The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.1.123
31(1), 123–137

Robust ridge regression for nonlinear mixed effects models with applications to quantitative high throughput screening assay data

Jiseon Yoo[a] · Changwon Lim[a,1]

[a]Department of Applied Statistics, Chung-Ang University

(Received November 16, 2017; Revised December 26, 2017; Accepted December 28, 2017)

Abstract

A nonlinear mixed effects model is mainly used to analyze repeated measurement data in various fields. A nonlinear mixed effects model consists of two stages: the first-stage individual-level model considers intra-individual variation and the second-stage population model considers inter-individual variation. The individual-level model, which is the first stage of the nonlinear mixed effects model, estimates the parameters of the nonlinear regression model. It is the same as the general nonlinear regression model, and usually estimates parameters using the least squares estimation method. However, the least squares estimation method may have a problem that the estimated value of the parameters and standard errors become extremely large if the assumed nonlinear function is not explicitly revealed by the data. In this paper, a new estimation method is proposed to solve this problem by introducing the ridge regression method recently proposed in the nonlinear regression model into the first-stage individual-level model of the nonlinear mixed effects model. The performance of the proposed estimator is compared with the performance with the standard estimator through a simulation study. The proposed methodology is also illustrated using quantitative high throughput screening data obtained from the US National Toxicology Program.

Keywords: dose-response study, toxicology, pharmacology, repeated measurement data, ridge regression

1. 서론

계층적 비선형 모형(hierarchical nonlinear model)이라고도 불리우는 비선형 혼합효과 모형(nonlinear mixed effects model)은 생물학 (Gerhard 등, 2014), 농업학 (Craig와 Schinckel, 2001; Aggrey, 2009), 환경과학 (Yeap 등, 2003), 의약학 (Morrell 등, 1995; Samson 등, 2006; Stirnemann 등, 2012; Nguyen 등, 2012; Rajeswaran와 Blackstone, 2017), 산림학 (Fang과 Bailey, 2001; Garber와 Maguire, 2003; Hall과 Clutter, 2004; Calegario 등, 2005; Zhao 등, 2005; Wang 등, 2007) 등 다양한 분야에서 반복 측정 자료를 분석할 때 주로 사용된다. 예를 들어, Yeap 등 (2003)은 오존에 장기간 노출된 쥐의 기도에 어떤 반응이 나타나는지를 연구한 실험자료를 비선형 혼합효과 모형을 사용하여 분석하였고, Craig와

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education (NRF-2017R1D1A1B03034509).
[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: clim@cau.ac.kr

Schinckel (2001)은 돼지의 성장함수에 대한 비선형 혼합효과 모형을 고려하였다. 또한, Samson 등 (2006)은 HIV 임상시험에서 치료 시작 후 바이러스 부하의 감소를 모델링하는데 비선형 혼합효과 모형을 사용하였고, Calegario 등 (2005)은 브라질 연안 지역에서의 유칼립투스의 높이 성장 패턴을 나타내기 위한 비선형 혼합효과 모형을 개발하였다. 최근 들어, Rajeswaran와 Blackstone (2017)은 수술 후 폐기능에 대한 단일 대 이중 폐 이식의 영향을 비교하기 위하여 다단계 비선형 혼합효과 모형을 제안하였다.

비선형 혼합효과 모형은 지난 수십년 동안 활발히 연구되어 왔다. Davidian과 Giltinan (1993a)은 개체 내 분산이 일정하지 않은 경우 분산함수를 가정하고 모든 개체의 정보를 결합하여 개체 내 분산구조를 추정하는 방법을 제안하였다. Davidian과 Giltinan (2003)은 비선형 혼합효과 모형의 전반적인 방법론과 이론, 그리고 그때까지의 관련 연구과 응용에 대하여 포괄적으로 기술하였다. Meza 등 (2012)은 비선형 혼합효과 모형에서 일반적으로 가정하는 정규분포가 아닌 다른 분포들, 특히 꼬리가 두꺼운 분포를 가정했을 경우의 모수 추정방법을 제안하였다. Williams 등 (2015)은 이상점에 로버스트한 비선형 혼합효과 모형에서의 추정방법에 대하여 제안하였다. 최근 들어, Bogacka 등 (2017)은 비선형 혼합효과 모형에서 공변량이 있는 경우 최적 설계를 제안하였다. 그 외에도, Lindstrom과 Bates (1990), Davidian과 Gallant (1993), Davidian과 Giltinan (1993b), Yeap과 Davidian (2001), Lee와 Xu (2004) 등을 참조하라.

비선형 혼합효과 모형은 개체 내 변동(intra-individual variation), 개체 간 변동(inter-individual variation)에 대해 고려하는 두 단계로 구성되어 있다. 첫 번째 단계는 개체 내 변동에 대해서 고려하며 이를 개별수준모델(individual-level model)이라 한다. 두 번째 단계는 개체 간 변동에 대해 고려하며 이를 개체군모델(population model)이라 한다. 비선형 혼합효과 모형의 첫 번째 단계인 개별수준모델은 비선형 회귀모형의 모수를 추정하는 것으로 일반적인 비선형 회귀모형과 같다. 비선형 회귀모형은 주로 보통최소제곱추정(ordinary least square estimator; OLSE) 방법을 사용하여 모수를 추정한다. 그러나 OLSE 방법은 자료가 전체적으로 가정된 비선형 함수를 따르지 않을 경우 등 여러 가지의 경우에서 모수의 추정 값을 매우 크게 추정하며, 이에 따라 모수 추정 값의 표준오차 또한 극단적으로 크게 추정되며 안정적인 결과를 보이지 못한다. 이렇듯 모수의 추정값과 그 표준오차가 극단적으로 크게 추정되는 경우 왈드 검정(Wald test)을 하는데 있어서 모형이 굉장히 유의한 경우임에도 귀무가설을 기각하지 못하는 경우가 발생하게 된다. 이러한 경우에 모수의 추정값과 그 표준오차를 안정화 시키는 것은 중요한 문제이다. 최근에, Lim (2015)은 선형 회귀모형에서 비슷한 경우인 다중공선성 문제를 해결하기 위해 제안된 능형회귀(ridge regression) 방법을 비선형 회귀모형에 도입함으로써 이 문제를 해결할 수 있는 새로운 추정량을 제안하였다.

비선형 혼합효과 모형에서 첫 번째 단계인 개별수준모델에서 어떤 개체에 대하여 위에서 기술한 것과 같은 문제가 발생할 경우 그 개체에 대한 개별 모수 추정 뿐만 아니라 전체 모집단에 대한 모수, 즉 고정효과의 추정에도 영향을 미칠 수 있다. 따라서, 우리는 본 연구에서 비선형 혼합효과모형의 첫 번째 단계인 개별수준모델에서 Lim (2015)이 제안한 능형회귀 방법을 사용하고 두 번째 단계인 개체군모델에서는 그 능형회귀 방법을 사용하여 구한 개별 모수의 추정값을 사용하여 고정효과를 추정하는 비선형 혼합효과 모형에서의 새로운 추정방법을 제안하고자 한다. 본 논문은 총 5장으로 구성되어 있다. 2장에서는 표준적인 비선형 혼합효과 모형에 대하여 기술하고, 기존의 추정방법들과 새롭게 제안하는 비선형 혼합효과 모형에서 능형회귀 방법을 설명하였다. 3장에서는 모의실험을 통하여 추정방법들의 성능을 비교하였다. 4장에서는 실제 자료를 통하여 추정방법들을 비교하였다. 마지막으로 5장에서는 본 연구를 종합적으로 정리하고 추후의 연구 주제에 대하여 논하였다.

2. 방법론

2.1. 현재의 방법들

본 논문에서 고려하는 비선형 회귀모형은 다음과 같다.

$$y_{ij} = f(x_{ij}, \theta_i) + \epsilon_{ij}, \quad i = 1, \ldots, m, \ j = 1, \ldots, n_i \tag{2.1}$$

여기에서, x_{ij}와 y_{ij}는 각각 i번째 개체에서의 j번째 공변량과 반응변수이고, $f(x_{ij}, \theta_i)$는 비선형 함수이고 θ_i는 모수들의 $p \times 1$ 벡터 $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ip})'$이고, ϵ_{ij}는 정규분포를 따르는 확률오차로 $E(\epsilon_{ij}|\theta_i) = 0$, $\mathrm{Var}(\epsilon_{ij}|\theta_i) = \sigma^2$을 만족하며 총 표본크기는 $n = \sum_{i=1}^{m} n_i$이다.

2.1.1. 비선형 혼합효과 모형 비선형 혼합효과 모형은 두 단계로 이루어진다. 첫 번째 단계는 개체 내 변동(intra-individual variation)을 고려하며 개별 모수의 추정값을 구하는 개별수준모델(individual-level model)이다. 두 번째 단계는 개체 간 변동(inter-individual variation)에 대해 고려하며 개별수준모델의 개별 모수들을 확률변수로 가정하고 첫 번째 단계에서 구한 개별 모수들의 추정값을 사용하여 모수의 분포를 추정하는 개체군모델(population model)이다 (Davidian과 Giltinan, 1995).

제 1단계: 개별수준모델 비선형 혼합효과 모형의 첫 번째 단계인 개별수준모델은 위의 비선형 회귀모형의 식 (2.1)과 같다. 이 단계에서는 m개의 개체들에 대해서 OLSE 방법 등을 이용하여 모수를 추정할 수 있으며 그 식은 다음와 같다:

$$\hat{\theta}_i = \mathrm{argmin} \left[\sum_{j=1}^{n_i} (y_{ij} - f(x_{ij}, \theta_i))^2 \right].$$

여기에서, $f(x_{ij}, \theta_i)$는 i번째 개체에서의 j번째 공변량 x_{ij}와 모수들의 벡터 θ_i의 비선형 함수이다.

위의 식으로부터 구한 $\hat{\theta}_i$는 θ_i가 주어졌을 때 점근적으로 다음 식과 같은 정규분포를 따르게 된다:

$$\hat{\theta}_i|\theta_i \sim N(\theta_i, C_i) \tag{2.2}$$

여기에서, C_i는 $\hat{\theta}_i$의 분산-공분산행렬로서 $C_i = \sigma^2(F_i'F_i)^{-1}$이고, F_i는 $(f(x_{i1}, \theta_i), \ldots, f(x_{in_i}, \theta_i))'$를 θ_i로 미분한 $n_i \times p$ 행렬로서 $F_i = \{\partial f(x_{ij}, \theta_i)/\partial\theta_{il}\}_{j,l}$이다. 비선형모형에서 C_i는 일반적으로 θ_i에 의존한다. 따라서, 제 2단계 개체군모델에서는 θ_i를 $\hat{\theta}_i$로 대체한 $\hat{C}_i$를 대신 사용한다.

제 2단계: 개체군모델 비선형 혼합효과 모형의 두 번째 단계인 개체군모델은 다음과 같다:

$$\theta_i = d(a_i, \theta, b_i), \quad i = 1, \ldots, m.$$

여기에서, d는 p-차원의 벡터값(vector-valued) 함수, a_i는 i번째 개체의 개별 특징을 나타내는 $a \times 1$ 공변량 벡터, θ는 $r \times 1$ 고정효과(fixed effect) 벡터, b_i는 $q \times 1$ 랜덤효과(random effect) 벡터를 나타낸다. 또한 b_i는 $q \times 1$ 벡터 0이 평균벡터이고 $q \times q$ 행렬 D_q가 분산-공분산행렬인 다변량정규분포를 따른다. 즉, 개체군모델에서는 θ_i가 어떤 알려진 분포를 따른다고 가정한다. 주로 다변량정규분포나 다변량로그정규분포를 가정하는데 여기에서는 다변량정규분포를 가정한다.

제 1단계인 개체수준모델 단계에서 개별 모수의 추정값을 구한 후에 제 2단계인 개체군모델에서 고정효과벡터 θ와 랜덤효과벡터 b_i의 분산-공분산 행렬인 D_q를 추정하게 되며 잘 알려진 추정방법은 표준 2단

계(standard two-stage; STS) 방법과 전체 2단계(global two-stage; GTS) 방법이 있다 (Davidian과 Giltinan, 1995). 이 두 방법에 대해 간략하게 기술하기 전에 먼저 제 2단계 개체군모델에서 d 함수가 다음과 같다고 가정한다:

$$\theta_i = A_i\theta + b_i. \tag{2.3}$$

여기에서, A_i는 i번째 개체의 $p\times r$ 공변량 행렬이고, b_i는 $p\times 1$ 랜덤효과벡터이다. 이 가정 하에서 θ_i는 $p\times 1$ 평균벡터가 $A_i\theta$이고, $p\times p$ 분산-공분산 행렬이 D_p인 다변량정규분포를 따르게 된다.

STS 방법에서는 $\hat{\theta}_i$를 마치 θ_i인 것처럼 가정하고 θ와 D_p를 추정한다. 따라서 그 추정량들은 다음과 같이 구할 수 있다:

$$\hat{\theta}_{\text{STS}} = \left(\sum_{i=1}^{m} A_i'A_i\right)^{-1}\left(\sum_{i=1}^{m} A_i'\hat{\theta}_i\right),$$
$$\hat{D}_{\text{STS}} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_i - A_i\hat{\theta}_{\text{STS}}\right)\left(\hat{\theta}_i - A_i\hat{\theta}_{\text{STS}}\right)'.$$

이 방법은 계산이 간단하고 다변량정규분포를 가정할 필요가 없다는 장점이 있지만, θ_i의 추정에 관한 불확실성(uncertainty)를 고려하지 않았기 때문에 바람직하지 않은 방법이다 (Davidian과 Giltinan, 1995). 식 (2.3)에 의하면 θ를 추정하기 위해서는 θ_i의 값을 알아야한다. 그러나, θ_i는 i번째 개체에 대한 모수이고 그 참값을 모르기 때문에 식 (2.3) 자체로는 θ를 추정할 수 없다. 직관적이고 간단한 방법은 θ_i의 추정값인 $\hat{\theta}_i$를 대신 사용하는 것이지만, $\hat{\theta}_i$는 추정값이기 때문에 참값과는 같지 않고 모수 추정과정에서 발생하는 그 차이만큼의 불확실성이 θ의 추정과정에 고려되어야 한다.

GTS 방법에서는 θ_i의 추정에 관한 불확실성을 고려하기 위하여 θ_i가 주어졌을 때의 $\hat{\theta}_i$의 점근적 분포를 사용한다. 제 1단계 개별수준모델에서 구한 $\hat{\theta}_i$은 근사적으로 식 (2.2)와 같은 분포를 따른다. 따라서, 이를 식 (2.3)의 가정과 함께 계산하면 $\hat{\theta}_i$의 주변확률분포는 근사적으로 평균벡터가 $A_i\theta$이고 공분산행렬이 $\hat{C}_i + D_p$인 다변량정규분포가 된다. 우리는 이 결과로부터 다음과 같은 대략적 모형식을 얻을 수 있다:

$$\hat{\theta}_i \approx A_i\theta + b_i + e_i \tag{2.4}$$

여기에서, e_i는 $p\times 1$ 평균벡터가 0이고 $p\times p$ 공분산행렬이 $\hat{C}_i$인 다변량정규분포를 따르는 $p\times 1$ 확률벡터이다. 이 대략적 모형으로부터 우리는 최대우도방법을 적용하여 θ와 D_p를 추정할 수 있다. 즉, 로그우도비 함수에 -2를 곱한

$$-2l_{\text{GTS}}(\theta, D_p) = \sum_{i=1}^{m}\log\left|\hat{C}_i + D_p\right| + \sum_{i=1}^{m}\left(\hat{\theta}_i - A_i\theta\right)'\left(\hat{C}_i + D_p\right)^{-1}\left(\hat{\theta}_i - A_i\theta\right)$$

를 최대화하는 θ와 D_p를 구하는 것이고 그 결과 추정량은 다음과 같다:

$$\hat{\theta} = \left(\sum_{i=1}^{m} A_i'\left(\hat{C}_i + \hat{D}_p\right)^{-1}A_i\right)^{-1}\sum_{i=1}^{m} A_i'\left(\hat{C}_i + \hat{D}_p\right)^{-1}\hat{\theta}_i,$$
$$\hat{D}_p = m^{-1}\sum_{i=1}^{m}\hat{c}_i\hat{c}_i' + m^{-1}\sum_{i=1}^{m}\left(\hat{C}_i^{-1} + \hat{D}_p^{-1}\right)^{-1}$$

여기에서, $\hat{c}_i = (\hat{C}_i^{-1} + \hat{D}_p^{-1})^{-1}\hat{C}_i^{-1}(\hat{\theta}_i - A_i\hat{\theta})$이다 (Davidian과 Giltinan, 1995).

2.1.2. 비선형 회귀모형에서의 능형 M-추정 방법 비선형 회귀모형 (2.1)에서 회귀계수벡터 θ_i를 추정하는 보통 능형 M-추정량(ordinary ridge M-estimator; ORME)은 다음과 같이 정의할 수 있다 (Lim, 2015):

$$\hat{\theta}_i^o(k) = \text{argmin}\left[\sum_{j=1}^{n_i} h^2(y_{ij} - f(x_{ij}, \theta_i)) + k\theta_i'\theta_i\right], \tag{2.5}$$

여기에서, h는 Huber score function (Huber, 1981)으로 아래와 같다:

$$h(u) = \begin{cases} \dfrac{u}{\sqrt{2}}, & \text{if } |u| < k_0, \\ k_0\left(|u| - \dfrac{k_0}{2}\right)^{\frac{1}{2}}, & \text{otherwise,} \end{cases}$$

여기에서, k_0는 미리 정해놓은 양의 상수로서 일반적으로 1과 2 사이의 값이 사용된다. 만약 $h(u) = u$이면 ORME은 보통능형추정량(ordinary ridge estimator; ORE)이 된다.

수식의 복잡함을 피하기 위해 ORME를 간단히 $\hat{\theta}_i$로 표시하면, 식 (2.5)에서 최소화 문제를 풀기위한 추정식은 아래와 같다:

$$\sum_{j=1}^{n_i} \psi\left(y_{ij} - f\left(x_{ij}, \hat{\theta}_i\right)\right) f_{\theta_i}\left(x_{ij}, \hat{\theta}_i\right) - k\hat{\theta}_i = 0$$

여기에서, $f_{\theta_i}(x_{ij}, \theta_i) = (\partial/\partial\theta_i)f(x_{ij}, \theta_i)$이고, $\psi(u) = (\partial/\partial u)h^2(u)$이다. 이 추정식으로부터 우리는 ORME의 점근적 선형성을 유도할 수 있고, 그로부터 다음과 같은 식을 얻을 수 있다:

$$\hat{\theta}_i^o(k) = \left(\gamma_{4i}F_i'F_i + kI_p\right)^{-1} F_i'\left(\gamma_{4i}F_i\theta_i + \psi(\theta_i)\right) + o_p\left(n_i^{-\frac{1}{2}}\right)$$

여기에서, $\gamma_4 = E\psi'(y_{i1} - f(x_{i1}, \theta_i))$이고, F_i는 앞에서 정의한 바와 같이 $n_i \times p$ 행렬으로 (j,l)번째 원소는 $(\partial/\partial\theta_{il})f(x_{ij}, \theta_i)$이며 $\hat{F}_i$는 F_i의 식에서 θ_i 대신 $\hat{\theta}_i$를 대입해서 계산한 행렬이다. 그리고 $\psi(\theta_i)$는 $n_i \times 1$ 벡터로 j번째 원소는 $\psi(y_{ij} - f(x_{ij}, \theta_i))$이다. 이로부터 우리는 다음과 같은 점근적 성질을 유도할 수 있다 (Lim, 2015):

(O1) $\hat{\theta}_i^o(k)$는 점근적으로 정규분포를 따른다.

(O2) $\lim_{k\to 0}\hat{\theta}_i^o(k) = \hat{\theta}_i^o(0)$로 보통 M-추정량(ordinary M-estimator; OME)이 된다.

(O3) $\text{Bias}(\hat{\theta}_i^o(k)) = E(\hat{\theta}_i^o(k)) - \theta_i = -kS_i(k)^{-1}\theta_i$, 여기에서, $S_i(k) = \gamma_{4i}F_i'F_i + kI_p$이다.

(O4) $\text{Var}(\hat{\theta}_i^o(k)) = \sigma_{\psi 3i}^2 S_i(k)^{-1}F_i'F_iS_i(k)^{-1}$, 여기에서, $\sigma_{\psi 3i}^2 = E\psi^2(y_{i1} - f(x_{i1}, \theta_i))$이다.

(O5) $\text{MSE}(\hat{\theta}_i^o(k)) = S_i(k)^{-1}(\sigma_{\psi 3i}^2 F_i'F_i + k^2\theta_i\theta_i')S_i(k)^{-1}$.

위의 식에서 알 수 있는 것과 같이 ORME의 mean square error (MSE)는 능형모수(ridge parameter)인 k에 의존하고 있다. 우리는 본 논문에서 다음과 같은 식으로 능형모수를 구하여 사용한다 (Lim, 2015):

$$\hat{k}_i^o = \frac{c_i\hat{\sigma}_{\psi 3i}^2}{\hat{\gamma}_{4i}^2\hat{\theta}_i'\hat{F}_i'\hat{F}_i\hat{\theta}_i}, \quad c_i = \frac{p-2}{n_i - p + 2} \tag{2.6}$$

여기에서, $\hat{\theta}_i$는 OME이며, $\hat{\gamma}_{4i}$, $\hat{\sigma}_{\psi 3i}^2$, $\hat{F}_i$는 OME를 사용하여 구한 값들이다.

비선형 회귀모형 (2.1)에서 확률오차 ϵ_{ij}가 등분산 가정을 만족하지 않을 때에는 식 (2.5)에서 정의된 ORME를 사용하는 것은 적절하지 않다. 이 경우 ϵ_{ij}의 분산을 $\mathrm{Var}(\epsilon_{ij}) = \{\sigma(z_{ij}, \tau_i)\}^2$라고 가정한다. 여기에서, σ는 알려져있는 함수이고, z_{ij}는 i번째 개체의 j번째 공변량이고, τ_i는 분산모수들의 $q_1 \times 1$ 벡터이다. 이 때 가중 능형 M-추정량(weighted ridge M-estimator; WRME)은 다음과 같이 정의할 수 있다 (Lim, 2015):

$$\begin{pmatrix} \hat{\theta}_i^w(k) \\ \hat{\tau}_i^w(k) \end{pmatrix} = \mathrm{argmin} \left[\sum_{j=1}^{n_i} \left\{ h^2 \left(\frac{y_{ij} - f(x_{ij}, \theta_i)}{\sigma(z_{ij}, \tau_i)} \right) + \log \sigma(z_{ij}, \tau_i) \right\} + k\theta_i' \theta_i \right]. \tag{2.7}$$

우리는 ORME에서와 비슷한 방법으로 다음과 같은 점근적 성질을 유도할 수 있다 (Lim, 2015):

(W1) $\hat{\theta}_i^w(k)$는 점근적으로 정규분포를 따른다.

(W2) $\lim_{k \to 0} \hat{\theta}_i^w(k) = \hat{\theta}_i^w(0)$로 가중 M-추정량(weighted M-estimator, WME)이 된다.

(W3) $\mathrm{Bias}(\hat{\theta}_i^w(k)) = E(\hat{\theta}_i^w(k)) - \theta_i = -kT_i(k)^{-1}\theta_i$, 여기에서, $T_i(k) = \gamma_{2i} F_i' W_i^2 F_i + kI_p$이고, $\gamma_{2i} = E\psi'(y_{i1} - f(x_{i1}, \theta_i)/\sigma(z_{i1}, \tau_i))$이고, W_i는 j번째 대각원소가 $1/\sigma(z_{ij}, \tau_i)$인 $n_i \times n_i$ 대각행렬이다.

(W4) $\mathrm{Var}(\hat{\theta}_i^w(k)) = \sigma_{\psi 1i}^2 T_i(k)^{-1} F_i' W_i^2 F_i T_i(k)^{-1}$, 여기에서, $\sigma_{\psi 1i}^2 = E\psi^2(y_{i1} - f(x_{i1}, \theta_i)/\sigma(z_{i1}, \tau_i))$이다.

(W5) $\mathrm{MSE}(\hat{\theta}_i^w(k)) = T_i(k)^{-1} \left(\sigma_{\psi 1i}^2 F_i' W_i^2 F_i + k^2 \theta_i \theta_i' \right) T_i(k)^{-1}$.

WRME를 구하기 위해 사용하는 능형모수는 다음과 같은 식으로 구한다 (Lim, 2015):

$$\hat{k}_i^w = \frac{c_i \hat{\sigma}_{\psi 1i}^2}{\hat{\gamma}_{2i}^2 \hat{\theta}_i' \hat{F}_i' \hat{W}_i^2 \hat{F}_i \hat{\theta}_i}, \quad c_i = \frac{p-2}{n_i - p + 2}, \tag{2.8}$$

여기에서, $\hat{\theta}_i$는 WME이며, $\hat{\gamma}_{2i}$, $\hat{\sigma}_{\psi 1i}^2$, $\hat{F}_i$, $\hat{W}_i$는 WME를 사용하여 구한 값들이다.

2.2. 제안된 추정방법

우리는 이제 비선형 혼합효과 모형에서의 두 종류의 능형 M-추정량을 제안한다. 먼저 비선형 혼합효과 모형에서의 ORME는 다음과 같이 정의된다:

제 1단계: 개별수준모델 제 1단계인 개별수준모델에서는 i번째($i = 1, \ldots, m$) 개체에 대하여 ORME 식 (2.5)를 이용하여 모수 θ_i를 추정한다. 모수 추정량 $\hat{\theta}_i^o(k)$는 θ_i가 주어졌을 때 식 (2.2)와 같은 정규분포를 따르게 된다. 여기에서 분산-공분산행렬을 $C_i^o(k)$라고 하고 ORME의 점근적 성질 (O4)로부터 $C_i^o(k) = \sigma_{\psi 3i}^2 S_i(k)^{-1} F_i' F_i S_i(k)^{-1}$이다. 제 2단계 개체군모델에서는 θ_i를 $\hat{\theta}_i^o(k)$로 대체한 $\hat{C}_i^o(k)$를 대신 사용한다. k 또한 $\hat{k}_i^o$로 대체하여 사용한다.

제 2단계: 개체군모델 제 2단계 개체군모델에서 식 (2.3)을 가정할 때 GTS 방법을 사용하여 고정효과벡터 θ와 랜덤효과벡터의 분산-공분산행렬 D_p의 추정량들을 다음과 같이 구할 수 있다:

$$\hat{\theta}^o = \left(\sum_{i=1}^m A_i' \left(\hat{C}_i^o \left(\hat{k}_i^o \right) + \hat{D}_p^o \right)^{-1} A_i \right)^{-1} \sum_{i=1}^m A_i' \left(\hat{C}_i^o \left(\hat{k}_i^o \right) + \hat{D}_p^o \right)^{-1} \hat{\theta}_i^o \left(\hat{k}_i^o \right),$$

$$\hat{D}_p^o = m^{-1} \sum_{i=1}^m \hat{c}_i \hat{c}_i' + m^{-1} \sum_{i=1}^m \left(\left(\hat{C}_i^o \left(\hat{k}_i^o \right) \right)^{-1} + \left(\hat{D}_p^o \right)^{-1} \right)^{-1}$$

여기에서, $\hat{c}_i = ((\hat{C}_i^o(\hat{k}_i^o))^{-1} + (\hat{D}_p^o)^{-1})^{-1}(\hat{C}_i^o(\hat{k}_i^o))^{-1}(\hat{\theta}_i^o(\hat{k}_i^o) - A_i\hat{\theta}^o)$이다.
비선형 혼합효과 모형에서의 WRME는 ORME와 같은 방식으로 다음과 같이 정의된다:

제 1단계: 개별수준모델 제 1단계인 개별수준모델에서는 i번째($i = 1, \ldots, m$) 개체에 대하여 WRME (2.7)를 이용하여 모수 θ_i를 추정한다. 모수 추정량 $\hat{\theta}_i^w(k)$는 θ_i가 주어졌을 때 식 (2.2)와 같은 정규분포를 따르게 된다. 여기에서 분산-공분산행렬을 $C_i^w(k)$라고 하고 WRME의 점근적 성질 (W4)로부터 $C_i^w(k) = \sigma_{\psi 1i}^2 T_i(k)^{-1} F_i' W_i^2 F_i T_i(k)^{-1}$이다. 제 2단계 개체군모델에서는 θ_i를 $\hat{\theta}_i^w(k)$로 대체한 $\hat{C}_i^w(k)$를 대신 사용한다. k 또한 $\hat{k}_i^w$로 대체하여 사용한다.

제 2단계: 개체군모델 제 2단계 개체군모델에서 식 (2.3)을 가정할 때 GTS 방법을 사용하여 고정효과벡터 θ와 랜덤효과벡터의 분산-공분산행렬 D_p의 추정량들을 다음과 같이 구할 수 있다:

$$\hat{\theta}^w = \left(\sum_{i=1}^m A_i' \left(\hat{C}_i^w \left(\hat{k}_i^o \right) + \hat{D}_p^w \right)^{-1} A_i \right)^{-1} \sum_{i=1}^m A_i' \left(\hat{C}_i^w \left(\hat{k}_i^w \right) + \hat{D}_p^w \right)^{-1} \hat{\theta}_i^w \left(\hat{k}_i^w \right),$$
$$\hat{D}_p^w = m^{-1} \sum_{i=1}^m \hat{c}_i \hat{c}_i' + m^{-1} \sum_{i=1}^m \left(\left(\hat{C}_i^w \left(\hat{k}_i^w \right) \right)^{-1} + \left(\hat{D}_p^w \right)^{-1} \right)^{-1},$$

여기에서, $\hat{c}_i = ((\hat{C}_i^w(\hat{k}_i^w))^{-1} + (\hat{D}_p^w)^{-1})^{-1}(\hat{C}_i^w(\hat{k}_i^w))^{-1}(\hat{\theta}_i^w(\hat{k}_i^w) - A_i\hat{\theta}^w)$이다.

2.3. 정량적 고속대량 스크리닝 자료 분석

비선형 혼합효과 모형에서의 능형회귀 추정방법은 뒤에서 자세히 설명할 정량적 고속대량 스크리닝(quantitative high throughput screening; qHTS) 자료를 분석하기 위하여 제안되었다. 지금까지 qHTS 자료는 비선형 (고정효과) 모형을 사용하여 분석되어 왔다 (Lim 등, 2013a, 2013b; Peddada, 2013; Lim, 2015). 비선형 혼합효과 모형을 적용하여 qHTS 자료를 분석한 연구는 지금까지는 이루어진 적이 없었고 본 논문에서 처음으로 고려되고 있다.

앞에서 기술한 것처럼 비선형 혼합효과 모형의 제 2단계인 개체군모델에서는 $p \times 1$ 벡터인 θ_i에 대하여 일반적으로 다변량정규분포 (또는 다변량로그정규분포)를 가정한다. 그러나 이 가정은 qHTS 자료를 비선형 혼합효과 모형으로 분석할 경우에는 성립하지 않는다. 우리는 실제 qHTS 자료로 제 1단계 개별수준모델을 적합시켰을 때 구한 θ_i의 일부 원소의 추정값들로 히스토그램을 그린 후에, 이 히스토그램들을 바탕으로 θ_i의 일부 원소들이 각각 두 정규분포의 혼합분포(mixture of two normal distributions)를 따른다고 가정하였다. 이렇게 비선형 혼합효과 모형의 제 2단계에서 다변량정규분포나 다변량로그정규분포를 가정하지 못하는 경우에는 모수 추정 과정이 더 복잡하게 된다. 따라서 본 논문에서는 제안된 추정방법인 비선형 혼합효과 모형에서의 능형회귀 추정방법으로 qHTS 자료를 분석하는 과정에 대한 설명을 위해, 오차의 분산이 같다는 등분산 가정을 하고 보통능형추정량(ORE)를 사용하고자 한다.

비선형 혼합효과모형에서의 ORE 방법은 제 1단계 개체수준모델에서 i번째 개체에서의 모수 θ_i를 다음과 같은 식으로 추정한다:

$$\hat{\theta}_i(k) = \text{argmin} \left[\sum_{j=1}^{n_i} (y_{ij} - f(x_{ij}, \theta_i))^2 + k\theta_i'\theta_i \right], \tag{2.9}$$

여기에서, k는 능형모수로 식 (2.6)을 사용하여 구하며, ORE의 경우, $h(u) = u$이기 때문에,

$$\hat{k}_i = \frac{c_i \hat{\sigma}^2}{\hat{\theta}_i' \hat{F}_i' \hat{F}_i \hat{\theta}_i}$$

이고 $\hat{\theta}_i$, $\hat{F}_i$, $\hat{\sigma}^2$은 OLS 방법을 사용하여 구한 추정값들이다. 이 때의 분산-공분산 행렬은 $C_i = \mathrm{Var}(\hat{\theta}_i) = \sigma^2 S_i^{-1} F_i^{'} F_i S_i^{-1}$로 구할 수 있으며 여기서 $S_i = F_i^{'} F_i + kI_p$이며 F_i는 앞에서 정의한 것과 같은 $(n_i \times p)$행렬이다. 앞에서 설명한 것과 같이 C_i는 θ_i에 의존하기 때문에 θ_i를 $\hat{\theta}_i(k)$로 대체한 $\hat{C}_i(k)$를 대신 사용한다.

혼합정규분포는 정규분포 여러 개가 선형결합 되어 있는 형태로 복수의 정규 확률밀도함수로 자료의 분포를 모델링할 수 있는 분포이다. 이론적으로 무한개의 정규분포가 혼합된 혼합정규분포도 존재하지만 본 논문에서는 유한개의 정규분포에 대한 혼합정규분포만을 고려하며 그 확률밀도함수는 다음과 같이 표현된다:

$$p\left(t \,|\, \lambda, \{\mu_j, \sigma_j^2\}\right) = \sum_{j=1}^{J} \lambda_j \phi\left(t \,|\, \mu_j, \sigma_j^2\right),$$

여기에서, t는 혼합정규분포를 따르는 확률변수의 관찰값이고, $\phi(\cdot\,|\mu_j, \sigma_j^2)$는 평균이 μ_j이고 분산이 σ_j^2인 정규분포의 확률밀도함수이고, $\lambda_j > 0$, $j = 1, \ldots, J$는 혼합계수이고 $\sum_{j=1}^{J} \lambda_j = 1$이다. 위의 식에서 알 수 있듯이 주어진 t값에 대해서 혼합정규분포의 확률밀도함수의 값은 J개의 정규분포의 확률밀도함수의 t에서의 값들의 가중평균으로 구해진다. 혼합정규분포의 모수들인 혼합계수 $\lambda = (\lambda_1, \ldots, \lambda_J)$와 각 정규분포의 평균과 분산인 $((\mu_1, \sigma_1^2), \ldots, (\mu_J, \sigma_J^2))$는 주어진 자료를 이용하여 EM 알고리즘을 통해서 추정한다.

EM 알고리즘은 반복 과정을 통하여 각 개체들이 혼합 모형에 속할 가능성을 조정하여 최적의 모델을 만드는 방법으로 Expectation step과 Maximization step으로 구성되어 있으며, 자료 $t_1, \ldots, t_m$이 주어져 있을 때 혼합정규분포의 모수 추정을 위한 EM 알고리즘의 순서는 다음과 같다 (Bishop, 2007):

(1) 혼합계수들과 각 정규분포의 평균과 분산들의 초기값을 구한다.

(2) E-Step: 각 모수들의 현재의 값에 대해서 t_i가 관측되었을 때 그것이 j번째 정규분포에서 발생했을 조건부 확률인 사후확률을 다음 식과 같이 구한다:

$$p(z_j = 1|t_i) = \frac{\lambda_j \phi\left(t_i \,|\, \mu_j, \sigma_j^2\right)}{\sum_{l=1}^{J} \lambda_l \phi\left(t_i \,|\, \mu_l, \sigma_l^2\right)}$$

여기에서, z_j, $j = 1, \ldots, J$는 관측값이 j번째 정규분포에서 발생했을 때에만 1이고 나머지는 0인 확률변수로서 결합확률분포는 확률이 각각 λ_j인 다항분포이다. 그리고 이때의 로그 우도함수는 아래와 같다:

$$\ln p\left(t_1, \ldots, t_m | \lambda, \{\mu_j, \sigma_j^2\}\right) = \sum_{i=1}^{m} \ln \sum_{j=1}^{J} \lambda_j \phi\left(t_i | \mu_j, \sigma_j^2\right).$$

(3) M-Step: E-step에서 계산한 로그 우도함수를 최대화하며 모수를 재추정하는 단계로서 그 추정량들의 식은 다음과 같다:

$$\hat{\mu}_j = \frac{1}{N_j} \sum_{i=1}^{m} P(z_j = 1|t_i) t_i, \quad \hat{\sigma}_j^2 = \frac{1}{N_j} \sum_{i=1}^{m} P(z_j = 1|t_i)(t_i - \mu_j)(t_i - \mu_j)', \quad \hat{\lambda}_j = \frac{N_j}{m}$$

여기에서, $N_j = \sum_{i=1}^{m} P(z_j = 1|t_i)$이다.

EM 알고리즘의 E-step과 M-step은 수렴조건이 만족할 때까지 반복된다. 그리고 최종적으로 수렴될 때의 값 $(\hat{\lambda}_1, \ldots, \hat{\lambda}_J)$와 $\{\hat{\mu}_j, \hat{\sigma}_j^2\}$을 혼합정규분포의 분포모수로 추정한다.

본 연구의 모의실험과 실제자료분석에서는 $J = 2$인 경우를 고려한다. 이러한 가정 하에서 비선형 혼합효과 모형의 제 2단계 개체군모델에서 혼합정규분포의 분포모수인 $(\lambda_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$은 유사 가능도(pseudo likelihood; PL) 방법을 사용하여 추정하며 그 식은 다음과 같다 (Arnold와 Strauss, 1991):

$$\left(\hat{\lambda}_1, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2\right) = \text{argmin} \sum_{i=1}^{m} \left[\frac{\left(\hat{\theta}_{il}\left(\hat{k}_i\right) - \theta_l\right)^2}{\sigma_{\theta_l}^2 + \hat{C}_{il}\left(\hat{k}_i\right)} + \log\left(\sigma_{\theta_l}^2 + \hat{C}_{il}\left(\hat{k}_i\right)\right) \right]$$

여기에서, $\theta_l = \sum_{j=1}^{2} \lambda_j \mu_j$, $\sigma_{\theta_l}^2 = \sum_{j=1}^{2} \lambda_j \left\{(\mu_j - \theta_l)^2 + \sigma_j^2\right\}$, $\lambda_2 = 1 - \lambda_1$이고, l은 추정하고자하는 θ 벡터의 l번째 원소인 모수를 의미하며, C_{il}는 C_i의 l번째 대각원소이다. PL 방법은 수렴할 때까지 반복하는 과정이기 때문에 모수의 초기값이 필요하며, 우리는 제 1단계인 개별수준모델에서 ORE 방법으로 추정한 모수의 값들을 기반으로 EM 알고리즘을 통해 추정한 값을 초기값으로 사용한다.

3. 모의실험 연구

3.1. 실험 계획

본 논문에서 우리는 qHTS 자료를 비선형 혼합효과 모형을 적용하여 분석하고자 하며 이 경우 식 (2.1)에서의 비선형 함수 f는 모수가 4개인 Hill model이 된다 (Xia 등, 2008; Lim 등, 2013a, 2013b; Peddada, 2013; Lim, 2015). Hill model은 생화학 분야에서 화합물의 생체 내(*in vivo*) 용량-반응 관계를 연구하는 데 주로 사용되는 비선형 모형이다 (Hill, 1910). 따라서 우리는 모의실험에서 자료를 생성할 때에도 다음과 같은 Hill model을 사용하였다:

$$y_{ij} = f(x_j, \theta) + \epsilon_{ij} = \theta_0 + \frac{\theta_1 x_j^{\theta_2}}{\theta_3^{\theta_2} + x_j^{\theta_2}} + \epsilon_{ij}, \quad i = 1, \ldots, 3,\ j = 1, \ldots, 14 \tag{3.1}$$

여기에서, 확률오차인 ϵ_{ij}는 평균이 0이고 분산이 σ^2인 정규분포를 따르고, x_j의 값은 qHTS 자료에서 실제 사용한 14가지의 농도값(0.59nM, 2.94nM, 14.7nM, 32.8nM, 73.4nM, 0.164μM, 0.367μM, 0.821μM, 1.835μM, 4.103μM, 9.175μM, 20.52μM, 45.87μM, 91.74μM)으로 정하였다. 모수 $(\theta_0, \theta_1, \theta_3, \theta_4)$에 대해 80개의 모수를 생성하였다. θ_1은 혼합정규분포 $0.6 \times N(35, 10^2) + 0.4 \times N(80, 10^2)$로부터 생성하였으며, θ_3는 혼합정규분포 $0.3 \times N(5, 1^2) + 0.7 \times N(30, 10^2)$로부터 생성하였다. θ_0는 $-\theta_1$으로 설정하였고, θ_2는 $1, 1.5, 2$에서 임의로 선택하여 사용하였다. 자료의 표준편차 σ는 $1, 2, 3, 4$에서 임의로 선택하였다. 위에 기술된 각각의 값은 실제 qHTS자료의 값들과 비슷하게 정하였다.

우리는 본 모의실험 연구에서 기존의 방법과 제안된 방법의 성능을 비교하였다. 기존의 방법은 비선형 혼합효과 모형의 제 1단계 개별수준 모델에서 등분산을 가정하고 최소제곱추정방법에 의해 모수를 추정하는 OLSE이고, 제안된 방법은 최소제곱추정방법 대신 능형회귀방법을 적용하는 ORE이다. 모의실험을 1,000번 반복하여 다음과 같은 두 가지의 기준치에 관하여 각 추정량의 성능을 비교하였다: (1) 모수 θ_1와 θ_3와 관련된 $(\lambda_1, \mu_1, \mu_2, \sigma_1, \sigma_2)$의 MSE, (2) 모수와 관련된 $(\lambda_1, \mu_1, \mu_2, \sigma_1, \sigma_2)$의 편향(bias).

3.2. 결과

모의실험 결과는 Tables 3.1과 3.2에 요약되어 있다. Table 3.1은 각 모수 추정량에 대한 MSE 값을 보여주고 있다. 전반적으로 제 1단계 모델에서 ORE를 적용한 제안된 추정방법을 사용했을 때가

Table 3.1. Simulation result (MSE) based on 1,000 replications for homoscedastic data using OLSE and ORE

Parameter	Method	λ_1	μ_1	μ_2	σ_1	σ_2
θ_1	OLSE	0.0114	15.8251	112.878	12.388	43.8635
	ORE	0.0107	11.4932	38.919	10.864	38.4776
θ_3	OLSE	0.2655	122.786	5448.82	51.511	2576.93
	ORE	0.1411	97.729	866.538	54.456	158.477

MSE = mean square error; OLSE = ordinary least square estimator; ORE = ordinary ridge estimator.

Table 3.2. Simulation result (Bias) based on 1000 replications for homoscedastic data using OLSE and ORE.

Parameter	Method	λ_1	μ_1	μ_2	σ_1	σ_2
θ_1	OLSE	−0.0100	1.3634	−2.2579	−0.1466	3.958
	ORE	−0.0329	0.0825	−1.2955	0.3176	3.825
θ_3	OLSE	0.4111	9.4372	41.4398	4.5153	29.802
	ORE	0.2061	6.2365	13.5717	4.3116	6.163

OLSE = ordinary least square estimator; ORE = ordinary ridge estimator.

OLSE에 기반한 추정방법보다 더 작은 MSE 값이 얻어짐을 확인할 수 있다. 특히, θ_1에서 μ_2의 MSE는 OLSE 방법에서는 112.878이지만, ORE 방법을 사용하게 되면 38.919로 훨씬 줄어들었음을 볼 수 있다. 또한, θ_3의 경우에는, μ_2와 σ_2의 MSE가 OLSE 방법에서는 각각 5448.82와 2576.93으로 굉장히 크게 나오는 반면, ORE 방법을 사용하게 되면 866.538과 158.477로 상대적으로 아주 작게 나오는 것을 알 수 있다. 이 결과를 바탕으로 우리가 제안한 비선형 혼합효과 모형에서의 ORE 방법이 기존의 비선형 혼합효과 모형에서의 OLSE 방법보다 훨씬 더 좋은 성능을 가지고 있음을 알 수 있다. MSE의 감소 정도가 굉장히 큰 것은 그만큼 모수 추정이 안정적이라는 뜻으로 해석할 수 있을 것이다.

Table 3.2는 각 모수 추정량에 대한 편향의 결과를 보여주고 있다. 비선형 혼합효과 모형에서 최소제곱 방법에 기반한 모수 추정량은 점근적으로 불편성을 가지고 있음이 알려져 있다 (Davidian과 Giltinan, 1995). 반면에 비선형 회귀모형에서 능형회귀 추정량은 점근적으로 약간의 편향이 있음이 알려져 있다 (Lim, 2015). 그러나 Table 3.2를 보면 OLSE와 ORE 방법 모두 약간의 편향이 존재함을 알 수 있다. 그리고 전반적으로 ORE의 편향의 값들이 OLSE의 편향의 값들보다 더 작음을 볼 수 있다. 특히, θ_1에서 μ_1과 μ_2의 편향은 각각 1.3634와 −2.2579이지만, ORE 방법을 사용하게 되면 0.0825와 −1.2955로 상당히 줄어들게 됨을 볼 수 있다. 또한, θ_3에서 μ_2와 σ_2의 편향은 각각 41.4398과 29.802로 상당히 크게 나오는 반면, ORE 방법을 사용하게 되면 13.5717과 6.1633으로 훨씬 작게 나오는 것을 알 수 있다.

모의실험 연구로부터 MSE와 편향의 결과를 종합적으로 봤을 때, 우리가 새롭게 제안한 비선형 혼합효과 모형에서의 ORE 방법이 기존의 방법보다 성능이 더 좋다는 것을 알 수 있다. 이는 ORE 방법을 적용함으로써 비선형 혼합효과 모형에서의 모수 추정이 안정화되기 때문이라고 할 수 있다.

4. 실제자료 예시

본 장에서는 제안된 방법을 qHTS 자료를 적용하여 분석을 진행하였다. qHTS 방법은 화합물의 잠재적 독성을 검사하기 위한 분석 방법으로서 미국 National Toxicology Program (NTP), NIH Chemical Genomic Center, National Center for Computational Toxicology, Environmetal Protection Agency, Food and Drug Administration 등이 협약을 맺고 Tox21이라는 공동연구 커뮤니티를 만들어서 개발한 방법이다 (Austin 등, 2008). qHTS 방법은 로봇을 이용한 자동화 시스템과 고감도 탐지기를 이용하여 7개 이상의 농도에서 60,000개가 넘은 화합물질에 대해 동시에 분석을 할 수 있다 (Xia 등, 2008).

Table 4.1. Estimates of the distribution parameters for qHTS data using OLSE and ORE

Parameter	Method	λ_1	μ_1	μ_2	σ_1	σ_2
θ_1	OLSE	0.32	43.03	94.99	14.42	14.91
	ORE	0.38	45.28	96.13	16.53	10.88
θ_3	OLSE	0.24	5.02	24.31	2.19	10.45
	ORE	0.24	4.33	25.07	2.27	10.54

qHTS = quantitative high throughput screening; OLSE = ordinary least square estimator; ORE = ordinary ridge estimator.

본 장에서 사용한 자료는 NTP에서 qHTS 방법을 사용하여 1,408개의 화합물에 대하여 분석을 실시한 자료이다 (Smith 등, 2007; Tice 등, 2007). 이 자료는 앞서 언급한 Hill model (3.1)을 사용하여 분석하였다. 여기서 x는 14가지의 농도(0.59nM, 2.94nM, 14.7nM, 32.8nM, 73.4nM, 0.164μM, 0.367μM, 0.821μM, 1.835μM, 4.103μM, 9.175μM, 20.52μM, 45.87μM, 91.74μM)이다. 그리고 y는 물질이 처리되는 세포의 세포내 adenosine triphospate (ATP) level을 측정한 후 positive control(-100%)과 vehicle control(0%)에 의해 정규화한 측정값이다 (Xia 등, 2008). 각 x의 값마다 3번 반복측정하여 총 표본의 크기는 42이다.

qHTS 자료를 분석하는데 있어서 독성이 있는 화합물과 독성이 없는 화합물은 서로 다른 분포에서 나온다고 가정을 하였으며, 독성을 가진 화합물에 대하여 분포 모수를 추정하기 위하여 비선형 혼합효과 모형을 가정하였다. 분석에서 사용하는 qHTS 자료는 일반적으로 비선형 혼합효과 모형을 적용하는 반복측정 자료라고 보기는 어렵다. 하지만 14가지의 농도에 대하여 3번 실험을 반복하여 얻은 자료의 구조를 고려할 때, 이를 비선형 혼합효과모형에 적용하는 것이 방법론 상 가능하다고 볼 수 있다.

제안된 방법을 적용하여 자료 분석을 수행하기 전에 간단한 가설검정을 통해 독성을 가진 화합물을 찾았으며 그 순서는 다음과 같다:

(1) H_0: $\mu = 0$에 대하여 t-검정을 실시하고 귀무가설을 기각하는 화합물을 찾는다. 여기에서, $\mu = E(y)$이다.

(2) (1)의 결과에서 귀무가설을 기각하는 화합물을 비선형 회귀모형에서 OLSE 방법으로 Hill model의 모수를 추정하고, 모든 모수에 대하여 그 추정값이 유의수준 0.05에서 유의한 화합물을 찾는다.

이러한 순서에 따라 86개의 화합물이 독성을 가진다고 판단하여 이를 비선형 혼합효과 모형에 적용해 분포모수를 추정하였다. 특히, θ_1과 θ_3에 대하여 앞에서 기술한 것과 같이 혼합정규분포를 가정하고 제안된 방법과 기존의 방법으로 분포모수 $(\lambda_1, \mu_1, \mu_2, \sigma_1, \sigma_2)$를 추정하였고, 그 결과는 Table 4.1에 요약되어 있다. 두 방법으로 분포모수를 추정한 결과 추정된 분포모수의 값이 전반적으로 비슷한 것을 확인할 수 있다. 그러나, θ_1에 대하여 OLSE 방법을 사용한 경우 μ_1과 μ_2의 추정값이 각각 43.03과 94.99이고, ORE 방법을 사용한 경우에는 45.28과 96.13로 전체적으로 1에서 2 정도 평균값이 증가하였음을 확인할 수 있다.

5. 결론

본 논문에서 우리는 비선형 혼합효과 모형에서의 보통 능형 M-추정량(ordinary ridge M-estimator; ORME)와 가중 능형 M-추정량(weighted ridge M-estimator; WRME)를 제안하였고, ORME의 특수한 경우인 보통능형추정량(ordinary ridge estimator; ORE)을 정량적 고속 대량 스크리닝(quantitative high throughput screening; qHTS) 자료 분석에 적용하였다. 비선형 혼합효과 모형에서의 표준적인

추정방법은 개별수준모델에서 가정된 비선형함수가 자료에 의해 명시적으로 드러나지 않는 경우 모수의 추정값과 그 표준오차가 극단적으로 커지는 문제가 발생할 수 있다. 그러나 본 논문에서 제안된 방법은 개별수준모델에서 능형회귀 추정방법을 사용하여 그러한 문제점을 해결하고 모수의 추정을 안정화시킬 수 있는 방법이다. 우리는 모의실험 연구와 qHTS 실제 자료 분석을 통하여 제안된 방법이 우수한 성능을 가지고 있음을 보였다.

본 논문에서는 qHTS 자료 분석에 응용하는 것을 목표로 하여 비선형 혼합효과 모형의 제 2단계인 개체군 모델에서 혼합정규분포를 가정하고 추정방법을 제안하였지만, 다른 자료를 분석하는데 있어서는 일반적으로 가정하는 것처럼 정규분포나 로그 정규분포를 가정하여 제안된 방법을 사용할 수 있다. 혼합정규분포를 qHTS 자료분석에 적용할 때, 제 1단계 개체수준모델에서 qHTS 자료의 각 화합물에 대한 모수를 추정한 후 그 추정값들로 그린 히스토그램을 바탕으로 혼합정규분포의 혼합 컴포넌트(mixture component)의 갯수를 2개로 고정시킨 후 분석하였다. 그러나 혼합정규분포의 모수를 추정하기 위해 EM 알고리즘을 적용할 때 혼합 컴포넌트의 갯수에 EM 알고리즘의 성능이 영향을 받는 것으로 알려져 있다 (Lee 등, 2006). 따라서 혼합 컴포넌트의 갯수를 고정시키지 않고 그 갯수를 추정하는 방법을 고려할 수 있다. 혼합정규분포에서 혼합 컴포넌트의 최적 갯수를 추정하는 방법은 여러가지가 있다. 그 중 Akaike information criteria (AIC), Bayesian information criteria (BIC) 등의 값을 사용하는 방법들은 모든 가능한 갯수에 대해서 정해진 기준에 부합되는지를 모두 검사해야 하기 때문에 계산량이 많다는 단점이 있다 (Lee 등, 2006). 반면에 Lee 등 (2006)은 그러한 단점을 해결하기 위하여 증분형(incremental) k-means에 기반한 새로운 방법을 제안하였고, 그 방법을 적용하여 혼합 컴포넌트의 최적 갯수를 추정하여 분석에 사용하는 것을 고려해볼 수 있다.

또한, 본 논문에서는 qHTS 자료를 분석할 때 자료의 등분산성을 가정하여 비선형 혼합효과 모형에서의 ORE 방법만을 고려하였는데, 추후에는 등분산 가정이 만족되지 않는 경우, 즉 자료가 이분산성을 갖는 경우에 대하여 가중능형회귀(weighted ridge regression) 방법을 사용하여 qHTS 자료를 분석하는 연구를 수행할 수 있다. 그 외에도, 이상점이나 영향점에 로버스트한 능형회귀 방법인 ORME나 WRME 방법을 사용하여 qHTS 자료를 분석하는 연구 또한 수행할 수 있을 것이다.

References

Aggrey, S. E. (2009). Logistic nonlinear mixed effects model for estimating growth parameters, *Poultry Science*, **88**, 276–280.

Arnold, B. C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples, *Sankhyā, The Indian Journal of Statistics, Series B*, **53**, 233–243.

Austin, C., Kavlock, R., and Tice, R. (2008). *Tox21: Putting a Lens on the Vision of Toxicity Testing in the 21st Century*, EPA Science Inventory.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning* (5th ed), Springer, New York.

Bogacka, B., Latif, M. A. H. M., Gilmour, S. G., and Youdim, K. (2017). Optimum designs for non-linear mixed effects models in the presence of covariates, *Biometrics*, **73**, 927-937.

Calegario, N., Daniels, R. F., Maestri, R., and Neiva, R. (2005). Modeling dominant height growth based on nonlinear mixed-effects model: a clonal Eucalyptus plantation case study, *Forest Ecology and Management*, **204**, 11–21.

Craig, B. A. and Schinckel, A. P. (2001). Nonlinear mixed effects model for swine growth, *The Professional Animal Scientist*, **17**, 256–260.

Davidian, M. and Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density, *Biometrika*, **80**, 475–488.

Davidian, M. and Giltinan, D. M. (1993a). Some simple methods for estimating intraindividual variability in nonlinear mixed effects models, *Biometrics*, **49**, 59–73.

Davidian, M. and Giltinan, D. M. (1993b). Some general estimation methods for nonlinear mixed-effects model, *Journal of Biopharmaceutical Statistics*, **3**, 23–55.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*, Champman & Hall, London.

Davidian, M. and Giltinan, D. M. (2003). Nonlinear models for repeated measurement data: an overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, **8**, 387–419.

Fang, Z. and Bailey, R. L. (2001). Nonlinear mixed effects modeling for slash pine dominant height growth following intensive silvicultural treatments, *Forest Science*, **47**, 287–300.

Garber, S. M. and Maguire, D. A. (2003). Modeling stem taper of three central Oregon species using nonlinear mixed effects models and autoregressive error structures, *Forest Ecology and Management*, **179**, 507–522.

Gerhard, D., Bremer, M., and Ritz, C. (2014). Estimating marginal properties of quantitative real-time PCR data using nonlinear mixed models, *Biometrics*, **70**, 247–254.

Hall, D. B. and Clutter, M. (2004). Multivariate multilevel nonlinear mixed effects models for timber yield predictions, *Biometrics*, **60**, 16–24.

Hill, A. V. (1910). The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves, *Journal of Physiology*, **40**, 4–7.

Huber, P. J. (1981). *Robust Statistics*, John Wiley & Sons, New York.

Lee, S. Y. and Xu, L. (2004). Influence analyses of nonlinear mixed-effects models, *Computational Statistics & Data Analysis*, **45**, 321–341.

Lee, Y., Lee, K. Y., and Lee, J. (2006). The estimating optimal number of Gaussian mixtures based on incremental k-means for speaker identification, *International Journal of Information Technology*, **12**, 13–21.

Lim, C. (2015). Robust ridge regression estimators for nonlinear models with applications to high throughput screening assay data, *Statistics in Medicine*, **34**, 1185–1198.

Lim, C., Sen, P. K., and Peddada, S. D. (2013a). Robust analysis of high throughput screening (HTS) assay data, *Technometrics*, **55**, 150–160.

Lim, C., Sen, P. K., and Peddada, S. D. (2013b). Robust nonlinear regression in applications, *Journal of the Indian Society of Agricultural Statistics*, **67**, 215–234.

Lindstrom, M. J., and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data, *Biometrics*, **46**, 673–687.

Meza, C., Osorio, F., and De la Cruz, R. (2012). Estimation in nonlinear mixed-effects models using heavy-tailed distributions, *Statistics and Computing*, **22**, 121–139.

Morrell, C. H., Pearson, J. D., Carter, H. B., and Brant, L. J. (1995). Estimating unknown transition times using a piecewise nonlinear mixed-effects model in men with prostate cancer, *Journal of the American Statistical Association*, **90**, 45–53.

Nguyen, T. T., Bazzoli, C., and Mentre, F. (2012). Design evaluation and optimisation in crossover pharmacokinetic studies analysed by nonlinear mixed effects models, *Statistics in Medicine*, **31**, 1043–1058.

Peddada, S. D. (2013). Statistical analysis of data from quantitative high throughput screening (qHTS) assays - Methods and challenges, *Journal of the Indian Society of Agricultural Statistics*, **67**, 141–150.

Rajeswaran, J. and Blackstone, E. H. (2017). A multiphase non-linear mixed effects model: an application to spirometry after lung transplantation, *Statistical Methods in Medical Research*, **26**, 21–42.

Samson, A., Lavielle, M., and Mentre, F. (2006). Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: application to HIV dynamics model, *Computational Statistics & Data Analysis*, **51**, 1562–1574.

Smith, C. S., Bucher, J., Dearry, A., Portier, C., Tice, R. R., Witt, K., and Collins, B. (2007). Chemical selection for NTP's high throughput screening initiative (Abstract), *Toxicologist*, **46**, 247.

Stirnemann, J. J., Samson, A., and Thalabard, J. C. (2012). Individual predictions based on nonlinear mixed modeling: application to prenatal twin growth, *Statistics in Medicine*, **31**, 1986–1999.

Tice, R. R., Fostel, J., Smith, C. S., Witt, K., Freedman, J. H., Portier, C. J., Dearry, A., and Bucher, J. (2007). The National Toxicology Program high throughput screeing initiative: current status and future directions (Abstract), *Toxicologist*, **46**, 246.

Wang, Y., LeMay, V. M., and Baker, T. G. (2007). Modelling and prediction of dominant height and site

index of Eucalyptus globulus plantations using a nonlinear mixed-effects model approach, *Canadian Journal of Forest Research*, **37**, 1390–1403.

Williams, J. D., Birch, J. B., and Abdel-Salam, A. S. G. (2015). Outlier robust nonlinear mixed model estimation, *Statistics in Medicine*, **34**, 1304–1316.

Xia, M., Huang, R., Witt, K. L., Southall, N., Fostel, J., Cho, M. H., Jadhav, A., Smith, C. S., Inglese, J., Portier, C. J., Tice, R. R., and Austin, C. P. (2008). Compound cytotoxicity profiling using quantitative high-throughput screening, *Environmental Health Perspectives*, **116**, 284–291.

Yeap, B. Y., Catalano, P. J., Ryan, L. M., and Davidian, M. (2003). Robust two-stage approach to repeated measurements analysis of chronic ozone exposure in rats, *Journal of Agricultural, Biological, and Environmental Statistics*, **8**, 438–454.

Yeap, B. Y. and Davidian, M. (2001). Robust two-stage estimation in hierarchical nonlinear models, *Biometrics*, **57**, 266–272.

Zhao, D., Wilson, M., and Borders, B. E. (2005). Modeling response curves and testing treatment effects in repeated measures experiments: a multilevel nonlinear mixed-effects model approach, *Canadian Journal of Forest Research*, **35**, 122–132.

비선형 혼합효과모형에서의 로버스트 능형회귀 방법과 정량적 고속 대량 스크리닝 자료에의 응용

유지선[a] · 임창원[a,1]

[a]중앙대학교 응용통계학과

(2017년 11월 16일 접수, 2017년 12월 26일 수정, 2017년 12월 28일 채택)

요 약

비선형 혼합효과 모형은 다양한 분야에서 반복 측정 자료를 분석할 때 주로 사용된다. 비선형 혼합효과 모형은 개체 내 변동(intra-individual variation)에 대해 고려하는 제 1단계 개별수준모델(individual-level model)과 개체간 변동(inter-individual variation)에 대해 고려하는 제 2단계 개체군모델(population model)의 두 단계로 구성되어 있다. 비선형 혼합효과 모형의 첫 번째 단계인 개별수준모델은 비선형 회귀모형의 모수를 추정하는 것으로 일반적인 비선형 회귀모형과 같고, 주로 보통최소제곱추정 방법을 사용하여 모수를 추정한다. 그러나 최소제곱추정방법은 가정된 비선형 함수가 자료에 의해 명시적으로 드러나지 않는 경우 모수의 추정값과 그 표준오차가 극단적으로 커지는 문제가 발생할 수 있다. 본 논문에서는 최근에 비선형 회귀모형에서 제안된 능형회귀(ridge regression) 방법을 비선형 혼합효과 모형의 제 1단계 개별수준모델에 도입함으로써 이러한 문제를 해결할 수 있는 새로운 추정방법을 제안하였다. 제안된 추정량은 모의실험 연구를 통하여 기존의 표준적인 추정량과 그 성능을 비교하였다. 또한 미국의 National Toxicology Program으로부터 얻어진 정량적 대량고속 스크리닝(quantitative high throughput screening) 실제 자료를 사용하여 추정 방법들을 비교하였다.

주요용어: 용량-반응 연구, 독성학, 약리학, 반복 측정 자료, 능형회귀

이 연구는 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2017R1D1A1B03034509).

[1]교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 경영경제대학 응용통계학과. E-mail: clim@cau.ac.kr