}_1^T\boldsymbol{\beta}_1 + \beta_2\widetilde{\boldsymbol{x}}_2 + \widetilde{\beta}_2\widetilde{\boldsymbol{x}}_2\boldsymbol{x}_1^T\boldsymbol{\beta}_1 \tag{3.6}$$

모형 (3.6)의 양변에 $(\boldsymbol{I}-\boldsymbol{P})$를 곱한 후 기대값을 취하면 다음과 같다.

$$\begin{aligned} E(\boldsymbol{e}) &= (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{D}(\widetilde{\boldsymbol{x}}_2)\boldsymbol{X}_1\boldsymbol{\beta}_1 \\ &= \widetilde{\beta}_2\boldsymbol{D}(\widetilde{\boldsymbol{x}}_2)\boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{P}\boldsymbol{D}(\widetilde{\boldsymbol{x}}_2)\boldsymbol{X}_1\boldsymbol{\beta}_1 \\ &\approx \widetilde{\beta}_2\boldsymbol{D}(\widetilde{\boldsymbol{x}}_2)\boldsymbol{X}_1\boldsymbol{\beta}_1 \end{aligned} \tag{3.7}$$

식 (3.7)에서 $\widetilde{\beta}_2\boldsymbol{D}(\widetilde{\boldsymbol{x}}_2)\boldsymbol{X}_1\boldsymbol{\beta}_1$를 $d_1 = \tilde{\beta}_2\boldsymbol{D}(\tilde{\boldsymbol{x}}_2)$과 $d_2 = \boldsymbol{X}_1\boldsymbol{\beta}_1$로 분해해보면 3차원산점도 $\{e_i, \widehat{\eta}_{i2.1}, \widehat{\eta}_{i1}\}$는 3차원산점도 $\{d_1d_2, d_1, d_2\}$의 형태가 됨을 알 수 있다. 따라서 $\boldsymbol{x}_1$과 $\boldsymbol{x}_2$의 교호작용의 효과가 존재하면 3차원 잔차산점도는 안장형 또는 U자형을 나타낼 것이다.

4. 예 제

설명변수 x_1은 $U(0, 0.1)$에서 101개의 값을 생성시키고 x_2는 0부터 1까지 0.01의 간격으로 얻어진 101개를 관찰값으로 한다. 회귀계수를 β_0=3, β_1=-5, β_2=-1, β_{12}=10으로 하고 모형 $\eta = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{12}x_1x_2$에 적용시켜 101개의 η를 계산하였다. 이때 식 (2.1)과 (2.2)를 이용하면 $\theta(\boldsymbol{x})$는 101개의 비율이 된다. $\theta(\boldsymbol{x})$를 특정사건이 발생할 확률로 하고 y는 특정사건의 발생 여부로 하는 베르누이 분포에 의해 생성하였다. 따라서 반응변수 y는 성공확률이 $\theta(\boldsymbol{x})$인 베르누이분포를 따른다고 할 수 있다. 모형 $\eta = \beta_0 + \beta_1x_1 + \beta_2x_2$에 대한 3차원 잔차산점도를 그리고 수직축을 중심으로 회전시키면 Figure 4.1과 같이 안장모양을 나타내고 있으며 교호작용의 효과가 있음을 확인할 수 있다. 즉 모형에 교호작용항을 추가되어야 함을 보여주고 있다. 이번에는 교호작용 효과가 없는 자료를 생성하기 위하여

$\beta_{12} = 0$으로 하여 $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$에 적용시켜 101개의 η를 계산하고 앞에서와 같은 방법으로 반응변수 y를 생성하였다. 3차원 잔차산점도를 그리고 수직축을 중심으로 회전시키면 Figure 4.2와 같이 안장모양이나 U자 형태가 나타나지 않는다.

이러한 방법으로 교호작용의 효과가 존재할 것으로 예상되는 자료와 존재하지 않을 것으로 예상되는 자료의 생성을 여러 번 수행하여 실험한 결과 앞에서의 두 가지 예에 나타난 것과 유사한 결과를 얻을 수 있었다.

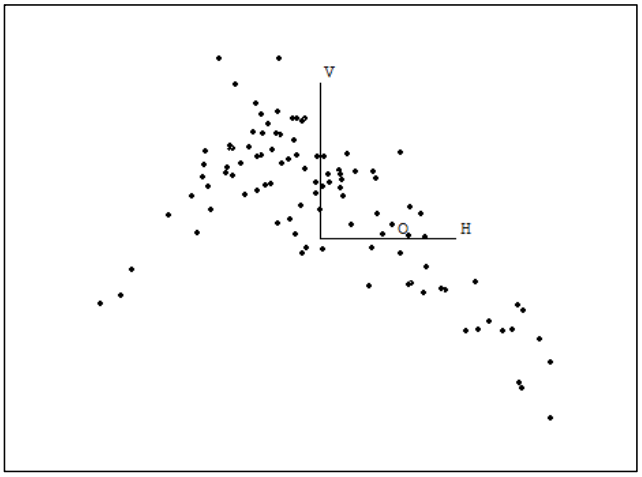

Figure 4.1 3-D residual plot of simulated data with interaction

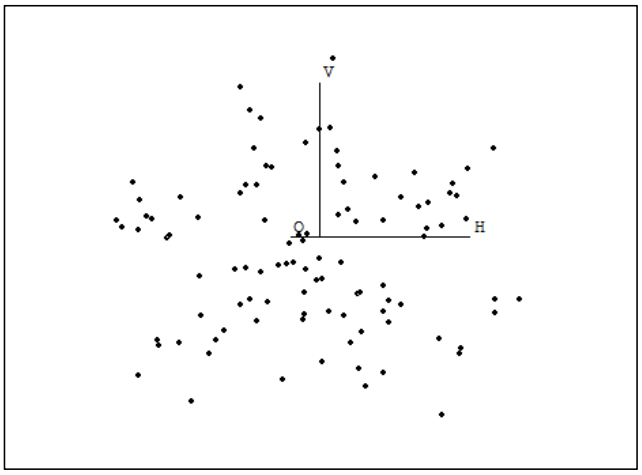

Figure 4.2 3-D residual plot of simulated data without interaction

다음은 Cook과 Weisberg (1994)에 제시된 오스트레일리아 스포츠 선수촌에서 훈련하는 운동선수 202명의 신체지수와 혈액검사 자료를 분석한다. 이 자료의 변수들은 성별 (Sex: 0=male, 1=female),

몸무게 (Wt), 키 (Ht), 적혈구수치 (RCC), 헤모글로빈수치 (Hg) 등이 있으며, 이 중 0과 1의 값을 갖는 성별을 반응변수로 사용하고, 여러 변수 중 Wt와 RCC를 설명변수로 사용한다.

이 자료에 대한 3차원 잔차산점도에서 수직축을 중심으로 회전시켜 보면 Figure 4.3과 같이 안장모양이나 U자 형태가 나타나지 않는다. 따라서 교호작용 항의 추가가 필요하지 않다고 할 수 있다. 실제로 이 자료에 로지스틱 모형을 적합 시킨 결과 교호작용항 $Wt \cdot RCC$의 p-값은 크게 나와 유의하지 않게 나타나고 이차항과 교호작용항 없이 일차항만으로도 충분한 설명력이 있다는 것을 알 수 있다.

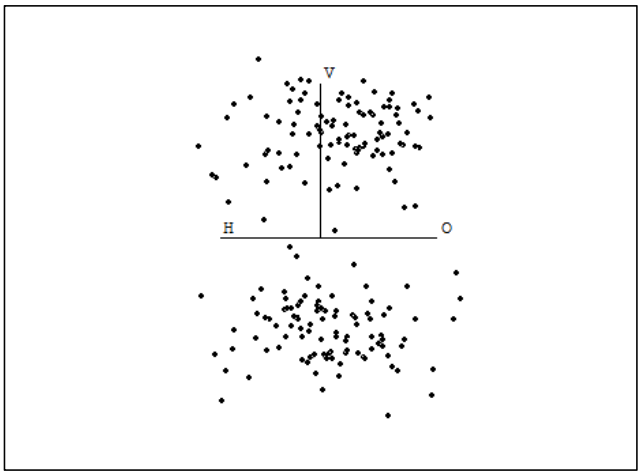

Figure 4.3 3-D residual plot of Australian athletes data

5. 결 론

설명변수가 두 개일 때 선형모형에서와 같이 로지스틱회귀모형에서도 일반적으로는 두 설명변수만 모형에 포함시킨다. 하지만 설명변수만으로는 충분히 설명이 되지 못하고 설명변수의 변환된 형태인 이차항 또는 교호작용항이 필요한 경우가 있다. 설명변수의 조건부 분포가 이변량 정규분포를 따르는 경우 두 분포의 분산과 상관계수에 따라 이차항과 교호작용 항이 필요하지 않게 되는데 분산이나 상관계수에 대한 정보는 그래프를 보고 대체적인 판단이 가능하다. 하지만 교호작용 항은 그래프에서 파악하기가 수월하지 않은 경우가 많다.

본 논문에서는 교호작용의 필요성을 판단할 수 있는 대안으로 3차원 잔차산점도를 제안하였는데 이를 통하여 로지스틱회귀모형에서 설명변수 간의 교호작용의 필요성의 탐색이 가능하였다.

References

Cook, R. D. and Weisberg, S. (1994). *An introduction to regression graphics*, Wiley, New York.

Cook, R. D. and Weisberg, S. (1999). *Applied regression including computing and graphics*, Wiley, New York.

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

Kahng, M. (2005). Exploring interaction in generalized linear models. *Journal of the Korean Data & Information Society*, **16**, 13-18.

Kahng, M., Kim, B. and Hong, J. (2010). Graphical regression and model assessment in logistic model. *Journal of the Korean Data & Information Society*, **21**, 21-32.

Kahng, M. and Shin, E. (2012). A study on log-density with log-odds graph for variable selection in logistic regression. *Journal of the Korean Data & Information Society*, **23**, 99-111.

Kahng, M. and Yoon, J. E. (2013). Log-density ratio with two predictors in logistic regression model. *The Korean Journal of Applied Statistics*, **26**, 141-149.

Kay, R. and Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, **74**, 495-501.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of Royal Statistical Society A*, **135**, 370-384.

Scrucca, L. and Weisberg, S. (2004). A simulation study to investigate the behavior of the log-density ratio under normality. *Communication in Statistics Simulation and Computation*, **33**, 159-178.

Journal of the Korean Data & Information Science Society
2014, **25**(1), 177–185

http://dx.doi.org/10.7465/jkdi.2014.25.1.177
한국데이터정보과학회지

Exploring interaction using 3-D residual plots in logistic regression model[†]

Myung-Wook Kahng[1]

[1]Department of Statistics, Sookmyung Women's University

Received 20 December 2013, revised 6 January 2014, accepted 12 January 2014

Abstract

Under bivariate normal distribution assumptions, the interaction and quadratic terms are needed in the logistic regression model with two predictors. However, depending on the correlation coefficient and the variances of two conditional distributions, the interaction and quadratic terms may not be necessary. Although the need for these terms can be determined by comparing the two scatter plots, it is not as useful for interaction terms. We explore the structure and usefulness of the 3-D residual plot as a tool for dealing with interaction in logistic regression models. If predictors have an interaction effect, a 3-D residual plot can show the effect. This is illustrated by simulated and real data.

Keywords: Binary regression, interaction, inverse regression, log-density ratio, logistic regression model, 3-D residual plot.

[†] This research was supported by the Sookmyung Women's University Research Grants 2012.
[1] Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea. E-mail: mwkahng@sm.ac.kr