The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.343
28(2), 343–351

SRC-Stat Package for Fitting Double Hierarchical Generalized Linear Models

Maengseok Noh[a,1] · Il Do Ha[a] · Youngjo Lee[b] · Johan Lim[b] · Jaeyong Lee[b] · Heeseok Oh[b] · Dongwan Shin[b] · Sanggoo Lee[b] · Jinuk Seo[b] · Yonhtae Park[b] · Sungzoon Cho[b] · Jonghun Park[b] · Youkyung Kim[b] · Kyungsang You[b]

[a]Department of Statistics, Pukyong National University
[b]Data Science for Knowledge Creation Research Center, Seoul National University

(Received March 24, 2015; Revised April 7, 2015; Accepted April 8, 2015)

Abstract

We introduce how to fit random effects models via a SRC-Stat statistical package. This package has been developed to fit double hierarchical generalized linear models where mean and dispersion parameters for the variance of random effects and residual variance (overdispersion) can be modeled as random-effect models. The estimates of fixed effects, random effects and variances are calculated by a hierarchical likelihood method. We illustrate the use of our package with practical data-sets.

Keywords: Disease mapping, double hierarchical generalized linear models, hierarchical likelihood, random effects, SRC-stat.

1. 서론

서울대학교『데이터과학과 지식창출연구센터』는 미래창조과학부와 한국연구재단이 추진하는 선도연구센터지원사업 및 에스이(랩)과 서울대학교 빅데이터센터 등의 지원으로 교육용 통계패키지인 SRC-STAT를 개발하고 국내 교육기관에서 무상으로 사용할 수 있도록 2013년 9월 베타 버전으로 보급하였다 (http://srcdsc.snu.ac.kr/srcstat/). SRC-STAT은 평균, 분산 등 기초통계부터 의학분야에서 사용되는 생존자료나 사회과학 분야에서 활용되는 다변량 자료, 나아가 금융분야에서 활용되는 시계열자료와 의학분야의 질병지도 작성 및 기상예측에 이르는 다양한 자료를 분석할 수 있는 통계 패키지이다. 특히 기존 상업적 통계 프로그램이 가진 기능 이외에 센터가 보유한 다단계우도기법에 기반한 통계기법과 계산 알고리즘을 구현하고 있어 기존 통계 패키지와는 차별화되고 있다.

최근 변량효과 모형(random effects models)은 통계학의 다양한 분야에서 적용되고 있는데, SRC-Stat은 Noh와 Lee (2011)가 제시한 dhglm R 패키지를 활용하여, 기존 패키지들이 제시하지 못하는 확장된 형태의 변량효과 모형을 다단계 우도(hierarchical likelihood) 접근법을 사용하여 사용자에게 적합결과를 제시한다. 이러한 확장된 형태의 변량효과 모형은 Lee와 Nelder (2006)가 제시한 이중 다단계

This research was supported by an NRF grant funded by Korea government (MSIP) (No. 2011-0030810).
[1]Corresponding author: Department of Statistics, Pukyong National University, 45 Yongso-ro, Nam-gu, Busan 608-737, Korea. E-mail: msnoh@pknu.ac.kr

일반화 선형모형(double hierarchical generalized linear models; DHGLMs)에 기반한다. DHGLMs은 일반화 선형모형(Generalized linear models; GLMs)에서 반응변수의 평균 부분에 관측되지 않은 변량효과를 고려한 다단계 일반화 선형모형(Hierarchical generalized linear models; HGLMs)을 확장하여 반응변수의 산포(dispersions)에도 변량효과를 고려한 모형이다. 이러한 DHGLMs은 t-분포와 같은 형태의 꼬리가 두터운 분포(heavy-tailed distribution)을 생성하여 극단값(outlier)에 강건한(robust) 추정치를 제시하는 등 다양한 통계적인 특성을 제시한다 (Noh와 Lee, 2007). 특히, SRC-Stat을 통해 변량효과에 시-공간적 상관성(spatial-temporal correlation)을 가지는 모형을 적합할 수 있어 질병지도(Disease mapping)에 활용할 수 있다.

본 논문에서는 실제 자료를 통해 DHGLMs의 적합 및 질병지도를 작성하기 위해 SRC-Stat를 어떻게 이용할 수 있는지에 대해서 소개하고자 한다. 이를 위해서는 SRC-stat에서 제시하는 DHGLMs은 다른 패키지들과 차별화되는 다음과 같은 확장된 형태의 변량효과를 제시한다.

i) 평균뿐만 아니라 산포 및 변량효과의 분산에도 변량효과의 모형.

ii) 선형 예측 변수에 서로 다른 분포를 가지는 변량효과.

iii) 다양한 형태의 시-공간 상관성을 가지는 변량효과 모형.

iv) 서로 다른 변량효과간의 상관성을 가지는 모형.

본 논문의 구성 체계는 다음과 같다. 2장에서는 DHGLMs을 소개하고 DHGLMs의 각 성분들이 SRC-stat의 메뉴에서 어떻게 설정되는지에 대해서 설명하며, 3장에서는 예제를 통해 SRC-stat을 통한 DHGLMs의 적합용방법에 대해서 다루고자 한다.

2. DHGLMs 및 SRC-stat에서의 설정

2.1. DHGLMs

μ, λ, ϕ를 반응변수의 평균, 변량효과의 분산, 반응변수의 산포라고 둔다. 이때, $u^{(\mu)}, u^{(\lambda)}$, $u^{(\phi)}$를 μ, λ, ϕ에 나타나는 변량효과라고 정의한다. 변량효과 $u^{(\mu,\lambda,\phi)} = (u^{(\mu)}, u^{(\lambda)}, u^{(\phi)})$가 조건부로 주어졌을 때, 반응변수 y는 정규, 이항, 포아송, 감마 분포와 같은 GLM 분포를 따르며, 그 조건부 평균과 분산은 다음 (2.1)과 같다. 이때, ϕ는 반응변수에 대한 산포모수이며, $V(\ )$는 분산함수를 나타낸다.

$$E\left(y|u^{(\mu,\lambda,\phi)}\right) = \mu, \qquad \text{var}\left(y|u^{(\mu,\lambda,\phi)}\right) = \phi V(\mu), \tag{2.1}$$

DHGLMs은 평균에 대한 DHGLM인 DHGLM(μ)과 산포에 대한 HGLM인 HGLM(ϕ) 두 개의 성분으로 이루어져 있다. α를 변량효과 $u^{(\lambda)}$에 대한 분산이라고 하였을 때, DHGLM(μ)는 μ, λ, α에 대한 선형 예측 변수는 다음 (2.2)–(2.4)와 같이 구성되어 있다. DHGLM(μ)는 반응변수의 평균 μ와 변량효과 $u^{(\mu)}$의 분산인 λ에 대해서도 변량효과를 고려하였음을 알 수 있다.

$$\eta^{(\mu)} = h^{(\mu)}(\mu) = X^{(\mu)}\beta^{(\mu)} + Z^{(\mu)}v^{(\mu)}, \tag{2.2}$$

$$\eta^{(\lambda)} = h^{(\lambda)}(\lambda) = X^{(\lambda)}\beta^{(\lambda)} + Z^{(\lambda)}v^{(\lambda)}, \tag{2.3}$$

$$\eta^{(\alpha)} = h^{(\alpha)}(\alpha) = X^{(\alpha)}\beta^{(\alpha)}, \tag{2.4}$$

여기에서 $h^{(\mu)}(\cdot)$, $h^{(\lambda)}(\cdot)$, $h^{(\alpha)}(\cdot)$는 연결함수(link function), $\beta^{(\mu)}$, $\beta^{(\lambda)}$, $\beta^{(\alpha)}$는 고정효과(fixed effects), $v^{(\mu)} = g^{(\mu)}(u^{(\mu)})$, $v^{(\lambda)} = g^{(\lambda)}(u^{(\lambda)})$는 변량효과, $g^{(\mu)}(\cdot)$, $g^{(\lambda)}(\cdot)$는 어떤 단조함수를 나타낸다.

Table 2.1. Component options specifying DHGLMs

Components	Option for models		
μ	DHGLM(μ)	HGLM(μ)	GLM(μ)
Link function			
$h^{(\mu)}(\cdot)$	identity, logit, probit, cloglog, log, inverse		
$h^{(\lambda)}(\cdot)$	log, inverse	log, inverse	NULL
$h^{(\alpha)}(\cdot)$	log, inverse	NULL	NULL
Linear predictor			
$\eta^{(\mu)}$	$X^{(\mu)}\beta^{(\mu)} + Z^{(\mu)}v^{(\mu)}$	$X^{(\mu)}\beta^{(\mu)} + Z^{(\mu)}v^{(\mu)}$	$X^{(\mu)}\beta^{(\mu)}$
$\eta^{(\lambda)}$	$X^{(\lambda)}\beta^{(\lambda)} + Z^{(\lambda)}v^{(\lambda)}$	$X^{(\lambda)}\beta^{(\lambda)}$	NULL
$\eta^{(\alpha)}$	$X^{(\alpha)}\beta^{(\alpha)}$	NULL	NULL
ϕ	HGLM(ϕ)	GLM(ϕ)	
Link function			
$h^{(\phi)}(\cdot)$	log, inverse	log, inverse	
$h^{(\tau)}(\cdot)$	log, inverse	NULL	
Linear predictor			
$\eta^{(\phi)}$	$X^{(\phi)}\beta^{(\phi)} + Z^{(\phi)}v^{(\phi)}$	$X^{(\phi)}\beta^{(\phi)}$	
$\eta^{(\tau)}$	$X^{(\tau)}\beta^{(\tau)}$	NULL	
Distributions of			
$y\|u^{(\mu,\lambda,\phi)}$	Gaussian, binomial, Poisson, gamma		
$u^{(\mu)}$, $u^{(\lambda)}$, $u^{(\phi)}$	Gaussian, beta, gamma, inverse-gamma		

DHGLM(μ)는 변량효과의 분포가정에 강건한 추정치를 제시하는 방법으로 처음 개발되었다 (Noh 등, 2005). 만약 식 (2.3)에서 변량효과인 $v^{(\lambda)}$ 성분이 없으면 Lee와 Nelder (1996, 2001)의 HGLM(μ)이 되며, HGLM(μ)에서 변량효과인 $v^{(\mu)}$ 성분이 없으면 GLM(μ)가 된다.

기존 패키지들은 산포모수 ϕ에 대해서 상수로 취급하지만, 품질관리 등 많은 분야에서 ϕ에 대한 모형화 연구들이 많이 진행되었다. 따라서, ϕ에 대한 변량효과 모형인 다음 식 (2.5), (2.6)와 같은 형태인 HGLM(ϕ)를 고려할 수 있다. 이때, τ는 $u^{(\phi)}$에 대한 분산을 의미한다.

$$\eta^{(\phi)} = h^{(\phi)}(\phi) = X^{(\phi)}\beta^{(\phi)} + Z^{(\phi)}v^{(\phi)}, \tag{2.5}$$

$$\eta^{(\tau)} = h^{(\tau)}(\tau) = X^{(\tau)}\beta^{(\tau)}, \tag{2.6}$$

여기에서 $h^{(\phi)}(\cdot)$, $h^{(\tau)}(\cdot)$는 연결함수, $\beta^{(\phi)}$, $\beta^{(\tau)}$는 고정효과 $v^{(\phi)} = g^{(\phi)}(u^{(\phi)})$는 변량효과를 나타내며, $g^{(\phi)}(\cdot)$는 어떤 단조함수이다. HGLM(ϕ)에서 변량효과인 $v^{(\phi)}$가 없으면 ϕ에 대한 고정효과만 고려한 GLM(ϕ)가 된다.

2.2. SRC-stat을 통한 DHGLMs의 설정

DHGLMs을 설정하기 위해서는 y에 대한 조건부 분포가 반드시 GLM 분포가 되어야 한다. SRC-stat에서는 μ에 대해서 GLM(μ), HGLM(μ), DHGLM(μ), ϕ에 대해서 GLM(ϕ), HGLM(ϕ)을 설정할 수 있으며 각 성분별 옵션은 다음 Table 2.1과 같다.

Figure 2.1은 이러한 각 성분들을 어떻게 메뉴 방식의 SRC-stat에서 설정하는 지에 대해서 나타내고 있다. SRC-stat의 DHGLMs을 위한 메뉴는 Figure 2.1에서 보는 바와 같이 크게 1) 평균모형, 2) 분산모형, 3) 세부 옵션으로 구성되어 있다. 평균모형은 i) model for mu, ii) model for lambda, iii) model for

Figure 2.1. Dialogue box for DHGLMs in the SRC-stat

alpha로 분산모형은 i) model for phi, ii) model for tau로 구성되어 있다. 평균모형의 model for mu에서는 반응변수 (y), 분포 ($y|u^{(\mu,\lambda,\phi)}$의 분포), 연결함수 ($h^{(\mu)}(\cdot)$), 고정효과에 대한 공변량 ($X^{(\mu)}$), 변량효과에 대한 공변량 ($Z^{(\mu)}$)을 통해 Table 2.1에서의 각 성분을 설정할 수 있다. 특히, 변량효과에 대한 공변량에서 상관형태를 지정해 주면 변량효과에 대한 다양한 시-공간 모형을 적합할 수 있다. 평균모형의 ii) model for lambda, iii) model for alpha와 분산모형의 i) model for phi, ii) model for tau에 대한 성분도 앞서 설명한 방식과 마찬가지로 설정해 주면 된다. 세부 옵션은 다음 7가지를 설정할 수 있으며, 자세한 설명은 다음과 같다.

i) Betafix: 평균치를 고정하는 옵션으로 수치를 넣을시 해당 수치로 평균값을 고정한다.

ii) Phifix: 잔차분산을 고정하는 옵션으로 수치를 넣을시 해당 수치로 잔차분산을 고정한다.

iii) Lamfix: 평균모형에 들어가는 변량효과의 분산을 고정하는 옵션으로 수치를 넣을시 해당수치로 변량효과의 분산을 고정한다.

iv) Mord: 평균모형에서 라플라스근사(Laplace Approximation)의 차수를 지정한다.

v) Dord: 분산모형에서 라플라스근사(Laplace Approximation)의 차수를 지정한다.

vi) Maximum Number of Iteration: 수렴하기까지 걸리는 반복의 최대 회수를 지정한다.

vii) The Criteria for Convergence: 알고리즘의 수렴여부를 판별하는 기준치를 설정한다.

3. 예제

3.1. 질병지도의 활용

2005년 행정구역 기준 서울지역 411개의 동별 질병사 및 사고사에 대한 지역별 차이에 대한 연구를 위해 반응변수로 2005–2008년 4년간 수집된 주요 사인별 사망자 수(질병사, 사고사), 설명변수는 박탈지수를 고려하였다. 동별 소지역에 따른 분석에 의미 있는 결과를 주기 위해서는 1년 사망 자료는 그 정보가 너무 희박하여 적어도 4년 동안 수집된 자료가 적절한 것으로 판단된다. 동별 박탈지수는 2005년 인구센서스 조사에서 10% 표본을 통해 나타난 지역의 사회적, 경제적 결핍 수준을 종합적으로 나타내는 지표이다. y_i를 $i\,(i=1,\ldots,n=411)$번째 소지역 단위인 2005년부터 2008년까지의 질병 및 사고에 대한 사망자 수라고 두었을 때, 다음과 같은 모형을 고려할 수 있다. 즉, 지역에 대한 변량 효과 v_i가 주어졌을 때 y_i는 조건부 기대도수 μ_i을 가지는 포아송 분포를 따른다고 가정한다. μ_i에 대한 연결함수(link

Table 3.1. Parameter estimates from MRF

Cause of Death	parameter	estimate	SE	t-value
Disease	β_0	−0.193	0.012	−15.9
	β_1	0.013	0.00084	14.9
	$\log\lambda$	−5.020	0.12000	−41.8
	ρ	0.150		
Accident	β_0	−0.376	0.016	−23.7
	β_1	0.020	0.0015	13.4
	$\log\lambda$	−4.770	0.18	−26.5
	ρ	0.146		

function)를 로그함수로 두고, 이때 η_i는 μ_i에 대한 선형 예측 변수가 된다.

$$y_i|v_i \sim \text{Poisson}(\mu_i), \tag{3.1}$$

$$\eta_i = \log\mu_i = \log(E_i) + \beta_0 + \beta_1 x_i + v_i, \tag{3.2}$$

여기에서 E_i는 i번째 소지역의 인구수에 대한 서울전체 의 사망률을 적용시킨 기대사망자 수, x_i는 해당 동의 박탈지수가 된다. 따라서, $\log(E_i)$는 오프셋(offset), x_i는 고정효과, v_i는 변량효과로 모형화된다. 소지역 효과 $v = (v_1, \ldots, v_n)^T$는 다변량 정규분포를 통해 공간적 상관 성을 고려한 MRF(Markov random field)를 고려할 수 있다. MRF 모형은 v의 분산-공분산 행렬의 역행렬이 다음 형태를 가지는 모형을 의미한다.

$$\Sigma^{-1} = [\text{var}(v)]^{-1} = \frac{I - \rho N}{\lambda}.$$

MRF에서 N은 서로 인접하면 1, 아니면 0인 인접행렬(neighborhood matrix)를 나타내며 $\rho = 0$이면 서로 독립인 모형이 된다. MRF 모형에서는 (i, j) 지점이 서로 인접할 경우에는 변량효과의 분산-공분산 행렬의 역행렬의 (i, j) 원소는 $-\rho/\lambda$의 값을 인접하지 않으면 0의 값을 가진다.

모형을 적합하기 위한 SRC-stat의 사용은 Figure 3.1과 같다. 질병사(obsin)를 반응변수로 두고 분포는 Poisson, 연결함수는 log을 선택한다. 오프셋은 기대도수의 로그값(expin), 고정효과는 박탈지수(town), 각 동별 효과를 변량효과(region)로 두었다. 만약 MRF 모형을 적합하기 위해서는 변량효과의 분포를 Gaussian으로 두고 Correlation Structure를 MRF로 지정하여 적합을 하면 된다. 적합결과 공간적 상관성을 나타내는 ρ는 통계적으로 유의하여 독립모형에 비하여 MRF 모형이 적절한 것으로 판단되며 박탈지수 역시 통계적으로 유의하였다. 즉, 서울지역 동별로 질병 및 사고사에 대한 지역별 차이를 분석한 결과 공간적으로 서로 인접한 지역 간의 상관성이 있으며 박탈지수가 높을수록 사망률도 높음을 말해주고 있다. Figure 3.2는 모형 추정 후 나타나는 변량효과 추정치를 바탕으로 공변량이 보정된 $\exp(v_i)$인 SMR(standardized mortality rate)를 나타내는데, 색이 진할수록 SMR이 높음을 의미한다. 그림에서 보는 바와 같이, 강남지역이 다른 지역에 비하여 상대적으로 질병 및 사고사망이 낮음을 알 수 있다.

3.2. 금속샘플의 균열 성장 자료

Lu와 Meeker (1993)가 제시한 21개의 금속샘플(metallic specimen)들의 균열(crack)의 성장에 대한 자료의 분석을 통해 DHGLMs의 유용성에 대해서 설명하고자 한다. 각각의 샘플은 총 120,000번의 하중주기을 견디면서 10,000번 주기마다 균열을 관측한다. 자료에서 변수명 “cycle”은 해서 주기번호에

Figure 3.1. Dialogue box for DHGLMs in the SRC-stat

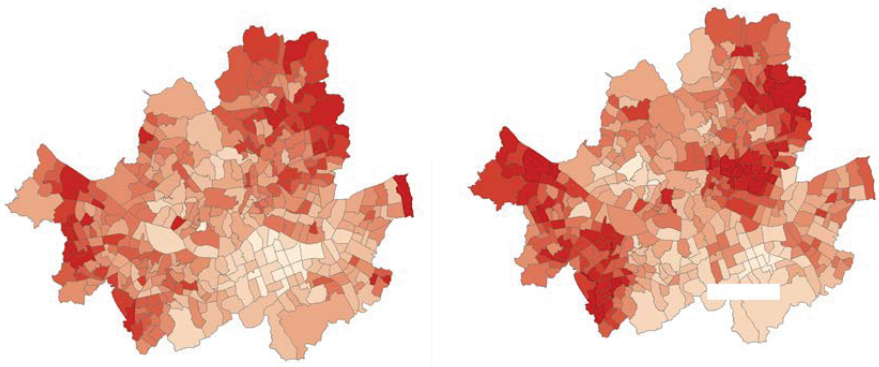

Figure 3.2. Estimated covariate adjusted SMR: disease (left), accident (right)

서 100을 나눈 $t_j = j/100$ $(j = 1, \ldots, 12)$에 대한 변수명을 의미한다. l_{ij}을 i번째 샘플의 j번째 관측값에 대한 균열크기로 반복이 증가할수록 커진다. 반응변수로 $y_{ij} = l_{ij} - l_{ij-1}$은 직전 균열과의 차이를 고려하며 항상 양수의 값을 가진다. 설명변수로는 직전 균열크기인 l_{ij-1}을 고려하고 변수명으로는 "crack0"로 두었다. 여기서, 반응변수 y_{ij}는 다음과 같은 조건부 평균과 분산을 가지는 감마분포를 가정하였다 (Lee 등, 2006).

$$E\left(y_{ij}|v_i^{(\mu)}, v_i^{(\phi)}\right) = \mu_{ij} \quad \text{and} \quad \text{var}\left(y_{ij}|v_i^{(\mu)}, v_i^{(\phi)}\right) = \phi_{ij}\mu_{ij}^2. \tag{3.3}$$

앞서 정의한 모형을 토대로 SRC-stat에서는 다음 모형 (3.4)–(3.8)와 같은 5가지의 모형을 적합할 수 있다. (i) GLM은 평균 (μ)에만 고정효과를 고려하고, 산포 (ϕ)는 상수를 가정한 모형이다. 평균뿐만 아니라 산포에도 고정효과를 고려한 모형은 (ii) JGLM(Joint GLM)이며, (iii) HGLM1은 평균에 고정

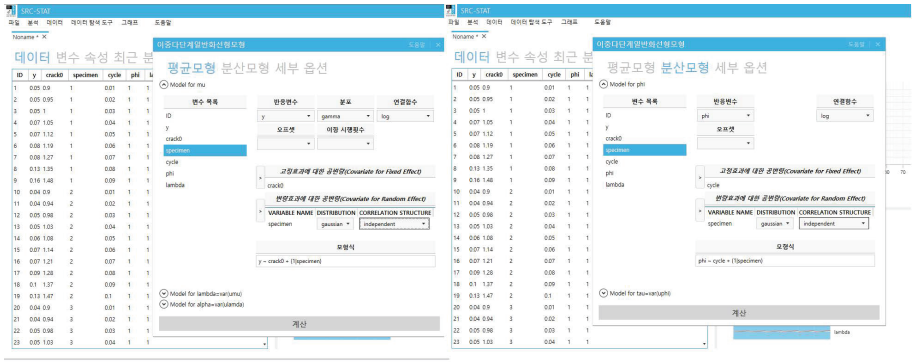

Figure 3.3. Dialogue box for DHGLMs fitting crack-growth data in the SRC-stat

및 변량효과, 산포는 상수를 (iv) HGLM2은 평균에 고정 및 변량효과, 산포는 고정효과를 고려한 모형들이다. 마지막, (v) DHGLM은 평균뿐만 아니라 산포에도 고정 및 변량효과를 고려한 모형이다. 평균에 대한 고정효과로는 직전 균열 크기를, 산포에 대한 고정효과로는 주기를 고려하였다. 즉, 직전 균열 크기가 클수록 평균은 증가 (혹은 감소), 주기가 커질수록 산포가 증가 (혹은 감소) 하는 모형을 나타내며, 변량효과는 각 샘플 효과를 나타낸다.

i) GLM {GLM(μ), constant}:

$$\eta_{ij}^{(\mu)} = \log \mu_{ij} = \beta_0^{(\mu)} + \beta_1^{(\mu)} l_{ij-1} \quad \text{and} \quad \eta_{ij}^{(\phi)} = \log \phi_{ij} = \beta_0^{(\phi)}. \tag{3.4}$$

ii) JGLM {GLM(μ), GLM(ϕ)}:

$$\eta_{ij}^{(\mu)} = \log \mu_{ij} = \beta_0^{(\mu)} + \beta_1^{(\mu)} l_{ij-1} \quad \text{and} \quad \eta_{ij}^{(\phi)} = \log \phi_{ij} = \beta_0^{(\phi)} + \beta_1^{(\phi)} t_j. \tag{3.5}$$

iii) HGLM1 {HGLM(μ), constant}:

$$\eta_{ij}^{(\mu)} = \log \mu_{ij} = \beta_0^{(\mu)} + \beta_1^{(\mu)} l_{ij-1} + v_i^{(\mu)} \quad \text{and} \quad \eta_{ij}^{(\phi)} = \log \phi_{ij} = \beta_0^{(\phi)}, \tag{3.6}$$

단, $v_i^{(\mu)} \sim N(0, \lambda)$, $\eta_i^{(\lambda)} = \log \lambda = \beta_0^{(\lambda)}$.

iv) HGLM2 {HGLM(μ), GLM(ϕ)}:

$$\eta_{ij}^{(\mu)} = \log \mu_{ij} = \beta_0^{(\mu)} + \beta_1^{(\mu)} l_{ij-1} + v_i^{(\mu)} \quad \text{and} \quad \eta_{ij}^{(\phi)} = \log \phi_{ij} = \beta_0^{(\phi)} + \beta_1^{(\phi)} t_j. \tag{3.7}$$

v) DHGLM {HGLM(μ), HGLM(ϕ)}:

$$\eta_{ij}^{(\mu)} = \log \mu_{ij} = \beta_0^{(\mu)} + \beta_1^{(\mu)} l_{ij-1} + v_i^{(\mu)} \quad \text{and} \quad \eta_{ij}^{(\phi)} = \log \phi_{ij} = \beta_0^{(\phi)} + \beta_1^{(\phi)} t_j + v_i^{(\phi)}, \tag{3.8}$$

단, $v_i^{(\phi)} \sim N(0, \alpha)$.

Figure 3.2는 위 5가지 모형에서 가장 복합한 (v) DHGLM 적합을 위한 SRC-stat에서의 다이알로그 화면이다. 평균 및 분산모형에 고정효과 및 변량효과를 모두 고려해 주면 된다. DHGLM 모형을 적합예시를 통해 그 하부 모형인 GLM, JGLM, HGLM 적합 또는 가능하리라 본다. Table 3.2는 SRC-stat을 통해 적합한 5개 모형의 모수 추정치, 표준오차 및 모형 선택을 위한 조건부 AIC(cAIC) 값을 나타내

Table 3.2. Estimates (SE) for the crack growth data

model	$\beta_0^{(\mu)}$	$\beta_1^{(\mu)}$	$\beta_0^{(\lambda)}$	$\beta_0^{(\phi)}$	$\beta_1^{(\phi)}$	cAIC
GLM	−5.85(0.11)	2.57(0.09)		−2.71(0.10)		−1418
JGLM	−5.94(0.10)	2.63(0.09)		−2.11(0.20)	−10.5(2.8)	−1432
HGLM1	−5.65(0.09)	2.38(0.07)	−3.37(0.37)	−3.40(0.12)		−1548
HGLM2	−5.69(0.09)	2.41(0.07)	−3.47(0.37)	−2.72(0.25)	−11.5(3.06)	−1561
DHGLM[a]	−5.62(0.07)	2.36(0.05)	−3.41(0.36)	−3.01(0.26)	−11.5(2.74)	−1621

[a] : the estimate (SE) for $\log(\alpha)$ is -0.41(0.15).

고 있다. Table 3.2에서 보는 바와 같이 5가지 모형 중 DHGLM이 가장 작은 cAIC 값을 가지고 있어, DHGLM이 자료에 가장 잘 맞는 모형이라고 할 수 있다. 즉, 평균뿐만 아니라 산포에도 각 샘플간의 차이를 고려해야 함을 의미한다. 평균에 대해서는 직전 균열 크기의 효과인 $\beta_1^{(\mu)}$은 양의 값, 산포에 대해서는 각 관측주기의 효과인 $\beta_1^{(\phi)}$는 음의 값으로 추정되며, 둘 다 통계적으로 유의하다. 즉, 직전 균일 클수록 균열의 차이에 대한 평균은 커지며, 주기가 높아질수록 산포는 작아짐을 의미하다.

4. 토론

본 연구에서는 SRC-stat 통계 패키지를 통해서 DHGLM 모형 적합이 가능함을 실제 예제를 통해서 보였다. DHGLM은 평균뿐만 아니라 산포에도 변량효과를 고려한 모형으로써 복잡하고 다양한 통계모형을 생성할 수 있음을 기존 연구들에서 보였다. 특히, 서로 상관이 있는 변량효과도 적합이 가능하여 질변지도 등에 응용될 수 있음을 알 수 있었다. 향후 DHGLM이 적용되리라 보는 의학, 유전학, 금융 자료 등에 SRC-stat의 유용한 적용이 기대된다.

References

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society B*, **58**, 619–678.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalized linear models: A synthesis of generalized linear models, random-effect models and structured dispersions, *Biometrika*, **88**, 987–1006.

Lee, Y. and Nelder, J. A. (2006). Double Hierarchical Generalized Linear Models (with discussion), *Applied Statistics*, **55**, 139–185.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via h-Likelihood*. Chapman & Hall, London.

Lu, C. J. and Meeker, W. Q. (1993). Using degeneration measurements to estimate a time-to-failure distribution, *Technometrics*, **35**, 161–174.

Noh, M. and Lee, Y. (2007). Robust modeling for inference from generalized linear model classes, *Journal of the American Statistical Association*, **102**, 1059–1072.

Noh, M. and Lee, Y. (2011). ***dhglm:*** *Double Hierarchical Generalized Linear Models.* R package version 1.0, URL http://CRAN.R-project.org/pkackage=dhglm.

Noh, M., Lee, Y. and Pawitan, Y. (2005). Robust ascertainment-adjusted parameter estimation, *Genetic Epidemiology*, **29**, 68–75.

이중 다단계 일반화 선형모형 적합을 위한 SRC-stat의 사용

노맹석[a,1] · 하일도[a] · 이영조[b] · 임요한[b] · 이재용[b] · 오희석[b] · 신동완[b] · 이상구[b] · 서진욱[b] · 박용태[b] · 조성준[b] · 박종헌[b] · 김유경[b] · 유경상[b]

[a]부경대학교 통계학과, [b]서울대학교 데이터과학과 지식창출 연구센터

(2015년 3월 24일 접수, 2015년 4월 7일 수정, 2015년 4월 8일 채택)

요 약

본 논문에서는 SRC-Stat 통계패키지를 이용하여 변량효과를 적합하는 방법에 대해서 소개하고자 한다. 본 패키지를 통하여 단변량 평균 뿐만 아니리 산포 및 분산에도 변량효과를 고려하는 이중 다단계 일반화 선형모형을 적합할 수 있다. 고정효과 및 변량효과의 추정치는 다단계 우도 방법을 이용하고 있으며, 실제 자료 적합을 통해 패키지의 사용법에 대해서 설명하고자 한다.

주요용어: 질병지도, 이중 다단계 일반화 선형 모형, 다단계 우도, 변량효과, SRC-stat.

이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2011-0030811).

[1]교신저자: (608-737) 부산광역시 남구 용소로 45, 부경대학교 통계학과. E-mail: msnoh@pknu.ac.kr