한국통계학회논문집
2012, 19권, 1호, 1–11

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.1.001

로지스틱회귀모형에서 로그-밀도비를 이용한 변수의 선택

강명욱[1,a], 신은영[a]

[a]숙명여자대학교 통계학과

요 약

로지스틱회귀모형에서 반응변수가 주어졌을 때 설명변수의 조건부 확률분포의 로그-밀도비는 어떤 설명변수가 어떻게 모형에 포함되는지에 대한 변수선택문제에서 유용한 정보를 제공한다. 설명변수의 조건부 확률분포가 좌우대칭이 아닌 경우 감마분포로 가정하는 것이 적절하다. 여러 가지 모의실험을 수행한 결과를 보면, $x|y=0$과 $x|y=1$의 두 분포가 겹치는 경우에서는 x항과 $\log(x)$항 모두 필요하다. 그리고 두 분포가 분리된 경우에는 x항 또는 $\log(x)$항 중 하나만 필요하다.

주요용어: 로그-밀도비, 로지스틱회귀모형, 역회귀, 이항반응변수, 쿨백-라이블러 발산.

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.

로지스틱회귀모형은 Nelder와 Wedderburn (1972)이 제안한 일반화선형모형의 한 형태이다. 일반화선형모형은 정규이론을 따르는 선형모형을 지수족(exponential family)과 연결함수(link function)를 이용해 다음과 같은 두 가지 과정으로 일반화 될 수 있다. 첫째, 반응변수의 기대값과 설명변수의 선형결합(linear predictor) 연결시키는 연결함수를 설정한다. 둘째, 오차의 분포는 정규분포를 포함하는 지수족의 여러 가지 분포를 사용한다.

이러한 일반화선형모형 중 하나인 로지스틱회귀모형에서 y는 시행 횟수가 m이고 성공확률이 p인 이항분포를 따르는 확률변수라 하고 y/m을 반응변수로 하는 다음과 같은 선형모형을 생각하자.

$$E\left(\frac{y}{m}\middle|\boldsymbol{x}\right) = p(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} = \sum_{j=1}^{p} \boldsymbol{x}_j\boldsymbol{\beta}_j. \tag{1.1}$$

이 경우 p는 0과 1사이의 값인데 $\boldsymbol{x}^T\boldsymbol{\beta}$는 $-\infty$와 ∞사이의 값을 갖게 되므로 모든 실수값을 취하게 된다. 따라서 선형모형은 적절하지 못하며, 최소제곱법을 사용하여 적합하면 모형 (1.1)에 있는 모수들의 유용성과 해석에 제한이 생긴다. 이것을 해결하기 위하여 다음과 같은 p의 로짓변환(logit transformation)을 생각할 수 있다.

$$\text{logit}(p(\boldsymbol{x})) = \log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right).$$

본 연구는 숙명여자대학교 2010년도 교내연구비 지원에 의해 수행되었음.
[1] 교신저자: (140-742) 서울 용산구 청파동, 숙명여자대학교 통계학과, 교수. E-mail: mwkahng@sm.ac.kr

로짓변환은 승산(odds), 즉 성공의 확률과 실패의 확률의 비에 로그를 취한 것이다. 이러한 로짓은 다음과 같은 성질을 가진다. 첫째, p가 증가함에 따라 logit(p)도 증가한다. 둘째, p가 0에서 1까지의 값만을 취하는 반면에 logit(p)는 실수값을 취한다. 로지스틱회귀모형은 다음과 같이 로짓으로 변환한 선형모형의 형태로 표현한다.

$$\text{logit}(p(\boldsymbol{x})) = \boldsymbol{x}^T\boldsymbol{\beta}.$$

이 식을 p에 대하여 정리하면 다음의 식을 얻는다.

$$E\left(\frac{y}{m}\middle|\boldsymbol{x}\right) = p(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{x}^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}^T\boldsymbol{\beta}\right)}. \tag{1.2}$$

성공횟수를 나타내는 y, 실행횟수 m과 p개의 설명변수 $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$의 n개 관찰값에 대해 생각하자. 우선 성공비율에 기대값을 취하여 다음과 같이 p_i라고 한다.

$$E\left(\frac{y_i}{m_i}\middle|\boldsymbol{x}_i\right) = p(\boldsymbol{x}_i).$$

변수 $p(\boldsymbol{x}_i)$의 로짓과 설명변수의 선형결합의 관계를 나타내는 다음의 모형을 생각하자.

$$\text{logit}(p(\boldsymbol{x}_i)) = \boldsymbol{x}_i^T\boldsymbol{\beta} = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p, \quad i = 1, \ldots, n. \tag{1.3}$$

식 (1.3)을 $p(\boldsymbol{x}_i)$에 대하여 정리하면 다음과 같다.

$$E\left(\frac{y_i}{m_i}\middle|\boldsymbol{x}_i\right) = p(\boldsymbol{x}_i) = \frac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}, \quad i = 1, \ldots, n,$$

여기서 $\boldsymbol{x}_i$는 p개의 설명변수의 i번째 관찰값들로 이루어진 차수 p인 설명변수벡터이고 $\boldsymbol{\beta}$는 차수가 p인 회귀계수벡터이다. 모형 (1.3)은 로지스틱회귀모형으로, 반응변수의 비선형함수로의 변환을 새로운 반응변수로 놓고 선형모형을 적용시킨 것이다. 그러므로 모형 (1.3)은 일반화선형모형의 한 형태이다.

2. 로지스틱 회귀모형에서 로그-밀도비

성공할 경우 "1"값을 실패할 경우 "0"값을 가지는 이항반응변수(binary response variable) y와 p개의 설명변수 $\boldsymbol{x} = (x_1, \ldots, x_p)^T$를 생각하자. 이항변수인 y의 조건부분포는 기댓값이 $E(y|\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$이 되는 베르누이분포를 따른다. 이항회귀모형은 일반화선형모형의 특별한 경우로 평균을 다음과 같이 쓸 수 있다.

$$E(y|\boldsymbol{x}) = g\left(\boldsymbol{\eta}^T\boldsymbol{u}\right), \tag{2.1}$$

여기서 $g(\ )$는 커널평균함수(kernel mean function)이고, 연결함수의 역함수이다. 모형 (2.1)에서의 커널함수는 모형 (1.2)에서와 같은 $\boldsymbol{x}$의 선형결합인 $\boldsymbol{x}^T\boldsymbol{\beta}$가 아닌 $\boldsymbol{u}$의 선형결합인 $\boldsymbol{\eta}^T\boldsymbol{u}$의 함수이다. $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$는 p개의 예측변수 $\boldsymbol{x}$로부터 구한 벡터이다. 일반적으로 $\boldsymbol{u}$는 $\boldsymbol{x}$의 함수로 구성된다. 로지스틱회귀모형에서는 로지스틱함수를 커널평균함수와 똑같이 사용한다.

정규오차와 함께 다중선형회귀에서 추정은 정확히 정규분포를 따른다. 그러나 로지스틱회귀에서의 추정은 근사적으로 정규분포를 따르게 되고 표본크기를 증가할수록 근사가 개선된다. 로지스틱회귀에서 계수는 유용한 해석이 가능하다. 우리는 로지스틱함수를 다음과 같이 쓸 수 있다.

$$p(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{\eta}^T\boldsymbol{u}\right)}{1+\exp\left(\boldsymbol{\eta}^T\boldsymbol{u}\right)}$$

그리고 이 $\boldsymbol{\eta}^T\boldsymbol{u}$에 대한 방정식을 풀면, 다음과 같이 쓸 수 있다.

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

로지스틱회귀에서는 연결함수 $\log(p(\boldsymbol{x})/(1-p(\boldsymbol{x})))$는 로짓이라 하고 $p(\boldsymbol{x})/(1-p(\boldsymbol{x}))$를 성공-오즈(odds of success)라 부른다. 반응변수의 분포가 특별한 경우에는 $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$를 적절하게 선택해야 한다. Cook과 Weisberg (1999)에 따르면, 정규분포일때는 $\boldsymbol{u}^T = (1, x, x^2)$를 사용하고 분산이 같을 때에는 $\boldsymbol{u}^T = (1, x)$를 사용한다. 그리고 감마분포에서는 $\boldsymbol{u}^T = (1, x, \log(x))$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 주변확률밀도함수라 하자. 반응변수가 이항변수이므로 로지스틱회귀에서의 평균함수 $E(y|\boldsymbol{x})$는 베이즈공식을 이용하면 다음과 같이 쓸 수 있다.

$$E(y|\boldsymbol{x}) = p(\boldsymbol{x}) = P(y=1|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|y=1)P(y=1)}{f(\boldsymbol{x})}. \tag{2.2}$$

식 (2.2)에서 $\boldsymbol{x}$가 주어졌을 때 $y=1$에 대한 확률 $p(\boldsymbol{x})$를 평균함수라 말할 수 있다. 또한 $\boldsymbol{x}$가 주어졌을 때 $y=0$에 대한 확률 $1-p(\boldsymbol{x})$는 다음과 같이 쓸 수 있다.

$$1-p(\boldsymbol{x}) = P(y=0|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|y=0)P(y=0)}{f(\boldsymbol{x})}. \tag{2.3}$$

식 (2.2)와 식 (2.3)의 두 값의 로그비를 취하면 다음과 같이 로그-오즈(log-odds)를 얻을 수 있다.

$$\begin{aligned}\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) &= \log\left(\frac{P(y=1)}{P(y=0)}\right) + \log\left(\frac{f(\boldsymbol{x}|y=1)}{f(\boldsymbol{x}|y=0)}\right)\\ &= \log\left(\frac{P(y=1)}{P(y=0)}\right) + h(\boldsymbol{x}).\end{aligned} \tag{2.4}$$

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. Cook과 Weisberg (1999), Scrucca (2003), 그리고 Scrucca와 Weisberg (2004)는 로그-밀도비와 관련된 조건부분포를 그래픽적으로 알아보았다.

만약 $f(x|y=j)$가 $j=\{0,1\}$에서 평균 μ_i와 분산 σ_j^2를 가지는 정규밀도함수라면, 우리는 로그-밀도비를 다음과 같이 쓸 수 있다.

$$h(\boldsymbol{x}) = \log\left(\frac{\sigma_0}{\sigma_1}\right) - \frac{\mu_1^2}{2\sigma_1^2} + \frac{\mu_0^2}{2\sigma_0^2} + \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}\right)x + \frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)x^2. \tag{2.5}$$

또한, 로그-오즈는 다음과 같다.

$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \log\left(\frac{P(y=1)}{P(y=0)}\right) + h(\boldsymbol{x}) = \eta_0 + \eta_1 x + \eta_2 x^2. \tag{2.6}$$

식 (2.6)에서 우리는 로그-오즈를 $\boldsymbol{x}|y$의 평균과 분산들에 의존하는 계수들에 대한 x와 x^2의 선형결합변수로 표현할 수 있다. 여기서 $k = \sigma_1^2/\sigma_0^2$를 분산들의 비라고 하자. 그러면 $\sigma = \sigma_0^2$, $\sigma_1^2 = k\sigma^2$이 된다. 반면, $c = (\mu_1 - \mu_0)/k$는 분산들 비에 의해서 조정된 평균들 사이의 차이를 말한다. 따라서 $\mu = m_0$, $\mu_1 = \mu + ck$라 쓸 수 있다. 식 (2.6)의 각 항들을 정리하면 다음과 같다.

$$\begin{aligned}
\eta_0 &= \log\left(\sqrt{k}\right)^{-1} + \frac{\mu^2(k-1) - ck(2\mu + ck)}{2k\sigma^2}, \\
\eta_1 &= \frac{\mu(1-k) + ck}{k\sigma^2}, \\
\eta_2 &= \frac{k-1}{2k\sigma^2}.
\end{aligned} \tag{2.7}$$

따라서 식 (2.6)에 식 (2.7)을 대입해 정리하면 다음과 같다.

$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \log\left(\frac{p}{1-p}\right) - \frac{\log(k)}{2} + \frac{\mu^2(k-1) - ck(2\mu+ck)}{2k\sigma^2} + \left(\frac{\mu(1-k)+ck}{k\sigma^2}\right)x + \left(\frac{k-1}{2k\sigma^2}\right)x^2.$$

$k = 1$이라면 이차항 x^2은 필요하지 않다. 반면에 $c = \mu(k-1)/k$일 때는 선형항 x가 필요하지 않다.

3. 비대칭자료에서의 로그-밀도비

일반적으로 자료가 좌우대칭이면 정규분포로 설명이 가능하다. 그러나 한쪽으로 치우친 비대칭 자료가 주어질 때, 감마분포가 정규분포보다 적절하다. 현실적으로 좌우대칭이 아닌 자료들이 많다. 이런 경우에는 변수변환 등을 이용해 정규분포의 형태로 만든다. 그러나 변수변환 등을 하여서도 좌우대칭형태로 만들 수 없다면, 정규분포를 이용하여 설명하는 것이 적절치 못하고 확장된 개념인 감마분포에 대한 로그-밀도비가 쓰이게 된다. 여기에서는 감마분포에서의 로지스틱회귀모형에서 로그-밀도비에 대해 알아본다.

만약 $f(x|y=j)$가 $j = \{0, 1\}$에서 형태모수(shape parameter) α_i와 척도모수(scale parameter) λ_j를 가지는 감마밀도함수라면 식 (2.4)에서와 같이 로그-밀도비를 다음과 같이 쓸 수 있다.

$$\begin{aligned}
h(x) &= \log\left(\frac{\frac{1}{\Gamma(\alpha_1)}\left(\frac{1}{\lambda_1}\right)^{\alpha_1} x^{\alpha_1-1}\exp\left(-\frac{x}{\lambda_1}\right)}{\frac{1}{\Gamma(\alpha_0)}\left(\frac{1}{\lambda_0}\right)^{\alpha_0} x^{\alpha_0-1}\exp\left(-\frac{x}{\lambda_0}\right)}\right) \\
&= \log\left(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)}\right) + [\alpha_0\log(\lambda_0) - \alpha_1\log(\lambda_1)] + \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right)x + (\alpha_1 - \alpha_0)\log(x).
\end{aligned} \tag{3.1}$$

식 (2.4)에 식 (3.1)을 대입해 정리하면, 로그-오즈는 다음과 같다.

$$\begin{aligned}
\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) &= \log\left(\frac{P(y=1)}{P(y=0)}\right) + h(x) \\
&= \eta_0 + \eta_1 x + \eta_2\log(x).
\end{aligned} \tag{3.2}$$

2장에서 설명한 바와 같이 $\boldsymbol{\eta} = (\eta_0, \eta_1, \eta_2)^T$, $\boldsymbol{u} = (1, x, \log(x))^T$로 쓸 수 있다. 그리고 회귀계수 η_0, η_1, η_2는 다음과 같다.

$$\begin{aligned}
\eta_0 &= \log\left(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)}\right) + \left[\alpha_0 \log(\lambda_0) - \alpha_1 \log(\lambda_1)\right], \\
\eta_1 &= \frac{1}{\lambda_0} - \frac{1}{\lambda_1}, \\
\eta_2 &= (\alpha_1 - \alpha_0).
\end{aligned} \tag{3.3}$$

식 (3.3)에서 우리는 로그-오즈를 $\boldsymbol{x}|y$의 형태모수 α와 척도모수 λ에 의존하는 계수들과 함께 선형결합변수 x와 $\log(x)$로서 다음과 같이 표현할 수 있다.

$$\log\left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right) = \eta_0 + \eta_1 x + \eta_2 \log(x) = \boldsymbol{\eta}^T \boldsymbol{u}.$$

식 (3.2)에서 $\alpha_0 = \alpha_1$이면 로그항 $\log(x)$는 필요하지 않다. 반면에 $\lambda_0 = \lambda_1$일 때는 선형항 x가 필요하지 않는다.

4장에서 다룰 모의실험 연구의 목표는 다음의 평균함수들을 비교하는 것이다.

$$\begin{aligned}
\text{모형 1} &: E(y|\boldsymbol{x}) = \frac{\exp(\eta_0 + \eta_1 x)}{1 + \exp(\eta_0 + \eta_1 x)} \\
\text{모형 1-1} &: E(y|\boldsymbol{x}) = \frac{\exp(\eta_0 + \eta_2 \log(x))}{1 + \exp(\eta_0 + \eta_2 \log(x))} \\
\text{모형 2} &: E(y|\boldsymbol{x}) = \frac{\exp(\eta_0 + \eta_1 x + \eta_2 \log(x))}{1 + \exp(\eta_0 + \eta_1 x + \eta_2 \log(x))}
\end{aligned}$$

Kullback (1959)가 제시한 쿨백-라이블러 발산(Kullback-Leibler divergence; KLD)은 두 확률분포의 차이를 계산하는 데에 사용하는 함수로, 어떤 이상적인 분포에 대해, 그 분포를 근사하는 다른 분포를 사용해 샘플링을 한다면 발생할 수 있는 정보 엔트로피 차이를 계산한다. 경쟁되는 두 모형의 비교는 KLD를 사용하여 수행할 수 있다. 연속형의 경우, KLD는 다음과 같이 정의된다.

$$I(g : f) = \int_{-\infty}^{\infty} \log \frac{f(x)}{g(x)} f(x) dx = E\left(\log \frac{f(x)}{g(x)}\right),$$

여기서 $f(x)$는 참모형(true model)이고 $g(x)$는 후보모형(candidate model)을 의미한다. 후보모형의 추정된 평균함수는 다음과 같이 표현할 수 있다.

$$\hat{E}(y|\boldsymbol{x}) = \hat{p}(\boldsymbol{x}) = \frac{\exp\left(\hat{\boldsymbol{\eta}}^T \boldsymbol{u}(\boldsymbol{x})\right)}{1 + \exp\left(\hat{\boldsymbol{\eta}}^T \boldsymbol{u}(\boldsymbol{x})\right)},$$

여기서 $\hat{\boldsymbol{\eta}}$는 최우추정량 벡터이고, $\boldsymbol{u}(\boldsymbol{x})$는 변수 $\boldsymbol{x}$로부터의 벡터항이다. 모형 1에서는 $\boldsymbol{u}^T(x) = (1, x)$이고, 모형 1-1에서는 $\boldsymbol{u}^T(x) = (1, \log(x))$이고, 모형 2에서는 $\boldsymbol{u}^T(x) = (1, x, \log(x))$이다. KLD를 사용하면 로지스틱모형으로부터 로그항 $\log(x)$ 또는 선형항 x가 탈락됐을 때 손실된 정보를 조사할 수 있다. 어떤 $\boldsymbol{x}$가 주어질 때, KLD를 다음과 같이 쓸 수 있다.

$$\begin{aligned}
I\left(\hat{p}(\boldsymbol{x}) : p(\boldsymbol{x})\right) &= \sum_{j \in 0,1} \log\left(\frac{P(y = j|\boldsymbol{x})}{\hat{P}(y = j|\boldsymbol{x})}\right) P(y = j|\boldsymbol{x}) \\
&= \log\left(\frac{p(\boldsymbol{x})}{\hat{p}(\boldsymbol{x})}\right) p(\boldsymbol{x}) + \log\left(\frac{1 - p(\boldsymbol{x})}{1 - \hat{p}(\boldsymbol{x})}\right)(1 - p(\boldsymbol{x})),
\end{aligned}$$

여기서 $p(\boldsymbol{x})$는 참확률이고, $\hat{p}(\boldsymbol{x})$는 식 (3.2)로부터의 추정확률이다. 표본의 크기 n에서, 독립적으로 관측된 전체 KLD 정보는 각 관측값으로부터 KLD의 합이 된다.

$$I(\hat{p} : p) = \sum_{i=1}^{n} I(\hat{p}(x_i) : p(x_i)).$$

$I(\hat{p} : p)$는 각각의 모형에 대해 계산될 수 있으며, 각각 모형 1 또는 모형 1-1과 모형 2의 $I(\hat{p} : p)$는 각각 I_1과 I_2로 표시할 수 있다. 일반적으로 모형 1 또는 모형 1-1로부터 $\hat{p}(\boldsymbol{x})$는 $p(\boldsymbol{x})$의 선형근사이기 때문에, 로그항이 필요할 때, I_1이 로그항이 포함된 I_2 보다 크다고 예상할 수 있다. 그 때의 차이 $(I_1 - I_2)$가 로지스틱회귀모형으로부터 로그항 또는 선형항이 탈락된 효과를 나타낸다고 할 수 있다.

4. 모의실험

모의실험에서 형태모수 α_1, α_2와 척도모수 λ_1, λ_2 그리고 감마분포의 평균값을 조정하는 b_i를 다음과 같이 다양하게 한다. $\alpha_1 \in \{2, 5, 10, 20\}$, $\alpha_2 \in \{2, 5, 10, 20\}$, $\lambda_1 \in \{2, 5, 10, 20\}$, $\lambda_2 \in \{2, 5, 10, 20\}$, $b_i \in \{0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5, 7.5, 10\}$.

4.1. 척도모수 λ변화

척도모수 λ의 변화에 따른 변수 선택의 과정을 알아보기 위해 다음과 같이 모의실험을 실시한다. 표본크기는 $n = 200$으로 하고 y를 Bernoulli(p)에서 200개를 생성한다. $y = 1$일 때의 x의 평균이 $y = 0$일 때의 x의 평균의 b_i배가 되도록 x를 생성한다. 이를 위하여 $x|(y = 0) \sim G(\alpha_1, \lambda_1)$과 $x|(y = 1) \sim G(\alpha_2, (\alpha_1\lambda_1/\alpha_2)b_i)$의 조건부 분포를 이용한다. 이 때, b_i는 0.1 ~ 10로 다양하게 사용한다. 각 표본들을 로지스틱회귀모형에 적합시키고, x항과 $\log(x)$항을 모두 포함한 모형과 $\log(x)$항이 없는 모형의 KLD를 계산하고 차이 $(I_1 - I_2)$를 구한다. 이와 같은 과정을 $m = 1000$번 반복하여 $(I_1 - I_2)$를 구한다.

예를 들면, 우선 $x|(y = 0) \sim G(2, 10)$로 하고 다양한 b_i값에 대해 $x|(y = 1) \sim G(2, 10b_i)$를 생성하고 로지스틱회귀모형에 적합시킨다. x, $\log(x)$항을 포함한 모형 2와 $\log(x)$항이 없는 모형 1의 KLD를 계산하고 차이 $(I_1 - I_2)$를 살펴본다. 또한, $x|(y = 1)$의 분포를 $G(5, 4b_i)$, $G(10, 2b_i)$, $G(20, b_i)$에 대해 다양한 b_i값을 대입하여 x를 생성하고 로지스틱회귀모형에 적합시킨다. x, $\log(x)$항을 포함한 모형 2와 $\log(x)$항이 없는 모형 1의 KLD를 계산한다. 그리고 모형 1과 모형 2의 KLD의 차이 $(I_1 - I_2)$를 구한다. 같은 방법으로 $x|(y = 1) \sim G(10, 2b_i)$, $x|(y = 1) \sim G(20, b_i)$ 순으로 각각 모형 1과 모형 2의 KLD를 계산하고 차이 $(I_1 - I_2)$를 구한다.

이제 $x|(y = 0) \sim G(5, 10)$와 $x|(y = 1) \sim G(2, 25b_i)$, $x|(y = 1) \sim G(5, 10b_i)$, $x|(y = 1) \sim G(10, 5b_i)$, $x|(y = 1) \sim G(20, 2.5b_i)$를 만든다. 역시 같은 방법으로 차례대로 모형 1과 모형 2의 KLD를 계산하고 그때 차이 $(I_1 - I_2)$를 구한다. 또한 $x|(y = 0) \sim G(10, 10)$와 $x|(y = 0) \sim G(20, 10)$에 대해서도 $(I_1 - I_2)$를 살펴본다.

4.2. 형태모수 α변화

척도모수 λ의 변화에 따른 변수 선택의 과정을 알아보기 위해 다음과 같이 모의실험을 실시한다. 표본크기는 $n = 200$으로 하고 y를 Bernoulli(p)에서 200개를 생성한다. $y = 1$일 때의 x의 평균이 $y = 0$일 때의 x의 평균의 b_i배가 되도록 x를 생성한다. 즉 $x|(y = 0) \sim G(\alpha_1, \lambda_1)$과 $x|(y = 1) \sim G((\alpha_1\lambda_1/\lambda_2)b_i, \lambda_2)$의 조건부 분포를 이용한다. 각 표본들을 로지스틱회귀모형에 적합시키고, x항과 $\log(x)$항을 모두 포함한 모형과 x항이 없는 모형의 KLD를 계산하자. 그리고 차이 $(I_1 - I_2)$를 구한다. 이와 같은 과정을 $m = 1000$번 반복하여 $(I_1 - I_2)$를 구한다.

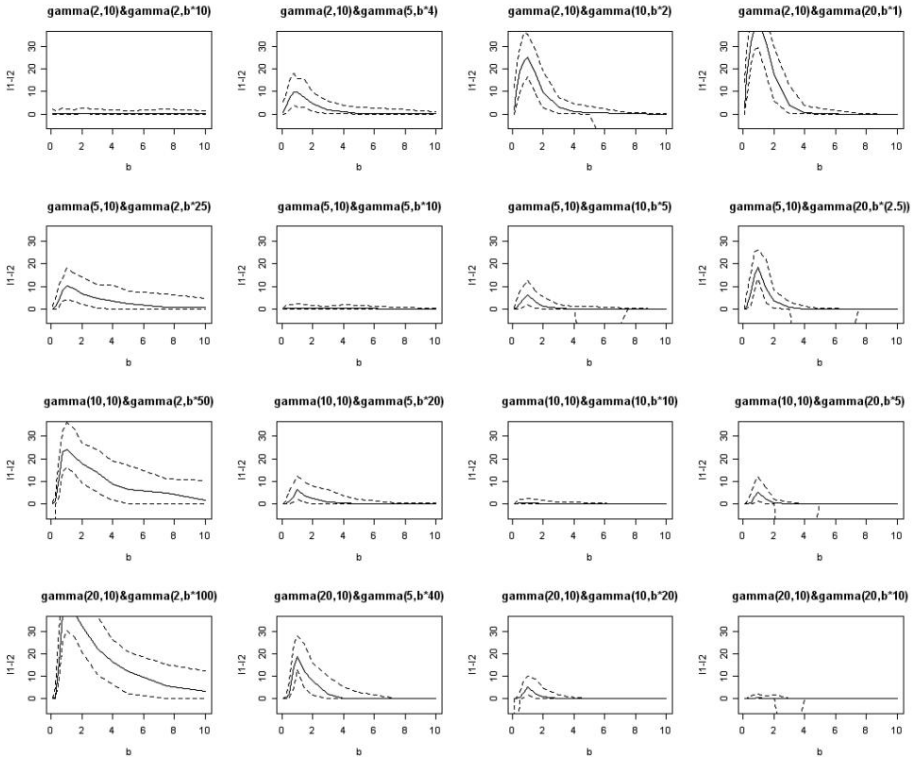

그림 1: *척도모수 λ에 대한 $(I_1 - I_2)$ 그래프*

예를 들면, 우선 $x|(y = 0) \sim G(2, 10)$로 하고 다양한 b_i값에 대해 $x|(y = 1) \sim G(10b_i, 2)$를 생성하고 로지스틱회귀모형에 적합시킨다. x, $\log(x)$항을 포함한 모형 2와 $\log(x)$항이 없는 모형 1-1의 KLD를 계산하고 차이 $(I_{1.1} - I_2)$를 살펴본다. 또한, $x|(y = 1)$의 분포를 $G(4b_i, 5)$, $G(2b_i, 10)$, $G(b_i, 20)$에 대해 다양한 b_i값을 대입하여 x를 생성하고 로지스틱회귀모형에 적합시킨다. x, $\log(x)$항을 포함한 모형 2와 $\log(x)$항만을 포함하는 모형 1-1의 KLD를 계산한다. 그리고 모형 1-1과 모형 2의 KLD의 차이 $(I_{1.1} - I_2)$를 구한다. 같은 방법으로 $x|(y = 1) \sim G(10, 2b_i)$, $x|(y = 1) \sim G(20, b_i)$순으로 각각 모형 1-1과 모형 2의 KLD를 계산하고 차이 $(I_1 - I_2)$를 구한다.

이제 $x|(y = 0) \sim G(5, 10)$와 $x|(y = 1) \sim G(25b_i, 2)$, $x|(y = 1) \sim G(10b_i, 5)$, $x|(y = 1) \sim G(5b_i, 10)$, $x|(y = 1) \sim G(2.5b_i, 20)$를 만든다. 역시 같은 방법으로 차례대로 모형 1-1과 모형 2의 KLD를 계산하고 그때 차이 $(I_{1.1} - I_2)$를 구한다. 또한 $x|(y = 0) \sim G(10, 10)$와 $x|(y = 0) \sim G(20, 10)$에 대해서도 $(I_{1.1} - I_2)$를 살펴본다.

4.3. 모의실험 결과

차이 $(I_1 - I_2)$로부터 로지스틱회귀모형의 적합에서 $\log(x)$의 추가여부를 파악할 수 있다. $p = 0.5$에 대한 결과는 다음 그림과 같이 요약되어진다. 그래프의 각 패널에서 $(I_1 - I_2)$와 b_i를 비교한 그림이다. 실선은 $(I_1 - I_2)$의 중앙값 표시하고, 반면에 파선은 5번째, 95번째 분위수들을 표시한 것이다. 또한, 파선은 $(I_1 - I_2)$의 변동을 그래픽적으로 나타낸다.

척도모수 λ에 대한 KLD의 차이인 $(I_1 - I_2)$ 그래프인 그림 1에서 평균을 조정하는 b_i값들이 커짐에

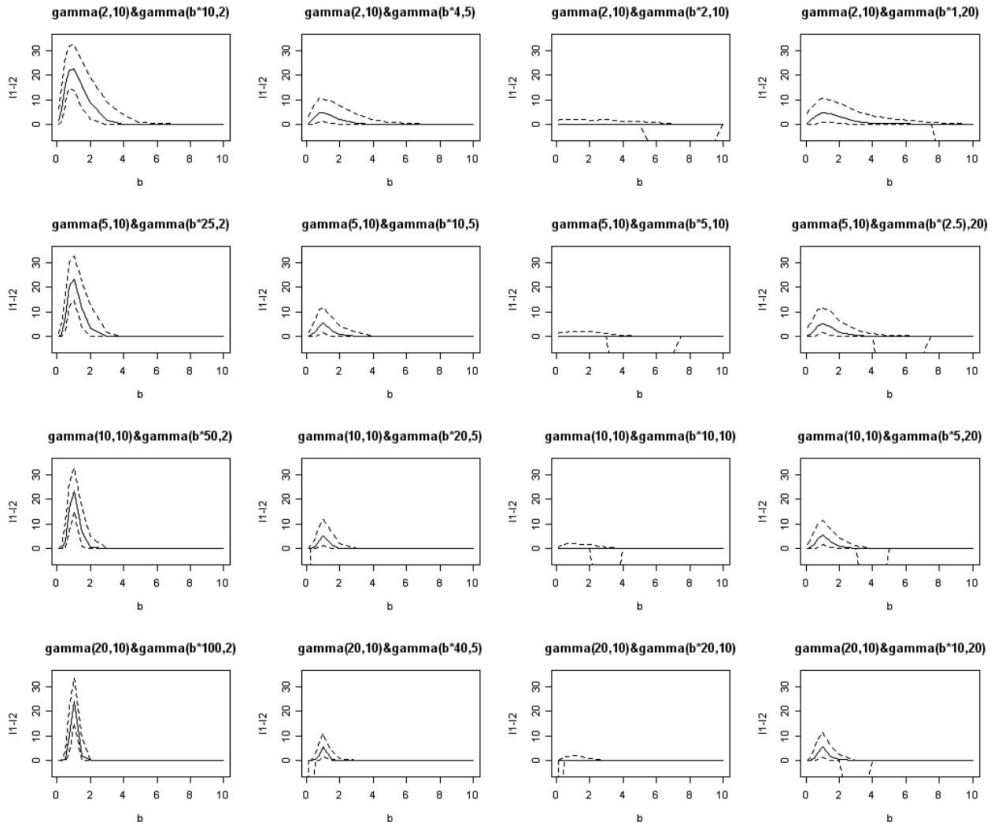

그림 2: *형태모수 α에 대한 $(I_{1.1} - I_2)$ 그래프*

따라 $(I_1 - I_2)$가 0이 되는 경향이 있다. b_i값이 매우 크거나 작은 경우에는 0이 되는 경향이 있다.

모양모수 $\alpha_1 = \alpha_2$일 때, b_i값에 상관없이 $\log(x)$가 필요 없다. b_i값이 1 근처는 평균이 같은 값을 가지는 경우를 말한다. 그리고 $\alpha_1 = \alpha_2$인 경우를 제외한 b_i값이 1 근처인 경우에는 $\log(x)$가 상당히 유의하다. 대략 전체적으로 $b = 4$이상인 경우에는 $(I_1 - I_2)$가 0이 되는 경향이 있다. 또한, $b = 0.25$이하인 경우에도 $(I_1 - I_2)$가 0이 되는 경향이 있다. 즉, 두 분포가 겹지는 부분이 없는 경우에 $(I_1 - I_2)$가 0이 되는 경향이 있다고 말할 수 있다. 그러나 $G(2, 10)$와 $G(5, 4b)$, $G(10, 2b)$, $G(20, b)$, $G(5, 10)$와 $G(2, 25b)$, $G(10, 10)$와 $G(2, 50b)$, $G(20, 10)$와 $G(2, 100b)$에서는 약간 예외적이다. b_i값이 커짐에 따라 KLD의 차이 $(I_1 - I_2)$가 줄어드는 경향이 있지만 $b = 10$이상인 경우에서 $(I_1 - I_2)$가 0에 가까워짐을 알 수 있다.

형태모수 α에 대한 KLD의 차이인 $(I_{1.1} - I_2)$ 그래프인 그림 2를 보면 평균을 조정하는 b_i값들이 커짐에 따라 $(I_{1.1} - I_2)$가 0이 되는 경향이 있다. b_i값이 매우 크거나 작은 경우에는 0이 되는 경향이 있다.

척도모수 $\lambda_1 = \lambda_2$인 경우인 세 번째 열 그래프일 때, b_i값에 상관없이 $(I_{1.1} - I_2)$가 0에 가까움을 알 수 있다. 다시 말해 $\log(x)$만으로도 설명이 가능하다. b_i값이 1 근처는 평균이 같은 값을 가지는 경우를 말한다. 그리고 척도모수 $\lambda_1 = \lambda_2$인 경우를 제외한 b_i값이 1 근처인 경우에는 x가 상당히 유의하다. 대략 전체적으로 $b = 4$이상인 경우에는 $(I_{1.1} - I_2)$가 0이 되는 경향이 있다. 또한, $b = 0.25$이하인 경우에도 $(I_{1.1} - I_2)$가 0이 되는 경향이 있다. 즉, 두 분포가 겹지는 부분이 없는 경우에 KLD의 차이 $(I_{1.1} - I_2)$가 0이 되는 경향이 있다고 말할 수 있다.

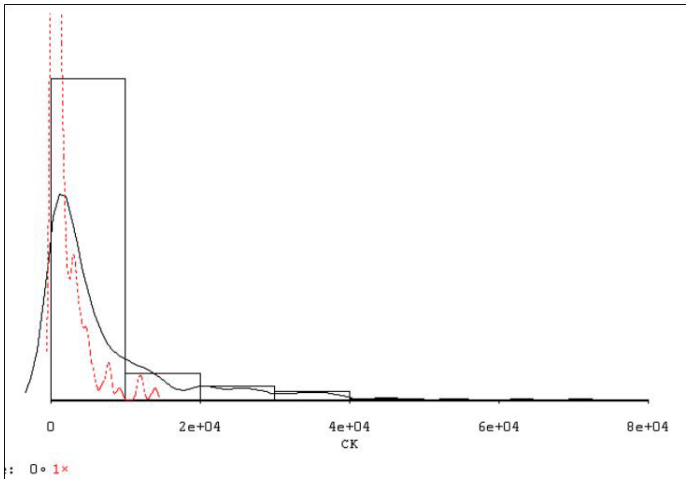

그림 3: *CK에 대한 히스토그램과 두 Outcome값에 대한 조건부확률밀도함수 그래프*

표 1: 젖소자료에 대한 CK와 log(CK)의 로지스틱회귀

Coefficients	Estimate	Std. Error	z value	Pr(> $\|z\|$)
Intercept	2.09917	0.845382	2.483	0.0130
CK	−0.000144639	0.0000594375	−2.433	0.0150
log(CK)	−0.288799	0.134007	−2.155	0.0312

5. 예제

자료는 Clark 등 (1987)에 제시된 1983~1984년 뉴질랜드의 Ruakura Animal Health Laboratory에서 500마리 이상의 젖소들의 혈액 표본을 조사한 것이다. 혈액에 대한 다양한 검사가 진행되었고 많은 동물들에 대한 생존, 사망 또는 다른 동물에 의한 사망의 결과들이 나왔다. 이 연구의 목적은 혈액분석 결과가 생존확률과 관계가 있는지를 결정하는 것이다. 그리고 분석으로부터 생존여부에 대한 예측이 가능한지를 알아보는 것이다. CK는 혈청 크레아킨 포스포키나아(serum creatine phosphokinase)이다. 반응변수는 결과들(outcomes)이다. 결과들은 생존할 경우에는 1로, 사망하거나 다른 동물에 의한 사망할 경우에는 0으로 표시한다.

그림 3은 두 분포 CK|(outcome = j), $j = \{0, 1\}$의 조건부확률밀도함수 그래프와 히스토그림이다. 히스토그램을 보면 왼쪽으로 크게 치우친다. 따라서 CK|(outcome = j)의 분포를 감마분포로 볼 수 있다. 따라서 3장에서 설명한 바와 같이, 감마분포를 따를 경우에는 설명변수를 CK 뿐만 아니라 log(CK)에 대해서도 생각한다. 그리고 두 분포의 조건부확률밀도함수 그래프를 보면 두 분포가 상당히 중복되게 보인다. 따라서, 감마분포에서의 로지스틱회귀모형분석에 의해 변수 CK와 log(CK)가 모두 유의적인 것으로 예상할 수 있다.

CK와 log(CK)를 설명변수로 하는 로지스틱 회귀분석을 하면 다음 표 1과 같다. CK의 p-값은 0.0150이므로 유의하고, log(CK)의 p-값도 0.0312로 유의하다. 따라서, 두 분포 CK|(outcome = j), $j = \{0, 1\}$가 치우진 경우에서는 변수 CK와 log(CK) 모두 필요하다고 말할 수 있다.

6. 결론

일반적으로 자료가 좌우대칭인 경우에는 정규분포로 설명이 가능하지만 한쪽으로 치우친 자료가

주어질 때, 감마분포가 정규분포보다 적절하다. 현실적으로 대부분의 자료는 좌우대칭이 아닌 경우가 많다. 그래서 그 자료들은 변수변환 등을 이용해 정규분포형태로 만든다. 그런데 변수변환 등을 하여서도 좌우대칭형태로 만들 수 없다면, 정규분포를 이용하여 설명하는 것이 적절치 못하다. 따라서 이 경우에는 확장된 개념인 감마분포에 대한 로그-밀도비가 쓰이게 된다. 본 논문에서는 감마분포에서의 로지스틱 회귀모형에서 로그-밀도비에 대해 알아보기 위해 여러 가지 모의실험을 수행하였다. 모의실험 결과를 보면, 로지스틱회귀모형에서 로그-밀도비를 이용해 변수를 적절히 선택할 수 있다.

우리는 $x|y = 0$과 $x|y = 1$의 두 분포가 겹치는 경우에서는 x항과 $\log(x)$항 모두 필요하다고 말할 수 있다. 그리고 두 분포가 분리된 경우에는 x항 또는 $\log(x)$항이 필요하다고 말할 수 있다. 형태모수 α의 변화는 두 분포의 확률밀도함수 그래프로 판단이 가능하다. 형태모수 α가 작을수록 그래프는 한쪽으로 치우치게 된다. 척도모수 λ의 변화를 두 분포의 확률밀도함수 그래프만으로 정확히 판단하기 힘들다. 따라서 정확한 판단을 위해서 로지스틱회귀모형에 대한 검정을 하거나 두 모형의 KLD의 차이 $(I_1 - I_2)$와 $(I_{1.1} - I_2)$를 살펴보면 어떤 변수가 필요한지를 판단할 수 있다.

참고 문헌

Clark, R. G., Henderson, H. V., Hoggard, G. K., Ellison, R. S. and Young, B. J. (1987). The ability of biochemical and haematological tests to predict recovery in periparturient recumbent cows, *New Zealand Veterinary Journal*, **35**, 126–133.

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

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

Kullback, S. (1959). *Information Theory and Statistics*, Wiley, New York.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **135**, 370–384.

Scrucca, L. (2003). Graphics for studying logistics regression models, *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.

2011년 11월 접수; 2011년 12월 수정; 2011년 12월 채택

Variable Selection with Log-Density in Logistic Regression Model

Myung Wook Kahng[1,a], Eun-Young Shin[a]

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

Abstract

We present methods to study the log-density ratio of the conditional densities of the predictors given the response variable in the logistic regression model. This allows us to select which predictors are needed and how they should be included in the model. If the conditional distributions are skewed, the distributions can be considered as gamma distributions. A simulation study shows that the linear and log terms are required in general. If the conditional distributions of $x|y$ for the two groups overlap significantly, we need both the linear and log terms; however, only the linear or log term is needed in the model if they are well separated.

Keywords: Binary response variable, inverse regression, Kullback-Leibler divergence, log-density ratio, logistic regression.

This research was supported by the Sookmyung Women's University Research Grants 2010.

[1] Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea. E-mail: mwkahng@sm.ac.kr