The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.1.141
26(1), 141–149

Log-density Ratio with Two Predictors in a Logistic Regression Model

Myung Wook Kahng[a,1] · Jae Eun Yoon[a]

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

(Received November 27, 2012; Revised December 31, 2012; Accepted January 18, 2013)

Abstract

We present methods for studying the log-density ratio that enables the selection of the predictors and the form to be included in the logistic regression model. Under bivariate normal distributional assumptions, we investigate the form of the log-density ratio as a function of two predictors. If two covariance matrices are equal, then the crossproduct and quadratic terms are not needed. If the variables are uncorrelated, we do not need the crossproduct terms, but we still need the linear and quadratic terms. We also explore other conditions in which the crossproduct and quadratic terms are not needed in the logistic regression model.

Keywords: Binary regression, inverse regression, kernel mean function, log-density ratio, log-odds, logistic regression.

1. 서론

일반적으로 회귀분석에서 반응변수는 특정한 구간 안에 있는 값이라면 어느 값이라도 취할 수 있는 연속형 자료이다. 그러나 회귀분석을 수행하는데 있어서 반응변수가 범주형인 경우를 흔히 볼 수 있다. 이 때에는 오차가 정규분포를 따른다는 가정을 할 수 없기 때문에 일반적인 정규선형모형(normal linear model)을 사용하는데 무리가 있다. 이것을 해결해 주는 방법 중 가장 일반적인 것이 로지스틱회귀모형(logistic regression model)이다.

Nelder와 Wedderburn (1972)이 제안한 일반화선형모형(generalized linear model)은 정규이론을 따르는 선형모형을 지수족(exponential family)과 연결함수(link function)를 이용하여 다음과 같은 두 가지 과정으로 일반화 될 수 있다. 첫째, 반응변수의 기댓값과 설명변수의 선형결합(linear predictor)을 연결시키는 연결함수를 설정한다. 둘째, 오차의 분포는 정규분포를 포함하는 지수족의 여러 가지 분포를 사용한다.

로지스틱회귀모형은 확률변수 y가 시행횟수가 m이고 성공확률이 θ인 이항분포를 따르는 경우 y/m를 반응변수로 하고 이것의 기댓값 θ와 주어진 설명변수들의 선형결합 $\boldsymbol{x}'\boldsymbol{\beta}$를 로짓(logit) 연결함수로 이어주는 일반화선형모형의 한 형태이다.

This research was supported by the Sookmyung Women's University Research Grants 2011.
[1]Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Yongsan-Gu, Seoul 142-742, Korea. E-mail: mwkahng@sookmyung.ac.kr

선형회귀모형에서 반응변수의 기댓값은 설명변수들의 선형결합 $\boldsymbol{x}'\boldsymbol{\beta}$이라고 가정한다. 로지스틱회귀모형에서도 역시 설명변수들의 선형결합이 이용된다. 하지만 설명변수의 선형결합만으로는 충분히 설명이 되지 못하고 설명변수의 변환된 형태 등의 추가적인 포함이 필요한 경우가 있다. 이러한 연구는 로그-밀도비(log-density ratio)를 근거로 하여 Kay와 Little (1987)에 의하여 시작되었고 설명변수가 하나인 경우 Scrucca (2003), Kahng과 Shin (2012)의 연구가 있다.

본 논문에서는 설명변수가 두 개일 때 설명변수의 선형결합만으로 반응변수를 충분히 설명할 수 있는지 아니면 추가적으로 설명변수의 변환된 요소가 필요한지를 알아보고자 한다. 2절에서는 성공-오즈(odds of success)에 로그를 취한 로그-오즈(log-odds)와 로그-밀도비에 대해 알아본다. 3절에서는 로지스틱회귀모형에서 설명변수가 두 개일 때 이변량 정규분포에 근거한 로그-밀도비를 알아보고 이를 이용하여 두 설명변수에 추가하여 이차항과 교차항이 필요한 조건을 알아본다. 4절에서는 실제 자료를 이용하여 3절에서 알아본 조건을 확인해본다.

2. 로지스틱회귀모형에서 로그-밀도비

성공횟수를 나타내는 확률변수 y가 시행횟수가 m이고 성공확률이 θ인 이항분포를 따른다고 하자. 반응변수를 성공비율인 y/m로 설명변수를 $\boldsymbol{x} = (x_1, \ldots, x_p)'$로 하는 이항회귀(binomial regression)의 모형은 $E(y/m) = \theta(\boldsymbol{x}) = g(\boldsymbol{x}'\boldsymbol{\beta})$로 표현된다.

이항회귀모형은 일반화선형모형의 한 형태로 $g(\,)$는 커널평균함수(kernel mean function)이고 연결함수의 역함수이다. 커널평균함수로 로지스틱함수(logistic function)를 사용하는 다음의 일반적인 로지스틱회귀모형을 생각하자.

$$E\left(\frac{y}{m}\Big|\boldsymbol{x}\right) = \theta(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})} = g(\boldsymbol{x}'\boldsymbol{\beta}). \tag{2.1}$$

모형 (2.1)은 로짓 연결함수를 통하여 다음과 같이 선형모형의 형태로 표현된다.

$$\text{logit}\left(\theta(\boldsymbol{x})\right) = \log\left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) = \boldsymbol{x}'\boldsymbol{\beta},$$

여기서 $\theta(\boldsymbol{x})/(1-\theta(\boldsymbol{x}))$를 성공-오즈라 부른다. 이제 성공할 경우 "1"을, 실패할 경우 "0"을 가지는 반응변수 y를 생각하자. y의 조건부분포는 기댓값이 $E(y|\boldsymbol{x}) = P(y=1|\boldsymbol{x}) = \theta(\boldsymbol{x})$가 되는 베르누이분포를 따른다. 이항회귀모형은 다음과 같이 쓸 수 있다.

$$E(y|\boldsymbol{x}) = g(\boldsymbol{\eta}'\boldsymbol{u}). \tag{2.2}$$

모형 (2.2)에서의 커널함수는 모형 (2.1)에서와 같은 $\boldsymbol{x}$의 선형결합인 $\boldsymbol{x}'\boldsymbol{\beta}$가 아닌 $\boldsymbol{u}$의 선형결합인 $\boldsymbol{\eta}'\boldsymbol{u}$의 함수이다. $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$는 p개의 설명변수 $\boldsymbol{x}$로부터 구한 벡터이다. 일반적으로 $\boldsymbol{u}$는 $\boldsymbol{x}$의 함수로 구성된다. 모형 (2.1)에서와 같이 로지스틱함수를 커널평균함수로 하면 모형 (2.2)는 다음과 같이 로지스틱회귀모형이 된다.

$$E(y|\boldsymbol{x}) = \theta(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\eta}'\boldsymbol{u})}{1+\exp(\boldsymbol{\eta}'\boldsymbol{u})} = g(\boldsymbol{\eta}'\boldsymbol{u}),$$

여기서 $\boldsymbol{\eta}'\boldsymbol{u}$에 대한 방정식을 풀면, 다음과 같이 쓸 수 있다.

$$\text{logit}\left(\theta(\boldsymbol{x})\right) = \log\left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) = \boldsymbol{\eta}'\boldsymbol{u}.$$

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)에 따르면, 설명변수가 하나이고 그 조건부분포가 정규분포라 하면 $\boldsymbol{u} = (1, x, x^2)'$를 사용하고 분산이 같을 때에는 $\boldsymbol{u} = (1, x)'$를 사용한다. 또한 Kahng과 Shin (2012)은 조건부분포가 좌우대칭이 아니면 감마분포로 보고 $\boldsymbol{u} = (1, x, \log(x))'$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 주변확률밀도함수라 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 평균함수라고 할 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-밀도비라고 한다.

만약 $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$의 몇 가지 변환집합에 대해 $h(\boldsymbol{x}) = \boldsymbol{\eta}'\boldsymbol{u}$과 같이 쓸 수 있다면, 연결함수가 로짓이 된다. 또한, 커널평균함수가 로지스틱이 되고, 예측변수가 $\boldsymbol{\eta}'\boldsymbol{u}$가 된다.

3. 이변량 정규분포에서의 로그-밀도비

일반적으로 회귀모형에서 설명변수에 대한 분포의 가정을 하지 않는다. 하지만 반응변수가 0으로 주어졌을 때 두 설명변수의 형태가 반응변수가 1로 주어졌을 때 두 설명변수의 형태와 유사한지 다른 지를 알아보기 위하여 두 형태를 모두 이변량 정규분포라고 보고 정규분포의 기댓값, 분산, 상관계수를 검토하여 차이가 나는지를 알아보고자한다. 이렇게 유사성 여부를 판단하고 포함되는 설명변수의 형태를 알아보고자 Cook과 Weisberg (1999), Scrucca (2003), Scrucca와 Weisberg (2004)는 조건부분포를 정규분포로 간주하였다. 좌우대칭이 아닌 경우 Kahng과 Shin (2012)에서는 감마분포를 사용하고 있다.

만약 $\boldsymbol{x} = (x_1, \ldots, x_p)'$가 $p \times 1$벡터이고 $f(\boldsymbol{x}|y = j)$, $j = 0, 1$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-밀도비 $h(\boldsymbol{x})$를 다음과 같이 쓸 수 있다.

$$h(\boldsymbol{x}) = \log\left(\frac{\frac{\left|\boldsymbol{\Sigma}_1^{-1}\right|^{\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)\right)}{\frac{\left|\boldsymbol{\Sigma}_0^{-1}\right|^{\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_0)\right)}\right)$$

$$= \frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|}\right) + \frac{1}{2}\left(\boldsymbol{\mu}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1\right)$$
$$+ \boldsymbol{x}'\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) + \frac{1}{2}\boldsymbol{x}'\left(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}\right)\boldsymbol{x}. \tag{3.1}$$

설명변수가 두 개이고 $\boldsymbol{x} = (x_1, x_2)'$가 이변량 정규분포를 따른다고 하자. 각각의 이변량 정규분포의 기댓값벡터는 $\boldsymbol{\mu}_0 = (\mu_{01}, \mu_{02})'$, $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12})'$이고 분산-공분산행렬 $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$은 다음과 같이 정의할 수 있다.

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \sigma_{01}^2 & \rho_0\sigma_{01}\sigma_{02} \\ \rho_0\sigma_{01}\sigma_{02} & \sigma_{02}^2 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_1 = \begin{pmatrix} \sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{12} \\ \rho_1\sigma_{11}\sigma_{12} & \sigma_{12}^2 \end{pmatrix}.$$

식 (3.1)에서, $h(\boldsymbol{x})$는 네 개의 항의 합으로 구성되어 있는데 그 중 두 항만 설명변수 $\boldsymbol{x}$의 영향을 받는다. 따라서 $\boldsymbol{x}'(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)$에서 일차항인 x_1과 x_2의 포함여부를 파악 할 수 있고 $\boldsymbol{x}'(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1})\boldsymbol{x}/2$에서 이차항과 교차항인 x_1^2, x_2^2, x_1x_2의 포함여부를 파악 할 수 있다.

이차항인 x_1^2, x_2^2, 교차항인 x_1x_2의 포함 여부를 알아보기 위하여 $\boldsymbol{x}'(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1})\boldsymbol{x}/2$에 대해 살펴보겠다. 분산-공분산행렬 $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$의 역행렬 $\boldsymbol{\Sigma}_0^{-1}$, $\boldsymbol{\Sigma}_1^{-1}$은 다음과 같다.

$$\boldsymbol{\Sigma}_0^{-1} = \frac{1}{(1-\rho_0^2)\sigma_{01}^2\sigma_{02}^2}\begin{pmatrix} \sigma_{02}^2 & -\rho_0\sigma_{01}\sigma_{02} \\ -\rho_0\sigma_{01}\sigma_{02} & \sigma_{01}^2 \end{pmatrix},$$
$$\boldsymbol{\Sigma}_1^{-1} = \frac{1}{(1-\rho_1^2)\sigma_{11}^2\sigma_{12}^2}\begin{pmatrix} \sigma_{12}^2 & -\rho_1\sigma_{11}\sigma_{12} \\ -\rho_1\sigma_{11}\sigma_{12} & \sigma_{11}^2 \end{pmatrix}, \tag{3.2}$$

여기서 $\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1} = \{a_{ij}\}$, $(i, j = 1, 2)$라 하면 이차항과 교차항은 다음과 같이 표현할 수 있다.

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

따라서 식 (3.2)을 이용하면 a_{11}, a_{22}, a_{12}는 다음과 같다.

$$a_{11} = \frac{\left(1-\rho_1^2\right)\sigma_{11}^2 - \left(1-\rho_0^2\right)\sigma_{01}^2}{\left(1-\rho_0^2\right)\left(1-\rho_1^2\right)\sigma_{01}^2\sigma_{11}^2}, \tag{3.3}$$

$$a_{22} = \frac{\left(1-\rho_1^2\right)\sigma_{12}^2 - \left(1-\rho_0^2\right)\sigma_{02}^2}{\left(1-\rho_0^2\right)\left(1-\rho_1^2\right)\sigma_{02}^2\sigma_{12}^2}, \tag{3.4}$$

$$a_{12} = \frac{\left(1-\rho_0^2\right)\rho_1\sigma_{01}\sigma_{02} - \left(1-\rho_1^2\right)\rho_0\sigma_{11}\sigma_{22}}{\left(1-\rho_0^2\right)\left(1-\rho_1^2\right)\sigma_{01}\sigma_{11}\sigma_{02}\sigma_{12}}. \tag{3.5}$$

만약 $y = 0$과 $y = 1$인 경우 x_1의 두 분산이 같고 x_2의 두 분산도 같으며 x_1과 x_2의 두 상관계수가 동일하면 두 개의 분산-공분산행렬이 같다. 즉 $\sigma_{01}^2 = \sigma_{11}^2$, $\sigma_{02}^2 = \sigma_{12}^2$, $\rho_0^2 = \rho_1^2$이면 $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$이 된다. 이 경우 이차항 x_1^2, x_2^2과 교차항인 x_1x_2는 밀도비에 포함되지 않는다. 따라서 일차항인 x_1과 x_2만으로 로지스틱회귀모형을 구성하면 된다.

하지만 $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$이 아닌 경우에도 이차항과 교차항의 일부나 모두가 필요하지 않을 수도 있다. 우선 두 상관계수의 제곱이 같은 경우를 생각해 보자. $\rho_0^2 = \rho_1^2 = 0$이면 식 (3.5)에서 분자는 0이 되고 교차항은 밀도비에 포함되지 않는다. 또한 $\sigma_{01}^2 = \sigma_{11}^2$이면 $a_{11} = 0$이고 이차항 x_1^2은 밀도비에 포함되지 않는다. $\sigma_{02}^2 = \sigma_{12}^2$이면 $a_{22} = 0$이고 x_2^2은 밀도비에 포함되지 않는다. 만약 $\rho_0^2 = \rho_1^2$이고 그 값이 0과 1 사이에

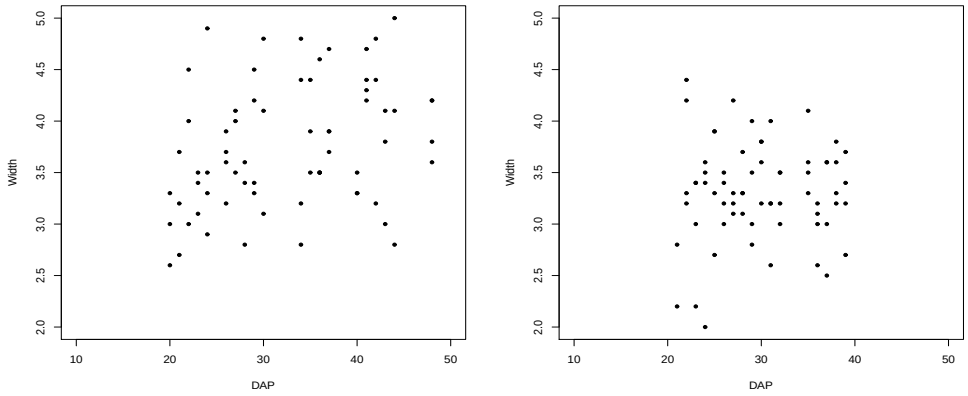

Figure 4.1. Scatter plots of DAP and Width (Year = 0, Year = 1)

Table 4.1. Fit of logistic regression for leaf areadata

Coefficients	Estimate	Std.Error	Chi-Square	Pr>ChiSq
Intercept	18.6429	7.8247	5.6767	0.0172
Width	0.3375	3.0123	0.0126	0.9108
DAP	−1.5334	0.4505	11.5866	0.0007
Width∗Width	−0.1740	0.4618	0.1419	0.7064
DAP∗DAP	0.0201	0.00592	11.4923	0.0007
Width∗DAP	0.0892	0.0678	1.7266	0.1888

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.

이제 상관계수의 제곱이 다른 경우를 생각해 보자. $\rho_0^2 \neq \rho_1^2$인 경우에도 식 (3.3), (3.4), (3.5)의 분자가 0이 되는 조건을 만족하면 이차항 또는 교차항이 밀도비에 포함되지 않을 수 있다. $\sigma_{01}^2/\sigma_{11}^2 = (1-\rho_1^2)/(1-\rho_0^2)$이면 $a_{11} = 0$이고 이차항 x_1^2은 밀도비에 포함되지 않는다. $\sigma_{02}^2/\sigma_{12}^2 = (1-\rho_1^2)/(1-\rho_0^2)$이면 $a_{22} = 0$이고 이차항 x_2^2은 밀도비에 포함되지 않는다. 또한 $(\sigma_{01}/\sigma_{11})(\sigma_{02}/\sigma_{12}) = (\rho_0/\rho_1)(1-\rho_1^2)/(1-\rho_0^2)$이면 $a_{12} = 0$이고 교차항 x_1x_2는 밀도비에 포함되지 않는다. 하지만 이러한 조건이 만족하는 것은 크게 기대하기 어려우므로 이차항과 교차항이 밀도비에 포함된다고 보아야 한다.

4. 예제

이 절에서는 실제자료를 이용하여 3절에서 살펴본 내용을 확인하여 본다. 먼저 Cook과 Weisberg (1999)에 제시된 콩 줄기 끝에 위치한 잎의 특징을 관측한 자료를 이용한다. 이 자료의 변수들은 관찰시기(Year: 0 = 1994년, 1 = 1995년), 잎의 면적(Area), 잎의 길이(Length), 잎의 너비(Width), 파종 후 경과일(DAP)이 있으며 이 중, 0과 1의 값을 갖는 Year를 반응변수로, Width와 DAP를 설명변수로 하였다. 관찰시기에 따른 DAP와 Width에 대한 산점도를 그려보았다.

Figure 4.1의 두 산점도에서 1994년과 1995년 경우 모두 Width와 DAP 사이의 상관관계가 없는 것으로 보인다. 또한 Width의 분산은 같아 보이지만, DAP의 분산은 같지 않아 보인다. 따라서 이차항인 Width^2과 교차항인 Width∗DAP은 필요하지 않고, DAP^2만 필요한 것으로 예상된다. 즉 두 산점도를 근거로 모형에 Width와 DAP 외에 추가적으로 DAP^2이 포함되어야 한다는 것을 알 수 있다.

Table 4.2. Fit of logistic regression for leaf area data without Width^2 and Width∗DAP

Coefficients	Estimate	Std.Error	Chi-Square	Pr>ChiSq
Intercept	12.4827	5.0634	6.0516	0.0139
Width	1.6695	0.4138	16.2753	<.0001
DAP	−1.2524	0.3582	12.2220	0.0005
DAP∗DAP	0.0204	0.00577	12.4947	0.0004

Table 4.3. Fit of logistic regression for leaf area data with Width and DAP

Coefficients	Estimate	Std.Error	Chi-Square	Pr>ChiSq
Intercept	−5.9979	1.4360	17.4457	<.0001
Width	1.4120	0.3738	14.2667	0.0002
DAP	0.0337	0.0274	1.5164	0.2181

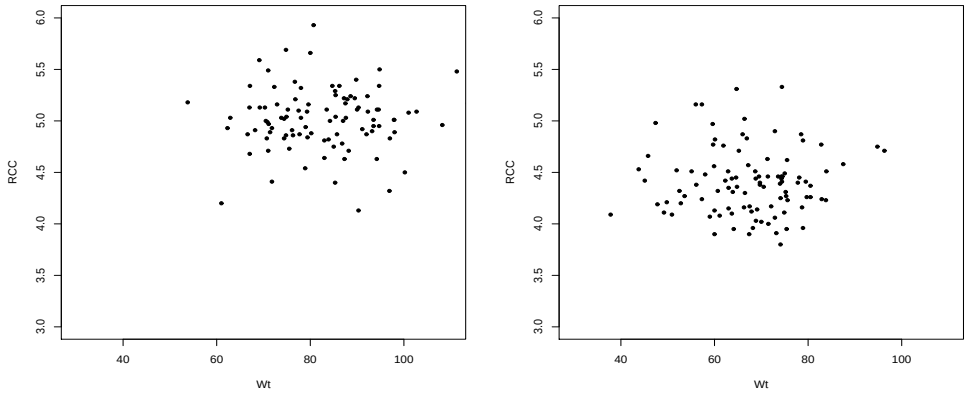

Figure 4.2. Scatter plots of Wt and RCC (Sex = 0, Sex = 1)

일차항, 이차항 그리고 교차항 모두를 설명변수로 하는 로지스틱회귀분석의 결과가 Table 4.1과 같다. Figure 4.1을 근거로 예상한바 대로 Width^2과 Width∗DAP의 p-값은 각각 0.7064와 0.1888로 모두 유의하지 않고 DAP^2의 p-값은 0.0007로 예상한대로 유의한 결과가 나왔다. 유의하지 않다고 판단되는 Width^2과 Width∗DAP을 제외하고 다시 적합한 결과가 Table 4.2에 있으며 일차항인 Width와 DAP 그리고 이차항 DAP^2가 모두 유의한 것으로 나타난다.

일반적으로 사용되어지는 설명변수 Width와 DAP, 즉 일차항만을 포함하는 모형 (2.1)을 적용해보자. 이 경우 로지스틱회귀분석의 결과는 Table 4.3과 같다. 이 때 DAP는 유의하지 않으므로 Width만을 포함하는 모형이 최종모형이 될 것이며 모형 (2.2)를 적용하여 얻은 결과와 다른 결론에 도달하게 된다.

다음은 Cook과 Weisberg (1994)에 제시된 오스트레일리아 스포츠 선수촌에서 훈련하는 운동선수 남녀 각각 100명의 신체지수와 혈액검사 자료를 분석한다. 변수들은 성별(Sex: 0 = male, 1 = female), 몸무게(Wt), 키(Ht), 적혈구수치(RCC), 헤모글로빈수치(Hg) 등이 있으며, 이 중 0과 1의 값을 갖는 Sex를 반응변수로 사용하고, Wt와 RCC를 설명변수로 사용한다. 먼저 남자인 경우와 여자인 경우를 나누어 Wt와 RCC의 산점도를 그려보았다.

Figure 4.2의 두 그림을 보면 남녀 두 경우에서 모두 Wt와 RCC는 상관관계가 없는 모습을 보인다. 남녀선수간의 RCC의 분산은 같아 보이고 남녀선수간의 Wt의 분산도 같아 보인다. 이런 경우에는 상관

Table 4.4. Fit of logistic regression for Australian athletes data

Coefficients	Estimate	Std.Error	Chi-Square	Pr>ChiSq
Intercept	−54.6046	72.9346	0.5605	0.4541
Wt	0.2614	0.6105	0.1833	0.6686
RCC	11.3967	25.7161	0.1964	0.6576
Wt∗Wt	−0.00110	0.00186	0.3480	0.5552
RCC∗RCC	−0.6632	2.4593	0.0727	0.7874
Wt∗RCC	0.00662	0.0932	0.0051	0.9433

Table 4.5. Fit of logistic regression for Australian athletes data without Wt^2, RCC^2, and Wt∗RCC

Coefficients	Estimate	Std.Error	Chi-Square	Pr>ChiSq
Intercept	−35.8789	5.0634	50.2107	<.0001
Wt	0.1275	0.0259	24.2403	<.0001
RCC	5.5769	0.8673	41.3424	<.0001

계수가 같고 두 분산이 같은 경우이므로 두 이차항과 교차항은 모두 필요하지 않고 Wt와 RCC만으로 로지스틱모형에서 구성하여도 된다.

실제로 이 자료에 로지스틱모형을 적합 시킨 결과가 Table 4.4에 있으며 Wt^2, RCC^2, Wt∗RCC의 p-값은 각각 0.5552, 0.7874, 0.9433으로 모두 유의하지 않게 나왔다. 이차항과 교차항을 모두 제외하고 일차항인 Wt와 RCC만으로 다시 적합한 결과가 Table 4.5에 있으며 Wt와 RCC는 유의한 결과가 나왔다. 이를 통해 이차항과 교차항 없이 일차항만으로도 충분한 설명력이 있다는 것을 알 수 있다.

5. 결론

설명변수가 두 개일 때 일반적으로 선형회귀모형과 로지스틱회귀모형에서 두 설명변수만 모형에 포함시킨다. 하지만 설명변수만으로는 충분히 설명이 되지 못하고 설명변수의 변환된 형태 즉 이차항과 교차항이 필요한 경우가 있다. 설명변수의 조건부분포가 이변량 정규분포를 따르는 경우 로지스틱모형에서는 기본적으로 일차항과 이차항, 교차항 모두 모형에 포함되어야 한다. 하지만 두 이변량 정규분포의 분산과 상관계수에 따라 이차항과 교차항이 필요하지 않게 되는 경우도 있다.

본 논문에서는 이변량 정규분포를 따르는 경우 로그-밀도비를 통해 각 항이 필요하게 되는 경우와 그렇지 않은 경우의 필요한 조건에 대해 알아보았다. 두 이변량 정규분포에서 분산-공분산행렬이 같으면, 즉 분산과 상관계수가 동일하면 이차항과 교차항이 모두 필요하지 않다. 상관계수가 같은 경우, 설명변수의 분산이 같으면 해당 이차항은 필요하지 않다. 특히 상관계수가 모두 0이면 교차항은 분산과 관계없이 항상 필요하지 않다. 상관계수가 같지 않은 경우에는 대부분의 경우 이차항과 교차항은 항상 필요하다.

자료의 분산이나 상관계수에 대한 정보는 그래프를 보고 대체적인 판단이 가능하다. $y = 0$인 경우와 $y = 1$인 경우에서 x_1과 x_2의 산점도를 통해 분산과 상관계수에 대한 정보를 얻고 그것을 바탕으로 어떤 항이 필요하게 될지를 판단할 수 있다. 만약 분산과 상관계수에 별 차이가 없다고 판단되면 일차항만으로 모형을 구성하여도 된다. 대부분의 자료에서 이러한 현상이 나타난다. 그러나 산점도를 통해 분산이나 상관계수의 차이가 의심된다면 이차항 또는 교차항을 포함시키는 것을 검토해 보아야한다. 상관계수가 같고 하나의 분산만 다르다면 해당 이차항만 포함된다. 상관계수가 0에 가깝다면 교차항은 필요없다. 따라서 이러한 산점도를 통해 설명변수의 선형항만으로 충분한지 아니면 이차항과 교차항이 필요한지에 대한 판단이 가능하다.

References

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

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*, John Wiley & Sons, New York.

Kahng, M. and Shin, E. (2012). Variable selection with log-density in logistic regression model, *Communications of the Korean Statistical Society*, **19**, 1–11.

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 the Royal Statistical Society, Series A*, **135**, 370–384.

Scrucca, L. (2003). Note on the consistency of the maximum likelihood estimate, *Statistical Methods and Applications*, **11**, 371–394.

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.

로지스틱 회귀모형에서 이변량 정규분포에 근거한 로그-밀도비

강명욱[a,1] · 윤재은[a]

[a]숙명여자대학교 통계학과

(2012년 11월 27일 접수, 2012년 12월 31일 수정, 2013년 1월 18일 채택)

요 약

로지스틱회귀모형에서 두 설명변수의 조건부 분포가 모두 이변량 정규분포라고 할 수 있다면 설명변수들의 함수로 표현되는 로그-밀도비를 통해 모형에 포함시켜야하는 항을 알 수 있다. 두개의 이변량 정규분포에서 분산-공분산행렬이 같은 경우에는 이차항과 교차항 없이 일차항만으로 충분하다. 상관계수가 모두 0이면 교차항은 설명변수의 분산과 관계없이 필요하지 않다. 또한 로지스틱회귀모형에서 로그-밀도비를 통해 이차항과 교차항이 필요하지 않게 되는 다른 조건들도 알아본다.

주요용어: 로그-밀도비, 로그-오즈, 로지스틱회귀, 역회귀, 이항회귀, 커널평균함수.

본 연구는 숙명여자대학교 2011년도 교내연구비 지원에 의해 수행되었음.

[1]교신저자: 서울특별시 용산구 청파로47길 100, 숙명여자대학교 통계학과, 교수.
E-mail: mwkahng@sookmyung.ac.kr