The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.269
28(2), 269–279

Likelihood-Based Inference of Random Effects and Application in Logistic Regression

Gwangsu Kim[a,1]

[a]Department of Statistics, Korea University

(Received March 17, 2015; Revised March 30, 2015; Accepted March 30, 2015)

Abstract

This paper considers inferences of random effects. We show that the proposed confidence distribution (CD) performs well in logistic regression for random intercepts with small samples. Real data analyses are also done to identify the subject effects clearly.

Keywords: Confidence distribution, generalized linear mixed effects model, logistic regression, predictive likelihood, prediction interval, random effects.

1. 서론

통계적 모형이 $f_\theta\,(y_1,\ldots,y_J)$로 주어져 있다고 하자. 여기에서 f_θ는 모수 θ를 가지는 확률밀도함수이며 $y_i \in \mathbb{R}$이다. 임의표본이 $(y_1,\ldots,y_J)^T$로 관측되고 추정하고자 하는 모수 θ의 차원이 p라고 하면 최대우도추정치(maximum likelihood estimate)는

$$\arg\max_\theta \log f_\theta(y_1,\ldots,y_J)$$

이 된다. 그리고 만약에 임의표본이 독립이며 동일한 분포에서 생성되었다고 한다면

$$\log f_\theta(y_1,\ldots,y_J) = \sum_{i=1}^{J} \log f_\theta(y_i)$$

로 되며, 최대우도추정량(maximum likelihood estimator)의 점근분포(J가 커지는데 따른 극한분포)는 적절한 조건에서 평균 벡터가 θ_0, 공분산 행렬이 $I(\theta_0)^{-1}/J$가 되는 다변량 정규분포가 된다는 것이 잘 알려져 있다. 여기에서 θ_0는 θ의 참값이며

$$I(\theta_0) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f_\theta(y)\right]\Bigg|_{\theta=\theta_0} \quad (y \sim f_{\theta_0})$$

이다. 하지만 주어진 자료가 독립이 아닌 경우가 자료분석에서 많이 나타나게 된다. 임의효과(random effects)는 주어진 자료 중 같은 개체로부터 생성된 자료가 있어 자료 생성에서 (독립이 아닌) 특정한 구

[1]Department of Statistics (BK21 plus Data Engineering Team for Complex Structured Data Analysis), Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 136-075, Korea. E-mail: gwangsu_2014@korea.ac.kr

조의 임의성이 개입하게 되는 경우 등에 사용하게 된다. 이런 임의효과는 잠재변수(latent variable)를 비롯한 다양한 이름으로 불리고 있으며 이에 대한 여러 연구들은 Skrondal과 Rabe-Hesketh (2007), Little과 Rubin (2002) 그리고 Rubin (2006) 등을 참고할 수 있다.

이 논문에서는 같은 개체 내지 동질성이 강한 집단으로부터 반복된 자료가 생성되어 임의효과가 있는 모형을 고려해 보려고 한다. 만약에 개체 내지 동질성이 강한 집단이 J개가 있어서 이로부터 각각 s개의 관측치가 관찰되어 자료의 크기가 $N = Js$라고 하자. 각각의 개체 내지 동질성이 강한 집단별로 임의효과 u_i가 있으며 여기에서 $\{y_{ij}\}_{j=1}^{s}$가 생성되고 공변량(covariates)들은 x_{ij}와 z_i라고 하자. 그러면 우리가 고려하는 모형은 특정한 우도함수(likelihood function) 하에서 다음과 같다.

$$E[y_{ij}|u_i] = \mu_{ij}, \quad g(\mu_{ij}) = \beta_0 + x_{ij}^T\beta + z_i^T u_i, \quad u_i \overset{i.i.d.}{\sim} f_\lambda, \tag{1.1}$$

여기에서 y_{ij}에 임의성을 주었으며 g는 연결함수(link function), β와 λ는 모수이다. 앞으로

$$\theta = \left(\beta_0, \beta^T, \lambda^T\right)^T$$

이며 y_i, y와 u는 각각 $(y_{i1}, \ldots, y_{is})^T$와 $(y_1^T, \ldots, y_J^T)^T$ 그리고 $(u_1^T, \ldots, u_J^T)^T$를 나타낸다고 하자. 모형 (1.1)은 Lee와 Nelder (1996)가 제안한 계층적 일반화선형모형(hierarchical generalized linear model)의 일종으로 볼 수 있으며 y_i가 u_i에 대해 조건부 독립이라고 한다면 확장우도함수(extended likelihood function)는

$$\prod_{i=1}^{J}\left[\prod_{j=1}^{s} f_\theta(y_{ij}|u_i)\right] f_\theta(u_i)$$

가 된다. 이런 우도함수는 모수들과 임의효과에 대한 추론을 가능하게 하는 정보를 담고 있다 (Birnbaum, 1962; Pawitan, 2001). 구체적으로 임의효과의 임의성에 대한 추론은

$$f_{\theta_0}\left(u_i|y_i\right)$$

를 이용하게 되며 여기에서 임의효과에 대한 확률적인 예측구간(prediction interval)을 구할 수 있게 된다. 그런데 문제는 θ_0를 모른다는 것인데 주어진 자료만을 이용해서

$$f\left(u_i|y_i\right)$$

를 만들고 이것이 $f_{\theta_0}(u_i|y_i)$로 수렴한다면 임의효과에 대한 추론이 가능할 것이다.

임의효과에 대한 추론은 실제로 관측이 되지 않았으나 실현된 값에 대한 추론 문제도 있으며 임의효과에 대한 두 가지 측면에 대해서는 Robinson (1991)을 참고할 수 있다. 본 논문에서는 실현된 값 자체에 대한 추론은 논외로 하고 임의효과의 임의성에 대한 추론에 초점을 맞추고자 한다.

2. 예측우도(predictive likelihood)와 예측구간(prediction interval)

예측우도는 주어진 자료로부터 미래의 관측되지 않은 확률변수의 분포를 알아내는 문제에 사용되었으며 관측된 모든 자료들과 미래의 값이 동일하고 독립인 확률분포에서 생성되었다고 가정하였다. 이런 예측우도를 관측되지 않은 값들에 대한 추론 문제로 확장하려는 것이 본 논문의 기본 목적이다. Bjørnstad (1990)의 정의를 따라 임의효과에 대한 예측우도를 정의하면

$$L\left(u_i|y_i\right)$$

의 형태로 표현되며 적분을 하였을 경우 1이 되는 것이다(예측우도는 예측에 사용되기에 예측분포, predictive distribution이라고도 불릴 수 있음에 유의). 물론 예측우도로는 적절한 조건을 만족하면서 $f_{\theta_0}(u_i|y_i)$로 수렴하는 것을 사용해야 할 것이다. 예측우도는 여러 방법으로 만들 수 있으며 완비 충분통계량(complete sufficient statistics)을 사용하는 경우를 제외한다면 예측우도는 다음과 같이 표현될 수 있다.

$$L(u_i|y_i) = \int f_\theta(u_i|y_i)g(\theta|y)d\theta.$$

예측우도는 $g(\theta|y)$를 무엇으로 사용하는 가에 따라 구분되어 진다. 만약에 디락 함수(Dirac function)

$$\delta_{\hat{\theta}}(\theta) \quad \left(\hat{\theta} = \arg\max_\theta \sum_{i=1}^{J} \log \int f_\theta(y_i|u_i)f_\theta(u_i)du_i\right) \tag{2.1}$$

를 사용한다면 플러그인(plug-in) 방법으로 예측우도가 만들어지는 것이며 사전분포(prior distribution) $\pi(\theta)$를 사용한다면 사후예측분포(posterior predictive distribution)로 예측우도가 만들어지는 것이다. 이런 방법 외에도 여러 가지 형태의 예측우도를 만들 수 있으며 예측우도가 $f_{\theta_0}(u_i|y_i)$에 대한 일치성을 보장하기 위해서는

$$\int |L(u_i|y_i) - f_{\theta_0}(u_i|y_i)|\, du_i$$

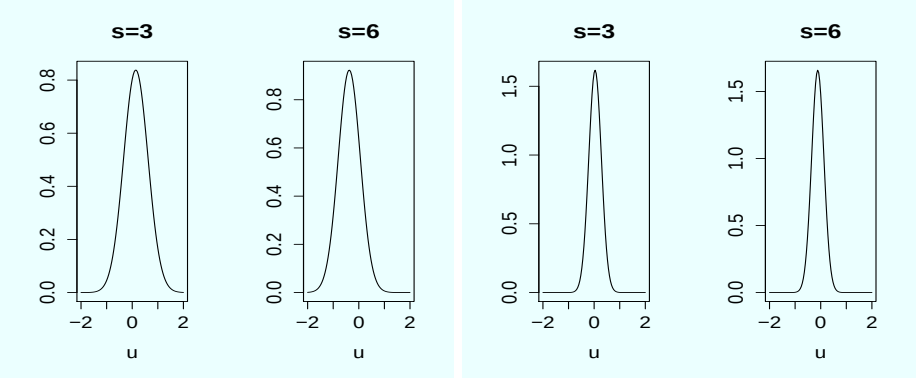
가 0으로 확률수렴해야 할 것이다. 여기에서는 y_i에 임의성을 주고 있다. 만약에 이런 확률수렴이 보장된다면 예측우도에서 만든 모든 예측구간은 $f_{\theta_0}(u_i|y_i)$로 만드는 모든 예측구간에 대해 균등하게 확률수렴이 됨을 쉽게 보일 수 있다. 즉 위의 확률수렴이 보장된다면 신뢰성 있는 예측구간을 만들 수 있는 것이다. 그런데 예측우도를 만드는데 완비 충분통계량의 사용은 사용할 수 있는 통계량에 제한이 있다는 단점이 있으며, 플러그인 방법은 추정에서의 분산을 고려하지 않고 있어서 지나치게 짧은 예측구간을 줄 수 있다는 단점이 있다. 사전 분포를 이용하는 것도 대안이 될 수 있으나 어떤 사전 분포를 사용하는 지에 대한 논점이 남아 있다. 그런 것으로 본 논문에서는 우도함수로부터 도출되는 정보만을 이용하려고 하며 신뢰분포(confidence distribution)를 주되게 사용하려고 한다. 이에 대한 여러 연구는 Xie와 Singh (2013)을 참고할 수 있으며 구체적으로 신뢰분포 $C(\theta|y)$는 p가 1인 경우 다음과 같이 정의된다.

$$\int_{-\infty}^{\theta} c(s|y)ds = P_\theta(T > t),$$

여기에서 T는 θ에 대한 일치추정량(consistent estimator)이며 t는 T의 관측치로 $P_\theta(T > t)$는 θ에 대한 단조증가함수가 되어야 한다. 그리고 p가 2이상인 경우는 데카르트 곱(Cartesian product)를 이용해서 정의한다. 만약에 (2.1)의 오른쪽에 있는 우도함수를 이용해서 구한 최대우도추정량을 T로 사용하고 그 점근분포를 이용한다면 $c(\theta|y)$는 평균 벡터가 최대우도추정치 $\hat{\theta}$, 공분산 행렬이 $\mathcal{I}(\hat{\theta})^{-1}$인 다변량 정규분포의 확률밀도함수가 된다. 여기에서

$$\mathcal{I}(\hat{\theta}) = -\frac{\partial^2}{\partial\theta\partial\theta^T}\sum_{i=1}^{J} \log f_\theta(y_i)|_{\theta=\hat{\theta}},\ f_\theta(y_i) = \int f_\theta(y_i|u_i)f_\theta(u_i)du_i$$

로 $\mathcal{I}(\hat{\theta})$는 양정치 행렬(positive definite matrix)임을 가정한다. 적절한 조건 하에서 신뢰분포가 가지는 점근적 성질로 인하여 예측구간의 일치성은 보장되며 신뢰분포와 베이지안 사후분포의 점근적 성질은 관련이 깊을 것으로 보인다. 한편 우리가 고려하는 있는 모형에서는 모든 i와 j에 대해서 y_i와 x_{ij}들이 같다면 $f_{\theta_0}(u_i|y_i)$들이 동일한 형태의 함수들이 되므로, $g(\theta|y) \to \delta_{\theta_0}(\theta)$에 의해서 보장되는 예측구간의 일치성은 모든 u_i들에 대한 균등한 일치성으로 확장된다.



(a) $\sigma_0^2 = 0.50^2$ (b) $\sigma_0^2 = 0.25^2$

Figure 3.1. Plots of $f_{\theta_0}(u_1|y_1)$ where $\theta_0 = (0.5, 1.5, \sigma_0^2)^T$. If $s = 3$, $x_1 = (1,1,0)^T$ and $y_1 = (1,1,1)^T$, and if $s = 6$, $x_1 = (0,0,0,0,1,0)^T$ and $y_1 = (0,0,0,1,1,0)^T$.

3. 로지스틱 회귀분석에의 적용

임의효과에 대한 구체적인 추론 문제로 로지스틱 회귀분석에서의 임의효과에 대한 추론 문제를 살펴보려고 한다. 여기에서 고려하는 모형은 계층적 일반화선형모형의 일종인 다음의 일반화선형혼합효과모형(generalized linear mixed effects model)이다.

$$\Pr\left(y_{ij} = 1|u_i, x_{ij}\right) = \mu_{ij}, \quad \log\frac{\mu_{ij}}{1-\mu_{ij}} = \beta_0 + x_{ij}^T\beta + u_i, \quad u_i \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right), \tag{3.1}$$

$y_{ij} \in \{0,1\}$, $i = 1, \ldots, J$ 그리고 $j = 1, \ldots, s$. 여기에서 x_{ij}는 $p-2$차원의 공변량이며 σ^2은 임의효과의 산포를 결정하는 모수이다. 임의절편(random intercept)를 가지는 로지스틱 회귀모형은 널리 사용되고 있고 임의효과의 분포로 정규분포를 가정할 수 있다. 이 모형의 경우 u_i에 대한 추론을 위해서 $f_{\beta_0,\beta,\sigma^2}(u_i|y_i)$를 구해보면

$$f_{\beta_0,\beta,\sigma^2}(u_i|y_i) \propto \left\{\prod_{j=1}^{s}\frac{\exp\left(\beta_0 + x_{ij}^T\beta + u_i\right)^{y_i}}{1+\exp\left(\beta_0 + x_{ij}^T\beta + u_i\right)}\right\}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{u_i^2}{2\sigma^2}\right) \tag{3.2}$$

이 되어 다루기가 쉽지 않다. Figure 3.1은 s가 3과 6인 경우의 $f_{\beta_0,\beta,\sigma^2}(u_i|y_i)$를 구해 본 것으로 확률밀도함수의 대칭성은 크게 깨지지 않지만 σ^2의 값이 분산에 미치는 영향이 큼을 알 수 있다. 만약에 임의효과의 산포를 과소추정하거나 과대추정하면 예측구간의 일치성에 적지 않은 영향을 준다는 것을 알 수 있다. 이것으로 모수의 추정 특히 산포에 대한 추정의 일치성이 중요한 문제로 됨을 알 수 있다. 그리고 예측구간을 구하기 위해서는 (3.2)에 대한 적분이 필요하며 이 적분은 수식으로 계산되지 않아 수치적분(numerical integration)이나 몬테칼로(Monte Carlo) 방법을 사용해야 한다. 본 논문에서는 마르코프 연쇄 몬테칼로(Markov chain Monte Carlo) 방법과 병행한 몬테칼로 방법을 기본으로 사용했다.

4. 모의실험

모의실험을 통한 연구에서는 신뢰분포를 이용한 방법을 기본으로 플러그인 방법 등을 같이 검토하여 보았다. 본 논문의 초점은 우도함수를 이용한 추론에 있어 사전분포의 선택이나 효율적인 알고리즘 문제 등의 논점을 가지는 베이지안(Bayesian) 방법은 비교에서 제외하였다. 모의실험에서는 각 방법의 효율

Table 4.1. Simulation results 1: means and standard deviations (in parentheses) of $q_{0.05}$ from 100 replications. True value of $(\beta_0, \beta_1, \sigma^2)$ is $(0.5, 1.5, 0.25^2)$, and PI, M and CD denote the plug-in method, maximization method and confidence distribution method, respectively.

Method	(J, s)		
	(100, 3)	(300, 3)	(500, 3)
PI (lme4)	0.520 (0.492)	0.460 (0.487)	0.430 (0.467)
M (lme4)	0.520 (0.492)	0.460 (0.487)	0.430 (0.467)
CD (dhglm)	0.004 (0.010)	0.029 (0.029)	0.049 (0.056)

Table 4.2. Simulation results 2: means and standard deviations (in parentheses) of $q_{0.05}$ from 100 replications. True value of $(\beta_0, \beta_1, \sigma^2)$ is $(0.5, 1.5, 0.50^2)$, and PI, M and CD denote the plug-in method, maximization method and confidence distribution method, respectively.

method	(J, s)		
	(50, 10)	(50, 30)	(50, 70)
PI (lme4)	0.236 (0.350)	0.094 (0.118)	0.070 (0.042)
M (lme4)	0.237 (0.349)	0.095 (0.117)	0.073 (0.043)
CD (dhglm)	0.070 (0.064)	0.060 (0.029)	0.055 (0.020)

성을 측정하기 위해서 다음과 같은 측도 $q_{0.05}$를 이용하였다.

$$q_{0.05} = 1 - \int_{a_{0.025}}^{a_{0.975}} f_{\theta_0}(y_1|y_1)du_1,$$

여기에서

$$0.025 = \int_{-\infty}^{a_{0.025}} L(u_1|y_1)du_1 = \int_{a_{0.975}}^{\infty} L(u_1|y_1)du_1$$

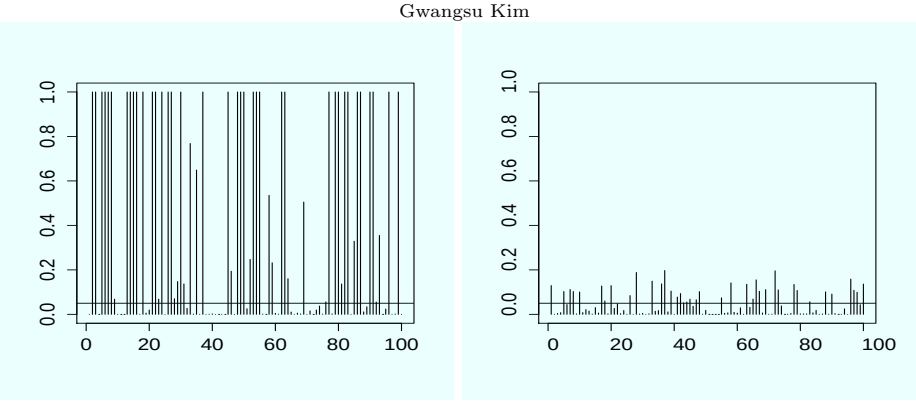
이 된다. 만약에 위의 값이 0.05에 가깝다면 예측우도에 의한 예측구간의 신뢰성은 높다고 하겠다. 신뢰분포를 사용하기 위해서는 $\hat{\theta}$과 그 로그 우도함수에 대한 헤시안 행렬(Hessian matrix)가 필요하게 된다. 이를 구하기 위한 계산 도구로는 Noh와 Lee (2011)의 DHGLM 패키지(R CRAN)을 사용하였다. 그리고 R의 패키지 중에 (일반화)선형혼합효과모형의 추론에 대한 여러 가지 기능을 제공해 주고 있는 lme4의 glmer 함수를 사용하여 플러그인 방법과 함께 이 함수가 제공하는 예측구간을 사용하여 보았다. 이 함수에서 제공하는 예측구간은

$$\sum_{i=1}^{J}\sum_{j=1}^{s}\left\{y_{ij}\left(\beta_0 + x_{ij}^T\beta + u_i\right) - \log\left(1 + \exp\left(\beta_0 + x_{ij}^T\beta + u_i\right)\right)\right\} - \sum_{i=1}^{J}\left\{\frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{u_i^2}{2\sigma^2}\right\}$$

을 최대로 하는 u_i와 그 미분값들에 기초한 것이다. 즉 임의효과에 대한 적분에 적용한 라플라스(Laplace) 방법에서 구한 것이다. 지금부터 이를 약칭해서 최대값을 이용한 방법이라고 하겠다.

4.1. 모의실험의 시행

신뢰분포를 이용하는 경우 이론적으로는 $s = 1$인 경우도 문제가 되지 않지만 실제 모의실험에서는 좀더 안정적인 결과를 얻기 위해서 기본을 $s = 3$으로 하였다. 여기에 대응한 J는 100, 300 그리고 500이며 σ^2의 참값은 0.25^2이다. 또한 J를 50으로 하고 σ^2의 참값을 0.50^2으로 하면서, s를 10, 30 그리고 70으로 키우면서 모의실험을 하였다. 즉 J와 s를 키워보면서 그 점근적 성질을 살펴보았다. 공변량들은 차원을 1로 놓고 이를 모수 0.5를 가지는 베르누이분포(Bernoulli distribution)에서 생성하였으며 모수의



(a) PI (lme4) (b) CD (dhglm)

Figure 4.1. Plots of $q_{0.05}$ in the simulation of the case $(J, s, \sigma^2) = (500, 3, 0.25^2)$ (PI: plug-in method, CD: confidence distribution method).

참값들은 Table 4.1과 Table 4.2에 명기되어 있다. 그리고 백분위수를 구하기 위한 모든 적분에는 마르코프 연쇄 몬테칼로 방법과 병행한 몬테칼로 방법을 적용하였다. 예를 들어서 신뢰분포의 $c(\theta|y)$를 이용해서 구한 상방 α% 백분위수 c_α는 다음 식에서 유도된다.

$$\int_{c_\alpha}^{\infty} \int f_\theta(u_i|y_i)c(\theta|y)d\theta du_i = \alpha.$$

4.2. 모의실험의 결과

Table 4.1를 보면 알 수 있듯이 J가 커지면 커질수록 $q_{0.05}$의 값은 0.05로 수렴하고 있으며 신뢰분포를 이용한 방법이 가장 좋은 결과를 보이고 있다. 플러그인 방법에서는 0.05보다 상당히 큰 값이 관측되며 lme4의 glmer 함수가 제공하는 예측구간도 비슷한 결과를 보여주고 있다. 플러그인 방법의 경우 값이 큰 것은 실제 예측구간보다 더 짧은 예측구간을 주기 때문이다. 이것은 임의효과의 분산을 과소추정하는 것에서 나오는 문제로 보인다. 실제 100번의 반복 결과에 대한 Figure 4.1를 보면 이것을 더 확연히 알 수 있다. 한편 σ^2의 참값을 키우고 J를 고정시킨 상태에서 s를 키울 때의 결과는 Table 4.2에 있다. 그 결과를 보면 신뢰분포를 이용한 방법은 s의 값이 작은 경우에도 일치성이 높은 결과를 주지만 플러그인 방법이나 최대값을 이용한 방법은 s가 70이 되어야 0.05에 근접하는 결과를 보여주고 있다. Lme4의 glmer 함수가 사용하는 최대값을 이용한 방법은 결국 라플라스 방법에 기초한 것이므로 s의 값이 커질 때 좋은 결과를 보여주는 것을 당연한 결과이다. 이에 반해 신뢰분포를 이용하는 경우 $N = Js$의 값이 커지면 일치성이 높아지는 결과를 보여주고 있다. 결과적으로 본다면 신뢰분포를 이용한 방법이 σ^2의 참값이 작은 경우, s의 값이 작은 경우에도 잘 작동하는 것을 보여준다. 이에 반해 최대값을 이용한 방법은 s의 값이 작은 경우 그 결과가 좋지 않다. 그런데 자료분석에서 s의 값이 작은 경우도 적지 않고 이에 대한 신뢰성 있는 추론이 요구되기에 신뢰분포를 이용한 방법이 더 우월하다고 할 수 있을 것이다.

5. 자료분석

5.1. 폐암 자료: 의사별 임의효과 분석

분석한 자료는 UCLA(University of California, Los Angeles) idre(institute for digital research and education)에서 공개한 자료로 http://www.ats.ucla.edu/stat/r/dae/melogit.htm을 통해서 접근할 수 있다. 이 자료는 폐암 자료로 표본의 크기는 8525이다. 이 중 분석에 사용할 변수들에 대한 설명이

Table 5.1. Description of lung cancer data: $x4_1, x4_2$ and $x4_3$ are dummy variables for cancer stages of II, III and IV (baseline: stage I), and summary statistics are means and standard deviations (in parentheses). Only response and CancerStage are discrete variables and others are continuous variables.

Variable	Type	Description	Notation	Summary Statistics
remission	response	0/1: fail or success of remission	R	0.296 (0.456)
Il6	predictor	status of patient	$x1$	4.017 (2.859)
CRP	predictor	status of patient	$x2$	4.973 (3.109)
LengthofStay	predictor	status of patient	$x3$	5.492 (1.050)
CancerStage	predictor	indicating the stage of cancer	$x4_1,\ x4_2,\ x4_3$	0.400, 0.200, 0.100
Experience	predictor	experience of doctor	$x5$	17.641 (4.075)
DID	random effects	number of each doctor	u	

Table 5.2. Results of estimated parameters in the model having random intercepts of doctors.

Variable	Estimate	Standard Error	Variable	Estimate	Standard Error
Intercept	-1.672	0.464	$x4_1$	-0.390	0.073
$x1$	-0.054	0.011	$x4_2$	-0.947	0.095
$x2$	-0.020	0.010	$x4_3$	-2.210	0.152
$x3$	-0.115	0.033	$x5$	0.105	0.024
$\log\sigma^2$	1.100	0.110			

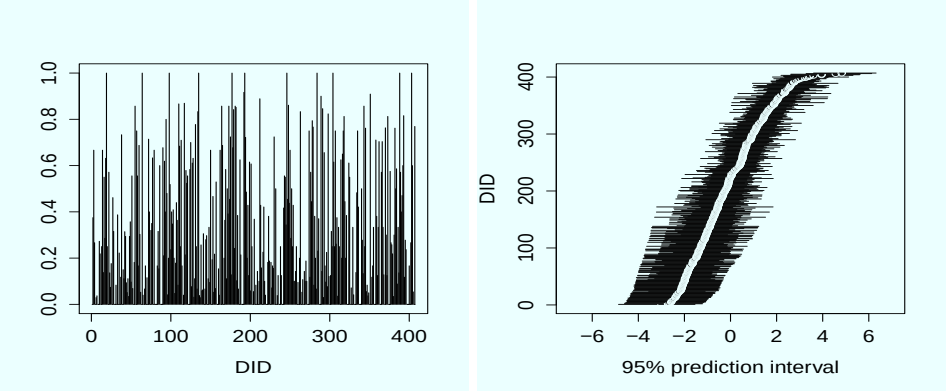
Table 5.1에 정리되어 있다. 치료효과가 나타나는 것에 영향을 주는 요인들로는 환자의 현 상태지표들, 의사의 숙련도 그리고 암의 진행단계가 될 것이다. 그런데 의사들이 가지는 특징이 치료효과가 나타나는 것에 영향을 줄 수가 있다. 예를 들어 비슷한 숙련도를 가진 의사들 간에도 능력이나 특징으로부터 나타나는 세밀한 차이들이 있을 수가 있다. 이 자료의 경우 의사 한사람이 치료한 환자 수의 평균은 20.96, 표준편차는 10.97(최소값 2, 최대값 40)으로 의사들의 치료 반복수가 많은 경우에 해당한다. 이 자료에 대해서 의사 개별의 효과를 고정효과로 모형화하고 추정하는 것이 불가능하지는 않다. 하지만 임의효과 모형을 이용한다면 모수의 차원이 줄어들어 모수에 대한 효율적인 추정을 가능하게 하고 의사들로 인한 효과에 대한 효율적인 확률적 추정을 가능하게 할 수 있다. 실제 의사별로 치료효과가 나타나는 비율(주변부 확률)을 보면 Figure 5.1(a)와 같다. 거의 0에 가까운 경우도 1에 가까운 경우도 있다. 이것은 여러 변수들의 영향을 제거하더라도 의사 개별의 차이가 상당히 강하게 존재할 수 있다는 것을 보여주고 있다.

실제 자료분석에는 다음과 같은 모형, 의사에 의해 발생하는 임의효과가 절편에 영향을 주는 임의절편모형을 적용하였다. 이 모형에서는 의사에 의한 효과와 다른 요인과의 상호작용은 고려하지 않았다.

$$\begin{aligned}
\mu_{ij} &= \Pr\left(R_{ij}=1|u_i, x_{ij}\right), \\
\log\frac{\mu_{ij}}{1-\mu_{ij}} &= \beta_0 + \beta_1 x1_{ij} + \beta_2 x2_{ij} + \beta_3 x3_{ij} + \sum_{c=1}^{3}\beta_{4_c} x4_{c\,ij} + \beta_5 x5_{ij} + u_i, \\
u_i &\overset{i.i.d.}{\sim} N\left(0,\sigma^2\right), \quad i=1,\ldots,407,\ j=1,\ldots,n_i,
\end{aligned} \tag{5.1}$$

여기에서 x_{ij}는 $(x1_{ij}, x2_{ij}, x3_{ij}, x4_{1\,ij}, x4_{2\,ij}, x4_{3\,ij}, x5_{ij})^T$이다. 자료분석에서는 모수에 대한 추정과 함께 신뢰분포를 이용하여 임의효과에 대한 95% 예측구간을 구하였다. 모수에 대한 추정 결과는 Table 5.2에 예측구간에 대한 결과는 Figure 5.1(b)에 요약되어 있다.

모수에 대한 추정 결과부터 보면 환자와 관련한 여러 변수들은 치료효과에 부정적인 작용을 의사의 숙련도는 긍정적인 작용을 하고 있다는것을 알 수 있다. 또한 진행단계가 높은 환자일수록 치료효과가 나타



(a) Ratios of remissions (b) 95% prediction intervals

Figure 5.1. Ratios of remissions and 95% prediction intervals of 407 doctors where white dots in the prediction intervals are the means from predictive likelihoods. Prediction intervals are sorted by means.

Table 5.3. Upper 5 prediction intervals and lower 5 prediction intervals, its lengths and means from predictive likelihood.

	Prediction interval	Length	Mean		Prediction interval	Length	Mean
Upper 1	(3.605, 6.317)	2.712	4.823	Lower 1	(−4.858, −1.543)	3.315	−2.999
Upper 2	(3.281, 6.127)	2.846	4.564	Lower 2	(−4.595, −1.217)	3.378	−2.712
Upper 3	(2.617, 5.628)	3.011	3.952	Lower 3	(−4.593, −1.231)	3.362	−2.711
Upper 4	(2.620, 5.600)	2.980	3.951	Lower 4	(−4.543, −1.165)	3.378	−2.655
Upper 5	(2.192, 5.313)	2.121	3.599	Lower 5	(−4.490, −1.090)	3.400	−2.579

날 확률이 낮아진다. 예측구간의 결과를 보면 의사별 차이가 적지 않은 것으로 나타난다. 상위 5개 구간과 하위 5개 구간이 정리된 Table 5.3를 보면 상위 5개는 모두 양수에 걸친 구간을 가지고 있고 하위 5개 모두 음수에 걸친 구간을 가지고 있다. 하위구간보다는 상위구간에 걸친 값들의 절대값이 크며 구간의 길이가 짧은 것으로 보아 특별히 높은 능력을 보일 것으로 예상되는 소수의 의사들이 나타나고 있음을 알 수 있다. 그리고 앞으로 관측된 자료들로부터도 이런 특징을 보일 것으로 예상할 수 있다.

5.2. 폐암 자료: 병원-의사에 대한 이중수준 임의효과 분석

다음으로는 병원과 의사 이중 임의효과에 대해서 분석하여 보았다. 의사들은 특정한 병원에 소속되어 있으며 특정 병원들 역시 임의효과를 가질 수 있다. 이런 이중수준의 임의효과에 대한 분석을 위해서 원 자료에서 임의로 6개의 병원 자료를 추출해(표본의 크기 1889) 분석하였다. 분석한 자료에 대한 모형은 아래와 같다.

$$\mu_{kij} = \Pr\left(R_{kij} = 1 | v_k, u_{ki}, x_{kij}\right),$$

$$\log \frac{\mu_{kij}}{1-\mu_{kij}} = \beta_0 + \beta_1 x1_{kij} + \beta_2 x2_{kij} + \beta_3 x3_{kij} + \sum_{c=1}^{3} \beta_{4_c} x4_{ckij} + \beta_5 x5_{kij} + v_k + u_{ki},$$

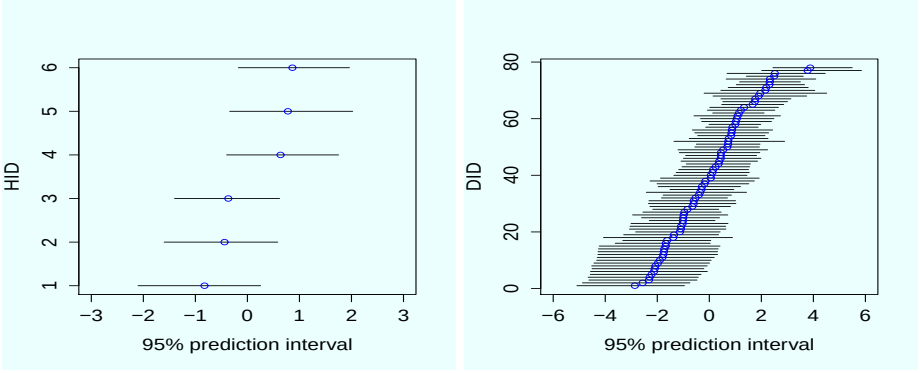
$$v_k \overset{i.i.d.}{\sim} N\left(0, \sigma_1^2\right), \quad u_{ki} \overset{i.i.d.}{\sim} N\left(0, \sigma_2^2\right)$$

$$v_k \perp u_{ki}, \quad k = 1, \ldots, 6,\ i = 1, \ldots, n_k,\ j = 1, \ldots, n_{ki}.$$

이 모형은 4절이나 5.1절에서 소개한 모형 (3.1)과 (5.1)보다 더 복잡한 형태를 띠는 구조화된 모형으로 소속된 병원의 효과를 보정한 순수한 의사에 의한 효과를 볼 수 있다는 장점이 있다. 신뢰분포를 이용한

Table 5.4. Results of estimated parameters in the model having random effects of hospitals and doctors.

Variable	Estimate	Standard Error	Variable	Estimate	Standard Error
Intercept	−1.679	1.051	$x4_1$	−0.410	0.153
$x1$	−0.052	0.023	$x4_2$	−0.936	0.204
$x2$	−0.028	0.021	$x4_3$	−2.509	0.353
$x3$	0.105	0.068	$x5$	0.028	0.051
$\log \sigma_1^2$	−0.225	1.031	$\log \sigma_2^2$	0.994	0.261



(a) Hospital (b) Doctor

Figure 5.2. Prediction intervals of random effects of hospitals and doctors where HID is for the identification of hospitals. Each dot in the prediction interval is the mean from predictive likelihood, and prediction intervals are sorted by means.

방법을 사용하는 경우 θ에 대한 최대우도추정량의 점근분포와 $f_\theta(v_k|y)$와 $f_\theta(u_{ki}|y)$만 알면 되기 때문에 분석은 어렵지 않게 진행할 수 있다. 한편 임의효과 모형으로 분석한 것과 각 의사별, 병원별 효과를 모수로 처리해서 분석한 것의 차이를 알아보기 위해서 일반화선형모형으로 자료를 처리해 보았으나 추정 알고리즘이 수렴이 되지 않고 추정치와 추정된 표준편차가 10^{16}을 넘어서는 경우도 발견되었다. 이 자료는 임의효과를 이용해서 분석하는 것이 모수에 대한 추정량의 분산을 작게 해주고 병원별, 의사별 효과에 대한 효율적인 추정을 가능하게 함을 알 수 있다.

모수에 대한 추정 결과부터 Table 5.4로 보면 5.1절의 결과와 크게 다르지 않으나 $x3$의 계수에 대한 추정치가 5.1절의 결과와 다른 부호를 가지고 있다. 하지만 이 모수가 0이라는 가설을 검정하기 위한 유의확률이 크기 때문에 유의한 차이는 아니다. Figure 5.2은 병원별, 의사별 예측구간을 보여준 것으로 병원보다는 의사로 인한 효과의 차이가 훨씬 크게 나타날 것으로 예측됨을 알 수가 있다. 실제 병원으로 인한 임의효과의 분산이 0이라는 가설을 검정하기 위한 유의확률의 값이 커 병원의 효과에 대한 통계적 유의성을 확신할 수 없으나 의사들로 인한 효과는 확실한 통계적 유의성을 보여주고 있다. 한편 의사들의 효과가 가지는 특징은 5.1절과 크게 다르지 않다.

6. 결론 및 토의

본 논문에서는 임의효과에 임의성에 대한 추론방법, 신뢰분포를 이용한 방법을 소개하고 이를 로지스틱 회귀분석에 적용하여 보았다. 신뢰분포를 이용한 방법은 플러그인 방법보다 훨씬 효율적이고 우도함수에 기초한 추론을 가능하게 하고 있다. 모의실험에서 신뢰분포는 매우 뛰어난 성능을 보여주었고 자료분석에도 적용될 수 있다는 것을 확인하였다. 한편 신뢰분포를 이용하기 위해서는 계산 차원에서의 부

담은 증가하며 실제 적지 않은 시간이 소요된다. 앞으로 신뢰분포가 가지는 효율성을 잃지 않으면서도 계산 속도가 빠른 방법의 개발, 조금더 다양한 임의효과의 분포나 구조화된 모형의 적용이 연구될 필요가 있다. 그리고 임의효과를 포함한 모형선택(공변량의 변수선택, 구조화된 임의효과나 임의효과에 공변량을 사용하는 모형에 대한 선택 등)은 DHGLM에서 제공하는 cAIC(conditional AIC) 값을 최소로 하는 방법 등이 있으며 이 방법과 신뢰분포를 이용한 임의효과에 대한 추론방법과의 관계에 대해서는 추후 연구과제로 남겨두려고 한다.

감사의 글

본 논문은 임의효과의 추론에 대한 서울대 이영조 교수님과의 공동연구에 힘입는 바 컸으며 이에 감사를 표합니다.

References

Birnbaum, A. (1962). On the foundations of statistical inference, *Journal of the American Statistical Association*, **57**, 269–306.

Bjørnstad, J. F. (1990). Predictive likelihood: A review, *Statistical Science*, **5**, 242–265.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society, Series B*, **58**, 791–806.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, Wiley.

Noh, M. and Lee, Y. (2011). dhglm: Double Hierarchical Generalized Linear Models (R package), http://cran.r-project.org/web/packages/dhglm/.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects, *Statistical Science*, **6**, 15–32.

Rubin, D. B. (2006). *Matched Sampling for Casual Effects*, Cambridge University Press.

Skrondal, A. and Rabe-Hesketh, S. (2007). Latent variable modelling: A survey, *Scandinavian Journal of Statistics*, **34**, 712–745.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, UK.

Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review, *International Statistical Review*, **81**, 3–39.

우도에 기반한 임의효과에 대한 추론과 로지스틱 회귀모형에서의 응용

김광수[a,1]

[a]고려대학교 통계학과

(2015년 3월 17일 접수, 2015년 3월 30일 수정, 2015년 3월 30일 채택)

요 약

본 논문에서는 임의효과에 대한 추론 문제가 다루어졌으며 이 추론에서 신뢰분포를 사용하는 것이 제안되었다. 신뢰분포를 이용한 방법은 표본의 크기가 작아도 임의절편들이 있는 로지스틱 회귀분석에서 좋은 결과를 보여주었으며, 자료분석을 통해서도 각 개체가 가지는 임의효과들에 대한 세밀한 분석이 가능함을 확인하였다.

주요용어: 신뢰분포, 일반화선형혼합효과모형, 로지스틱 회귀귀분석, 예측우도, 예측구간, 임의효과.

[1](136-701) 서울시 성북구 안암로 145, 고려대학교 통계학과(BK21플러스 데이터공학팀).
E-mail: gwangsu_2014@korea.ac.kr