The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.2.277
27(2), 277–290

Comparison of Bias Correction Methods for the Rare Event Logistic Regression

Hyungwoo Kim[a] · Taeseok Ko[a] · No-Wook Park[b] · Woojoo Lee[a,1]

[a]Department of Statistics, Inha University
[b]Department of Geoinformatic Engineering, Inha University

(Received December 26, 2013; Revised March 25, 2014; Accepted March 31, 2014)

Abstract

We analyzed binary landslide data from the Boeun area with logistic regression. Since the number of landslide occurrences is only 9 out of 5000 observations, this can be regarded as a rare event data. The main issue of logistic regression with the rare event data is a serious bias problem in regression coefficient estimates. Two bias correction methods were proposed before and we quantitatively compared them via simulation. Firth (1993)'s approach outperformed and provided the most stable results for analyzing the rare-event binary data.

Keywords: Rare event, logistic regression, bias correction.

1. 서론

특정 지역에서 미래의 산사태에 취약한 지역을 찾는 산사태 취약성 분석(landslide susceptibility analysis)은 대상 지역의 토지 이용 및 관리를 위해 필수적이다. 이러한 산사태 취약성 분석은 과거 산사태 발생 지역과 산사태 발생과 관련이 있는 환경 변수와의 연관성 분석을 통해 수행된다. 이러한 연관성 분석에는 대표적으로 로지스틱 회귀(logistic regression) 모형이 사용될 수 있다. 산사태 취약성 분석에 사용되는 과거 산사태 발생 자료의 개수는 일반적으로 연구 지역 전체 영역에 비해 매우 작기 때문에, 산사태 발생 지역을 1로 표현했을 때 발생 자료(1)와 미발생 자료(0)의 비율이 매우 불균형적인 것이 특징이다. 가장 널리 쓰이는 분석 방법으로 로지스틱 회귀분석 모형이 고려되어왔지만, 1의 비율이 매우 낮은 희귀사건(rare event)의 경우 로지스틱 회귀분석모형에서의 최대가능도 추정량(maximum likelihood estimator)에 심각한 편의(bias) 문제가 발생할 수있음이 이미 알려져 있다 (King과 Zeng, 2001). 따라서 본 논문에서는 이 편의 문제를 다루기 위해 앞으로 소개될 두 가지 방법을 시뮬레이션된 자료를 통해 비교연구를 수행한 후에, 편의가 작은 방법을 이용하여 충청북도 보은 지역에서 랜덤하게 얻어진 5000개 지역을 대상으로 산사태 발생 여부에 영향을 주는 환경 변수를 분석하고자 하였다.

The first two authors contribute equally.

The work by Woojoo Lee was supported by INHA UNIVERSITY Research Grant (INHA-47275-01), and the work by No-Wook Park was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2012R1A1A1005024).

[1]Corresponding author: Department of Statistics, Inha University, 235 Yonghyun-Dong, Nam-Gu, Incheon 402-751, Korea. E-mail: lwj221@gmail.com

로지스틱 회귀분석은 이항 반응변수(y)와 설명변수 사이의 연관성을 분석하는 통계 모형으로 널리 사용되고 있다. 연구자가 흥미있어 하는 사건이 발생하는 경우 $y = 1$이라고 하고, 반대의 경우를 $y = 0$이라고 표현했을 때, 1의 비율이 0의 비율에 비해 매우 낮은 경우를 우리는 보통 희귀 사건이라고 한다. King과 Zeng (2001)은 희귀 사건에 의해 발생하는 로지스틱 회귀분석 모형의 통계적인 문제로 회귀계수 추정치의 편의문제를 지적하였다. 먼저 다음의 단순한 로지스틱 모형

$$\pi_i = P(y_i = 1|x_i) = \frac{\exp(\beta_0 + x_i)}{1 + \exp(\beta_0 + x_i)}$$

을 생각해보자. 여기서 P는 확률을 의미하고, 설명변수 x_i에 대한 회귀계수는 1로 고정된 것이며 n개의 랜덤샘플이 있다고 가정한다. 이 때 King과 Zeng (2001)은 β_0의 최대 가능도 추정량의 편의가

$$E\left(\widehat{\beta}_0 - \beta_0\right) \approx \frac{\bar{\pi} - 0.5}{n\bar{\pi}(1 - \bar{\pi})}$$

으로 주어진다는 것을 밝혔다. 여기서 $\bar{\pi} = 1/n \sum_{i=1}^{n} \hat{\pi}_i$이다. 위의 식으로 부터 $\bar{\pi}$의 값이 0에 매우 가깝다면 $\bar{\pi}(1 - \bar{\pi})$의 값이 0에 매우 가깝게되고 따라서 편의가 매우 커지게 되어 문제가 심각해질 수 있음을 보여준다. 따라서 로지스틱 회귀분석의 최대가능도 추정량이 갖는 편의의 문제는 희귀사건에서 중요하게 다루어 져야 할 주제가 된다. 이 방법에 대해 King과 Zeng (2001)은 최대가능도 추정량으로부터 $O(1/n)$에 해당하는 편의항을 수정한 추정량을 제안하였고, 이 수정된 추정량이 편의뿐만 아니라 분산도 더 작게 해주는 좋은 결과가 있음을 밝혔다.

한편 Firth (1993)는 일반적인 문제에서 최대가능도 추정량이 갖는 $O(1/n)$ 편의항을 제거해줄 수 있는 벌칙 가능도 방법을 제안하였다. 이 방법은 로지스틱 회귀분석 모형에 적용될 수 있고, 또한 King과 Zeng (2001)에서 제안된 방법과 마찬가지로 $O(1/n)$ 편의항을 제거해주므로 서로 경쟁적인 대안관계로 생각된다. Heinze와 Schemper (2002)에서는 이 벌칙 가능도를 이용해서 로지스틱 회귀분석에서 어떠한 설명변수에 의해 0과 1이 완전히 분리(separation)가 되는 경우 회귀계수의 추정치가 발산하는 것을 막아주는 방법으로 활용하였다. 이 논문에서는 희귀사건을 다루는 로지스틱 회귀분석 모형에서 편의문제를 해결할 수 있는 두 가지 방법 - King과 Zeng (2001)의 방법과 Firth (1993)의 방법을 시뮬레이션 연구를 통해 비교분석 해 보고 어떠한 방법의 성능이 더 좋은지 판단하고자 한다. 앞으로 편의상 전자를 King 방법, 후자를 Firth 방법이라 명명하여 사용하겠다.

본 논문의 순서는 2절에서는 Firth의 방법과 King의 방법의 핵심적인 아이디어를 쉽게 전달하고, 3절에서 여러 시뮬레이션 세팅하에 로지스틱 회귀분석에서 편의를 줄여주는 두 가지 방법에 대해 (1) 제공하는 추정치의 분포의 특성, (2) 1종 오류, (3) 검정력의 관점에서 비교하는 연구를 진행할 것이다. 4절에서 실제 보은 지역의 산사태 자료를 이용하여 분석할 것이고, 5절에서 결론을 살펴 볼 것이다.

2. 희귀 사건 로지스틱 회귀계수의 편의를 수정하는 두 가지 방법

Firth의 방법과 King의 방법 모두 이론적으로는 로지스틱 회귀모형에서 최대가능도 추정량의 $O(1/n)$ 편의를 제거해주는 것으로 알려져 있기 때문에 성능이 비슷할 것으로 기대되지만 실제 희귀 사건 자료에 적용해 보게 되면 알고리즘의 안정성 측면에서 Firth의 방법이 훨씬 더 좋다는 것을 알 수 있다. 이를 이해하기 위해서 각 방법이 어떠한 방식으로 편의를 제거하는지에 대해 자세히 살펴보도록 하겠다.

2.1. Firth의 방법

Firth (1993)의 방법은 최대가능도 추정량의 $O(1/n)$크기의 편의를 제거하는 일반적인 방법으로 제안되었고, 스코어 함수의 기하적인 성질을 이용한 방법으로 이 절에서 그 아이디어를 설명해 보도록 하겠다.

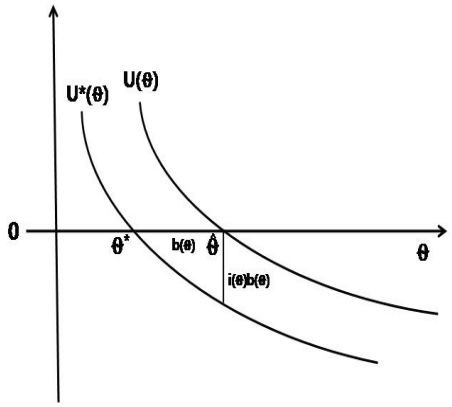

Figure 2.1. The relationship between the score function ($U(\theta)$) and the modified score function ($U^*(\theta)$)

먼저 θ를 관심있는 스칼라 모수라고 할때, 우리는 최대 가능도 추정치를 로그 가능도 함수의 일차 미분에 해당하는 스코어 함수

$$U(\theta) = 0$$

을 θ에 대해 풀어서 얻는다. 이때 얻어지는 값을 $\widehat{\theta}$라고 한다면 이는 참값 θ^*와 일반적으로 $O(1/n)$에 해당하는 편의 $b(\theta)$를 가지게 된다. 이는 Figure 2.1에 표현되어 있다.

그림으로부터 $\widehat{\theta}$의 편의 $b(\theta)$의 없애기 위해서 $U(\theta)$의 그래프를 아래쪽으로 조금 이동시켜서 스코어 함수가 θ축과 만나는 점이 참값이 되도록 맞추는 것이 Firth의 핵심 아이디어이다. 이를 구현하기 위해 먼저 점 θ에서 스코어 함수의 접선의 기울기의 절대값이 관측된 피셔 정보량(observed Fisher information)인 $i(\theta) = -U'(\theta)$임을 주목해보면, 스코어 함수가 편의 $b(\theta)$만큼 θ축과 평행하게 이동하기 위해서는 스코어 함수는 아래쪽으로 $b(\theta)i(\theta)$ 정도 이동하면 된다는 것을 것을 알수 있다. 따라서 원래의 스코어 함수에 $-i(\theta)b(\theta)$에 해당하는 양을 더해주어 다음의 수정된 스코어 함수를 정의한다.

$$U^*(\theta) = U(\theta) - i(\theta)b(\theta).$$

Firth (1993)에서는 이러한 수정된 스코어 함수로 부터 얻어진 추정량이 원래의 최대가능도 추정량이 갖는 $O(1/n)$의 편의를 제거할 수 있음을 밝혔다.

특히 이 논문의 관심사인 이항반응변수에 대한 로지스틱 회귀 분석의 경우 Firth의 방법이 어떻게 적용되는지를 살펴보도록 하겠다. 먼저 설명변수 x_i가 주어졌을 때 이항반응변수 $y_i = 1$의 확률을

$$\pi_i = P(y_i = 1|x_i)$$

으로 정의하면, 로지스틱 회귀 모형은

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\theta$$

으로 주어진다. 여기서 θ가 우리의 관심있는 모수가 된다.

McCullagh와 Nelder (1989)는 일반화 선형 모형군에 대해 다음의 편의 공식

$$b(\theta) = \left(X^T W X\right)^{-1} X^T W \xi \tag{2.1}$$

을 제공하였다. 여기서 표본의 수는 n, W는 $n \times n$ 대각행렬으로 i번째 성분이 $\pi_i(1-\pi_i)$이고, X는 i번째 행이 x_i^T으로 채워져 있는 행렬에 해당한다. $W\xi$는 그것의 i번째 원소가 $h_i(\pi_i - 1/2)$인데 h_i는

$$H = W^{\frac{1}{2}} X \left(X^T W X\right)^{-1} X^T W^{\frac{1}{2}}$$

의 i번째 대각선 성분을 뜻한다.

따라서 $U^*(\theta)$ 식에 $b(\theta)$와 $i(\theta) = X^T W X$를 넣어 보면

$$\begin{aligned} U^*(\theta) &= U(\theta) - \left(X^T W X\right)\left(X^T W X\right)^{-1} X^T W \xi \\ &= U(\theta) - X^T W \xi \end{aligned}$$

이 된다. 로지스틱 회귀분석에서

$$U(\theta) = X^T (y - \pi)$$

이므로

$$U^*(\theta) = X^T \left(y - \pi - h\left(\pi - \frac{1}{2}\right)\right) = X^T \left(y + \frac{h}{2} - (1+h)\pi\right)$$

가 된다. 이것은 결과적으로 보면 원래의 일반화선형모형의 회귀계수 θ를 구하는 식에서 y_i 대신에 $y_i + h_i/2$, 이항분포의 시행횟수인 1 대신 $1 + h_i$를 사용하는 것과 같으므로 기존의 피셔-스코어링 알고리즘을 약간의 수정만 하면 계속 활용할 수 있다는 장점을 가지고 있다. 추정치에 대한 표준 오차에 대해서 Firth (1993)은 원래 스코어 함수의 미분을 통해서 얻어지는 관측된 피셔 정보량의 역행렬의 대각성분에 제곱근을 취한 양을 사용할 것을 이야기하였다. 여기서 피셔 정보량을 $U^*(\theta) = 0$을 만족하는 점에서 계산한다는 것이 차이점이다. 이는 수정된 스코어 함수의 미분으로부터 얻어지는 표준오차의 좋은 근사로서 사용될 수 있음을 밝혔다. 본 논문에서도 Firth의 제안을 따르는 것으로 하였고, 이는 오픈 소스인 R에서 "logistf"라는 패키지에 구현되어 있다.

2.2. King의 방법

King과 Zeng (2001)에서는 사회과학 분야에서 나오는 희귀 사건을 분석하기 위해 본인들이 로지스틱 모형을 써본 결과, 편의의 문제가 심각함을 다양한 시뮬레이션 연구를 통해 밝혔다. 그리고 이러한 편의 문제를 해결하기 위한 방법으로 위에서 주어진 McCullagh와 Nelder (1989)의 편의공식을 최대가능도 추정치에서 직접 빼서 사용하는 것을 제안하였다. 즉,

$$\widehat{\theta}^{King} = \widehat{\theta} - b(\theta)$$

이 된다. 실제 구현을 위해서는 먼저 로지스틱 회귀분석이 가능한 일반적인 패키지를 이용하여 자료를 적합한다. 이 때 얻어진 최대가능도 추정치 $\widehat{\theta}$을 이용하여 식 (2.1)에 있는 W와 ξ를 계산한 후 편의 수정항 $b(\theta)$를 얻는다. 이를 원래 최대가능도 추정치에서 빼줌으로써 $\widehat{\theta}^{King}$을 쉽게 구할 수 있다. King과 Zeng (2001)에서는 무작위 추출법(random sampling) 뿐 아니라 환자군-대조군 연구(case-control study)와 같은 표본 기법에 대한 희귀사건 로지스틱 회귀분석 문제를 같이 다루었다. 특히 환자

군-대조군 연구를 통해 얻어지는 자료에 로지스틱 회귀모형이 적합되는 경우 절편에 큰 편의가 생길 수 있는데, 이 절편이 일치 추정량이 되도록 수정하는 방법을 해당 논문에서 제안하였다. 또한 사회과학에서 쓰이는 다양한 표본 추출 방식을 고려하였는데, 이 때 로그 가능도함수에 가중치를 고려하여 로지스틱 회귀분석의 모수를 올바르게 추정하는 방식에 대해서도 상세히 논하였다. 이 방법은 현재 R 패키지인 Zelig에 구현되어있다.

이론적으로는 Firth의 방법과 King의 방법 모두 최대우도 추정치의 $O(1/n)$ 편의를 제거해 주는 방법에 해당하므로 두 방법간의 차이가 크지 않을 것이라고 생각할 수 있으나 희귀 사건을 분석하는 경우에는 두 방법간 차이가 있을 수 있음에 주의하여야 한다. King의 방법은 추정치에 편의를 수정하는 방법이므로 최대가능도 추정치 자체가 먼저 존재해야 한다. 그러나 희귀 사건의 경우, 1과 0이 완벽히 분리되는 경우가 생길 수 있고 이때에는 최대가능도 추정치 자체가 존재하지 않게 되는데 이러한 경우는 편의 수정을 한다는 것 자체가 어렵게 된다. 반면, Firth의 방법은 스코어 함수 자체에 대한 수정을 이야기 하므로 1과 0이 완벽히 분리되는 경우라도 유한한 최대가능도 추정치를 제공하게 되어 더 안정적인 결과를 제공하는 알고리즘으로 생각할 수 있다. 따라서 두 방법은 최대가능도 추정치가 안정적으로 구해지는 경우에는 비슷한 추정치를 제공하겠지만, 최대가능도 추정치 자체가 안정적으로 구해지지 않는 경우에는 Firth의 방법이 훨씬 바람직한 성질을 가지고 있다고 볼 수 있다. 이러한 지적은 Heinze와 Schemper (2002)에서도 나타나고 있는데, 로지스틱 회귀분석의 경우 0과 1인 완전히 분리가 되는 경우, 가능도 함수가 단조증가하거나 감소하는 형태를 갖게되고 따라서 최대 가능도 추정치는 잘 정의가 되지 않음을 밝힌 후, 이러한 문제점을 보완하는 방법으로 Firth의 방법을 고려하였다. 그러나 Heinze와 Schemper (2002)의 연구에서는 희귀 사건에서 대해서 Firth의 방법이 어떤 면이 우수한지에 대한 정확한 논의가 결여되어있다. 본 논문에서는 먼저 희귀 사건 시뮬레이션 자료에 대해 실제 편의의 수준이 어느 정도인지, 제 1종 오류와 검정력은 어떤지 자세히 살펴보도록 하겠다.

3. 시뮬레이션 연구

희귀 사건에 대해서 위의 두가지 방법 및 일반적인 로지스틱 회귀분석의 성능을 비교하기 위해 시뮬레이션 연구를 진행하였다. 먼저 희귀사건을 생성하기 위한 시뮬레이션 세팅은 다음과 같다.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + x_i\beta_1,$$

여기서 $\pi_i = P(y_i = 1|x_i)$, y_i는 i번째 이항 반응변수, x_i는 i번째 설명변수이다. x_i는 표준정규분포에서 생성하였다. β_1이 우리의 관심사가 되는 모수이고, 귀무가설이 성립하는 경우와 대립가설이 성립하는 경우를 각각 살펴보기 위해 $\beta_1 = 0$인 세팅과 $\beta_1 = 1$인 세팅을 살펴보는 것으로 하였다. β_0는 1이 나오는 주변확률(marginal probability)이 각각

- $P(y_i = 1) = 0.01$
- $P(y_i = 1) = 0.05$
- $P(y_i = 1) = 0.1$

이 되도록 정하여 주었는데, 순서대로 그 값이 $\beta_1 = 0$일 때 $\beta_0 = -4.593, -2.943, -2.198$이 사용되었고, $\beta_1 = 1$일 때에는 $\beta_0 = -5.057, -3.351, -2.575$가 사용되었다. 일반적인 로지스틱 회귀분석은 R에서 glm 함수를 이용하였고, Firth의 방법과 King의 방법은 각각 logistf, Zelig 패키지를 이용하여 계산되었다. 전체 500번의 시뮬레이션을 실행한 결과를 보고하였고, 매회 사용된 표본의 수는 100개이다. 따라서 $P(y_i = 1) = 0.1$인 경우에는 평균적으로 10개 정도의 1이 나타나는 것으로 해석할 수 있다.

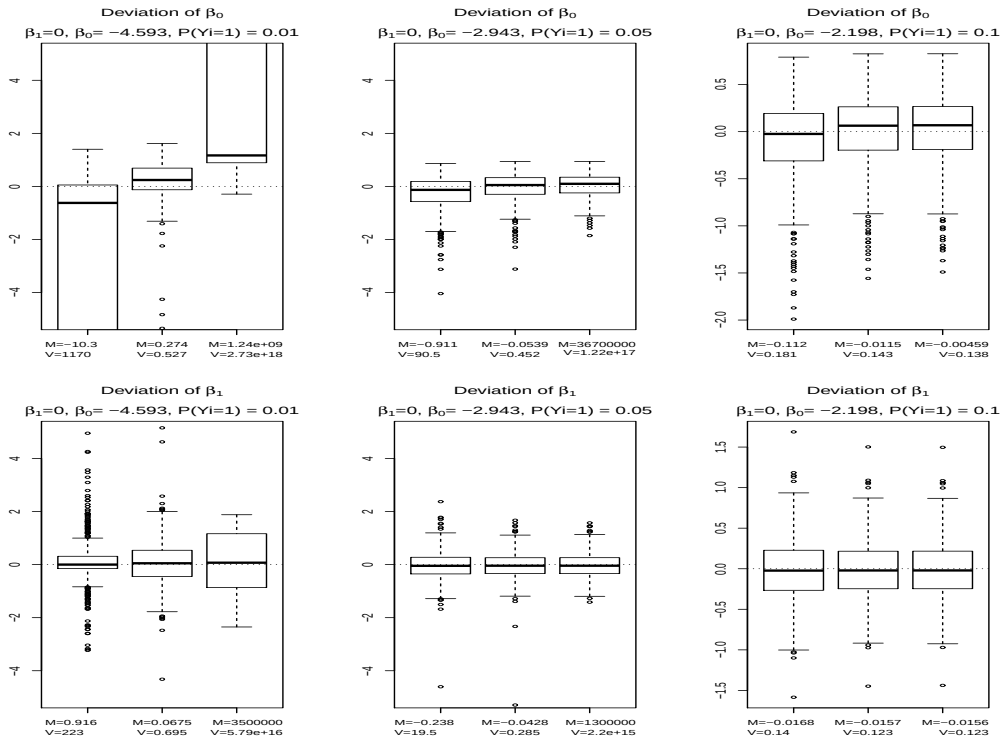

Figure 3.1. Boxplot for the difference between estimates from each method and the true value when $\beta_1 = 0$. From the leftmost case, the marginal probabilities of the event are given by 0.01, 0.05 and 0.1, respectively. In each plot, the results from ordinay logistic regression, Firth's method and King's method are shown in the order named.

첫 번째 시뮬레이션 결과는 $\beta_1 = 0$일 때 이고, Figure 3.1에 상자 그림으로 주어져있다. 각 상자그림 안에서는 순서대로 일반적인 로지스틱, Firth 방법, King 방법의 결과가 주어져있다. 그리고 맨 왼쪽 그림부터 순서대로 $P(y_i = 1)$이 0.01, 0.05, 0.1인 경우의 결과가 주어져있다. 각 그림의 아래에는 500번의 실험을 이행 후 회귀계수 추정치들의 평균(M)과 분산(V)을 나타내었다. 상자그림은 추정치에서 참값을 빼준 값에 대한 것으로 0에서 많이 벗어날 수록 편의가 큰 추정치가 얻어진 것으로 해석할 수 있다. 예상대로 로지스틱 회귀분석은 매우 희귀한 사건의 경우 매우 큰 편의와 분산을 보이고 있다. King의 방법은 로지스틱 회귀분석의 결과보다도 훨씬 안 좋은 결과를 보이고 있다. 얻어진 추정치 가운데에 이상치(outlier)가 너무 많아서 상자그림 전체를 보이게 되면 Firth 방법의 결과가 나타나지 않기 때문에 편의상 상자그림의 y축 범위를 -5에서 5로 제한하게 되었다. 따라서 $P(y_i = 1) = 0.05$인 경우를 볼 때 Firth의 방법과 King의 방법이 보이고 있는 상자 그림은 비슷해 보이지만, King의 경우에는 상자그림에서 나타나지 않는 매우 극단적인 이상치들이 여러개가 있으므로 유의해야한다. 이는 회귀 계수 추정치의 분산이 매우 큰 값을 통해 확인해 볼 수 있다. $P(y_i = 1) = 0.1$인 경우에도 상자그림에서 로지스틱 회귀분석의 중앙값이 다른 방법에 비해 나은 것처럼 보이지만 평균과 분산을 비교해보면 실제로는 그렇지 않음을 확인해 볼 수 있다.

β_0의 경우 과소추정되는 이상치가 많이 나타났는데, 그 이유는 이 논문에서 희귀사건을 고려하고 있기 때문이다. 극단적인 경우로 모든 y_i가 0으로 나타난 경우를 생각해 본 다면 β_0는 $-\infty$로 추정된다. 같

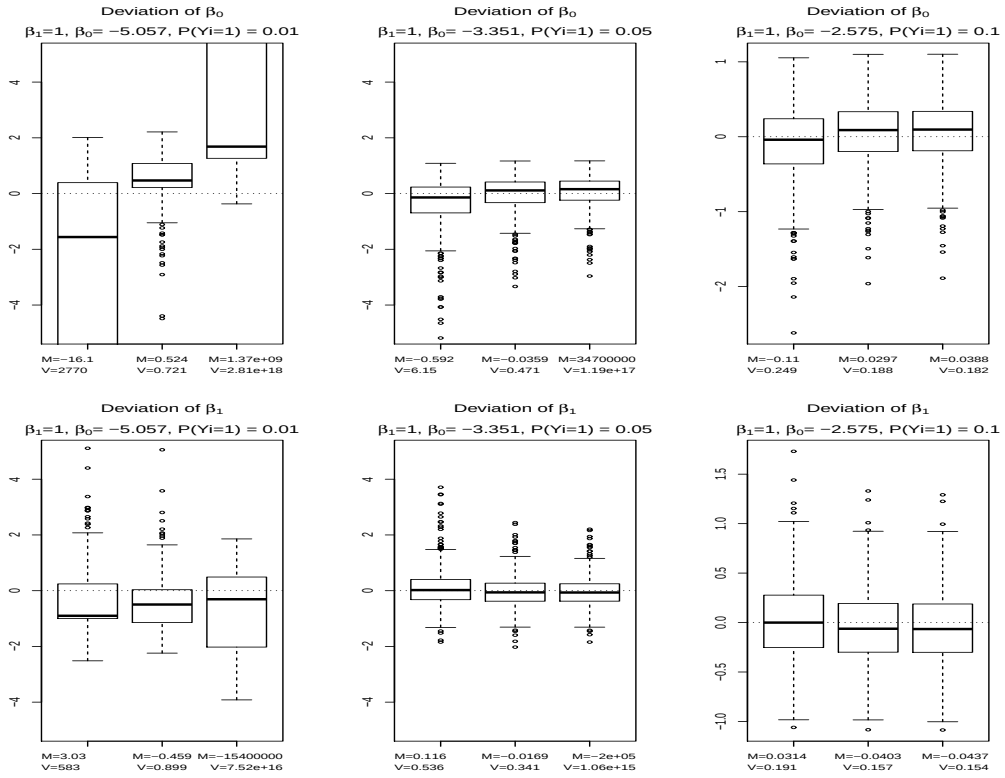

Figure 3.2. Boxplot for the difference between estimates from each method and the true value when $\beta_1 = 1$. From the leftmost case, the marginal probabilities of the event are given by 0.01, 0.05 and 0.1, respectively. In each plot, the results from ordinay logistic regression, Firth's method and King's method are shown in the order named.

은 선상에서 보면 $y_i = 1$의 개수가 매우 적은 경우에는 β_0의 추정치는 매우 큰 음수가 나오게 되므로 자연스럽게 과소추정되는 이상치가 많아 지는 것으로 이해할 수 있다.

전체적으로 Firth의 방법이 다른 두 방법에 비해 분산의 관점에서 훨씬 안정적이고 편의도 적은 것으로 나타나고 있다. 1이 나오는 확률을 더 키운 후 시뮬레이션을 진행하였을 때에는 방법 간의 차이가 많이 사라졌는데, 이 때는 점점 희귀사건에서 멀어지는 경우이므로 우리의 관심사가 아니기 때문에 자세한 내용은 생략하는 것으로 하였다.

두 번째 시뮬레이션 결과는 $\beta_1 = 1$일 때 이고, 첫 번째와 마찬가지로 추정치에서 참값을 뺀 값에 대해 상자 그림을 그린 후 Figure 3.2에 보고하였다. $P(y_i = 1)$이 0.01인 경우 로지스틱 방법과 King의 방법은 거의 사용할수가 없는 방법이 된다. 그러나 Firth의 방법은 다른 두 가지 방법에 비해 매우 안정적인 추정치를 주는 것으로 나타났다. 이러한 패턴은 대부분의 경우 유지되는 것으로 나타났고, 사건의 발생빈도가 높아지게 되면 방법간의 차이가 작아지는 것으로 나타났다. 전체적인 경향은 $\beta_1 = 0$일 때와 유사한 것으로 판단된다.

위의 시뮬레이션 결과를 이용하여 방법 간 제 1종 오류와 검정력의 관점에서 차이가 얼마나 나타나는지를 확인하였다. Figure 3.3을 살펴보면 위 패널의 세 그림은 귀무가설($\beta_1 = 0$) 하에서 500번의 실험

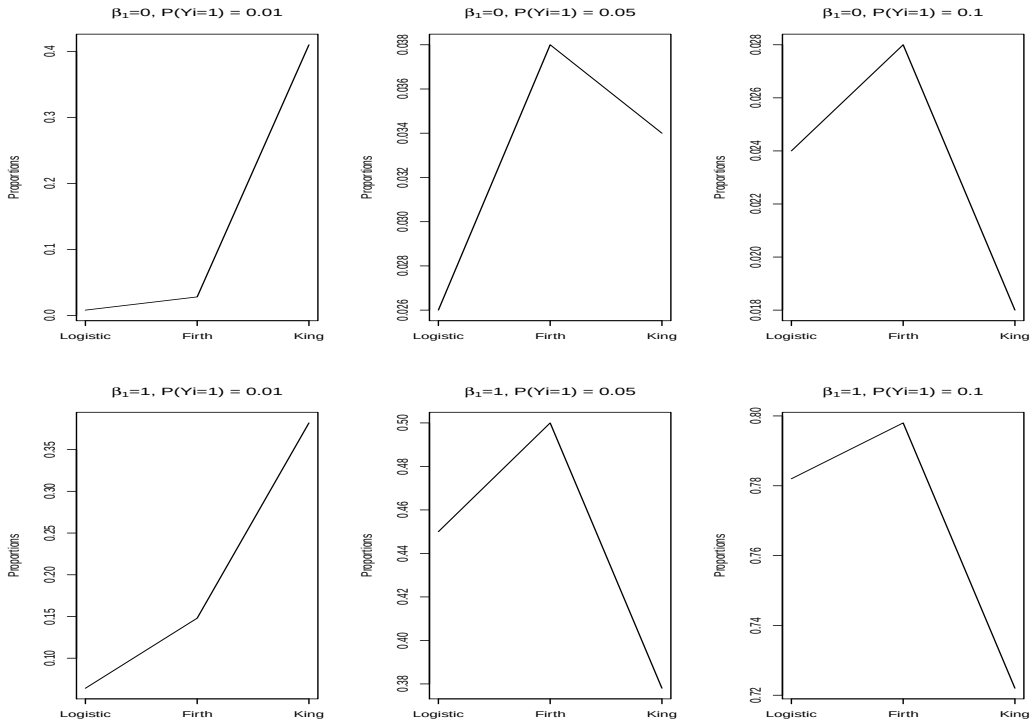

Figure 3.3. These figures show the type I error and power for each method when $n = 100$. The three plots in the upper panel are obtained when $\beta_1 = 0$, those in the lower panel are obtained from $\beta_1 = 1$. Out of 500 simulations, the proportion of the p-value less than 0.05 is shown, and in each plot, the results from ordinay logistic regression, Firth's method and King's method are shown in the order named. From the leftmost figure, $P(y_i = 1)$ is given by 0.01, 0.05 and 0.1, respectively.

가운데 유의확률(p-value) 값이 0.05보다 작은 경우의 비율이 어느 정도 인지를 나타내 주고있다. 먼저 Firth의 방법은 다른 두 방법에 비해 그 비율이 5%에 더 가까운 값을 유지하는 것으로 나타났고 로지스틱 회귀분석에 비해 덜 보수적인 판단을 하는 것으로 나타났다. King의 방법은 $P(y_i = 1)$이 0.01일때는 전체의 40%가 넘게 기각하였다. 따라서 심각하게 1종 오류가 크므로, 희귀사건에 사용할 때 매우 주의해야할 필요가 있는 것으로 보인다. $P(y_i = 1)$인 0.05나 0.1일때는 기각의 횟수가 Firth의 방법에 비해 적은 것으로 나타났으므로, 보수적인 판단을 한다고 볼 수 있다. 이러한 역전 현상을 보이는 이유를 파악하기 위해 시뮬레이션 결과를 자세히 살펴보니, King의 방법은 모든 y_i 값이 0으로 주어진 경우 매우 극단적이 추정치가 나오고 이 때 거의 대부분의 경우 귀무가설 $H_0 : \beta_1 = 0$을 기각하는 것으로 밝혀졌다. 예를들어 $P(y_i = 1)$이 0.01인 경우 모든 y_i가 0인 경우는 전체 500회의 시뮬레이션 중 197회로 나타났고, 이 때 197번 모두 $H_0 : \beta_1 = 0$을 기각하였다. 모든 y_i가 0인 경우는 $P(y_i = 1)$의 확률이 커지면서 급속히 줄기 때문에 King의 방법의 기각하는 횟수가 급격히 줄게되어 $P(y_i = 1)$이 0.05거나 0.1일 때에는 보수적으로 판단하는 것으로 나타났다.

아래 그림 세개는 $\beta_1 = 1$일 때 기각된 비율을 나타낸 것으로 검정력을 확인해 볼 수 있다. $P(y_i = 1)$이 0.01일때 King의 방법이 가장 높은 검정력을 보이나 이는 1종 오류를 심각하게 키우고 있는 상황에서 얻어진 값이므로 의미있게 해석할 수 없다. 그외의 세팅에서는 일관적으로 Firth의 방법이 가장 높은

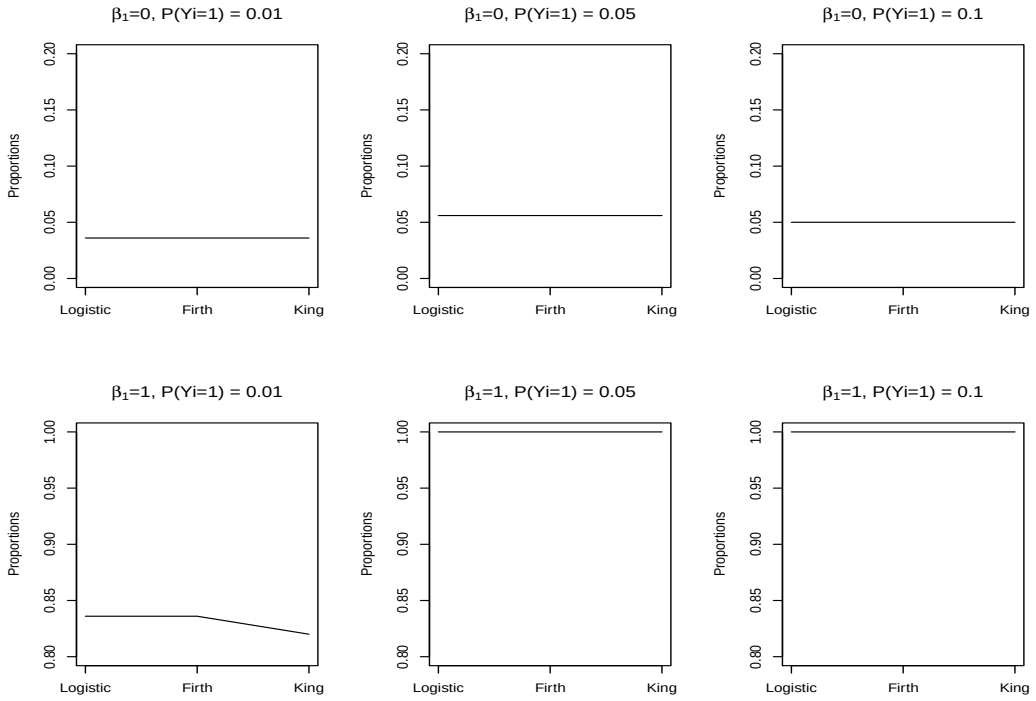

Figure 3.4. These figures show the type I error and power for each method when $n = 1000$. The three plots in the upper panel are obtained when $\beta_1 = 0$, those in the lower panel are obtained from $\beta_1 = 1$. Out of 500 simulations, the proportion of the p-value less than 0.05 is shown, and in each plot, the results from ordinay logistic regression, Firth's method and King's method are shown in the order named. From the leftmost figure, $P(y_i = 1)$ is given by 0.01, 0.05 and 0.1, respectively.

검정력을 보이고 있고, King의 방법이 가장 낮은 검정력을 보여주는 것으로 나타났다.

표본의 수가 클 때 방법간 차이가 있는지의 여부를 살펴보기 위해 n을 1000으로 늘린 후 같은 시뮬레이션 세팅을 반복한 후, 그 결과를 Figure 3.4에 보고하였다. 표본의 수가 커지게 되면 모든 y_i가 0인 경우는 발생하지 않게 되므로 방법 간의 큰 차이는 나타나지 않았으며, Figure 3.4에서 보이는 것처럼 실제로 세가지 방법은 서로 거의 유사한 성능을 보이는 것으로 나타났다.

시뮬레이션 연구로부터 우리는 다음을 확인하였다.

- 희귀사건 이항자료를 다룰 때 고려된 분석 방법 중에서 Firth의 방법이 추정치의 편의와 분산을 줄이는 성능이 매우 우수하다.
- Firth의 방법이 제 1종 오류를 가장 잘 조절하고 있으며, 이 조건하에서 검정력이 가장 높다.
- 방법 간 성능의 차이는 표본의 개수가 작을 때 특히 두드러지나, 표본의 수가 커짐에 따라 그 차이는 사라지게 된다.

아울러, Heinze와 Schemper (2002)가 지적한것처럼 Firth의 방법은 다른 방법에 비해 회귀 계수의 이상치를 훨씬 더 적게 주고있다.

Table 4.1. Descriptive statistics for the continuous covariates

	Slope	Elevation	Distance
Min	0	139	0
Q1	2	162	3
Median	11	195	7
Mean	12.29	211.5	11.17
Q3	20	245	14
Max	56	494	86

Table 4.2. Frequency table for each category of the nominal variables

forest	1	2	3	4	5	6	7	8	9	10	11	12	13
Observations	2534	14	142	352	34	2	865	26	118	87	3	818	5
geology	1	2	3	4	5								
Observations	1366	523	29	207	2875								
soil	1	2	3	4	5	6							
Observations	171	775	469	1440	2005	140							

4. 사례연구 - 보은 지역의 산사태 자료

이 연구에서는 앞의 실험 결과를 바탕으로 1998년 여름 집중호우로 인한 산사태 발생으로 많은 피해를 입은 충청북도 보은 지역에서 임의로 선택된 5000개 지역을 대상으로 사례연구를 수행하였다. 5000개 지역 중 산사태 발생 지역은 총 9곳으로 1의 비율이 0.0018에 불과하다. 대상 지역에서는 집중호우로 인해 산사태가 유발되었는데, 동일한 강우량에도 불구하고 특정 지역에서만 산사태가 발생했다는 사실은 산사태 발생이 강우 이외의 다른 환경 요인에 의해서도 발생한다는 것을 의미한다. 이러한 산사태 발생 환경 요인으로 이 연구에서는 선행 연구 (Park 등, 2003; Lee 등, 2004)를 고려하여 총 6개의 환경 변수를 분석에 사용하였다. 연속형 변수로는 고도(elevation), 사면 경사(slope), 선구조로부터의 거리(distance)의 3가지를 고려하였다. 일반적으로 사면 경사가 급하고, 풍화 정도가 심한 선구조로부터 가까운 지역이 산사태 발생에 더 취약하다. 범주형 변수로는 지질(geology), 임상(forest), 토양 배수(soil drainage)의 3가지를 고려하였다. 여기서 지질은 5개, 임상은 13개, 토양 배수는 총 6개의 범주로 구성되어 있고 모두 명목형 변수로 고려하였다. 각 범주의 이름과 의미는 맨 아래쪽의 부록에 상세하게 제시하였다. 특정 지질 종류, 임상 종류에 따라 산사태 발생 빈도가 달라지는데, 대상 지역에서는 화강암 지역 및 침엽수림에서 산사태가 많이 발생하였다. 토양 배수의 경우, 배수가 잘 될수록 산사태 발생 빈도가 높아지는 것으로 알려져 있다 (Lee 등, 2004). 위의 설명 변수들에 대한 기술 통계량은 Table 4.1과 Table 4.2에 주어져있다. 연속형 변수에 대해서는 최소값, 최대값, 중앙값 등을 보고하였고, 범주형 변수에 대해서는 각 범주 별 관측치 수를 보고하였다.

이 산사태 자료를 분석하기 위해 우리는 두가지 방법 - 일반적인 로지스틱 회귀분석과 Firth의 방법을 고려하였다. 이미 시뮬레이션 연구에서 King의 방법은 Firth의 방법에 비해 충분한 경쟁력을 갖추고 있지 못한 것으로 판명되었으므로 여기서는 고려하지 않도록 하겠다. 분석 결과는 Table 4.3에 정리하여 제시하였다. 각 분석 방법을 통해 얻어진 추정치(estimate), 표준오차(se), 유의확률(p-value)를 보고하였다. 유의하지 않은 변수가 많기 때문에 단계적 변수 선택법(stepwise variable selection)을 이용한 결과를 함께 살펴보았다. 단계적 변수선택법에서 변수 선택 기준은 유의수준 0.05로 하였다.

먼저 변수선택 이전에 두 방법으로 부터 얻어진 결과를 살펴보자. 로지스틱 회귀분석의 경우 유의수준 0.05하에서 통계적으로 유의하다고 판단되는 변수는 임상의 10번째와 12번째 범주 두 개 뿐이었다. 반

Table 4.3. Analysis result for Boeun landslide data

Variable name	logistic regression				Firth method			
	Estimate	Se	P-value	Stepwise(se)	Estimate	Se	P-value	Stepwise(se)
slope	0.000	0.000	0.4699	x	0.018	0.033	0.677	x
elevation	0.000	0.000	0.1513	−0.000(0.000)	−0.012	0.008	0.288	x
distance	0.000	0.000	0.100	x	−0.298	0.094	0.007	−0.298(0.112)
geology(2)	−0.003	0.003	0.356	x	−2.327	1.756	0.280	x
geology(3)	−0.006	0.008	0.455	x	−0.716	1.615	0.676	x
geology(4)	−0.002	0.003	0.570	x	−1.977	1.025	0.115	x
geology(5)	−0.003	0.002	0.179	x	−2.103	0.809	0.060	x
forest(2)	−0.000	0.011	0.950	x	4.144	1.671	0.096	x
forest(3)	−0.000	0.004	0.875	x	1.882	1.576	0.360	x
forest(4)	0.000	0.003	0.876	x	1.214	1.524	0.535	x
forest(5)	0.002	0.008	0.801	x	4.275	1.640	0.091	x
forest(6)	−0.000	0.030	0.998	x	7.260	2.377	0.018	6.348(2.287)
forest(7)	0.003	0.002	0.206	0.004(0.002)	2.443	1.077	0.075	x
forest(8)	0.000	0.009	0.956	x	4.034	1.706	0.103	x
forest(9)	0.008	0.004	0.059	0.009(0.004)	2.789	1.125	0.048	2.597(1.147)
forest(10)	0.012	0.005	0.010	0.013(0.005)	4.535	1.204	0.006	3.357(1.153)
forest(11)	0.005	0.025	0.854	x	11.066	2.774	0.004	8.332(2.366)
forest(12)	0.005	0.002	0.037	0.006(0.002)	2.437	1.014	0.042	1.840(0.928)
forest(13)	−0.002	0.019	0.933	x	4.929	1.908	0.061	x
soil(2)	0.000	0.004	0.946	x	−1.381	1.872	0.507	x
soil(3)	0.001	0.004	0.792	x	−0.200	1.830	0.921	x
soil(4)	0.002	0.004	0.630	x	−0.458	1.551	0.793	x
soil(5)	0.003	0.004	0.483	x	−0.163	1.587	0.929	x
soil(6)	−0.000	0.005	0.966	x	−0.094	1.953	0.963	x

면, Firth의 방법은 연속형 변수인 선구조로부터의 거리와 범주형 변수 임상에서 5개의 범주, 총 6개가 유의하게 나타났다. 이것이 의미하는 바는 두 가지 분석 방법이 단순히 추정치에 있어 약간의 차이가 나는 것을 넘어서 결론에 어떠한 변수가 영향을 주는지 판단할 때 질적인 해석의 차이가 나타날 수 있는 가능성을 보여준다. 변수 선택 이후에서도 여전히 두 방법간에는 큰 차이가 나타나는데, 전체적으로 Firth의 방법이 훨씬 유의한 결과를 주는 것으로 보인다. Firth의 방법에서 선구조로부터의 거리에 해당하는 회귀계수는 −0.298로 distance가 한 단위 커질수록 산사태가 발생할 오즈(odds)는 다른 변수가 고정되었을 때 $\exp(-0.298) = 0.742$배가 되는 것으로 나타났다. 이는 선구조로부터 가까운 지역이 산사태 발생에 더 취약하다고 하는 지질학적 사실과 일치한다 (Park 등, 2003). 임상과 관련하여 다양한 범주가 유의한 것으로 나타났는데, 두 방법간 추정치의 큰 차이가 나타나고 있음을 Table 4.3에서 살펴볼 수 있었다. 세부적인 해석은 위에서 언급하였듯이 오즈의 관점에서 쉽게 가능하므로 논문의 간결성을 위해 생략하였다.

5. 결론

본 논문에서는 희귀 사건을 다룰 때 로지스틱 회귀분석에서 나타날 수 있는 편의 문제를 다루는 두 가지 방법에 대해 살펴보았다. 시뮬레이션 연구를 통해 살펴본 바에 따르면, 추정치가 아니라 스코어 함수를 수정함으로써 추정치를 구하는 알고리즘이기 때문에 큰 안정성을 갖는 Firth의 방법이 실제 훨씬

성능이 좋음을 알 수 있었다. 또한 보은 지역의 산사태 자료를 분석하면서 방법 간 큰 차이가 나타났고, 따라서 Firth의 방법에서 얻어진 결과를 해석하는 것이 더 신뢰할 수 있는 것으로 판단된다. 비록 아직 제한적인 시뮬레이션 연구, 특정한 자료를 분석한 것이라는 제약이 있기 때문에 좀 더 일반적인 결론을 얻기 위해서는 희귀 사건을 좀 더 다양한 각도에서 검토할 필요가 있는 것으로 사료된다. 특히 얼마나 $y_i = 1$의 비율이 낮아야 희귀사건으로 볼 수 있는가에 대한 합의가 아직 분명하지 않기 때문에 실제 자료의 분석에서 언제 Firth의 방법을 적용해야 이득을 얻을 수 있는가에 대한 논의가 좀 더 분명하게 이루어지는 것이 중요하다고 판단된다. 한편 희귀 사건을 비율이 아니라 1의 개수로 정의하는 것이 더 합리적인 접근 방식일 수 있으므로, 이에 대한 별도의 검토도 필요할 것으로 생각된다.

부록: 범주형 설명변수의 각 범주별 이름 및 의미

Geology	1	Alluvium	충적층
	2	Hwanggangri formation	황강리퇴적층
	3	Acidic dyke	산성암맥
	4	Two mica admellite	죽전리운모화강암
	5	Biotite granite	흑운모화강암
Soil drainage	1	etc	기타
	2	Somewhat poorly drained	배수 약간 불량
	3	Moderately well drained	배수 중간
	4	Well drained	배수 양호
	5	Excessively drained	배수 매우 양호
	6	Poorly drained	배수 불량
Forest	1	Non-forest	비산림
	2	Rigida pine	리기다 소나무
	3	Pine	소나무
	4	Neelde and broad-leaf	침활혼효림
	5	Artificially afforested broad-leaf tree	인공 활엽수
	6	Korea nut pine	잣나무
	7	Larch	낙엽송
	8	Broad-leaf tree	활엽수
	9	Field	초지
	10	Cultivated land	경작지
	11	Chestnut tree	밤나무
	12	Poplar	포플라
	13	Ranch	제지

References

Firth, D. (1993). Bias reduction of maximum likelihood estimates, *Biometrika*, **80**, 27–38.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression, *Statistics in Medicine*, **21**, 2409–2419.

King, G. and Zeng, L. (2001). Logistic regression in rare event data, *Political Analysis*, **9**, 137–163.

Lee, S., Choi, J. and Min, K. (2004). Probabilistic landslide hazard mapping using GIS and remote sensing data at Boeun, Korea. *International Journal of Remote Sensing*, **25**, 2037–2052.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2nd ed, Chapman and Hall, London.

Park, N. W., Chi, K. H., Chung, C. F. and Kwon, B. D. (2003). GIS-based data-driven geological data integration using fuzzy logic: theory and application, *Economic and Environmental Geology*, **36**, 243–255.

희귀 사건 로지스틱 회귀분석을 위한 편의 수정 방법 비교 연구

김형우[a] · 고태석[a] · 박노욱[b] · 이우주[a,1]

[a]인하대학교 통계학과, [b]인하대학교 지리정보공학과

(2013년 12월 26일 접수, 2014년 3월 25일 수정, 2014년 3월 31일 채택)

요 약

본 연구에서는 로지스틱 회귀 모형을 이용하여 보은 지방의 산사태 자료를 분석하였다. 5000 지역의 관측치 가운데 단 9개만이 산사태 발생 지역이므로 이 자료는 희귀 사건 자료로 간주될 수 있다. 로지스틱 회귀 분석 모형이 희귀 사건 자료에 적용될 때 주요 이슈는 회귀 계수 추정치에 심각한 편의 문제가 생길 수 있다는 것이다. 기존에 두 가지의 편의 수정 방법이 제안되었는데, 본 논문에서는 시뮬레이션을 통해 정량적으로 비교 연구를 진행하였다. Firth (1993)의 방식이 다른 방법에 비해 우수한 성능을 보였으며, 이항 희귀 사건을 분석하는 데 있어서 매우 안정된 결과를 보여주었다.

주요용어: 희귀 사건, 로지스틱 회귀모형, 편의 수정.

처음 두 명의 저자는 동등하게 논문에 참여하였습니다.

이 논문에서 이우주의 참여는 인하대학교 교내연구비의 지원(INHA-47275-01)을 받아 수행되었으며, 박노욱의 참여는 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단 기초연구사업(NRF-2012R1A1A1005024)의 지원을 받았습니다.

[1]교신저자: (402-751) 인천광역시 남구 용현동 235, 인하대학교 통계학과. E-mail: lwj221@gmail.com