The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.6.1121
28(6), 1121–1131

Semiparametric Approach to Logistic Model with Random Intercept

Mijeong Kim[a,1]

[a]Department of Statistics, Ewha Womans University

(Received November 24, 2015; Revised December 2, 2015; Accepted December 8, 2015)

Abstract

Logistic models with a random intercept are useful to analyze longitudinal binary data. Traditionally, the random intercept of the logistic model is assumed to be parametric (such as normal distribution) and is also assumed to be independent to variables. Such assumptions are very strong and restricted for application to real data. Recently, Garcia and Ma (2015) derived semiparametric efficient estimators for logistic model with a random intercept without these assumptions. Their estimator shows the consistency where we do not assume any parametric form for the random intercept. In addition, the method is computationally simple. In this paper, we apply this method to analyze toenail infection data. We compare the semiparametric estimator with maximum likelihood estimator, penalized quasi-likelihood estimator and hierarchical generalized linear estimator.

Keywords: semiparametric method, logistic model, random intercept, longitudinal data

1. 서론

랜덤 절편를 갖는 로지스틱 모형은 한 개체에 대해서 반복 측정된 이항형(binary) 데이터 분석에 적합하다. 다양한 분야 중에서도 반복 측정 데이터를 구하기 용이한 의학이나 사회과학에서 자주 쓰이고 있다. 지금까지는 랜덤 절편이 정규분포와 같은 특정 분포를 따르고 랜덤 절편과 설명변수가 독립이라는 가정 하에 분석이 이루어진 경우가 많았다. 그러나 이러한 경우는 현실적으로 아주 드물며, 그러한 가정 하에서 이루어진 분석 결과는 잘못되었을 가능성이 크다. 최근에 Garcia와 Ma (2015)는 랜덤 절편에 대한 분포 가정을 하지 않고, 또한 랜덤 절편과 설명변수 독립이 아닐 때에도 적용 가능한 로지스틱 모형에 대한 연구를 하였다. 설명변수에 대한 모수를 갖고, 랜덤 절편에 대해서는 분포 가정을 하지 않았으므로 준모수적 방법(semiparametric method) (Tsiatis, 2006)이 주로 쓰였다. 이 연구에서는 Garcia와 Ma (2015)의 주된 방법인 비모수 방법에 대해 설명하고, Garcia와 Ma (2015)의 추정량과 정규분포 가정에 근거한 최대우도추정량(normal-based maximum likelihood estimator), 벌점 편우도 방법(penalized quasi-likelihood) (Schall, 1991; Breslow과 Clayton, 1993)의 추정량과 비교하도록 한다. 또한 널리 활용되고 있는 계층적 일반화 선형 모형(Hierarchical generalized Linear Model)에 대한 설명도 추가하겠다. 케냐의 초등학생들의 영양 섭취와 말라리아 발병을 조사한 데이터에 각각의 방법을 적용하고 비교하도록 한다. 준모수적 방법을 이용한 모형에서는 고정 절편을 가정할 수 없는 반면, 계층적 일반화 선형 모형은 항상 고정 절편을 갖는 모형이므로, 이 두 방법에 대한 직접 비교는 불가능함을 언급한다.

This research is supported by grants from Ewha Womans University.

[1]Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: m.kim@ewha.ac.kr

2. 랜덤 절편 로지스틱 모형에 대한 가정

이 연구에서는 각 그룹 또는 개체에 대해서 여러 시간에 걸쳐 반복 측정된 자료의 종속 변수가 이항형(binary)인 로지스틱 모형을 고려한다. 그룹 i $(i = 1, \ldots, n)$ 내에서 측정된 횟수를 $j = 1, \ldots, m_i$라고 할 때, 다음과 같은 모형을 가정한다.

$$\text{logit}\, P\left(Y_{ij} = 1 | \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, R_i = r_i\right) = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + r_i = \sum_{k=1}^{p} \beta_k x_{kij} + r_i. \tag{2.1}$$

여기서 $\text{logit}(p) = \log\{p/(1-p)\}$이며, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$는 설명변수의 계수이고 $\boldsymbol{\beta}$는 p차원 설명변수 벡터 $\boldsymbol{x}$에 해당하는 p차원 모수이다. 예를 들어, 정신 분열증의 발작의 유무와 성별간의 연관성을 찾고자 할 때 각각의 사람들로부터 여러 번 측정된 데이터가 있는 경우 위와 같은 모형을 이용할 수 있다. 여기서 랜덤 절편 $\boldsymbol{R} = (R_1, \ldots, R_n)^{\mathrm{T}}$는 그룹 내에서 반복측정으로 인한 그룹 내 효과로 볼 수 있다. 이러한 랜덤 절편 R_i에 정규분포와 같은 특정 분포를 가정하는 것이 가장 널리 쓰여 왔던 방법이었다. 또한 $\boldsymbol{R}$과 설명변수 $\boldsymbol{x}$는 독립이라는 가정도 함께 하는 경우가 많았으나 이러한 가정은 극히 드문 경우에만 적용 가능하기 때문에 이러한 가정으로 인해 잘못된 결과가 도출될 수 있다. 설명변수 $\boldsymbol{x}$와 랜덤 절편 $\boldsymbol{R}$이 독립이 아닌 경우는 다양하게 존재한다. 예를 들어, 자녀가 있는 여성 집단과 없는 여성 집단을 나눈 후 여성의 취업 유무를 조사한다고 하자. 자녀가 있는 그룹과 없는 그룹은 각각 다른 특성을 갖고 있기 때문에, 설명변수로 자녀의 유무를 고려하여 로지스틱 모형으로 분석할 경우 절편이 항상 같을 수는 없다. 따라서 고정된 절편을 가정하는 것보다 랜덤 절편을 가정한 모형이 합리적이며, 설명변수에 따라 절편이 달라지므로 설명변수와 절편이 독립이라고 볼 수 없다. 설명 변수와 랜덤 절편이 독립이 아닌 경우를 가정한 시뮬레이션 결과는 Garcia와 Ma (2015)의 3장 Table 3, 4를 참고한다.

Garcia와 Ma (2015)에서는 랜덤절편 $\boldsymbol{R}$에 대한 어떠한 가정도 하지 않고, 랜덤 절편 $\boldsymbol{R}$과 설명변수 $\boldsymbol{X}$가 독립이 아닐 경우에도 적용 가능한 연구 결과를 제시하였다. Garcia와 Ma (2015)의 연구의 목표는 $\boldsymbol{\beta}$ 를 가장 잘 추정하는 방법을 찾는 것이다. $\boldsymbol{\beta}$는 관심 모수(parameters of interest)이고 $\boldsymbol{R}$에 대한 어떠한 가정도 하지 않았다는 것은 $\boldsymbol{R}$을 비모수(infinite dimesional nuisance parameter)로 두고 접근했다고 볼 수 있다. 이와 같이 모수와 비모수가 한 모형에 안에 함께 존재하는 모형을 준모수적 모형(semiparametric model)이라고 한다.

3. 기존에 제시된 방법

랜덤 절편 로지스틱 모형에 대해 기존에 제시된 방법을 소개하고자 한다. 최대 우도 추정량, 벌점 편우도(penalized quasi-likelihood) 추정량, 계층적 일반화 선형 모형(hierarchical generalized linear model) 등이 있다.

3.1. 최대 우도 추정량

최대 우도 추정 방법은 통계학에서 전형적으로 자주 쓰이는 방법이다. 식 (4.1)에서 랜덤 절편 $\boldsymbol{R}$과 설명변수 $\boldsymbol{X}$가 독립이고, 랜덤 절편 R_i가 정규분포 $N(0, \sigma^2)$을 따른다는 가정을 한다. 이러한 가정 하에서 다음과 같은 우도(likelihood)를 얻을 수 있다.

$$\begin{aligned} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\beta}) &= \int f_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{y}_i|\boldsymbol{x}_i, r_i;\boldsymbol{\beta}\right) f_{\boldsymbol{R}}(r_i) d\mu(r_i) \\ &= \int \prod_{j=1}^{m_i} \exp\left[y_{ij}\left(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + r_i\right) - \log\left\{1 + \exp\left(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + r_i\right)\right\}\right] f_{\boldsymbol{R}|\boldsymbol{X}}(r_i) d\mu(r_i). \end{aligned} \tag{3.1}$$

위의 식 (3.1)은 절편 R과 설명변수 $\boldsymbol{X}$가 독립이므로 $f_{\boldsymbol{R}|\boldsymbol{X}}(r_i|\boldsymbol{x}_i) = f_{\boldsymbol{R}}(r_i)$이고, 이런 점에서 식 (4.1)과 차이가 있다. 관심 모수 $\boldsymbol{\beta}$에 대한 최대우도 추정량은 식 (3.1)를 최대로 하는 값이다. 랜덤 절편 $\boldsymbol{R}$과 $\boldsymbol{X}$가 독립이라는 강한 가정 때문에 최대우도 추정량의 분산이 준모수적 추정량보다 더 작을 수 있다. 하지만 그러한 가정은 극히 드문 경우에만 적용 가능하기 때문에 최대 우도 방법에 의해 잘못된 추정량을 구하게 될 가능성이 크다. 또한 계산과정의 문제점은 식 (3.1)에 대한 최대우도 추정량에 대한 닫힌 형식(closed form)이 없다는 것이다. 따라서 수치해석적으로 접근해야 하며, 계산 과정이 불안정(computationally unstable)하다.

3.2. 벌점 편우도 추정량

벌점 편우도 방법 또한 식 (4.1)에서 설명변수 $\boldsymbol{X}$와 랜덤 절편 R이 독립이며, 랜덤 절편 R_i는 정규분포 $N(0, \sigma^2)$를 따른다는 가정을 한다. 식 (4.1)을 일반화 선형 혼합 모형 관점에서 다음과 같이 표현할 수 있다.

$$E(Y_{ij}|r_i) = \mu_{ij}^{\boldsymbol{R}} = h\left(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + z_i r_i\right), \qquad \mathrm{Var}(Y_{ij}|r_i) = \frac{\phi}{a_{ij}} V\left(\mu_{ij}^{\boldsymbol{R}}\right).$$

위의 식에서 $g = h^{-1}$는 연결 함수(link function)인 로짓 함수(logit function), ϕ는 산포 모수(dispersion parameter), a_{ij}는 사전 가중치(prior weight), $V(\cdot)$은 분산 함수이다. 이 경우, Jang과 Lim (2006)에 따르면 편우도 함수(quasi-likelihood)는 다음과 같다.

$$L = \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left[-\frac{1}{2\phi}\sum_{i=1}^{n}\sum_{j=1}^{m_i} d_{ij}\left(y_{ij}, \mu_{ij}^{\boldsymbol{R}}\right) - \frac{1}{2}\sigma^2 r^2\right] db.$$

위의 식에서 $d_{i,j}(y,\mu) = -2a_{i,j}\int_y^{\mu} (y-\mu)/v(y)\, du$이다. 이 식은 모수 추정에 대한 닫힌 형식(closed form)이 없으므로, 라플라스 근사를 이용하여 다음과 같은 벌점 편우도 추정량(Penalized Quasi-likelihood estimator; PQL)을 구할 수 있다.

$$\mathrm{PQL}(\boldsymbol{\beta}, r) = -\frac{1}{2\phi}\sum_{i=1}^{n}\sum_{j=1}^{m_i} d_{i,j}\left(y_{i,j}, \mu_{ij}^{\boldsymbol{R}}\right) - \frac{1}{2}\sigma^2 r^2.$$

최대 우도 추정량은 닫힌 형식(closed form)이 없기 때문에 우도값과 편미분의 계산이 어려운데 반해, 벌점 편우도 추정량은 라플라스 근사를 이용하여 계산이 수월한 장점이 있다. 하지만, Breslow와 Clayton (1993)에 따르면, 이진 데이터의 분석의 경우에는 불편 추정량을 얻을 수 없다는 단점이 있다.

3.3. 계층적 일반화 선형 모형(Hierarchical generalized Linear Model)

Raudenbush와 Bryk (2002)는 랜덤 절편 모형에 대해 계층적 일반화 선형 모형을 제시했다. 1단계 모형은 다음과 같다.

$$\begin{aligned} &y_{ij} \sim \mathrm{Bernoulli}(p_{ij}), \qquad \mathrm{logit}(p_{ij}) = \eta_{ij}, \\ &\eta_{ij} = \beta_{0i} + \beta_1 x_{1ij} + \cdots + \beta_p x_{1ij} + \epsilon_{ij}. \end{aligned} \tag{3.2}$$

식 (3.2)에서 $\beta_1, \ldots, \beta_p$는 고정변수 $\boldsymbol{X}$에 대한 계수이고, β_{0i}는 i번째 그룹 또는 개체에 대한 랜덤효과가 포함된 절편이다. 2단계 모형은 다음과 같다.

$$\beta_{0i} = \beta_0 + r_i.$$

2단계 모형에서 절편 β_{0i}를 고정 절편 β_0와 랜덤 효과 r_i로 나누어서 설명할 수 있다. 1단계에서는 최대우도 추정 방법 또는 단순 로지스틱 회귀분석 방법을 이용하여 $\beta_1, \ldots, \beta_p$에 대한 추정치를 구하고, $\epsilon_{ij} \sim N(0, \sigma_e^2)$이라는 가정 하에 1단계 오차 ϵ_{ij}의 분산에 대한 추정량을 구할 수 있다. 2단계에서는 $r_i \sim N(0, \sigma_R^2)$이라는 가정 하에 랜덤 절편 r_i에 대한 분산을 추정하는 방법이다. 이 때, 개체간의 변동은 σ_R^2이고, 2단계 모형은 분산성분모형(variance component model)이라고 할 수 있다.

Skrondal과 Rabe-Hesketh (2009)는 베이지안 방법으로 계층적 일반화 선형모형에 대한 연구를 하였다. 1단계 일반화 선형 모형은 다음과 같다.

$$\boldsymbol{h}^{-1}\left\{E(y_{ij}|\boldsymbol{\zeta}_i, \boldsymbol{x}_{ij}, z_{ij})\right\} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + z_{ij}\zeta_i \equiv \eta_{ij}$$

랜덤 효과 $\zeta_i|\boldsymbol{X}_i$, $Z_i \sim N(0, \psi)$이고, 1단계 모형의 오차 ϵ_i는 $\epsilon_i|\boldsymbol{\zeta}_i$, $\boldsymbol{X}_i, Z_i \sim N(0, \theta)$이라는 가정을 한다. 위의 식에서 $\boldsymbol{h}^{-1}$은 연결 함수이고, η_{ij}는 선형 예측치이다. 1단계에서 일반화 선형 모델을 이용하여 $\boldsymbol{\beta}$의 추정량을 구할 수 있다. 2단계에서 ζ_i에 대한 사후분포는 베이지안 방법에 의해 반복적으로 업데이트하여 구하고, 랜덤 효과 ζ_i는 경험적 사후분포의 평균값으로 계산할 수 있다.

$$\hat{\zeta}_i^{EB} = E\left(\zeta_i|\boldsymbol{X}_i, Z_i; \hat{\theta}\right) = \int \zeta_i w\left(\zeta_i|\boldsymbol{X}_i, Z_i; \hat{\theta}\right) d\zeta_i.$$

이 때, $w(\zeta_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \hat{\theta})$는 ζ_i에 대한 사후분포이다. $\hat{\zeta}_i^{EB}$을 추정한 후 랜덤 효과 ζ의 분산 ψ를 추정한다.

4. 준모수적 모형 분석

4.1. 준모수적 방법

2절에서 언급했듯이, 관심 모수는 $\boldsymbol{\beta}$이고, 랜덤 절편 $\boldsymbol{R}$에 대해서는 비모수 가정을 하였다. 모수와 비모수를 동시에 포함한 모형인 준모수적 모형을 고려한다. 이러한 준모수적 모형에 적합한 추정량으로써 Newey와 Powell (1990)은 regular asymptotic linear estimator(RAL)를 소개했다. RAL 추정량 $\hat{\boldsymbol{\beta}}$은 다음과 같은 관계식을 따른다.

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}\left(\boldsymbol{X}_i, \boldsymbol{\beta}_0\right) + o_{p(1)},$$

여기에서 $\boldsymbol{\beta}$는 관심 모수이고, $\boldsymbol{\beta}_0$는 참값이다. $\boldsymbol{\psi}(\boldsymbol{X}_i, \boldsymbol{\beta}_0)$은 i번째 influence 함수로써 $E(\boldsymbol{\psi}) = \mathbf{0}$이고, $E(\boldsymbol{\psi}\boldsymbol{\psi}^{\mathrm{T}})$는 유한한 값을 갖는다. 이 식으로부터 $\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{X}_i, \boldsymbol{\beta}_0) = \mathbf{0}$의 근이 근사적으로 $\hat{\boldsymbol{\beta}}$가 된다는 것을 알수 있다. Semiparametric 모형에서 RAL 추정량 $\hat{\boldsymbol{\beta}}$을 구하기 위해서는 influence 함수 $\boldsymbol{\psi}(\boldsymbol{X}_i, \boldsymbol{\beta})$ 를 찾아야 한다. Bickel 등 (1993)에 의하여 영향력 함수(influence function)에 대한 기하학적인 접근 또한 이루어졌다. 평균이 $\mathbf{0}$인 p차원 랜덤 벡터를 원소로 하는 힐베르트 공간(Hilbert space)에서 두 개의 벡터의 내적(inner product)은 두 함수간의 공분산이 된다. 이 공간에서 어떤 함수의 노름(norm)은 그 랜덤 벡터의 분산이다. 또한 힐베르트 공간에서 nuisance tangent 공간을 정의할 수 있다. nuisance tangent 공간을 설명하기 위해서는 parametric submodel에 대해서 알아야 한다. parametric submodel이란 원래 모형인 준모수적 모형에 포함되는 형태의 모수 모형(parametric model)이고, 준모수적 모형에서의 관심 모수의 참값을 포함하면서 닫힘(closure)의 성격을 갖고 있는 모형을 뜻한다. 이 parametric submodel에서 nuisance parameter의 점수 함수(score function)를 생성하여(spanned) 얻어진 공간을 nuisance tangent 공간이라고 한다. 예를 들어, 어떤 준모수적 모형의 parametric submodel이 $p(\boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\alpha})$라고 하면, $p(\boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\alpha})$에 특정값을 대입하여 원래의 준모수적 모형을 만들 수 있으며, nuisance 스코어 함수 $\boldsymbol{S}_{\boldsymbol{\alpha}} = \partial \log p(\boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha}|_{\boldsymbol{\alpha}_0}$으로 생성된(spanned) 공간을

nuisance tangent 공간이라고 한다. 이 nuisance tangent space의 직교 여공간(orthogonal complement space) 안에 영향력 함수 $\boldsymbol{\psi}(\boldsymbol{X}_i, \boldsymbol{\beta}_0)$가 포함되어 있다. $\boldsymbol{S}_{eff}$을 $\boldsymbol{S}_{\boldsymbol{\beta}} = \partial \log p(\boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)/\partial \boldsymbol{\beta}|_{\boldsymbol{\beta}_0}$을 orthogonal nuisance tangent 공간에 사영(projection)한 벡터라고 할 때, $\boldsymbol{S}_{eff}$을 통해 최소 분산을 갖는 영향력 함수 $\boldsymbol{\psi}_{eff} = \{E(\boldsymbol{S}_{eff}\boldsymbol{S}_{eff}^{\mathrm{T}})\}^{-1}\boldsymbol{S}_{eff}$을 구할 수 있다. 이처럼 준모수적 방법으로 최대효율추정량을 찾는 것이 가능하다.

4.2. 랜덤 절편 로지스틱 모형에서의 준모수적 추정량

의학에서는 어떤 현상의 유무, 즉 두 가지 결과를 갖는 자료가 많다. 예를 들면, 정신분열증의 발작했는지에 관심을 두고 연구를 하는 경우가 있는데, 이 경우 로지스틱 모형이 적합하고, 특히 각각의 사람에 대해서 반복 측정되었다면 절편을 랜덤 변수로 가정하는 것이 합리적이다. 이 때의 모형은 (2.1)과 같고, 절편에 해당하는 R_i에 어떠한 가정을 하지 않는다면 R_i는 모수로 설명되지 않는 비모수 방법으로 설명할 수 있다. 따라서 식 (2.1) 모형은 관심 모수 $\boldsymbol{\beta}$와 비모수 R_i를 갖고 있는 준모수적 모형이다. 준모수적 관점에서 이 모형을 다음과 같이 표현할 수 있다.

$$\begin{aligned} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\beta}) &= \int f_{\boldsymbol{Y}|\boldsymbol{X},R}(\boldsymbol{y}_i|\boldsymbol{x}_i, r_i;\boldsymbol{\beta}) f_{\boldsymbol{R}|\boldsymbol{X}}(r_i|\boldsymbol{x}_i) d\mu(r_i) \\ &= \int \prod_{j=1}^{m_i} \exp\left[y_{ij}\left(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + r_i\right) - \log\left\{1 + \exp\left(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + r_i\right)\right\}\right] f_{\boldsymbol{R}|\boldsymbol{X}}(r_i|\boldsymbol{x}_i) d\mu(r_i). \end{aligned} \tag{4.1}$$

위의 식에서 $\mu(\cdot)$은 dominating 측도이다. 설명변수 $\boldsymbol{X}$와 랜덤절편 $\boldsymbol{R}$이 독립이 아니므로 $f_{\boldsymbol{R}|\boldsymbol{X}}(r_i|\boldsymbol{x}_i) = f_{\boldsymbol{R}}(r_i)$는 성립하지 않는다. Garcia와 Ma (2015)는 랜덤 절편 로지스틱 모형에서 준모수적 방법을 이용하여 다음과 같은 결과를 얻었다. 각각의 개체에 대해서 힐베르트 공간은 $H_i = \{\boldsymbol{h}(\boldsymbol{Y}_i, \boldsymbol{X}_i) : E(\boldsymbol{h}) = \boldsymbol{0}, \mathrm{var}(\boldsymbol{h}) < \infty\}$이고, 여기서 $\boldsymbol{h}$는 p차원 벡터 함수이다. 식 (2.1)을 고려하여 도출해 낸 nuisance tangent 공간 Λ_i와 orthogonal nuisance tangent 공간 Λ_i^{T}는 다음과 같다.

$$\begin{aligned} \Lambda_i &= \left\{E\{\boldsymbol{h}(\boldsymbol{X}_i, R_i)|\boldsymbol{Y}_i, \boldsymbol{X}_i\} : E(\boldsymbol{h}) = \boldsymbol{0}, E\left(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h}\right) < \infty\right\}, \\ \Lambda_i^{\mathrm{T}} &= \left\{\boldsymbol{g}(\boldsymbol{Y}_i, \boldsymbol{X}_i) : E\{\boldsymbol{g}(\boldsymbol{Y}_i, \boldsymbol{X}_i)|\boldsymbol{X}_i, R_i\} = \boldsymbol{0}, E\left(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h}\right) < \infty\right\}, \end{aligned}$$

여기에서 $\boldsymbol{h}$와 $\boldsymbol{g}$는 p차원 벡터이다. 공간 Λ_i^{T}에 $\boldsymbol{\beta}$의 점수 함수를 사영하면 다음과 같은 효율적인(efficient) 점수 벡터를 얻을 수 있다.

$$\boldsymbol{S}_{eff}(\boldsymbol{Y}_i, \boldsymbol{X}_i;\boldsymbol{\beta}) = \boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i;\boldsymbol{\beta}) - E\{\boldsymbol{h}(\boldsymbol{X}_i, R_i)|\boldsymbol{Y}_i, \boldsymbol{X}_i\}, \tag{4.2}$$

위의 식에서 $\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i;\boldsymbol{\beta}) = \partial \log f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y}_i|\boldsymbol{X}_i;\boldsymbol{\beta})/\partial\boldsymbol{\beta} = E\{\partial \log f_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{R}}(\boldsymbol{Y}_i|\boldsymbol{X}_i, R_i;\boldsymbol{\beta})/\partial\boldsymbol{\beta}|\boldsymbol{Y}_i, \boldsymbol{X}_i\}$이고, h는 p차원 벡터 함수로써 $E\{\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X})|\boldsymbol{X}_i, R_i\} = E\left[E\{\boldsymbol{h}(\boldsymbol{X}_i, R_i)|\boldsymbol{Y}_i, \boldsymbol{X}_i\}|\boldsymbol{X}_i, R_i\right]$를 만족한다. 식 (4.2)를 이용하여 $\sum_{i=1}^{n}\boldsymbol{S}_{eff}(\boldsymbol{Y}_i, \boldsymbol{X}_i;\boldsymbol{\beta}) = \boldsymbol{0}$의 근을 찾으면 $\boldsymbol{\beta}$에 대한 점근적 유효추정량(asymptotically efficient esitmator)를 구할 수 있고, 이 추정량은 일치성(consistency)를 갖는다. 그러나, 식 (4.2)를 만족하는 벡터 함수 $\boldsymbol{h}$에 대해 명시적 형태(explicit form)로 나타내기 어렵기 때문에 위의 식을 통해 최대효율추정량을 찾는 것은 어려운 일이다.

식 (4.2)을 간단하게 만들기 위해 Garcia와 Ma (2015)는 충분통계량(sufficient statistic)과 완전통계량(complete statistic)을 찾았다. 새로운 변수 $\boldsymbol{W}$와 $\boldsymbol{V}$를 다음과 같이 정의한다.

$$\boldsymbol{W}_i = \sum_{j=1}^{m_i} Y_{ij}, \qquad \boldsymbol{V}_i = (Y_{i2}, \ldots, Y_{im_i})^{\mathrm{T}}.$$

$\boldsymbol{W}$와 $\boldsymbol{V}$를 이용하여 다음과 같은 관계를 증명하였다.

$$f_{\boldsymbol{V}|\boldsymbol{W},\boldsymbol{X},\boldsymbol{R}}(\boldsymbol{v}_i|w_i,\boldsymbol{x}_i,r_i) = f_{\boldsymbol{v}|\boldsymbol{W},\boldsymbol{X}}(\boldsymbol{v}_i|w_i,\boldsymbol{x}_i),$$
$$f_{\boldsymbol{R}|\boldsymbol{W},\boldsymbol{X},\boldsymbol{V}}(r_i|w_i,\boldsymbol{x}_i,\boldsymbol{v}_i) = f_{\boldsymbol{R}|\boldsymbol{W},\boldsymbol{X}}(r_i|w_i,\boldsymbol{x}_i).$$

즉, $(W_i, \boldsymbol{X}_i)$가 주어졌을 때, $\boldsymbol{V}_i$와 R_i는 독립이다. 또한 $E\{\boldsymbol{g}(W_i,\boldsymbol{X}_i)|\boldsymbol{x}_i,r_i\} = \boldsymbol{0}$이면 반드시 $\boldsymbol{g}(W_i,\boldsymbol{X}_i) = \boldsymbol{0}$임을 보였다. 변수 $\boldsymbol{W}$와 $\boldsymbol{V}$를 이용하여 식 (4.2)은 다음과 같이 쓸 수 있다.

$$\boldsymbol{S}_{eff}(\boldsymbol{Y}_i,\boldsymbol{X}_i;\boldsymbol{\beta}) = \boldsymbol{X}_i\boldsymbol{A}_i^{-1}\left\{0, \boldsymbol{V}_i^{\mathrm{T}} - E\left(\boldsymbol{V}_i^{\mathrm{T}}|W_i,\boldsymbol{X}_i;\boldsymbol{\beta}\right)\right\}^{\mathrm{T}}.$$

즉, R_i 대신에 $\boldsymbol{W}$와 $\boldsymbol{V}$를 이용하여 효율적인(efficient) 점수 벡터를 표현했다. 따라서 다음 식의 근이 준모수적 최대효율추정량이 된다.

$$\sum_{i=1}^{n}\boldsymbol{S}_{eff}(\boldsymbol{Y}_i,\boldsymbol{X}_i;\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=2}^{m_i}(\boldsymbol{X}_{ij}-\boldsymbol{X}_{i1})\{V_{i,j} - E(V_{i,j-1}|W_i,\boldsymbol{X}_i,\boldsymbol{\beta})\} = \boldsymbol{0}. \tag{4.3}$$

위의 식 (4.3)를 통해 $\boldsymbol{\beta}$의 추정량을 계산할 수 있다. 여기에서 $E(V_{i,j-1}|W_i,\boldsymbol{X}_i,\boldsymbol{\beta})$는 i번째 개체의 $\boldsymbol{V}$에 대한 조건부 기대값이다. 예를 들어, i번째 개체에 대해 5번 반복 측정 되었고, 그 중 두 번째, 세 번째 측정된 자료가 결과값 $y=1$을 갖는다고 하자. 이 경우에 y값은 $(0,1,1,0,0)$이므로 $m_i = 5$, $w_i = 2$이다. $m_i = 5$, $w_i = 2$일 때, 5개 중 2개를 1로 갖고 나머지는 0인 경우의 수는 $\binom{5}{2}$이다. 즉, 이 경우 가능한 $\boldsymbol{v}$는 10가지이다. Garcia와 Ma (2015)에서는 $E(V_{i,j-1}|W_i,\boldsymbol{X}_i,\boldsymbol{\beta})$를 계산하는 방법을 다음과 같이 제시하고 있다. 각각의 m_i, w_i에 대해서 가능한 $\boldsymbol{v}_i$의 집합을 $\mathcal{R}(\boldsymbol{v}_i)$라고 할 때, 이 때 $\boldsymbol{v}$의 조건부 확률은 다음과 같이 구할 수 있다.

$$f_{\boldsymbol{V}|W,\boldsymbol{X}}(\boldsymbol{v}_i|w_i,\boldsymbol{x}_i;\boldsymbol{\beta}) = \frac{\exp\left\{\sum_{j=2}^{m_i}(\boldsymbol{x}_{ij}-\boldsymbol{x}_{i1})^{\mathrm{T}}\boldsymbol{\beta}v_{i,j-1}\right\}}{\sum_{\mathcal{R}(\boldsymbol{v}_i)}\exp\left\{\sum_{j=2}^{m_i}(\boldsymbol{x}_{ij}-\boldsymbol{x}_{i1})^{\mathrm{T}}\boldsymbol{\beta}v_{i,j-1}\right\}}.$$

위의 확률질량함수를 이용하여 $E(V_{i,j-1}|W_i,\boldsymbol{X}_i,\boldsymbol{\beta})$을 계산하는 것이 가능해진다. 따라서 $E(V_{i,j-1}|W_i,\boldsymbol{X}_i,\boldsymbol{\beta})$에 대한 계산은 일반화 선형 혼합 모형에서 최대우도 추정량을 구하거나 벌점 편우도 추정량을 구하는 경우보다는 간단하고, 수치적으로 안정적이라고 할 수 있다. 그러나 단점이 있다면, 준모수 방법을 이용할 경우 코딩의 문제가 생길 수 있다. 왜냐하면 이 계산은 v_i의 가능한 모든 경우의 수 즉, $\binom{m_i}{w_i}$를 계산해야 하는데, 이 값이 크다면 컴퓨터를 이용한 계산이 불가능해질 수 있다. 예를 들어, $\binom{40}{20}$ = 137,846,528,820이다. 이 경우 가능한 모든 v를 행으로 갖는 행렬을 만드는 작업은 불가능할 수 있다. 현실적으로는 한 개체에 대해 40번까지 반복측정하는 경우가 매우 드물기 때문에, 반복 측정 횟수가 적은 다양한 경우에 이 방법이 유용하게 이용될 수 있을 것으로 보인다. 그러나 개체가 아닌 그룹별로 같은 랜덤 절편을 가정하는 경우에는, 한 그룹 안에 데이터 수가 비교적 클 가능성이 있으므로 준모수적 방법을 이용한 랜덤 절편 로지스틱 모형을 쓰는 것은 코딩 문제로 인해 적절하지 않을 수 있다.

5. 데이터 분석

5.1. 시뮬레이션

설명변수와 랜덤 절편이 독립이 아닌 경우에 준모수적 방법, 최대우도 추정 방법, 벌점화 모형 방법, 계층적 일반화 선형 모형을 비교하였다. 모형 (2.1)에서 $p=4$, 모수는 $\boldsymbol{\beta} = (-0.3, -0.5, 0.2, 0.7)^{\mathrm{T}}$, 그룹의 수는 500 ($i = 1,\ldots,500$), 각 그룹 안에서 반복 측정 횟수는 다섯번 ($m_i = 5$)으로 설정하였다. 각

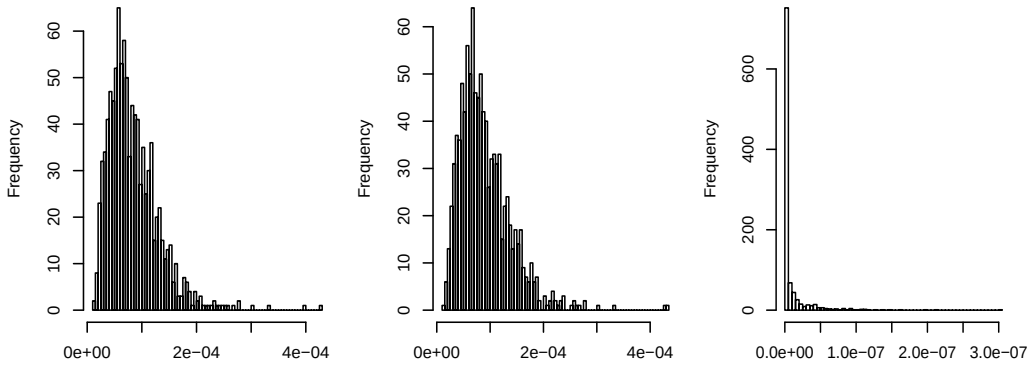
Table 5.1. Simulation results for the semiparametric estimation, the maximum likelihood estimation and the penalized quasi-likelihood estimation. The median of estimators and estimated standard errors of 1000 simulations are presented for each model.

			준모수적 추정 방법	최대우도 추정 방법	벌점 편우도 추정 방법
		계수	−0.2981	−0.2343	−0.2332
$\beta_1 = -0.3$	$\hat{\beta}_1$	표준 오차1	(0.0510)	(0.0389)	(0.0386)
		표준 오차2	(0.0507)	(0.0382)	(0.0380)
		계수	−0.5019	−0.3945	0.3927
$\beta_2 = -0.5$	$\hat{\beta}_2$	표준 오차1	(0.0538)	(0.0406)	(0.0401)
		표준 오차2	(0.0529)	(0.0400)	(0.0399)
		계수	0.2010	0.1589	0.1580
$\beta_3 = 0.2$	$\hat{\beta}_3$	표준 오차1	(0.0502)	(0.0384)	(0.0381)
		표준 오차2	(0.0495)	(0.0378)	(0.0376)
		계수	0.7010	0.5503	0.5488
$\beta_4 = 0.7$	$\hat{\beta}_4$	표준 오차1	(0.0573)	(0.0426)	(0.0420)
		표준 오차2	(0.0561)	(0.0427)	(0.0422)

그룹의 랜덤 절편 R_i는 Unif$(-2, 2)$에서 랜덤하게 발생하였으며, 설명변수 $\boldsymbol{X}_{ij}$는 $10 + 3r_i$를 올림으로 한 자연수를 모수로 갖는 t분포에서 난수를 랜덤하게 발생하였다. 즉, $\boldsymbol{X}_{ij}$는 r_i에 따라 모수가 다른 t 분포$(t(5), \ldots, t(17))$에서 난수가 발생되었으므로, r_i와 $\boldsymbol{X}_{ij}$값이 독립이 아닌 구조이다. 1,000번의 시뮬레이션을 수행하였으며, 시뮬레이션 결과는 Table 5.1에서 확인할 수 있다. 로버스트한 추정값을 얻기 위해 각 방법을 이용해서 구한 1,000개의 β의 추정치의 중간값을 표시하였다. Table 5.1에서 준모수적 추정치가 원래 β와의 편차가 다른 방법보다 더 작음을 알 수 있다. Hausman (1978)에 의하면 두 추정치 θ_0, θ_1이 일치성을 보이는가에 대한 검정은 $H = (\hat{\theta}_1 - \hat{\theta}_0)^{\mathrm{T}}\{\mathrm{var}(\hat{\theta}_1) - \mathrm{var}(\hat{\theta}_0)\}^{-1}(\hat{\theta}_1 - \hat{\theta}_0)$을 계산하여 자유도가 $\mathrm{rank}\{\mathrm{var}(\hat{\theta}_1) - \mathrm{var}(\hat{\theta}_0)\}$인 카이제곱 검정를 해 보는 것이다. 이 때 귀무가설은 두 추정치가 일치성을 갖는다는 가설이다. Garcia와 Ma (2015)의 Table 5에서 $\boldsymbol{X}$와 R이 독립이 아니고, 상관성이 높을수록 Hausman 검정 결과가 귀무가설을 기각한다는 것을 시뮬레이션을 통해 확인하였다. 그러나, 이 연구에서는 $\boldsymbol{X}$와 R이 독립이 아니고, 상관관계가 없는 경우, Hausman 검정 결과가 귀무가설을 항상 기각하는 것은 아니라는 것을 예를 통해 확인할 수 있다. Figure 5.1에 1,000번의 시뮬레이션의 Hausman 검정값을 히스토그램으로 나타내었다. 이 연구의 시뮬레이션 결과에서는 1,000번의 시뮬레이션 중 준모수방법 추정치, 최대우도 추정치와 벌점 편우도 추정치가 일치성을 보인다는 귀무가설을 한 번도 기각하지 못하였다. $\chi^2(0.05,\ \mathrm{df} = 3) = 0.997071$인 점을 유의하기 바란다. 또한 (a)과 (b)의 검정값이 (c)보다 더 큰 것을 알 수 있다. $\boldsymbol{X}$와 R이 독립이고, 랜덤 절편이 정규분포이라면, 준모수 방법과 최대우도 추정치가 일치성을 보일 것이다. 즉, Hausman 검정 결과, 귀무가설은 기각되지 않을 것이다. 하지만, 역으로 귀무가설이 기각되지 않았다고 하여 $\boldsymbol{X}$와 R이 반드시 독립이거나 랜덤절편이 정규분포임을 입증하기는 어렵다는 것을 이 예를 통해 알 수 있다.

5.2. 케냐 학생들의 영양 섭취 및 질병의 발병 여부에 대한 데이터 분석

이 연구에서는 Neumann 등 (2003)에 의해 소개된 코호트 조사 데이터를 분석하였다. 이 데이터는 1998년부터 2001년까지 케냐의 12개 초등학교의 학생들을 대상으로 학교에서 주기적으로 음식을 제공하고, 매 시점마다 건강 상태(열, 두통, 목 아픔, 발진 등)을 확인하고, 말라리아, 장염의 발병 상태를 조사한 데이터이다. 음식을 제공하는 방법은 총 세가지(칼로리 보충, 우유, 육류 제공)이며 음식을 제공하지 않은 대조군 집단이 있다. 총 12개의 학교에 대해 각각의 방법을 랜덤하게 지정하였다. 즉, 대조



(a) SE vs. MLE (b) SE vs. PQL (c) MLE vs. PQL

Figure 5.1. Histograms of Hausman test statistic of 1000 simulations. Pairwise comparisons for three models were carried out, which are (1) the semiparametric estimator versus the maximum likelihood estimator, (2) the semiparametric estimator versus the penalized quasi-likelihood estimator, and (3) the maximum likelihood estimator versus the penalized quasi-likelihood estimator.

Table 5.2. Analysis of malaria data with the random intercept logistic model

		준모수적 추정 방법		최대우도 추정 방법		벌점 편우도 추정 방법	
		계수	표준오차	계수	표준오차	계수	표준오차
Time×Control	$\hat{\beta}_1$	−0.0695	(0.0126)	−0.1745	(0.0121)	−0.1716	(0.0111)
Time×Calorie	$\hat{\beta}_2$	−0.0856	(0.0124)	−0.1821	(0.0114)	−0.1759	(0.0105)
Time×Milk	$\hat{\beta}_3$	−0.0617	(0.0134)	−0.1776	(0.0119)	−0.1745	(0.0108)
Time×Meat	$\hat{\beta}_4$	−0.0993	(0.0137)	−0.1893	(0.0123)	−0.1813	(0.0114)
sin(π·year)	$\hat{\beta}_5$	−0.1532	(0.0686)	−0.5790	(0.0621)	−0.5701	(0.0592)
cos(π·year)	$\hat{\beta}_6$	0.2032	(0.0579)	0.3314	(0.0554)	0.3079	(0.0542)

군을 포함한 4가지 그룹에 대해 각각 세 학교가 랜덤하게 선택되었다. 분석에 이용된 데이터는 Robert Weiss가 웹사이트에 공개한 505명의 초등학생의 데이터 중 반복측정이 된 502명의 데이터이다. 한 개체 또는 그룹에 대해서 반복 측정 자료가 있어야 준모수적 추정 방법을 이용할 수 있기 때문에 한 번의 측정 데이터만 있는 세 명의 학생의 데이터는 분석에서 제외하였다.

이 데이터를 이용하여 말라리아 발병 여부를 랜덤 절편 로지스틱 모형으로 분석하였다. Neumann 등 (2013)에 의하면 질병 발병률은 개인차가 크므로, 랜덤 절편 로지스틱 모형이 적합하다. 모형에 포함한 설명 변수는 (1) 나이(개월)×대조군 여부, (2) 나이(개월)×칼로리 제공 여부, (3) 나이(개월)×우유 제공 여부, (4) 나이(개월)×육류 제공 여부, (5) sin(π·시간), (6) cos(π·시간)이다. Weiss (2005)에 따르면, 말라리아 발병률은 계절의 영향을 받기 때문에 sin(π·시간)과 cos(π·시간)을 설명변수로 포함시키는 것이 적절하다.

Table 5.2는 고정 절편이 없는 모형에서 준모수적 추정 방법 및 최대우도 추정 방법, 벌점 편우도 추정 방법으로 얻은 결과이다. 고정 절편을 가정할 경우, 고정절편에 해당하는 계수가 β_1이라고 한다면, 설명변수 $\boldsymbol{X}$에서 첫 번째 요소를 $\boldsymbol{X}_{ij1} = 1$로 둘 수 있다. 식 (4.3)에서 모든 j에 대해 $\boldsymbol{X}_{ij1} = \boldsymbol{X}_{i11}$이면, $\sum_{i=1}^{n} \boldsymbol{S}_{eff}(\boldsymbol{Y}_i, \boldsymbol{X}_i; \boldsymbol{\beta})$ 벡터의 첫번째 요소가 β_1의 값에 관계 없이 항상 0이 되기 때문에 식 (4.3)을 이용해서 고정 절편에 해당하는 계수 β_1를 추정하는 것이 불가능해진다. 따라서, 고정 절편을 갖는 모형에 대해서는 준모수적 방법을 이용할 수 없다. 같은 이유로 인해, 각 그룹에서 같은 값을 갖는 변수

Table 5.3. Analysis of malaria data with the logistic model which includes both a fixed intercept and a random intercept

		최대우도 추정 방법		벌점 편우도 추정 방법		계층적 일반화 선형 방법	
		계수	표준오차	계수	표준오차	계수	표준오차
고정 절편	$\hat{\beta}_0$	−1.6124	(0.0927)	−1.5079	(0.0801)	0.1813	(0.0008)
Time×Control	$\hat{\beta}_1$	−0.0756	(0.0109)	−0.0734	(0.0097)	−0.0063	(0.0008)
Time×Calorie	$\hat{\beta}_2$	−0.0754	(0.0106)	−0.0750	(0.0094)	−0.0064	(0.0008)
Time×Milk	$\hat{\beta}_3$	−0.0760	(0.0107)	−0.0736	(0.0095)	−0.0062	(0.0009)
Time×Meat	$\hat{\beta}_4$	−0.0795	(0.0114)	−0.0804	(0.0101)	−0.0067	(0.0057)
sin(π·year)	$\hat{\beta}_5$	−0.1532	(0.0661)	−0.1505	(0.0574)	−0.0112	(0.0051)
cos(π·year)	$\hat{\beta}_6$	0.2012	(0.0579)	0.1972	(0.0502)	0.0194	(0.0084)

를 모형에 포함시킬 수 없다. 이 경우 한 학생의 반복측정된 데이터를 하나의 그룹으로 두고 분석하였는데, 이 그룹에서 성별은 고정된 하나의 값을 갖는다. 그러므로 성별에 따른 말라리아 발병률은 이 모형으로 예측할 수 없다. 하지만, 그룹을 다르게 설정하여 성별의 효과를 구할 수 있다. Table 5.2을 통해 준모수 추정치가 다른 두 방법과는 다른 값을 갖는 것을 알 수 있다. 귀무가설을 두 추정치의 차이가 없다는 가설이라고 할 때, Hausman 검정 결과로는 귀무가설을 기각할 수 없다. (1) 준모수 추정치와 최대우도 추정치를 비교한 Hausman 검정값은 0.000100, (2) 준모수 추정치와 벌점 우도 추정치를 비교한 Hausman 검정값은 0.000192이고, (3) 최대우도 추정치와 벌점 우도 추정치를 비교한 Hausman 검정값은 1.4215e−7이므로, 세가지 경우 모두 $\chi^2(0.05, \mathrm{df} = 6) = 0.999997$보다 작은 값을 갖는다. 하지만, 이 Hausman 검정 결과만으로 설명변수와 랜덤절편이 독립, 랜덤 절편이 정규분포를 따르는지 검정하기 어렵다. 시뮬레이션 결과에서 명백히 독립이 아닌 경우에, Hausman 검정으로 귀무가설을 기각할 수 없지만 (1), (2)의 값이 (3)의 값보다 비교적 큰 값을 가졌다는 것을 고려했을 때, 이 데이터의 결과도 시뮬레이션 결과와 비슷한 양상을 띄는 것으로 보아 설명변수와 랜덤 절편이 독립이 아닐 가능성이 있다고 생각한다. 설명변수와 랜덤 절편의 독립성을 검정하는 효과적인 방법은 추후에 연구되어야 할 부분이다. Table 5.3은 고정 절편과 랜덤 절편을 갖고 있는 로지스틱 모형이다. 최대우도 추정량은 R에서 glmer 함수를 이용하고, 벌점 편우도 추정량은 R에서 glmmPQL 함수를 이용하였다. Skrondal과 Rabe-Hesketh (2009)이 제시한 계측적 일반화 선형 방법을 통한 추정치는 Stata에서 gllamm 함수를 이용하여 구하였다.

준모수적 추정량은 일치성(consistency)를 갖고, 최대효율 추정량이라는 것이 Newey와 Powell (1990)에 의해 증명되었다. 그러나, Table 5.2를 통해 다른 두 가지 방법과 비교하여, 준모수적 추정량의 분산이 항상 더 작은 것은 아님을 알 수 있다. 최대 우도 추정량과 준모수적 추정량은 설명변수와 랜덤 절편이 독립이라는 가정 하에 계산된 추정량이고, 준모수적 추정량은 그러한 가정을 하지 않고 계산된 추정량인 점을 고려해야 한다.

6. 결론

반복 측정된 이진 자료를 분석할 수 있는 방법으로 랜점 절편을 포함한 로지스틱 모형에 준모수적 방법을 적용한 연구를 소개하였다. 각각의 설명변수와 랜덤 절편이 독립이라는 다소 강한 가정과 랜덤 절편이 정규분포와 같은 특정 분포를 따른다는 가정이 일반적이었다. 그러한 가정 하에서 최대 우도 추정 방법이나 벌점 편우도 추정 방법을 이용하거나 고정 절편과 랜덤 절편을 동시에 포함하는 모형에서는 계층적 일반화 선형 모형이 주로 이용되었다. 하지만, 설명 변수와 랜덤 절편이 독립이 아니거나 랜덤 절편이 특정 분포를 따르지 않을 경우에는 그러한 방법을 이용하는 것은 문제가 있다. 또한 최대 우도 추

정 방법을 이용할 경우에는 추정량에 대한 방정식이 간단한 형태가 아니기 때문에 근을 찾는 계산을 할 때 수치해석적 문제가 생기기도 한다. 한편, 벌점 편우도 추정량은 일치성을 갖지 않는 문제가 있다. 반면에, 준모수적 추정 방법을 이용하면 설명 변수와 랜덤 절편이 독립이 아닌 경우나 랜덤 절편을 비모수로 가정한 경우에도 적용가능하고, 준모수적 추정량은 일치성(consistency)을 갖는다. 충분 통계량과 완전 통계량을 이용하여 추정량에 대한 점수 함수를 간단하게 할 수 있기 때문에 추정량에 대한 방정식이 간단해지고, 계산 또한 수월해진다는 점이 최대우도 추정 방법에 비해 큰 장점이라고 할 수 있다. 준모수적 랜덤 절편 로지스틱 모형의 단점으로는, 같은 개체 또는 그룹 내에서 같은 값을 갖는 변수가 있다면 그 변수에 대한 계수를 추정할 수 없다는 점이다. 특히, 고정절편을 가정한 모형 로지스틱 모형을 쓸 수 없다. 계층적 일반화 선형 모형에서는 항상 고정 절편을 가정하기 때문에, 준모수 방법과 직접 비교가 어렵다. 또 다른 단점으로는, 개체 또는 그룹 내에서 반복 횟수를 m이라고 하고, 이 중 종속 변수가 1인 관측치의 개수가 w일 때, 컴퓨터 코딩할 경우 $m \times \binom{m}{w}$ 크기의 행렬을 선언해야 하는데, $\binom{m}{w}$이 크다면 코딩이 불가능할 수 있다. 또한, 반복 측정을 하지 않은 자료는 분석에 포함시킬 수 없다. 랜덤 절편 로지스틱 모형은 랜덤 절편이 관찰되는 값이 아니므로, 설명변수와 랜덤절편간의 독립성을 증명하는 것은 어려운 일이다. Garcia와 Ma (2015)에서는 Hausman 검정 방법을 이용하여 설명변수와 랜덤 절편의 독립성을 확인 가능한 예를 들었으나, 이 연구에서는 설명변수와 랜덤 절편이 독립이 아니지만 Hausman 검정 방법으로 그 사실을 증명할 수 없는 사례를 예로 들었다. 따라서 설명변수와 랜덤 절편의 독립성을 확인할 수 있는 구체적인 검정 방법에 대한 연구가 필요하다.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, The Johns Hopkins University Press, Baltimore.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 9–25.

Garcia, T. P. and Ma, Y. (2015). Optimal estimator for logistic model with distribution-free random intercept, *Scandinavian Journal of Statistics*, in press.

Hausman, J. A. (1978). Specification tests in econometrics, *Econometrica*, **46**, 1251–1271.

Jang, W. and Lim, J. (2006). PQL estimation biases in generalized linear mixed models, *Institute of Statistics and Decision Sciences*, Duke University Springer-Verlag, Durham, NC, USA, 5–21.

Newey, W. and Powell, J. L. (1990). Efficient estimation of linear and type I censored regression models under conditional quantile restrictions, *Econometric Theory*, **6**, 295–317.

Neumann, C. G., Bwibo, N. O., Murphy, S. P., Sigman, M., Guthrie, D., Weiss, R. E., Allen, L. H. and Demment, M. W. (2003). Animal source foods improve dietary quality, micronutrient status, growth and cognitive function in Kenyan school children: background, study design and baseline findings, *The Journal of Nutrition*, **133**, 3941S–3949S.

Neumann, C. G., Bwibo, N. O., Jiang, L. and Weiss, R. E. (2013). School snacks decrease morbidity in Kenyan schoolchildren: A cluster randomized, controlled feeding intervention trial, *Public Health Nutrition*, **16**, 1593–1604.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models*, 2nd Ed., Sage Publications, California.

Schall, R. (1991). Estimation in generalized linear models with random effects, *Biometrika*, **78**, 719–727.

Skrondal, A. and Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **172**, 659–687.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*, Springer, New York.

Weiss, R. E. (2005). *Modeling Longitudinal Data*, Springer-Verlag, New York.

준모수적 방법을 이용한 랜덤 절편 로지스틱 모형 분석

김미정[a,1]

[a]이화여자대학교 통계학과

(2015년 11월 24일 접수, 2015년 12월 2일 수정, 2015년 12월 8일 채택)

Abstract

의학이나 사회과학에서 이진 데이터 분석 시 랜덤 절편(random intercept)을 갖는 로지스틱 모형이 유용하게 쓰이고 있다. 지금까지는 이러한 로지스틱 모형에서 랜덤 절편이 정규분포와 같은 모수 모형(parametric model)을 따른다는 가정과 설명변수와 랜덤 절편이 독립이라는 가정 하에 실행된 데이터 분석이 전반적이었다. 그러나 이러한 두 가지 가정은 다소 무리가 있다. 이 연구에서는 설명 변수와 랜덤 절편의 독립성을 가정하지 않고, 비모수 랜덤 절편을 따르는 로지스틱 모형의 방법론을 기존에 널리 쓰인 방법과 비교하여 설명하도록 한다. 케냐의 초등학생들의 영양 섭취 및 질병의 발병을 조사한 데이터에 이 방법을 적용하였다.

주요용어: 로지스틱 모형, 랜덤 절편, 준모수적 방법, 반복 측정 자료

이 논문은 2015년도 이화여자대학교의 연구비에 의하여 수행되었음.

[1](03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과. E-mail: m.kim@ewha.ac.kr