*The Korean Journal of Applied Statistics* (2017) DOI: http://dx.doi.org/10.5351/KJAS.2017.30.6.957
**30**(6), 957–981

# Maximum likelihood estimation of Logistic random effects model

Minah Kim[a] · Minjung Kyung[b,1]

[a b]Department of Statistics, Duksung Women's University

(Received October 10, 2017; Revised November 29, 2017; Accepted November 30, 2017)

**Abstract**

A generalized linear mixed model is an extension of a generalized linear model that allows random effect as well as provides flexibility in developing a suitable model when observations are correlated or when there are other underlying phenomena that contribute to resulting variability. We describe maximum likelihood estimation methods for logistic regression models that include random effects - the Laplace approximation, Gauss-Hermite quadrature, adaptive Gauss-Hermite quadrature, and pseudo-likelihood. Applications are provided with social science problems by analyzing the effect of mental health and life satisfaction on volunteer activities from Korean welfare panel data; in addition, we observe that the inclusion of random effects in the model leads to improved analyses with more reasonable inferences.

Keywords: logistic regression, generalized linear mixed model, random effect, maximum likelihood estimation

## 1. 서론

일반화 선형 모형(generalized linear model)은 다양한 오차 구조를 포함한 이산형 또는 범주형 반응변수 $\mathbf{y} = (y_1, \ldots, y_n)$를 모형화 할 수 있는 유연한 프레임 워크를 제공함으로 다양한 분야에서 사용되는 통계적 모형이다. 일반적으로 선형 모형에 대한 가정은 선형성, 오차항의 정규성, 오차항의 독립성, 등분산성이 있는데, 이러한 가정을 적용하지 못하는 경우의 자료에는 일반화 선형 모형을 사용한다. Nelder와 Wedderburn (1972) 그리고 McCullagh와 Nelder (1989) 등의 고전적인 논문들은 일반화 선형 모형에 대하여 지수족 형태에 의해 제공되는 맥락에서 비선형 회귀모형을 선형 모형으로 변환하여 고려했기 때문에, 우도함수의 재구성은 링크함수의 형태와 그 결과 발생하는 오차 구조의 유형과 같이 모형의 구조적 구성 요소를 나타낸다. 그러므로 Breslow와 Clayton (1993)은 일반화 선형 모형을 다양한 결과의 측정을 위해 회귀분석의 우도 기반의 접근을 모두 종합한 것이라고 정의하였다. 일반화 선형 모형에 대한 자세한 설명은 McCullagh와 Nelder (1989), Fahrmeir와 Tutz (2001) 등에서 찾을 수 있다.

### 1.1. 일반화 선형 혼합 모형

기술의 발전과 함께 자료의 저장, 처리속도가 빨라지면서 다양한 데이터가 등장하고, 그로 인해 예전과는 달리 데이터가 복잡해지고 있다. 다양하고 복잡한 데이터를 수용할 수 있는 일반화 선형 혼합 모

This work was supported by a Duksung Women's University research grants (No. 3000002744).
[1]Corresponding author: Department of Statistics, Duksung Women's University, 33 Samyangro 144-gil, Seoul 01369, Korea. E-mail: mkyung@duksung.ac.kr

형(generalized linear mixed model; GLMM)은 통계 분야에서 널리 사용되고 있다. 적합성 부족을 설명하기 위해서 임의효과(random effect)가 포함되었기 때문이다. 이러한 임의효과는 알지 못하는 작은 변화와 실제 데이터에서 직면하기 쉬운 극단값(outlier)을 허용할 수 있기 때문에 이항분포, 포아송 분포 등 따르는 회귀모형들을 바탕으로 상관되고 과대 산포된 데이터에 많이 사용된다.

일반화 선형 모형에서 선형 예측치에 임의적인 용어인 임의효과가 포함된 모형을 일반화 선형 혼합 모형이라 한다. 임의효과와 고정효과가 임의로 상관된 혼합에서 결과 변수를 조건부로 수용하도록 명시할 수 있다 (Breslow와 Clayton, 1993; Buonaccorsi, 1996; Wang 등, 1998; Wolfinger와 O'Connell, 1993). McCulloch와 Searle (2001)는 비선형 모형 안에 임의 효과가 포함되면서 상호 연관된 데이터를 수용하는 모형을 만들거나, 모집단을 추론하기 위해 수준의 모집단으로부터 선택된 요소의 수준을 고려하기 위해서 일반화 선형 혼합 모형을 사용한다고 언급하였다.

임의효과는 기댓값과 분산에 대한 약간의 가정만 하고, 임의효과 분포에 대한 구체적인 형태는 필요 없다. 고정효과 모형보다 임의효과 모형이 더 선호되는 세 가지 이유가 있다. 첫 번째는 그룹 수준에서 공분산의 효과를 추정하기 위하여 임의효과 모형을 사용한다, 고정효과 모형에서는 집단 수준의 공분산의 효과로부터 집단의 효과를 분리할 수 없기 때문에 임의효과를 사용하게 된다. 두 번째로 임의효과 모형은 집단의 모집단으로부터 온 임의표본으로서 집단을 취급한다. 고정효과 모형을 사용하면 추론은 표본 안에 있는 집단들을 넘어서서 할 수 없다. 세 번째는 통계적 추론이 잘못될 수 있기 때문에 사용된다. 전통적인 회귀 기술은 다수준의 구조를 인식하지 못하기 때문에 잘못 추정된 회귀 계수의 표준 오차를 유발한다. 이 잘못 추정된 회귀 계수의 표준 오차는 더 높거나 낮은 수준의 공분산 계수에 대해서 통계적 유의성의 과장이나 과소를 이끌 수 있다. 이 같은 이유로 고정효과만 있는 모형보다 임의효과를 추가한 모형을 사용하는 것이 좋다 (Li 등, 2011).

### 1.2. 로지스틱 회귀모형

일반화 선형 모형의 종류는 지수족 형태를 바탕으로 선형 모형으로 구성할 수 있기 때문에 매우 다양하다. 그 많은 모형 중 우리는 이 논문에서 로지스틱 회귀모형(logistic regression)을 고려한다. 로지스틱 회귀모형은 반응변수가 이분적으로 나타나는 반응변수에 대하여 이항분포를 가정하고, 어떤 사건이 일어날 확률을 추정하기 위해 설명변수들의 선형성을 모형화한 일반화 선형 모형이다.

로지스틱 회귀모형은 Verhulst (1838, 1845)에 의해 기하학적으로 증가하는 인구에 대한 모형을 만들어 연구하면서 처음 제안되었다. Verhulst 외의 다른 저자들도 인구 증가에 대한 분석을 하기 위해 로지스틱 모형을 사용하였으며, 자세한 설명은 Pearl와 Reed (1920), Pearl 등 (1940), Schultz (1930) 등에서 찾을 수 있다. Agresti (1990)는 로지스틱 회귀모형이 가장 보편화된 일반화 선형 모형이라고 언급하였다. 로지스틱 회귀모형은 범주형 반응 자료에 대한 가장 중요한 모형이라고 할 수 있고, 현재는 인구학뿐만 아니라 금융, 의학, 생물학, 사회학, 인구학, 마케팅 등 다양한 범위에서 사용이 증가하고 있다.

이 논문에서는 둘로 나누어진 결과를 가지는 이항반응을 가지는 데이터를 이용하며, 결과에서 개인 또는 집단의 특성이 요구되기 때문에 임의효과가 포함된 로지스틱 모형을 고려한다. 이러한 이항반응은 결과가 발생할 경우가 2가지만 있기 때문에 변동을 나타내기 쉽지 않고 복잡하지만 많은 분야에서 사용되고 있다 (Kim 등, 2013).

### 1.3. 로지스틱 선형 혼합 모형의 최대우도 추정법

임의효과가 포함된 로지스틱 회귀모형에 대한 통계적 추론으로 최대우도법(maximum likelihood estimation; MLE), 베이즈 방법 등 여러 가지의 방법론이 있다. 그 중 베이즈 방법은 임의효과의 공분산

행렬이 포함되어 있어 사전분포를 구체화하기가 복잡하기 때문에 모의실험에서 최대우도법보다 효율이 낮아질 수 있다. 그래서 이 연구에서는 선형 혼합 모형의 모수를 추정하기 위하여 최대우도법을 사용하고자 한다. 하지만 임의효과가 포함된 로지스틱 모형은 최대우도추정량을 구하기 위해서 사용하는 우도함수가 정보행렬(information matrix)과 높은 차원의 적분을 포함하는 닫힌 형식(closed form)을 가지고 있기 때문에 계산적으로 복잡함과 어려움이 따른다. 이를 해결할 수 있는 방법이 근사추정이며, 변량효과가 포함된 모형에 대한 최대우도의 추정에 대한 어려움을 해결하기 위해 Schall (1991), Breslow와 Clayton (1993), Wolfinger와 O'Cconnell (1993) 등의 학자들이 다양한 추론 방법을 다루었다.

대표적인 방법으로, Laplace (1986)는 본래 라플라스 근사법(Laplace approximation)에 대한 원리를 제안하였다. 그 후에 Solomon과 Cox (1992) 그리고 Liu와 Pierce (1994)은 적분된 우도에 대한 라플라스 근사를 제안하였다. Wolfinger (1993)는 적분에 대한 라플라스 근사법이 비선형에서 고정효과, 임의효과 둘 다 있는 데이터의 주변분포에 적용되고, 이 접근은 위에서 언급했던 모델의 적합(fitting)을 위한 최근 알고리즘들의 대안적 도출을 제공한다고 언급했다. 그리고 Breslow와 Clayton (1993)은 일반화 선형 혼합 모형의 넓고 다양한 적용을 위해 라플라스 근사 확장을 위한 몇 가지 수정점을 제시하였다 (Breslow와 Lin, 1995). 라플라스 근사법은 테일러 임의의 함수를 특정 위치에서 정규분포로 근사시키는 근사 방법이다. 함수를 한 위치에 근사하기 위하여 테일러 급수를 사용하였으며, 테일러 급수는 보통 2차 형태까지 적용한다. 확률분포가 차수가 높거나 복잡하여 계산하기 힘들 경우, 그 분포와 가장 비슷한 정규분포의 함수를 찾아 그 함수로 대신하여 사용하는 것이다 (Bishop, 2006). 또한 라플라스 근사법의 정확성은 표본 수에 의존하고, 가우스-에르미트 구적법(Gauss-Hermite quadrature)과 적응 가우스-에르미트 구적법보다 계산적으로는 더 간단하지만 작은 표본을 가지는 데이터에 대하여 덜 정확한 추정을 하는 경향이 있다 (Kim 등, 2013). 라플라스 근사법의 경우 여러 가지 주의할 점들이 포함되어 있지만 사용하기 간편하기 때문에 현재 가장 보편적으로 사용하는 근사 방법이다.

가우스-에르미트 구적법은 다양하게 제시된 가우스 구적법(가우스-에르미트 구적법, 가우스-르장드르 구적법, 체비셰프-가우스 구적법 등) 중 하나이다. 가우스 구적법은 적분하는 간격에 따라 사용되는 구적법이 다르다. 그중에서도 가우스-에르미트 구적법은 정규 밀도 함수의 모양을 가지는 또 다른 함수의 곱인 함수 $f(\cdot)$의 적분으로 근사하는 방법이다. 적분 간격은 모든 실수 구간에서 고려할 수 있다. 적분은 특정한 점에서 함수를 평가하는 가중된 합에 근사한다. Kim 등 (2013)은 근사의 정확성은 구적 점과 그에 상응하는 가중치에 의해 나타내지는 면적에 의해 의존한다고 주장하였다. 임의효과가 여러 개가 포함된 모형인 경우 임의효과가 $k$개가 존재한다면 가우스-에르미트 구적법을 진행하기 위해서 1개의 임의효과 당 $Q$개의 점을 사용하기 때문에 근사에 필요한 구적 점은 $Q^k$개로 기하급수적으로 증가한다는 것을 볼 수 있으며 (Lesaffre와 Spiessens, 2001; Hedeker와 Gibbons 2006), 구적 점의 개수가 많아질수록 근사가 정확하게 된다. 하지만 임의효과가 증가할 경우 계산이 복잡해지는 문제가 생길 수 있으므로 필요한 개수만큼 포함시키는 것이 좋다. 가우스-에르미트 구적법 외에도 가우스-에르미트 구적법을 보완한 적응 가우스-에르미트 구적법도 언급되고 있으며, 가우스-에르미트 구적법은 미리 설정된 구적 점을 사용하지만 적응 가우스-에르미트 구적법은 분포의 모양에 따라 근사하는 모형을 중심으로 구적 점을 찾는다 (Lesaffre와 Spiessens, 2001; Hedeker와 Gibbons, 2006). 필요한 적당한 구적 점을 찾아주기 때문에 일반적으로 사용되는 가우스-에르미트 구적법보다 적응 가우스-에르미트는 근사하기 위해 더 적은 수의 구적 점이 필요로 하므로 더 효율적일 수 있다 (Kim 등, 2013).

이러한 최대우도법을 사용하여 일반화 선형 혼합 모형의 모수를 추정할 수 있는 통계적 패키지로는 R의 lme4, SAS의 PROC NLMIXED와 PROC GLIMMIX 등이 있다. 최근에는 SAS의 프로시저 중 PROC NLMIXED보다 PROC GLIMMIX가 더 많이 사용되며, 이러한 R과 SAS의 패키지는 벌점준가능도(penalized quasi-likelihood) 근사, 적분의 근사 방법을 이용한 라플라스 근사와 가우스-에르

미트 적분 근사 중 적어도 한 가지의 방법을 제공하고 있다. 특히 두 패키지 안에 있는 우도 근사 방법은 PROC GLIMMIX의 경우 라플라스 우도 근사, 유사가능도 우도 근사, 적응 가우스-에르미트 구적법의 우도 근사 등이 제공되고 있고, R의 lme4 패키지에서 사용되는 glmer 함수의 경우 라플라스 우도 근사, 적응 가우스-에르미트 구적법의 우도 근사가 제공하고 있다 (Bolker 등, 2009). 최근에 두 통계적 패키지를 비교한 Li 등 (2011)에 의하면 R에서의 lme4와 SAS에서의 PROC GLIMMIX는 고정효과와 임의효과의 공분산에 대하여 거의 같은 결과를 제공한다고 설명하였다. 이를 확인하기 위해서 이 논문에서는 다양한 우도함수 근사법 중에서도 lme4 패키지와 PROC GLIMMIX가 동시에 포함하고 있는 라플라스 근사법과 적응 가우스-에르미트 구적법을 수행하며, 반복적으로 최대 우도 추정치가 수렴할 때까지 갱신하여 구체적인 추정량을 얻을 수 있는 뉴튼-라프슨 방법(Newton-Raphson method)을 통하여 로지스틱 선형 혼합모형의 모수를 추정한다.

### 1.4. 논문구성

본 논문의 구성은 다음과 같다. 2절에서는 임의효과가 포함된 로지스틱 모형에 대한 가정과 우도함수를 제시하고, 3절에서는 임의효과가 포함된 로지스틱 모형의 최대 우도 추정량을 구하는 여러 가지 방법론들에 대해 설명할 것이다. 4절에서는 분석에 사용된 복지 데이터에 대한 자세한 설명과 R lme4 패키지의 glmer 함수와 SAS에서 PROC GLIMMIX의 두 패키지를 중심으로 이 주제의 방법론인 최대우도법을 이용하여 데이터를 분석한다. 마지막으로 5절에서는 임의효과가 포함된 로지스틱 회귀모형에 적합한 모형의 분석 결과 및 연구에 대한 결론을 제시하고자 한다.

## 2. 로지스틱 임의효과 모형

로지스틱 임의효과 모형은 일반적인 로지스틱 회귀모형을 확장한 모형이다. 그러므로 로지스틱 임의효과 모형을 살피기 전 로지스틱 회귀모형의 우도 함수 및 이론적인 모수 추정법에 대해 살펴본다.

### 2.1. 로지스틱 회귀모형

베르누이 반응변수 $Y_i$ ($y_i = 1$ 또는 0)에 대하여 랜덤성분은 이항분포로 가정하고, 설명변수들의 선형관계성을 고려한 성공의 확률 $\Pr(Y_i = 1|\mathbf{X}_i) = \pi(\mathbf{X}_i)$에 대한 모형으로 다음 모형을 가정한다.

$$\log\left(\frac{\pi(\mathbf{X}_i)}{1-\pi(\mathbf{X}_i)}\right) = \mathbf{X}_i\boldsymbol{\beta} = \sum_{j=0}^{p}\beta_j x_{ij} \Longleftrightarrow \pi(\mathbf{X}_i) = \frac{\exp\left(\sum_{j=0}^{p}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{p}\beta_j x_{ij}\right)}. \tag{2.1}$$

$\mathbf{Y} = \{Y_1, \ldots, Y_n\}$은 이항반응 변수들의 벡터이고 $\pi(\mathbf{X}_i) = E[Y_i]$이며, $\mathbf{X}_i = (1, X_{i1}, \ldots, X_{ip})$는 $p$개의 설명변수를 가지는 설계벡터이다 ($i = 1, \ldots, n$). 독립인 반응변수들의 우도함수에 식 (2.1)에서 정의한 모형을 사용하면, 로지스틱 회귀모형의 우도함수는 다음과 같이 표현할 수 있다.

$$\begin{aligned} l(\boldsymbol{\beta}) &= \prod_{i=1}^{N}\pi(X_i)^{y_i}(1-\pi(X_i))^{n_i-y_i} = \prod_{i=1}^{n}\left\{\frac{\pi(X_i)}{1-\pi(X_i)}\right\}^{y_i}\prod_{i=1}^{n}(1-\pi(X_i))^{n_i} \\ &= \prod_{i=1}^{n}\left[\exp\left\{\log\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right)^{y_i}\right\}\right]\prod_{i=1}^{n}(1-\pi(X_i))^{n_i} \\ &= \exp\left\{\sum_j\left(\sum_i y_i x_{ij}\right)\beta_j\right\}\prod_{i=1}^{n}\left\{1+\exp\left(\sum_j\beta_j x_{ij}\right)\right\}^{-1}. \end{aligned} \tag{2.2}$$

식 (2.2)와 같이 표현된 우도함수에서 로그를 적용한 형태인 로그 우도함수는

$$L(\boldsymbol{\beta}) = \sum_j \left( \sum_i y_i x_{ij} \right) \beta_j - \sum_i n_i \log \left( 1 + \exp \left( \sum_j \beta_j x_{ij} \right) \right)$$

이다. 모수 $\beta$에 대한 최대우도 추정량을 구하기 위해 먼저 로그 우도함수를 $\beta_j$로 미분하여, 일차 미분함수를 0으로 놓고 얻은 $\hat{\pi}_i$의 최대우도추정량은 다음과 같다.

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_i y_i x_{ij} - \sum_i n_i x_{ij} \frac{\exp\left(\sum_j \beta_j x_{ij}\right)}{1 + \exp\left(\sum_j \beta_j x_{ij}\right)} = 0 \\ &\Rightarrow \sum_i y_i x_{ij} - \sum_i n_i x_{ij} \hat{\pi}_i = 0. \end{aligned} \tag{2.3}$$

식 (2.3)의 방정식은 $\hat{\pi}_i = \exp\left(\sum_j \beta_j x_{ij}\right) / \{1 + \exp\left(\sum_j \beta_j x_{ij}\right)\}$가 $\pi(X_i)$의 최대우도추정량일 때의 우도방정식이며, 이항의 일반화 선형 모형에서 비선형이고, 반복된 처리를 해야 하므로 특별한 경우이다. 최대우도의 추정량인 $\hat{\boldsymbol{\beta}}$에 대한 미분방정식은 닫힌 형식을 가지고 있지 않기 때문에 이를 구하기 위해서는 수리적으로 계속 계수를 수정하면서 모델을 개선해야 하는지 확인하는 과정을 거쳐야 한다. 최대우도 추정량의 성질에 의해 $\hat{\boldsymbol{\beta}}$은 정보행렬의 역행렬과 동일한 공분산행렬을 가지는 큰 표본의 정규분포를 가진다. 이러한 정보행렬에 필요한 요소인 음이차 미분방정식 $-\partial L^2(\boldsymbol{\beta})/(\partial\beta_a \partial\beta_b)$을 계산하는 데 필요한 세 개의 식은 다음과 같다.

1) $\dfrac{d}{dx} e^{u(x)} = e^{u(x)} \dfrac{d}{dx} u(x)$,

2) $\left( \dfrac{f}{g} \right)'(a) = \dfrac{f'(a)g(a) - f(a)g'(a)}{\{g(a)\}^2}$,

3) $\dfrac{d}{dx} \dfrac{e^{u(x)}}{1 + e^{u(x)}} = \dfrac{\frac{d}{dx}u(x) e^{u(x)} \left(1 + e^{u(x)}\right) - e^{u(x)} \frac{d}{dx} e^{u(x)}}{\left(1 + e^{u(x)}\right)^2} = \dfrac{\frac{d}{dx}u(x) e^{u(x)}}{\left(1 + e^{u(x)}\right)^2}$.

이 세 개의 식을 이용하여 구한 $\beta_a$와 $\beta_b$의 음이차 미분식 $-\partial L^2(\boldsymbol{\beta})/(\partial\beta_a\partial\beta_b)$은 다음과 같다.

$$\begin{aligned} -\frac{\partial L^2(\boldsymbol{\beta})}{\partial\beta_a \partial\beta_b} &= \sum_i n_i x_{ia} x_{ib} \frac{\exp\left(\sum_j \beta_j x_{ij}\right)}{\left[1 + \exp\left(\sum_j \beta_j x_{ij}\right)\right]^2} \\ &= \sum_i n_i x_{ia} x_{ib} \frac{1}{1 + \exp\left(\sum_j \beta_j x_{ij}\right)} \frac{\exp\left(\sum_j \beta_j x_{ij}\right)}{1 + \exp\left(\sum_j \beta_j x_{ij}\right)} \\ &= \sum_i n_i x_{ia} x_{ib} \pi(x_i)(1 - \pi(x_i)). \end{aligned}$$

모든 모수에 대하여 위의 방법을 적용하여 구한 최대우도의 추정량인 $\hat{\boldsymbol{\beta}}$의 추정된 공분산행렬은 다음과 같다.

$$\widehat{\mathrm{cov}}\left(\hat{\boldsymbol{\beta}}\right) = \left\{ \mathbf{X}' \mathrm{diag}[n_i \hat{\pi}_i (1 - \hat{\pi}_i)] \mathbf{X} \right\}^{-1}$$

여기서 $\mathrm{diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)]$는 주대각선이 $n_i\hat{\pi}_i(1-\hat{\pi}_i)$인 $N \times N$ 대각행렬이며, 추정량 $\hat{\pi}_i$으로 나타낸 분산으로 추정된 공분산행렬이라는 특징이 있다.

### 2.2. 로지스틱 임의효과 모형

이항 반응변수를 가지고 있는 일반적인 로지스틱 모형을 확장한 임의효과가 포함된 로지스틱 모형은 일반화 선형 혼합 모형의 특별한 형태 중 하나이다. 임의효과는 표본에 적용되는 효과를 말하며, 알지 못하는 설명변수에 대한 분산 구조나 관측되지 않은 숨겨진 분산에 대해 고려할 때 사용하는 모형에 가정되는 효과이다. 일반적으로 일반화 선형 혼합 모형은 다음과 같이 정의한다.

$$g(E(\mathbf{y}|\mathbf{u})) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \quad \mathbf{y} \sim F,$$
$$E(\mathbf{y}|\mathbf{u}) = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) = H(\boldsymbol{\eta}) = \boldsymbol{\mu},$$

여기서

$\mathbf{y}$ : 반응변수의 $n \times 1$ 벡터, $F$-분포로부터 온 반응변수
$\mathbf{X}$ : $\boldsymbol{\beta}$의 계획행렬 $n \times p$ 벡터
$\boldsymbol{\beta}$ : 고정효과 $p \times 1$ 벡터
$\mathbf{Z}$ : $\mathbf{u}$의 계획행렬 $n \times q$ 벡터
$\mathbf{u}$ : 임의효과 $q \times 1$ 벡터
$g(\cdot)$ : 연결함수

이때 임의효과는 $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I_n})$로 가정하고, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$는 선형 예측치라고 한다. 추정치 $\hat{\sigma}_u^2$는 그룹 간의 변동과 설명변수의 누락으로 인한 변동을 나타낸다. $\mathbf{u}$가 주어진 $\mathbf{y}$의 평균과, 연결함수로 로짓 연결을 사용한 로지스틱 모형은 식 (2.4)와 같이 정의한다.

$$\text{logit}\{E(\mathbf{y})\} = \log \frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}. \tag{2.4}$$

특히 식 (2.4)의 모형을 M개의 독립적인 그룹을 가지고 있는 그룹화 된 자료의 경우라면 아래와 같이 다시 표현할 수 있다.

$$E(\mathbf{y_j}|\mathbf{u_j}) = g^{-1}(\mathbf{X_j}\boldsymbol{\beta} + \mathbf{Z_j}\mathbf{u_j}), \quad j = 1, \ldots, M, \ i = 1, \ldots, n_j,$$

여기서

$\mathbf{y_j}$ : 반응변수의 $n_j \times 1$ 벡터, $F$-분포로부터 온 반응변수
$\mathbf{X_j}$ : $\boldsymbol{\beta}$의 계획행렬 $n_j \times p$ 벡터
$\boldsymbol{\beta}$ : 고정효과 $p \times 1$ 벡터
$\mathbf{Z_j}$ : $\mathbf{u}$의 계획행렬 $n_j \times q$ 벡터
$\mathbf{u_j}$ : 임의효과 $q \times 1$ 벡터
$g(\cdot)$ : 연결함수

$\mathbf{u_j} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I_{n_j}})$라고 할 때, $j$번째 집단의 우도함수는 다음과 같다.

$$\begin{aligned} l_j\left(\boldsymbol{\beta}, \sigma_u^2\right) &= \int f(\mathbf{y_j}|\eta_\mathbf{j}) f\left(\mathbf{u_j}|\sigma_u^2\right) d\mathbf{u_j} \\ &= \int f(\mathbf{y_j}|\eta_\mathbf{j})(2\pi)^{-\frac{q}{2}} \left(\sigma_u^2\right)^{-\frac{q}{2}} \exp\left(-\frac{1}{2\sigma_u^2}\mathbf{u_j'}\mathbf{u_j}\right) d\mathbf{u_j} \end{aligned}$$

$$= (2\pi)^{-\frac{q}{2}} \left(\sigma_u^2\right)^{-\frac{q}{2}} \int f(\mathbf{y_j}|\eta_\mathbf{j}) \exp\left(-\frac{1}{2\sigma_u^2}\mathbf{u_j'u_j}\right) d\mathbf{u_j}$$
$$= (2\pi)^{-\frac{q}{2}} \left(\sigma_u^2\right)^{-\frac{q}{2}} \int \exp\left\{\log f(y_j|\eta_j) - \frac{1}{2\sigma_u^2}\mathbf{u_j}'\mathbf{u_j}\right\} d\mathbf{u_j}.$$

$j$번째 집단의 우도함수에 로그를 취하여 식 (2.5)의 로그 우도함수를 얻을 수 있다.

$$L_j\left(\boldsymbol{\beta}, \sigma_u^2\right) = -\frac{q}{2}\log(2\pi) - \frac{q}{2}\log\left(\sigma_u^2\right) + \int \left\{\log f(y_j|\eta_j) - \frac{1}{2\sigma_u^2}\mathbf{u_j}'\mathbf{u_j}\right\} d\mathbf{u_j}. \tag{2.5}$$

이러한 그룹화 된 자료의 경우를 관측치 각각의 특징을 고려한 개별적 임의효과 모형으로 확장하고 로그 우도함수를 쉽게 구할 수 있다.

식 (2.5)의 우도함수는 닫힌 형식이 아니기 때문에 계산이 복잡하고 정확한 최대우도 추정량을 구할 수 없다. 그러므로 이를 위하여 식 (2.5)의 함수를 대신하는 근사 된 우도함수를 구하고 근사 된 우도함수를 통하여 뉴튼-라프슨 방법과 같은 반복적인 처리를 하여 최대우도 추정치를 구하는 방법을 적용할 수 있다. 우도함수 근사 방법으로는 라플라스 근사법, 가우스-에르미트 구적법, 유사가능도, 적응 가우스-에르미트 구적법 등이 있으며 이는 다음 절에서 자세히 논의한다.

## 3. 로지스틱 임의효과 모형의 최대우도 추정법

$\mathbf{Y} = \{Y_1, \ldots, Y_n\}$이 독립적이고 동일한 분포 $f_{\boldsymbol{\beta},\sigma_u^2}(\mathbf{y})$를 따르는 (independent identically distributed) 표본이라면, 일반적으로 우도함수 $l(\boldsymbol{\beta}, \sigma_u^2)$와 로그 우도함수 $L(\boldsymbol{\beta}, \sigma_u^2)$는 다음과 같다.

$$l\left(\boldsymbol{\beta}, \sigma_u^2\right) = \prod_{i=1}^{n} f_{\boldsymbol{\beta},\sigma_u^2}(y_i), \qquad L\left(\boldsymbol{\beta}, \sigma_u^2\right) = \sum \log f_{\boldsymbol{\beta},\sigma_u^2}(y_i).$$

우도함수와 로그 우도함수에 포함된 모수의 최대값을 구하기 위해서 주어진 함수를 모수에 대하여 1차 미분한 스코어 벡터를 구해야 한다. 선형모수 $\boldsymbol{\beta}$와 $\sigma_u^2$의 스코어 벡터는 식 (2.4)의 로그우도를 $\boldsymbol{\beta}$, $\sigma_u^2$에 대하여 한 번 미분한 $(p+1)\times 1$ 벡터이며 식 (3.1)과 같다.

$$u(\boldsymbol{\beta}, \sigma_u^2) = \begin{pmatrix} \dfrac{\partial L\left(\boldsymbol{\beta}, \sigma_u^2\right)}{\partial \beta_1} \\ \vdots \\ \dfrac{\partial L\left(\boldsymbol{\beta}, \sigma_u^2\right)}{\partial \beta_p} \\ \dfrac{\partial L\left(\boldsymbol{\beta}, \sigma_u^2\right)}{\partial \sigma_u^2} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial L\left(\boldsymbol{\beta}, \sigma_u^2\right)}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial L\left(\boldsymbol{\beta}, \sigma_u^2\right)}{\partial \sigma_u^2} \end{pmatrix}. \tag{3.1}$$

스코어 벡터의 적률은 두 가지의 성질을 만족한다.

1) 스코어 벡터의 기댓값은 0이다.

2) 스코어 벡터의 분산은 $L(\boldsymbol{\beta}, \sigma_u^2)$를 $\boldsymbol{\beta}$에 대해 두 번 미분한 함수의 음의 기댓값이다.

$$\mathrm{Var}\left(u\left(\boldsymbol{\beta}, \sigma_u^2\right)\right) = E\left[u\left(\boldsymbol{\beta}, \sigma_u^2\right) u\left(\boldsymbol{\beta}, \sigma_u^2\right)'\right] = -E\left[\begin{pmatrix} \dfrac{\partial^2 L\left(\boldsymbol{\beta}, \sigma_u^2\right)}{\partial \beta_j \partial \beta_k} & \dfrac{\partial^2 L\left(\boldsymbol{\beta}, \sigma_u^2\right)}{\partial \beta_j \partial \sigma_u^2} \\ \dfrac{\partial^2 L\left(\boldsymbol{\beta}, \sigma_u^2\right)}{\partial \sigma_u^2 \partial \beta_k} & \dfrac{\partial^2 L\left(\boldsymbol{\beta}, \sigma_u^2\right)}{\partial^2 \sigma_u^2} \end{pmatrix}\right]. \tag{3.2}$$

식 (3.2)의 행렬을 정보행렬이라고 하고 $I(\boldsymbol{\beta})$으로 표기한다. 정보행렬은 확률변수의 관측 값으로부터 확률변수 분포의 매개변수에 대해 유추할 수 있는 정보의 양이다. 위의 스코어 벡터 적률의 두 가지 성질을 이용하면 최대우도 추정량 $\hat{\boldsymbol{\beta}}$의 분포는 근사적으로 $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathbf{I}(\boldsymbol{\beta})^{-1})$를 따른다.

### 3.1. 뉴튼-라프슨 방법

뉴튼-라프슨 방법은 비선형 방정식을 풀기 위한 반복적인 방법이다. 최대우도추정량의 근사적인 해를 찾을 수 있다. 시작점부터 시작하여 반복적으로 최대값을 갱신하면서 함수의 극대값을 가지는 점을 결정하게 된다. 함수가 적절하거나 시작점이 좋을 때 최대값의 위치가 더 빨리 수렴되기 쉽다.

$$\mathbf{u}' = \begin{pmatrix} \frac{\partial L(\boldsymbol{\beta}, \sigma_u^2)}{\partial \beta_1} \\ \vdots \\ \frac{\partial L(\boldsymbol{\beta}, \sigma_u^2)}{\partial \beta_p} \\ \frac{\partial L(\boldsymbol{\beta}, \sigma_u^2)}{\partial \sigma_u^2} \end{pmatrix}, \qquad \mathbf{H} = \begin{pmatrix} \frac{\partial^2 L(\boldsymbol{\beta}, \sigma_u^2)}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 L(\boldsymbol{\beta}, \sigma_u^2)}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta}, \sigma_u^2)}{\partial \beta_1 \partial \sigma_u^2} \\ \frac{\partial^2 L(\boldsymbol{\beta}, \sigma_u^2)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 L(\boldsymbol{\beta}, \sigma_u^2)}{\partial \beta_2 \partial \beta_2} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta}, \sigma_u^2)}{\partial \beta_2 \partial \sigma_u^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 L(\boldsymbol{\beta}, \sigma_u^2)}{\partial \sigma_u^2 \partial \beta_1} & \frac{\partial^2 L(\boldsymbol{\beta}, \sigma_u^2)}{\partial \sigma_u^2 \partial \beta_2} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta}, \sigma_u^2)}{\partial^2 \sigma_u^2} \end{pmatrix}$$

$\boldsymbol{\beta}$와 $\sigma_u^2$의 모수 벡터를 $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta} \\ \sigma_u^2 \end{pmatrix}$라고 하고, $\boldsymbol{\theta}$에 대해서 $t$번 반복한 $\boldsymbol{\theta}^{(t)} = \begin{pmatrix} \boldsymbol{\beta}^{(t)} \\ \sigma_u^{2\,(t)} \end{pmatrix}$에서 평가된 $\mathbf{u}'$와 $\mathbf{H}$를 $\mathbf{u}^{(t)}$, $\mathbf{H}^{(t)}$라고 정의한다. $t$번 반복은 테일러 급수에서 두 번째 순서까지의 항에 의하여 $\boldsymbol{\theta}^{(t)}$근처의 $L(\boldsymbol{\beta}, \sigma_u^2)$를 극대화한다.

$$\begin{aligned} L\left(\boldsymbol{\beta}, \sigma_u^2\right) &\approx L\left(\boldsymbol{\beta}^{(t)}, \sigma_u^2\right) + \mathbf{u}^{(t)}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}\right) + \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}\right)' \mathbf{H}^{(\mathbf{t})}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}\right) \\ &= L\left(\boldsymbol{\beta}^{(t)}, \sigma_u^2\right) + \mathbf{u}^{(t)}\boldsymbol{\theta} - \mathbf{u}^{(t)}\boldsymbol{\theta}^{(t)} + \frac{1}{2}\left(\boldsymbol{\theta}'\mathbf{H}^{(t)}\boldsymbol{\theta} - 2\boldsymbol{\theta}'\mathbf{H}^{(t)}\boldsymbol{\theta}^{(t)} + \boldsymbol{\theta}^{(t)\prime}\mathbf{H}^{(t)}\boldsymbol{\theta}^{(t)}\right), \\ \frac{\partial L\left(\boldsymbol{\beta}, \sigma_u^2\right)}{\partial \boldsymbol{\theta}} &= \mathbf{u}^{(t)} + \frac{1}{2}\left(2\mathbf{H}^{(t)}\boldsymbol{\theta} - 2\mathbf{H}^{(t)}\boldsymbol{\theta}^{(t)}\right) \\ &= \mathbf{u}^{(t)} + \mathbf{H}^{(t)}\boldsymbol{\theta} - \mathbf{H}^{(t)}\boldsymbol{\theta}^{(t)} = \mathbf{u}^{(t)} + \mathbf{H}^{(t)}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}\right). \end{aligned}$$

다음 $t+1$번째 $\boldsymbol{\theta}$에 대해서 $\partial L(\boldsymbol{\beta}, \sigma_u^2)/\partial \boldsymbol{\theta} = 0$을 계산하면 다음과 같이 표현된다.

$$\begin{aligned} \frac{\partial L\left(\boldsymbol{\beta}, \sigma_u^2\right)}{\partial \boldsymbol{\theta}} &\approx \mathbf{u}^{(t)} + \mathbf{H}^{(t)}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}\right) = 0, \\ \boldsymbol{\theta} - \boldsymbol{\theta}^{(t)} &= -\left\{\mathbf{H}^{(t)}\right\}^{-1}\mathbf{u}^{(t)}, \\ \boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \left\{\mathbf{H}^{(t)}\right\}^{-1}\mathbf{u}^{(t)}, \end{aligned}$$

여기서 $\mathbf{H}^{(t)}$는 정칙행렬이라고 가정한다.

$\hat{\boldsymbol{\theta}}$의 근사적인 공분산행렬인 $\mathrm{cov}(\hat{\boldsymbol{\theta}})$은 정보행렬의 역행렬이다. 정보행렬의 요소는 $-E(\partial^2 L(\boldsymbol{\beta}, \sigma_u^2)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}')$와 같다. 큰 곡률은 $\hat{\boldsymbol{\theta}}$에서 $\boldsymbol{\theta}$로 가면서 로그 우도함수가 빠르게 떨어지는 것을 암시하기 때문에 정보를 찾기 쉬움을 의미한다.

반복은 연속된 과정에서 $L(\boldsymbol{\theta}^{(t)})$의 변화가 충분히 작아질 때 까지 진행된다. $t \to \infty$일 때, 최대우도추정량은 $\boldsymbol{\theta}^{(t)}$에 수렴한다. 하지만 뉴튼-라프슨 방법은 $L(\boldsymbol{\beta}, \sigma_u^2)$의 도함수가 0이 되는 다른 극댓값이 존재한다면 다른 극댓값을 최대우도추정량으로 선택할 수 있기 때문에 주의해야 한다. 이 때문에 시작점을 선택하는 것이 중요하고, 시작점을 여러 개로 설정하여 진행하는 것이 좋다.

### 3.2. 라플라스 근사법

다음은 닫힌 형태에 대한 우도를 근사하기 위한 라플라스 근사법을 이용하여 임의효과가 포함된 로지스틱 모형을 고려하였다. 식 (2.5)에서 지수함수의 식을 정리하면 식 (3.3)과 같다.

$$h\left(\boldsymbol{\beta}, \sigma_u^2, \mathbf{u_j}\right) = \log f(\mathbf{y_j}|\eta_\mathbf{j}) - \frac{1}{2\sigma_u^2}\mathbf{u_j'u_j}. \tag{3.3}$$

라플라스 근사법은 식 (3.3)의 테일러 전개를 기초로 한다. 테일러 전개를 적용하기 위해 식 (3.3)의 1차 미분 함수와 2차 미분 함수를 구한다.

$$h'\left(\boldsymbol{\beta}, \sigma_u^2, \mathbf{u_j}\right) = \frac{\partial h\left(\boldsymbol{\beta}, \sigma_u^2, \mathbf{u_j}\right)}{\partial \mathbf{u_j}} = \mathbf{Z_j'}\frac{\partial \log f(\mathbf{y_j}|\eta_\mathbf{j})}{\partial \eta_\mathbf{j}} - \frac{1}{\sigma_u^2}\mathbf{u_j}, \tag{3.4}$$

$$h''\left(\boldsymbol{\beta}, \sigma_u^2, \mathbf{u_j}\right) = \frac{\partial^2 h\left(\boldsymbol{\beta}, \sigma_u^2, \mathbf{u_j}\right)}{\partial \mathbf{u_j}\partial \mathbf{u_j'}} = \mathbf{Z_j'}\frac{\partial^2 \log f(\mathbf{y_j}|\eta_\mathbf{j})}{\partial \eta_\mathbf{j}\partial \eta_\mathbf{j}'}\mathbf{Z_j} - \frac{1}{\sigma_u^2}, \tag{3.5}$$

$$h'\left(\boldsymbol{\beta}, \sigma_u^2, \mathbf{u_j}\right) = \mathbf{Z_j'}\frac{\partial^2 \log f(\mathbf{y_j}|\eta_\mathbf{j})}{\partial \eta_\mathbf{j}} - \frac{1}{\sigma_u^2}\mathbf{u_j} = 0,$$

$$\hat{\mathbf{u}}_\mathbf{j} = \sigma_u^2 \mathbf{Z_j}\frac{\partial \log f(\mathbf{y_j}|\eta_\mathbf{j})}{\partial \eta_\mathbf{j}}. \tag{3.6}$$

식 (3.4)가 0인 점에서 최대값을 갖기 때문에, 이를 활용하여 식 (3.6)의 형태를 가지고 있는 $\hat{\mathbf{u}}_\mathbf{j}$을 구할 수 있으며, 테일러 전개를 이용한 근사식에서는 식 (3.4)가 제거되면서 식 (3.5)의 모형만 남는다. 그래서 테일러 전개가 적용된 식 (3.3)은 2차식이 되고, 가우스 분포의 형태로 근사할 수 있다.

$$\begin{aligned} h\left(\boldsymbol{\beta}, \sigma_u^2, \mathbf{u_j}\right) &\approx h\left(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j}\right) + h'\left(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j}\right)\left(\mathbf{u_j} - \hat{\mathbf{u}}_\mathbf{j}\right)' + \frac{1}{2}\left(\mathbf{u_j} - \hat{\mathbf{u}}_\mathbf{j}\right)' h''\left(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j}\right)\left(\mathbf{u_j} - \hat{\mathbf{u}}_\mathbf{j}\right) \\ &= h\left(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j}\right) + \frac{1}{2}\left(\mathbf{u_j} - \hat{\mathbf{u}}_\mathbf{j}\right)' h''\left(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j}\right)\left(\mathbf{u_j} - \hat{\mathbf{u}}_\mathbf{j}\right). \end{aligned}$$

적분에 대한 근사는

$$\int \exp h\left(\boldsymbol{\beta}, \sigma_u^2, \mathbf{u_j}\right) d\mathbf{u_j} \approx \left|-h''\left(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j}\right)\right|^{-\frac{1}{2}} \exp\left\{h\left(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j}\right)\right\}$$

이며, 라플라스 근사 된 로그 우도함수는 식 (3.7)과 같다. 미분하여 최대 우도 추정량을 구하는 경우, 상수는 통계량에 변화를 주지 않으므로 계산의 편의를 위해 식에서 상수를 생략하였다.

$$L_j\left(\boldsymbol{\beta}, \sigma_u^2\right) = -\frac{1}{2}\log\left|-h''\left(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j}\right)\right| + h\left(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j}\right). \tag{3.7}$$

식 (3.7)에서의 $h''(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j})$는 최대값이 되기 위해서 2차 미분 값이 음수의 형태로 나타나야 하므로 0보다 큰 값을 가져야 한다. $L(\boldsymbol{\beta}, \sigma_u^2)$는 식 (3.7) $j = 1$에서 $M$까지 함수의 합이며 ($L(\boldsymbol{\beta}, \sigma_u^2) = \sum_{j=1}^{M} L_j(\boldsymbol{\beta}, \sigma_u^2)$), $L(\boldsymbol{\beta}, \sigma_u^2)$의 최대화는 $\boldsymbol{\theta}$에 대해 수행된다. 최대화하여 얻은 추정량을 사용하여 최적의 $\hat{\boldsymbol{\beta}}$와 $\hat{\sigma}_u^2$를 구하기 위한 반복적인 방법은 뉴튼-라프슨 방법을 사용한다. 뉴튼-라프슨 방법을 적용하면

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left\{\frac{\partial^2 L\left(\boldsymbol{\beta}, \hat{\sigma}_u^2\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right\}^{-1}\left\{\frac{\partial L\left(\boldsymbol{\beta}, \hat{\sigma}_u^2\right)}{\partial \boldsymbol{\theta}}\right\}, \tag{3.8}$$

$$\operatorname{cov}\left(\boldsymbol{\theta}^{(t+1)}\right) = -\left\{\frac{\partial^2 L(\boldsymbol{\theta}^{(t+1)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right\}. \tag{3.9}$$

식 (3.8)과 식 (3.9)를 얻을 수 있다. 식 (3.8)과 식 (3.9)를 반복적으로 계산하여, 값의 변화가 적당히 작아지면 반복을 멈춘다. 라플라스 근사법에서 특정한 점에서 근사하는 근사값을 찾고, 근사값과 가까운 극대값을 얻기 때문에 항상 최대값이 아님을 주의해야 한다. 그래서 뉴튼-라프슨 방법처럼 시작점을 여러 개 두고 시작하는 것이 좋다. 그리고 과정을 진행하기 위해서 라플라스 근사의 조건이 있는데, 첫 번째는 식 (3.5)가 필요하기 때문에 함수가 두 번 이상 미분할 수 있어야 한다. 두 번째로 식 (3.3)이 최대값을 가지고 있어야 한다. 그리고 마지막으로 라플라스 근사는 가우스 분포에 맞도록 근사시키기 때문에 실수의 범위를 가지고 있어야 한다. 자세한 설명은 Agresti (2002), StataCorp (2013) 등을 참고하여라.

### 3.3. 가우스-에르미트 구적법

가우스-에르미트 구적법의 형태는 다음과 같다.

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx \approx \sum_{i=1}^{n} w_i f(x_i)$$

$n$은 사용된 표본점의 수, $x_i$는 구적 점이며, $w_i$는 구적 가중치이다. 구적 점과 구적 가중치를 구하는 방법은 Abramowitz와 Stegun (1972)을 참고한다. 이를 토대로 임의효과가 포함된 로지스틱 모형의 우도함수를 적용하면 다음과 같다. 우도를 결정하는 적분은 임의효과 구조에 의존하는 차원을 가지고 있다.

$$\begin{aligned} l_j\left(\boldsymbol{\beta}, \sigma_u^2\right) &= (2\pi)^{-\frac{q}{2}} \left(\sigma_u^2\right)^{-\frac{q}{2}} \int \exp\left\{\log f(\mathbf{y_j}|\eta_\mathbf{j}) - \frac{1}{2\sigma_u^2}\mathbf{u_j'u_j}\right\} d\mathbf{u_j} \\ &= (2\pi)^{-\frac{q}{2}} (\sigma_u^2)^{-\frac{q}{2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left\{\log f(\mathbf{y_j}|\eta_\mathbf{j}) - \frac{1}{2\sigma_u^2}\sum_{k=1}^{q} u_{jk}^2\right\} du_{j1}, \ldots, u_{jq}. \end{aligned}$$

가우스-에르미트 구적법을 이용한 근사된 우도함수는 식 (3.10)으로 모형화 될 수 있다.

$$\begin{aligned} l_j\left(\boldsymbol{\beta}, \sigma_u^2\right) &= \sum_{k_1=1}^{r} \cdots \sum_{k_q=1}^{r} \left[\exp\left\{\log f(\mathbf{y_j}|\eta_\mathbf{j})\right\} \prod_{p=1}^{q} w_{jk_p}\right] \\ &= \sum_{k_1=1}^{r} \cdots \sum_{k_q=1}^{r} \left[\exp\left\{\sum_{i=1}^{n_j} \log f(y_{ij}|\eta_{ijk})\right\} \prod_{p=1}^{q} w_{jk_p}\right], \end{aligned} \tag{3.10}$$

여기서 $\boldsymbol{\eta_{jk}}$는 $k$번째 점에서의 $\boldsymbol{\eta_{jk}} = \mathbf{X_j}\boldsymbol{\beta} + \mathbf{Z_j u_j}$이고, $\eta_{ijk}$는 $\boldsymbol{\eta_{jk}}$의 $i$번째 요소이다. 각 차원에서 $r$개의 구적 점을 가진다. 식 (3.10)은 반복된 과정, 예를 들어 뉴튼-라프슨 방법과 같은 알고리즘을 통해 최대화를 할 수 있다. 여기서 관측된 정보행렬은 표준오차의 최대우도추정량을 제공한다.

### 3.4. 적응 가우스-에르미트 구적법

$\mathbf{z} = (z_1, \ldots, z_{k_q})$이 가우스-에르미트 구적법에 대한 표준 가로좌표라고 하면, $\mathbf{z_j^*} = (z_{j_1}, \ldots, z_{k_{j_r}})$는 $r$차원의 구적 격자망(grid)에서의 점이라고 하자. 집중되어 크기가 조정된 가로좌표 $\mathbf{a_j^*}$는

$$\mathbf{a_j^*} = \hat{\mathbf{u}}_\mathbf{j} + \sqrt{2} h''\left(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j}\right)^{-\frac{1}{2}} \mathbf{z_j^*}$$

여기서 $f''$는 임의효과에 대해서 두 번 미분한 행렬이다.

$$h''\left(\boldsymbol{\beta}, \sigma_u^2, \hat{\mathbf{u}}_\mathbf{j}\right) = \frac{\partial f\left(\mathbf{y_j}|\eta_\mathbf{j}\right)}{\partial \mathbf{u_j} \partial \mathbf{u_j'}}|_{\hat{\mathbf{u}}_\mathbf{j}}.$$

적응 가우스-에르미트 구적법(adaptive Gauss-Hermite quadrature)은 가우스-에르미트 구적의 가중치인 $\mathbf{w} = (w_{k_1}, \ldots, w_{k_q})$에 따라 집중되고 크기가 조정된 가로 좌표는 일차원 규칙의 순서에 의해 $r$차원의 적분을 구성하는 데 사용된다. 적응 가우스-에르미트 구적법을 이용한 근사 우도함수는 식 (3.11)로 표현될 수 있다.

$$l_j\left(\boldsymbol{\beta}, \sigma_u^2\right) = \sum_{k_1=1}^{n_1} \cdots \sum_{k_r=1}^{n_j} \left[ \exp\{\log f(\mathbf{y_j}|\eta_{\mathbf{jk}})\} \prod_{p=1}^{r} w_{jk_p^*} \right] \tag{3.11}$$
$$\boldsymbol{\eta_{jk}} = \mathbf{X_j}\boldsymbol{\beta} + \mathbf{Z_j}\mathbf{a_j^*}.$$

$\boldsymbol{\mu_j}$와 $\sigma_u^2$는 두 가지 방법을 사용하여 계산된다.

1) 식 (3.11)을 이용하여 사후 확률 질량 함수의 적률을 갱신하여 반복적으로 계산한다.

2) $\mathbf{u_j}$에 대해서 적분을 최적화하여 $\boldsymbol{\mu_j}$와 $\sigma_u^2$를 계산한다.

가우스-에르미트 구적법보다 효과적으로 적분을 근사하는 데 필요한 구적 점의 수가 줄어들기 때문에 효율성이 있는 방법이다.

### 3.5. 유사가능도 우도

일반화 선형혼합모형의 개념을 다시 불러오면

$$E(\mathbf{y}|\mathbf{u}) = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) = g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu}$$
$$\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{I_n}), \qquad \mathrm{Var}(\mathbf{y}|\mathbf{u}) = \sigma^2\mathbf{I}$$

이고, 분산함수는 평균의 함수로서 반응변수의 분산을 표현한다. Wolfinger와 O'Connell (1993)에 따라 $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}$에 대해 $\mu$의 첫째항을 이용한 테일러 정리로 함수를 근사할 수 있다. $\hat{\boldsymbol{\Delta}} = (\partial g^{-1}(\boldsymbol{\eta})/\partial\boldsymbol{\eta})|_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}$인 대각행렬을 이용하여 유사가능도 우도(pseudo-likelihood)를 구하기 위한 모델을 정의하자.

$$\begin{aligned} g^{-1}(\eta) &= g^{-1}(\hat{\eta}) + \hat{\boldsymbol{\Delta}}\mathbf{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) + \hat{\boldsymbol{\Delta}}\mathbf{Z}(\mathbf{u} - \hat{\mathbf{u}}) \\ &= g^{-1}(\hat{\eta}) + \hat{\boldsymbol{\Delta}}\left(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{u} - \mathbf{Z}\hat{\mathbf{u}}\right) \\ \hat{\boldsymbol{\Delta}}^{-1}\left(g^{-1}(\eta) - g^{-1}(\hat{\eta})\right) &+ \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}. \end{aligned}$$

$\mathbf{P} \equiv \hat{\boldsymbol{\Delta}}^{-1}(g^{-1}(\eta) - g^{-1}(\hat{\eta}))$는 $\mathbf{u}$가 통제된 예측값이다. $\mathrm{Var}(\mathbf{P}|\mathbf{u}) = \sigma^2\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{\Delta}}^{-1}$이고, 따라서 $\mathbf{P} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$의 모형을 고려할 수 있다. $\mathrm{Var}(\boldsymbol{\epsilon}) = \mathrm{Var}(\mathbf{P}|\mathbf{u})$이다.

$$V(\boldsymbol{\theta}) = \sigma_u^2\mathbf{Z}\mathbf{Z}' + \sigma^2\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{\Delta}}^{-1}.$$

$V(\boldsymbol{\theta})$는 주변함수의 분산이다. $\boldsymbol{\theta}$는 $\sigma_u^2$와 $R$에서 알지 못하는 것이 모두 포함된 $q \times 1$모수 벡터이며, 여기서 SAS 프로그램의 PROC GLIMMIX에서 $\boldsymbol{\epsilon}$는 정규분포를 따른다고 가정한다.

$$L(\boldsymbol{\theta}, p) = -\frac{1}{2}|V(\boldsymbol{\theta})| - \frac{1}{2}\mathbf{r}'V(\boldsymbol{\theta})\mathbf{r} - \frac{f}{2}\log 2\pi,$$
$$L_R(\boldsymbol{\theta}, p) = -\frac{1}{2}|V(\boldsymbol{\theta})| - \frac{1}{2}\mathbf{r}'V(\boldsymbol{\theta})\mathbf{r} - \frac{1}{2}\log\{\mathbf{X}'V(\boldsymbol{\theta})\mathbf{X}\} - \frac{f-k}{2}\log\{2\pi\},$$

여기서

$$f : \text{분석에서 사용된 빈도수의 합}$$
$$k : \text{행렬 } \mathbf{X}\text{의 계수}$$
$$L(\boldsymbol{\theta}, p) : \text{유사가능도 로그 우도함수의 최대값}$$
$$L_R(\boldsymbol{\theta}, p) : \text{잔차 유사가능도 로그 우도함수의 최대값}$$

$\mathbf{r} = \mathbf{P} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{P}$이다. 유사가능도 로그 우도함수가 수렴하게 되면 프로파일 함수의 모수가 추정되고, 임의효과가 예측된다.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'V\left(\hat{\boldsymbol{\theta}}\right)^{-1}\mathbf{X}\right)^{-}\mathbf{X}'V\left(\hat{\boldsymbol{\theta}}\right)^{-1}\mathbf{P},$$
$$\hat{\mathbf{u}} = \hat{\sigma}_u^2\mathbf{Z}'V\left(\hat{\boldsymbol{\theta}}\right)^{-1}\hat{\mathbf{r}}.$$

값을 대입하여 다음 값을 반복적으로 계산하며, 값에 대한 변화가 적당히 작아지면 과정을 멈춘다. 유사가능도 우도는 임의변수가 있는 결합 확률 분포의 근사치이다. 변수가 많을 때 주변화를 요구하는 경우 추정에 대한 계산을 쉽게 해주며, 모수의 정확한 추정을 제공한다. 실제 우도함수를 대신하여 유사가능도 로그 우도함수를 사용하는 것은 좋은 추정량이 될 수 있다. 하지만 추정 불확실성에 대한 정보를 유도하는 보통 우도의 간단한 적용에서는 부정확할 수 있다.

## 4. 로지스틱 임의효과 모형의 활용

### 4.1. 데이터 설명 및 기초분석

한국보건사회연구원에서 주관하는 2016년 11차 한국 복지 패널 조사를 이용하여 로지스틱 임의효과 모형을 적용하고자 한다. 한국 복지 패널 조사는 매년 진행하는 조사이며, 11차 자료는 3월 2일부터 6월 8일까지 총 99일 동안 진행되었다. 17개 시·도, 209개의 시·군·구를 대상으로 진행한 전국을 포괄하는 조사이며, 국민의 생활실태, 복지 욕구 외에도 다양한 분야에 대한 조사내용을 포함하고 있기 때문에 대표성을 띠는 패널 데이터이다. 개별가구에 대한 정보와 각 가구원에 대한 정보가 있기 때문에 임의효과에 대해 분석하기에 적합한 자료이다. 2016년 11차 조사의 가구 단위 관측치는 총 6,723가구이며, 개인 단위 관측치는 16,664명이다. 가구원의 일반사항 관련 변수가 없는 경우는 연구에서 제외되었다.

본 연구에서는 통계 소프트웨어 중에서도 SAS의 GLIMMIX와 R의 lme4을 사용하여 정신건강과 생활만족도가 자원봉사활동에 미치는 영향을 분석하였다. 연구에서 사용할 변수는 개별 가구원들의 일반사항 문항인 성별($X_1$), 연령대($X_2$), 거주 지역($X_3$), 소득($X_4$), 최종학력($X_5$)과 자원봉사활동 여부($y$), 정신건강 관련 변수(21개 문항), 생활실태·만족도 변수(8개 문항)이다. 거주 지역은 7개 권역 (서울/인천·경기/부산·경남·울산/대구·경북/대전·충남/강원·충북/광주·전남·전북·제주도)으로 구분되어 있고, 성별은 남자는 1, 여자는 2로 코딩된 가변수이다. 연령대는 출생연도에서 나이로 변환 후 연령대로 코딩 하였으며 소득 변수는 연속형 변수인 총 소득(gross income)을 범주화하였다. 최종학력은 교육수준 변수에서 졸업을 제외한 모든 응답을 졸업하지 않았다고 간주하고 변수를 다시 범주화하였다. 일반사항 문항을 알아보기 위해서 빈도분석을 한 결과는 Table 4.1에 제시된다.

자원봉사활동 여부는 그렇다 1, 아니다 2로 코딩되어있는 반응이 이분형인 변수이다. 총 12,305명의 응답자 중 1,054명은 자원봉사활동을 한 경험이 있다고 응답했으며, 나머지 11,251명은 자원봉사활동의 경험이 없다고 답하였다. 정신건강 관련 변수는 우울의 관한 인식과 자아존중감에 대한 인식으로 구분

**Table 4.1.** Data description and frequency of subdivision

| Variable | Subdivision | Notation | Frequency (rate) |
|---|---|---|---|
| 성별 ($X_1$) | 남성 | $X_1$_1 | 5287 (42.97) |
| | 여성 | $X_1$_2 | 7018 (57.03) |
| 연령대 ($X_2$) | 10대 이하 | $X_2$_1 | 8 (0.07) |
| | 20대 | $X_2$_2 | 1132 (9.20) |
| | 30대 | $X_2$_3 | 1417 (11.52) |
| | 40대 | $X_2$_4 | 2060 (16.74) |
| | 50대 | $X_2$_5 | 1892 (15.38) |
| | 60대 이상 | $X_2$_6 | 5796 (47.10) |
| 거주지역 ($X_3$) | 서울 | $X_3$_1 | 1793 (14.57) |
| | 인천/경기 | $X_3$_2 | 2633 (21.40) |
| | 부산/경남/울산 | $X_3$_3 | 2093 (17.01) |
| | 대구/경북 | $X_3$_4 | 1532 (12.45) |
| | 대전/충남 | $X_3$_5 | 1056 (8.58) |
| | 강원/충북 | $X_3$_6 | 935 (7.60) |
| | 광주/전남/전북/제주 | $X_3$_7 | 2263 (18.39) |
| 수입 ($X_4$) | 2000만원 미만 | $X_4$_1 | 3521 (28.61) |
| | 2000만원 이상–4000만원 미만 | $X_4$_2 | 2895 (23.53) |
| | 4000만원 이상–6000만원 미만 | $X_4$_3 | 2407 (19.56) |
| | 6000만원 이상–8000만원 미만 | $X_4$_4 | 1547 (12.57) |
| | 8000만원 이상 | $X_4$_5 | 1935 (15.73) |
| 최종학력 ($X_5$) | 초등학교 졸업 이하 | $X_5$_1 | 3812 (30.98) |
| | 중학교 졸업 | $X_5$_2 | 1438 (11.69) |
| | 고등학교 졸업 | $X_5$_3 | 4003 (32.53) |
| | 대학교(4년제, 전문대학) 졸업 | $X_5$_4 | 2825 (22.96) |
| | 대학원 이상 | $X_5$_5 | 227 (1.84) |

**Table 4.2.** Catalog of mental-healty, living condition and satisfaction variables

| 우울의 관한 인식 ($X_6$) | 자아존중감에 대한 인식 ($X_7$) | 생활실태·만족도 ($X_8$) |
|---|---|---|
| 1. 식욕이 없음 | 1. 나는 가치있는 사람이다 | 1. 건강 만족도 |
| 2. 비교적 잘 지냈다 | 2. 나는 좋은 성품을 지녔다 | 2. 가족의 수입 만족도 |
| 3. 상당히 우울 | 3. 나는 실패한 사람이라는 느낌이 든다 | 3. 주거 환경 만족도 |
| 4. 모든 일이 힘들게 느껴짐 | 4. 다른 사람들과 같이 일을 잘 할 수 있다 | 4. 가족관계 만족도 |
| 5. 잠을 설침 | 5. 자랑할 것이 별로 없다 | 5. 직업 만족도 |
| 6. 외로움 | 6. 긍정적인 태도를 가졌다 | 6. 사회적 친분관계 만족도 |
| 7. 불만없이 생활 | 7. 대체로 만족 | 7. 여가생활 만족도 |
| 8. 사람들이 차갑게 대하는 것 같은 느낌 | 8. 내 자신을 존경할 수 있으면 좋겠다 | 8. 전반적 만족도 |
| 9. 마음이 슬펐다 | 9. 내 자신이 쓸모없는 사람이라는 느낌 | |
| 10. 사람들이 나를 싫어하는 것 같은 느낌 | 10. 내가 좋지 않은 사람이라고 생각한다 | |
| 11. 뭘 해 나갈 엄두가 나지 않음 | | |

된다. 정신건강 관련 변수와 생활실태·만족도 변수는 범주 값이 클수록 부정에서 긍정적임을 나타낸다. 그래서 값이 커질수록 긍정에서 부정적인 일부 문항은 역코딩을 하였다. 정신건강 관련 변수와 생활실태·만족도 변수에 대한 자세한 설명은 Table 4.2에 있다.

우울의 관한 의식(11개 문항)과 자아존중감에 대한 인식(10개 문항), 그리고 생활실태·만족도(8개 문항)는 구분된 문항들 사이의 비슷한 특징이 있다. 그 경우에는 항목을 대표하는 종합척도를 사용하는 것이 좋다. 종합척도를 사용할 수 있는지 알아보기 위해 신뢰도분석을 하였다. 우울의 관한 의식 문항들

**Table 4.3.** Correlation between mental-healty, living condition and satisfaction variables

| Variable | $X_6$ | $X_7$ | $X_8$ |
|---|---|---|---|
| $X_6$ | 1 | 0.53881(< 0.0001) | 0.49706(< 0.0001) |
| $X_7$ | 0.53881(< 0.0001) | 1 | 0.56381(< 0.0001) |
| $X_8$ | 0.49706(< 0.0001) | 0.56381(< 0.0001) | 1 |

**Table 4.4.** Variable inflation factors (VIF) between independent variables

| Variable | VIF |
|---|---|
| $X_1$ | 1.07218 |
| $X_2$ | 1.73425 |
| $X_3$ | 1.04624 |
| $X_4$ | 1.68964 |
| $X_5$ | 2.11085 |
| $X_6$ | 1.56241 |
| $X_7$ | 1.80584 |
| $X_8$ | 1.68556 |

의 크론바흐의 알파 값은 0.885867, 자아존중감에 대한 인식은 0.802986, 생활실태·만족도에 대한 문항의 크론바흐의 알파 값은 0.834547로, 전부 1에 가까운 값을 보인다. 이를 통하여 우울의 관한 인식, 자아존중감에 대한 인식, 생활실태·만족도 구분으로 나누어져 있는 변수들은 종합척도로 사용될 수 있음을 알 수 있다. 종합척도는 개별 항목의 척도를 모두 더한 후 문항의 수로 나누어 주었다. 본 연구에서는 구분된 항목의 개별 척도를 사용하지 않고 종합척도를 사용한다.

종합척도를 상관분석 한 결과는 Table 4.3에서 보여주고 있다. 상관분석에 대한 가설은 다음과 같다.

$H_0$ : 두 변수 사이에 상관관계가 존재하지 않는다. vs. $H_1$ : 두 변수 사이에 상관관계가 존재한다.

각 변수 간의 상관계수 값이 의미가 있는지 확인하는 검정에서 모든 변수에서 유의한 것으로 보아 Table 4.3의 종합척도 간의 상관관계는 모두 의미가 있는 것으로 볼 수 있다. 우울함에 대한 인식, 자아존중감에 대한 인식, 생활실태·만족도 문항 사이의 상관계수가 약 0.5로 양의 상관관계가 존재하고 있다. 우울한 인식을 할수록 자아존중감이 낮은 경향을 보이고 생활만족도가 낮다. 반면 자아존중감이 높으면 생활만족도가 증가하는 경향을 보인다. 상관관계가 있는 경우 변수들의 설명력이 떨어지는 다중공선성의 문제가 생길 수 있어 확인이 필요하다.

각 변수 간의 다중공선성이 있는지 확인하기 위한 척도로 분산 팽창 인수(variance inflation factor; VIF)가 있으며, 분산 팽창 인수가 1에 가까우면 변수들 간의 다중공선성이 없다고 할 수 있다. Table 4.4에서는 8개 변수 대부분의 분산 팽창 인수가 1에 가까운 수를 가지고 있다. 그 이유로 변수 사이의 다중공선성의 문제가 보이지 않는다고 보았다.

Table 4.5는 반응변수와 명목형 독립변수(성별, 연령대, 거주 지역, 소득, 최종학력)의 카이제곱 검정 결과를 나타내고 있다. 카이제곱 검정에 대한 가설은 다음과 같다.

$H_0$ : 두 변수가 서로 상관성이 없고 독립이다. vs. $H_1$ : 두 변수가 서로 상관성이 있고 독립이 아니다.

Table 4.5의 카이제곱 검정의 결과로 보아 성별을 제외한 연령대, 거주 지역, 소득, 최종학력은 유의확률이 0.0001보다 작기 때문에 자원봉사활동 여부와 독립이 아닌 연관된 변수임을 알 수 있었다.

Table 4.6은 반응변수 (자원봉사활동 여부)와 연속형 독립변수 (우울의 관한 인식, 자아존중감에 대한 인식, 생활실태·만족도)가 통계적인 차이가 있는지 없는지에 대한 $t$ 검정의 결과를 나타내고 있다. 독립

**Table 4.5.** Chi-square tests between $y$ and categorical independent variables

| Variable | df | $\chi^2$ statistic | $p$-value |
|---|---|---|---|
| $X_1$ | 1 | 2.0748 | 0.1498 |
| $X_2$ | 5 | 470.5694 | $< 0.0001$ |
| $X_3$ | 6 | 55.8580 | $< 0.0001$ |
| $X_4$ | 4 | 573.9967 | $< 0.0001$ |
| $X_5$ | 4 | 767.0751 | $< 0.0001$ |

df = degree of freedom.

**Table 4.6.** $t$ test between $y$ and continuous independent variables

| Variable | $t$ test method | Homogeneity of variance test | $t$ statistic |
|---|---|---|---|
| 우울의 관한 인식 ($X_6$) | Satterthwaite | 1.86 ($< 0.0001$) | 11.94 ($< 0.0001$) |
| 자아존중감에 대한 인식 ($X_7$) | Satterthwaite | 1.43 ($< 0.0001$) | 22.97 ($< 0.0001$) |
| 생활실태·만족도 ($X_8$) | Satterthwaite | 1.41 ($< 0.0001$) | 19.69 ($< 0.0001$) |

성 검정에 앞서 등분산 검정을 하였다. 등분산 검정에 대한 가설은 다음과 같다.

$$H_0 : \text{두 집단의 분산이 같다.} \quad \text{vs.} \quad H_1 : \text{두 집단의 분산이 다르다.}$$

세 변수 모두 등분산 검정의 귀무가설을 기각하기 때문에 반응변수와 독립변수 간의 분산과 다른 것을 확인할 수 있다. 그래서 두 집단의 분산이 다를 때 사용하는 Satterthwaite 방법으로 $t$ 검정을 하였다. $t$ 검정의 가설은 다음과 같다.

$$H_0 : \text{두 집단의 평균 차이가 없다.} \quad \text{vs.} \quad H_1 : \text{두 집단의 평균 차이가 있다.}$$

$X_6$, $X_7$, $X_8$ 변수 모두 $t$ 검정의 귀무가설을 기각한다. 각 변수는 반응변수와 평균 차이가 있다고 할 수 있다. 우울의 관한 인식, 자아존중감에 대한 인식, 생활실태·만족도는 봉사활동 여부에 영향을 주는 변수라고 할 수 있다.

기초분석의 목적은 자료의 일반적인 특징분포를 알고자 하는 것이 아닌, 최종모형인 로지스틱 임의효과 회귀모형에 포함되는 설명변수들의 특징을 알고자 하는 것이다. 그러므로 위의 변수들을 사용하여 자원봉사활동 여부의 확률에 대한 모형을 적용한다.

### 4.2. 로지스틱 회귀모형 분석

로지스틱 임의효과 모형을 적합하기 이전에, 먼저 임의효과가 없는 로지스틱 모형에 적합하였다. 로지스틱 모형의 설명변수는 성별($X_1$), 연령대($X_2$), 거주 지역($X_3$), 소득($X_4$), 최종학력($X_5$)과 우울의 관한 인식($X_6$), 자아존중감에 대한 인식($X_7$), 생활실태·만족도($X_8$)로 총 8개의 변수이며, 자원봉사활동 여부($y$)는 종속변수로 사용되었다.

Table 4.7의 결과는 SAS의 PROC LOGISTIC과 R의 glm 함수를 이용하여 로지스틱 모형의 모수 추정치와 통계량, 유의수준을 나타낸 결과이다. 두 모형의 추정 결과는 작은 차이를 보이고 있다. PROC LOGISTIC의 경우에는 로그 우도함수의 일차 미분 방정식을 이용하여 추정값을 구하지만, glm 함수의 경우 편차의 상대적 변화에 기반하여 추정값을 구한다. 즉, 변수의 계수를 추정하는 반복 알고리즘은 각 프로그램마다 다른 수렴 기준을 가지기 때문에 프로그램간의 결과가 차이를 보이는 것이다.

**Table 4.7.** Parameter estimates in logistic regression model

| Variable | SAS (PROC GLIMMIX) | | | | R (lme4) | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimation | SD | $\chi^2$ statistic | $p$-value | Estimation | SD | $z$ statistic | $p$-value |
| intercept | −18.2273 | 218.9000 | 0.0069 | 0.9336 | −18.6182 | 161.4427 | −0.115 | 0.9082 |
| $X_1$_2 | 0.2442 | 0.0706 | 11.9612 | 0.0005 | 0.2443 | 0.0706 | 3.458 | 0.0005 |
| $X_2$_2 | 9.8513 | 218.9000 | 0.0020 | 0.9641 | 10.2421 | 161.4423 | 0.063 | 0.9494 |
| $X_2$_3 | 10.7765 | 218.9000 | 0.0024 | 0.9607 | 11.1673 | 161.4423 | 0.069 | 0.9449 |
| $X_2$_4 | 11.3500 | 218.9000 | 0.0027 | 0.9587 | 11.7409 | 161.4423 | 0.073 | 0.9420 |
| $X_2$_5 | 11.4556 | 218.9000 | 0.0027 | 0.9583 | 11.8464 | 161.4422 | 0.073 | 0.9415 |
| $X_2$_6 | 11.0957 | 218.9000 | 0.0026 | 0.9596 | 11.4866 | 161.4422 | 0.071 | 0.9433 |
| $X_3$_2 | −0.2869 | 0.1065 | 7.2584 | 0.0071 | −0.2869 | 0.1065 | −2.694 | 0.0071 |
| $X_3$_3 | −0.0648 | 0.1120 | 0.3352 | 0.5626 | −0.0648 | 0.1120 | −0.579 | 0.5626 |
| $X_3$_4 | −0.0924 | 0.1307 | 0.4996 | 0.4797 | −0.0924 | 0.1307 | −0.707 | 0.4797 |
| $X_3$_5 | −0.5032 | 0.1495 | 11.3356 | 0.0008 | −0.5032 | 0.1495 | −3.367 | 0.0008 |
| $X_3$_6 | −0.1164 | 0.1482 | 0.6163 | 0.4324 | −0.1164 | 0.1482 | −0.785 | 0.4324 |
| $X_3$_7 | −0.5779 | 0.1229 | 22.1067 | < 0.0001 | −0.5779 | 0.1229 | −4.702 | < 0.0001 |
| $X_4$_2 | 0.3336 | 0.1526 | 4.7761 | 0.0289 | 0.3336 | 0.1526 | 2.185 | 0.0289 |
| $X_4$_3 | 0.4624 | 0.1558 | 8.8133 | 0.0030 | 0.4624 | 0.1558 | 2.969 | 0.0030 |
| $X_4$_4 | 0.5950 | 0.1636 | 13.2313 | 0.0003 | 0.5950 | 0.1636 | 3.637 | 0.0003 |
| $X_4$_5 | 0.9140 | 0.1585 | 33.2637 | < 0.0001 | 0.9140 | 0.1585 | 5.767 | < 0.0001 |
| $X_5$_2 | 1.0226 | 0.1895 | 29.1100 | < 0.0001 | 1.0226 | 0.1895 | 5.395 | < 0.0001 |
| $X_5$_3 | 1.4213 | 0.1727 | 67.7346 | < 0.0001 | 1.4213 | 0.1727 | 8.230 | < 0.0001 |
| $X_5$_4 | 2.0740 | 0.1792 | 133.9845 | < 0.0001 | 2.0740 | 0.1792 | 11.575 | < 0.0001 |
| $X_5$_5 | 2.6726 | 0.2247 | 141.4573 | < 0.0001 | 2.6726 | 0.2247 | 11.894 | < 0.0001 |
| $X_6$ | −0.5380 | 0.1175 | 20.9693 | < 0.0001 | −0.5380 | 0.1175 | −4.579 | < 0.0001 |
| $X_7$ | 0.9984 | 0.1154 | 74.8267 | < 0.0001 | 0.9984 | 0.1154 | 8.650 | < 0.0001 |
| $X_8$ | 0.4906 | 0.0883 | 30.8642 | < 0.0001 | 0.4906 | 0.0883 | 5.556 | < 0.0001 |

SD = standard deviation.

카이제곱 검정의 귀무가설은 '변수가 모형에 효과가 없다', 대립가설은 '변수가 모형에 효과가 있다'이다. 유의수준이 0.05일 때 $p$-값이 0.05보다 큰 변수는 모든 연령대와 부산/경남/울산, 대구/경북, 강원/충북 거주 지역이다. 즉 연령대와 거주 지역 변수는 모형에 영향을 주지 않는 범주가 포함되어 있다고 볼 수 있다. 임의효과가 추가된 모형은 유의하지 않는 변수를 추가하는 경우, 계수에 전체적으로 영향을 주며, 특히 절편 추정량이 크게 증가한다. 또한 절편과 유의하지 않는 변수인 연령대 변수의 표준편차의 값이 과도하게 높아지기 때문에 모형에서 연령대 변수의 사용을 고려해야 한다. 또한 유사가능도 함수, 잔차 유사가능도 함수로 근을 근사하기 위해 뉴튼-라프슨 방법을 사용하는 경우 해가 수렴하지 않으므로 추정값이 추정되지 않는다. 그 이유로 로지스틱 임의효과 모형에서는 연령대의 변수를 제거하였다. 하지만 거주 지역 변수는 일부 범주에서 모형에 영향을 주어 해석에 필요하기 때문에 모형에 포함하였다.

회귀 모형의 추정 계수가 성별이 여자이거나 소득이 높을수록, 그리고 최종학력이 높을수록 증가하므로 자원봉사활동 참여에 긍정적인 영향을 주는 것을 알 수 있다. 또한 자아존중감에 대한 인식이 높고 생활실태 만족도가 높은 응답자들 또한 자원봉사활동의 참여가 더 많았다. 반대로 우울함에 대한 인식이 높은 응답자는 자원봉사활동 참여에 부정적인 영향을 준다. 서울 지역에서 거주하는 응답자의 경우 자원봉사활동을 많이 하는 편이지만 대전/충남과 광주/전남/전북/제주 지역에 거주하는 응답자의 경우 자원봉사활동을 비교적 적게 하는 것을 알 수 있다.

### 4.3. 개체별 임의효과가 포함된 로지스틱 회귀모형 분석

개체별 임의효과가 포함된 로지스틱 회귀모형을 분석하고자 한다. 여기서 지역은 4.1에서 시와 도별로 구분된 거주 지역 변수($X_3$)로 분류되었다. SAS와 R이 동시에 제공하는 우도 근사 방법인 라플라스 근사법과 가우스-에르미트 구적법을 중심으로 분석을 하였고, SAS에서 추가로 분석할 수 있는 유사가능도 우도와 잔차 유사가능도 우도를 비교하였다.

통계 프로그램은 함수의 최대화 방법의 다양한 옵션들을 제공한다. SAS의 PROC GLIMMIX는 뉴튼-라프슨 방법, 준 뉴튼 방법, 능선을 이용한 뉴튼-라프슨 방법, 그리고 쌍대 준 뉴튼 방법 등에 대한 방법들을 사용할 수 있다. 반면 R의 lme4는 가우스-뉴튼 방법으로 최대화하는 알고리즘을 사용하고 있다. 두 프로그램 모두 시작점은 고정효과 모형을 적합하여 얻은 추정값을 사용하는데 특히 lme4는 사용자들이 직접 시작점을 지정하는 것이 가능한 장점이 있다.

먼저 라플라스 근사법을 이용하여 개체별 임의효과가 포함된 로지스틱 회귀모형을 우도 근사한 결과는 Table 4.8의 결과이다. GLIMMIX에서는 뉴튼-라프슨 방법을 대안하는 쌍대 준 뉴튼 방법으로, lme4에서는 뉴튼-라프슨 방법을 이용하여 최대값을 구하였다. 준 뉴튼 방법과 뉴튼-라프슨 방법에 대한 자세한 내용은 SAS Institute Inc (2008)에 있다. 보통 모형을 적합한 경우 프로그램과 패키지가 달라도 고정효과의 계수와 임의효과의 분산에 대한 비슷한 결과를 얻을 수 있다. 하지만 R의 lme4에서는 임의효과 분산(표준편차)이 각 2356.31(406.58), 955.7(30.91)으로 너무 크기 때문에 의미가 있는 값을 얻지 못하였다. 그 이유는 분산의 표집분포에서 시작이 매우 왜곡되었기 때문에 무의미한 표준오차를 만들기 때문이다. 패키지에 대한 자세한 내용은 개발자에 의해서 제시된 Bates와 Maechler (2009)에 있다. 그렇기 때문에 lme4에서 회귀계수가 유의한지에 대한 검정에서 유의확률의 값이 크게 측정되었고 GLIMMIX와 lme4에서 얻은 추정된 계수에 대한 차이를 보였다. 또한 고정효과에 대한 검정의 유의수준이 결측값으로 출력되었다. 개체별 임의효과 모형에서는 관측치 갯수만큼의 임의효과를 추정하여, 변수별 유의수준을 출력을 한다. 약 15,000개의 임의효과를 추정하고 있으며, 이 때 자유도가 크게 증가하게 되면서 고정효과 검정에 대한 유의수준이 파생되지 않는다.

Table 4.8의 결과 GLIMMIX와 lme4에서 최종학력이 높아질수록 추정된 계수가 각 $(0, 1.0903, 1.3821, 2.0178, 2.8951)$, $(0, 3.5242, 3.7165, 3.0798, 9.7835)$으로 점차 높아진다. 그리고 성별 변수의 경우 추정된 계수는 $(0, 0.1829)$, $(0, 0.2557)$이다. 여성이며 최종 학력이 향상할수록 자원봉사활동을 많이 하는 경향이 있다. 반면 GLIMMIX에서는 광주/전남/전북/제주 지역의 계수가 $-0.5555$, lme4에서는 강원/충북 지역의 계수가 $-11.2201$로 가장 낮았다. 광주/전남/전북/제주와 강원/충북 지역이 비교적 봉사활동을 적게 한다고 볼 수 있다. $(X_6, X_7, X_8)$은 추정된 계수가 GLIMMIX에서 $(-0.4484, 0.9726, 0.4576)$, lme4에서 $(-1.8371, 0.135, 0.5183)$으로 우울함이 적을수록, 자아존중감이 높고 생활실태의 만족도가 클수록 자원봉사활동을 더 많이 하였다.

다음으로 가우스-에르미트 구적법을 이용하여 개체별 임의효과가 포함된 로지스틱 회귀모형을 우도 근사 한 결과를 서술한다. 라플라스 근사법과 같이 최적화하는 방법으로 SAS에서는 뉴튼-라프슨 방법을 대안하는 쌍대 준 뉴튼 방법, R에서는 뉴튼-라프슨 방법을 사용하였다. 모형에서 사용될 구적 점의 개수는 10개로 설정하였다. 프로그램 시행 결과, 가우스-에르미트 구적법을 사용한 뒤 두 프로그램에 대한 추정 계수의 차이가 라플라스 근사법보다 줄었다. 그리고 임의효과의 분산(표준편차)은 GLIMMIX와 lme4에서 각 6.0882(2.6123), 0.3763(0.6135)으로 추정되어 상당히 높았던 라플라스 근사법보다 의미가 있는 분산을 얻을 수 있다. 또한 R의 lme4에서 $t$ 검정에 대한 유의확률이 대부분 낮아졌다.

성별의 추정된 계수는 GLIMMIX와 lme4에서 $(0, 0.3084)$, $(0, 0.1925)$이며, 최종학력은 $(0, 1.4256, 1.8184, 2.7986, 4.166)$, $(0, 1.0248, 1.3105, 1.9377, 2.7670)$이다. 성별이 여자인 응답자와 대학원 이상 응

**Table 4.8.** Parameter estimation applied to Laplace approximation and Gauss-Hermite quadrature by subjects in logistic random effect model

| Variable | SAS (PROC GLIMMIX) | | | | R (lme4) | | | |
|---|---|---|---|---|---|---|---|---|
| | Laplace approximation | | Gauss-Hermite quadrature | | Laplace approximation | | Gauss-Hermite quadrature | |
| | Estimation (SD) | $t$ statistic ($p$-value) | Estimation (SD) | $t$ statistic ($p$-value) | Estimation (SD) | $t$ statistic ($p$-value) | Estimation (SD) | $t$ statistic ($p$-value) |
| Intercept | −17.7473 (2.3708) | −7.49 (< 0.0001) | −11.1186 (1.6297) | −6.82 (< 0.0001) | −12.5639 (11.6492) | −1.079 (0.2808) | −7.1474 (1.05708) | −6.761 (< 0.0001) |
| $X_1$_2 | 0.1829 (0.3598) | 0.51 (·) | 0.3084 (0.1190) | 2.59 (·) | 0.2557 (1.8621) | 0.137 (0.8908) | 0.1925 (0.08016) | 2.402 (0.0163) |
| $X_3$_2 | −0.2368 (0.5414) | −0.44 (·) | −0.4044 (0.1806) | −2.24 (·) | −2.0285 (2.6117) | −0.777 (0.4373) | −0.2717 (0.11959) | −2.272 (0.0231) |
| $X_3$_3 | −0.04169 (0.5675) | −0.07 (·) | −0.1352 (0.1833) | −0.74 (·) | −6.1012 (2.8954) | −2.107 (0.0351) | −0.0853 (0.11832) | −0.721 (0.4708) |
| $X_3$_4 | −0.0429 (0.6711) | −0.06 (·) | −0.1324 (0.2098) | −0.63 (·) | −6.1304 (3.5107) | −1.746 (0.0808) | −0.0814 (0.13719) | −0.594 (0.5528) |
| $X_3$_5 | −0.4996 (0.7697) | −0.65 (·) | −0.7613 (0.2574) | −2.96 (·) | −8.1378 (4.0162) | −2.026 (0.0427) | −0.5109 (0.17143) | −2.980 (0.0029) |
| $X_3$_6 | −0.04503 (0.7543) | −0.06 (·) | −0.1576 (0.2390) | −0.66 (·) | −11.2201 (3.9343) | −2.852 (0.0044) | −0.1087 (0.15612) | −0.696 (0.4862) |
| $X_3$_7 | −0.5555 (0.6344) | −0.88 (·) | −0.8795 (0.2294) | −3.83 (·) | −9.8223 (3.3057) | −2.971 (0.0030) | −0.5831 (0.15575) | −3.744 (0.0002) |
| $X_4$_2 | 0.3235 (0.8320) | 0.39 (·) | 0.4357 (0.2222) | 1.96 (·) | 3.5730 (4.4562) | 0.802 (0.4227) | 0.3086 (0.15444) | 1.998 (0.0457) |
| $X_4$_3 | 0.4699 (0.8316) | 0.57 (·) | 0.6838 (0.2358) | 2.90 (·) | 3.6815 (4.4236) | 0.832 (0.4053) | 0.4761 (0.15901) | 2.994 (0.0028) |
| $X_4$_4 | 0.6453 (0.8589) | 0.75 (·) | 0.9338 (0.2621) | 3.56 (·) | −0.3951 (4.6429) | −0.085 (0.9322) | 0.6473 (0.17210) | 3.761 (0.0002) |
| $X_4$_5 | 1.0015 (0.8263) | 1.21 (·) | 1.5122 (0.3043) | 4.97 (·) | 5.7941 (4.3632) | 1.328 (0.1842) | 0.9945 (0.20028) | 4.965 (< 0.0001) |
| $X_5$_2 | 1.0903 (1.0503) | 1.04 (·) | 1.4256 (0.3042) | 4.69 (·) | 3.5242 (5.6073) | 0.628 (0.5297) | 1.0248 (0.19517) | 5.251 (< 0.0001) |
| $X_5$_3 | 1.3821 (0.9227) | 1.50 (·) | 1.8184 (0.3034) | 5.99 (·) | 3.7165 (4.9233) | 0.755 (0.4503) | 1.3105 (0.17862) | 7.337 (< 0.0001) |
| $X_5$_4 | 2.0178 (0.9256) | 2.18 (·) | 2.7986 (0.4149) | 6.75 (·) | 3.0798 (4.9685) | 0.620 (0.5353) | 1.9377 (0.23584) | 8.216 (< 0.0001) |
| $X_5$_5 | 2.8951 (1.1334) | 2.55 (·) | 4.1660 (0.6426) | 6.48 (·) | 9.7835 (5.6366) | 1.736 (0.0826) | 2.7670 (0.4193) | 6.599 (< 0.0001) |
| $X_6$ | −0.4484 (0.5938) | −0.76 (·) | −0.7056 (0.2043) | −3.45 (·) | −1.8371 (3.0568) | −0.601 (0.5478) | −0.4609 (0.14109) | −3.248 (0.0012) |
| $X_7$ | 0.9726 (0.5848) | 1.66 (·) | 1.4466 (0.2616) | 5.53 (·) | 0.1350 (2.9797) | 0.045 (0.9639) | 0.9442 (0.17261) | 5.470 (< 0.0001) |
| $X_8$ | 0.4576 (0.4442) | 1.03 (·) | 0.7115 (0.1699) | 4.19 (·) | 0.5183 (2.4035) | 0.216 (0.8293) | 0.4462 (0.11265) | 3.961 (< 0.0001) |

SD = standard deviation.

**Table 4.9.** Parameter estimation applied to pseudo-likelihood and residuals pseudo-likelihood by subjects in logistic random effect model

| Variable | SAS (PROC GLIMMIX) | | | | R (lme4) | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimation | SD | $t$ statistics | $p$-value | Estimation | SD | $t$ statistics | $p$-value |
| Intercept | −7.0033 | 0.4434 | −15.80 | $< 0.0001$ | −7.0032 | 0.4437 | −15.78 | $< 0.0001$ |
| $X_1$_2 | 0.1854 | 0.06919 | 2.68 | · | 0.1855 | 0.06928 | 2.68 | · |
| $X_3$_2 | −0.2423 | 0.1050 | −2.31 | · | −0.2423 | 0.1052 | −2.30 | · |
| $X_3$_3 | −0.0646 | 0.1107 | −0.58 | · | −0.0647 | 0.1108 | −0.58 | · |
| $X_3$_4 | −0.0579 | 0.1291 | −0.45 | · | −0.0580 | 0.1293 | −0.45 | · |
| $X_3$_5 | −0.4740 | 0.1475 | −3.21 | · | −0.4739 | 0.1477 | −3.21 | · |
| $X_3$_6 | −0.0802 | 0.1462 | −0.55 | · | −0.0803 | 0.1464 | −0.55 | · |
| $X_3$_7 | −0.5391 | 0.1214 | −4.44 | · | −0.5392 | 0.1215 | −4.44 | · |
| $X_4$_2 | 0.3113 | 0.1514 | 2.06 | · | 0.3112 | 0.1515 | 2.05 | · |
| $X_4$_3 | 0.4708 | 0.1521 | 3.09 | · | 0.4708 | 0.1522 | 3.09 | · |
| $X_4$_4 | 0.6287 | 0.1583 | 3.97 | · | 0.6288 | 0.1584 | 3.97 | · |
| $X_4$_5 | 0.9578 | 0.1525 | 6.28 | · | 0.9579 | 0.1526 | 6.28 | · |
| $X_5$_2 | 1.0681 | 0.1883 | 5.67 | · | 1.0681 | 0.1884 | 5.67 | · |
| $X_5$_3 | 1.3408 | 0.1649 | 8.13 | · | 1.3408 | 0.1650 | 8.13 | · |
| $X_5$_4 | 1.9345 | 0.1664 | 11.62 | · | 1.9345 | 0.1665 | 11.62 | · |
| $X_5$_5 | 2.6978 | 0.2178 | 12.39 | · | 2.6978 | 0.2180 | 12.37 | · |
| $X_6$ | −0.4345 | 0.1150 | −3.78 | · | −0.4347 | 0.1151 | −3.78 | · |
| $X_7$ | 0.9181 | 0.1124 | 8.17 | · | 0.9181 | 0.1125 | 8.16 | · |
| $X_8$ | 0.4337 | 0.08645 | 5.02 | · | 0.4338 | 0.08656 | 5.01 | · |

* SD = standard deviation.

답자가 자원봉사활동에 적극적인 것을 확인할 수 있다. 소득 또한 $(0, 0.4357, 0.6838, 0.9338, 1.5122)$, $(0, 0.3086, 0.4761, 0.6473, 0.9945)$으로 소득이 증가함에 따라 자원봉사활동을 많이 한다. 반면 광주/전남/전북/제주 지역의 계수가 각 −0.8795, −0.5831으로 자원봉사활동에 비교적 부정적인 영향을 받는 지역이다. $(X_6, X_7, X_8)$은 추정된 계수가 GLIMMIX에서 $(-0.7056, 1.4466, 0.7115)$, lme4에서 $(-0.4609, 0.9442, 0.4462)$으로 우울함이 적고 자아존중감이 높을수록, 생활실태의 만족도가 큰 경우 자원봉사활동에 긍정적이다.

Table 4.8에서 lme4 패키지를 이용한 가우스-에르미트 구적법 추정치의 표준편차가 SAS에 비해서 상대적으로 크다. 그 이유로 SAS의 방법이 안정적이며, 더 좋은 방법이라고 여겨진다.

SAS에서는 유도 가능도 우도와 잔차 유도 가능도 우도의 기능을 추가로 제공하고 있다. 유도 가능도 우도의 최적화는 데이터의 초기 설정으로 시작되는 반면, 잔차 유도 가능도 우도는 공분산 모수만이 최적화하는데 포함된다는 차이점이 있다. 잔차 유도 가능도 우도는 분산에 대한 해가 양수인 것이 제약 조건이며, 다른 방법들에 비해 제약 조건의 경계가 낮은 특징이 있다. 최적화 방법은 유사 가능도 우도의 두 방법 모두 능선을 이용한 뉴튼-라프슨 방법을 적용하였다. 능선을 이용한 뉴튼-라프슨 방법은 때에 따라 능선을 사용하여 약간 편차를 증가시키면서 더 큰 분산을 감소시키기 때문에 모형의 최소제곱오차를 더 작게 만들 수 있음을 활용한 방법이다. 능선을 이용한 뉴튼-라프슨 방법은 뉴튼-라프슨 방법에 비해 $\mathbf{X}'\mathbf{X}$이 0에 수렴하여 역행렬이 존재하지 않는 경우에도 사용할 수 있기 때문에 훨씬 유연하다.

유도 가능도 우도와 잔차 유도 가능도 우도를 이용하여 모수를 추정한 결과는 Table 4.9에 나타냈다. 두 방법을 사용하여 추정된 고정효과의 계수가 대부분 같다. 임의효과의 분산(표준편차)도 유도 가능도 우도가 0.01708(0.1360), 잔차 유도 가능도 우도가 0.03687(0.1356)으로 크게 차이가 없는 것으

로 나타났다. 여자인 경우, 소득이 증가할수록, 최종학력이 높을수록 추정된 고정효과의 계수가 커지기 때문에 자원봉사활동에 긍정적인 영향을 주는 것을 알 수 있다. 변수 $(X_6, X_7, X_8)$의 추정치는 유사가능도 함수를 이용한 방법에서 $(-0.4345, 0.9181, 0.4337)$, 잔차 유사가능도 함수를 이용한 경우는 $(-0.4347, 0.9181, 0.4338)$으로 우울함이 적을수록, 자아존중감이 높고 생활실태의 만족도가 클수록 자원봉사활동을 더 많이 함을 나타냈다.

### 4.4. 지역별 임의효과가 포함된 로지스틱 회귀모형 분석

다음은 지역별 임의효과가 포함된 로지스틱 회귀모형을 분석하고자 한다. 여기서 지역은 4.1에서 시·도별로 구분된 거주 지역 변수$(X_3)$로 분류되었다. 개체별 임의효과가 포함된 로지스틱 회귀모형과 같이 우도 근사 방법인 라플라스 근사법과 가우스-에르미트 구적법을 중심으로 분석을 하였다. 또한 GLIMMIX에서 추가로 분석할 수 있는 유사가능도 우도와 잔차 유사가능도 우도를 비교했다. 지역별 로지스틱 임의효과 모형은 개체별 로지스틱 임의효과 모형과 비교하였을 때, 프로그램의 실행 속도가 더 빨랐다. 이유는 지역별 임의효과 행렬이 개체별 임의효과보다 더 작고 간단한 행렬을 가지고 있어서 실행이 빨라지기 때문이다.

프로그램 시행 결과는 Table 4.10에 있으며, GLIMMIX와 lme4 결과가 대부분 비슷하였다. 개체별 임의효과가 포함되었을 때와 다르게 두 모형이 비슷한 이유는 집단의 수가 적어지면서 추정된 임의효과의 분산이 안정화되었기 때문이다.

GLIMMIX에서 라플라스 근사법을 적용한 결과 각 지역의 임의효과 추정치(표준편차)가 (0.1713 (0.0974), $-0.0312$(0.0922), 0.1165(0.097), 0.1088(0.1077), $-0.1783$(0.1183), 0.0855(0.1147), $-0.2601$ (0.1065))을 얻었다. 또한 가우스-에르미트 구적법을 적용한 결과는 (0.1714(0.0974), $-0.0312$(0.0923), 0.1165(0.097), 0.1089(0.1077), $-0.1783$(0.1183), 0.0855(0.1147), $-0.2601$(0.1065))로 추정되었다. lme4에서는 지역별 임의효과에 대해서 지역 전체적인 분산이 추정된다. 임의효과 추정치는 라플라스 근사법이 0.03022(0.1738), 가우스-에르미트 구적법이 0.03024(0.1739)이다. SAS와 R에서 임의효과의 영향이 비슷함을 알 수 있다.

GLIMMIX와 lme4에서 변수의 추정된 계수는 성별이 $(0, 0.1866)$, $(0, 0.1867)$이며, 최종학력은 $(0, 1.0691, 1.3387, 1.9358, 2.6943)$, $(0, 1.0715, 1.3412, 1.9383, 2.6978)$이다. 성별이 여자이며, 최종학력이 높을수록 자원봉사활동을 많이 하는 경향을 보인다. 그리고 소득의 계수가 (0, 0.3132, 0.4748, 0.6306, 0.9643), $(0, 0.3115, 0.4725, 0.6285, 0.9622)$으로 점차 높아진다. 이는 소득이 높을수록 자원봉사활동에 긍정적이라고 할 수 있다. $(X_6, X_7, X_8)$은 GLIMMIX에서 $(-0.4319, 0.9208, 0.419)$, lme4에서 $(-0.4317, 0.9207, 0.4191)$으로 우울함이 적을수록, 자아존중감과 생활실태의 만족도가 클수록 자원봉사활동을 더 많이 하였다.

다음으로 가우스-에르미트 구적법을 이용하여 구적 점을 10개로 설정한 후 지역별 임의효과가 포함된 로지스틱 회귀모형을 우도 근사하였다. 라플라스 근사법과 같이 GLIMMIX에서는 쌍대 준 뉴튼 방법으로, lme4에서는 뉴튼-라프슨 방법을 사용하였다. 가우스-에르미트 구적법 또한 두 패키지에서 추정된 계수 값은 거의 비슷하다.

성별의 추정 계수는 GLIMMIX와 lme4에서 $(0, 0.1866)$, $(0, 0.1867)$이며 최종학력은 $(0, 1.0695, 1.339, 1.936, 2.6951)$, $(0, 1.0714, 1.3411, 1.9383, 2.6977)$으로 나타났다. 여성이며 대학원 이상인 응답자가 가장 자원봉사활동에 긍정적인 영향을 주는 것을 알 수 있다. 소득의 추정계수는 $(0, 0.3135, 0.4742, 0.6306, 0.9645)$, $(0, 0.3116, 0.4726, 0.6286, 0.9624)$이다. 이는 소득이 커질수록 자원봉사활동에 비교적 참여를 많이 함을 뜻한다. $(X_6, X_7, X_8)$은 GLIMMIX에서 $(-0.4317, 0.9207, 0.4190)$, lme4에서

**Table 4.10.** Parameter estimation applied to Laplace approximation and Gauss-Hermite quadrature by individual in logistic random effect model

| Variable | SAS (PROC GLIMMIX) | | | | R (lme4) | | | |
|---|---|---|---|---|---|---|---|---|
| | Laplace approximation | | Gauss-Hermite quadrature | | Laplace approximation | | Gauss-Hermite quadrature | |
| | Estimation (SD) | $t$ statistic ($p$-value) | Estimation (SD) | $t$ statistic ($p$-value) | Estimation (SD) | $t$ statistic ($p$-value) | Estimation (SD) | $t$ statistic ($p$-value) |
| Intercept | −7.1773 (0.4452) | −16.12 (< 0.0001) | −7.1786 (0.4452) | −16.12 (< 0.0001) | −7.1791 (0.4450) | −16.133 (< 0.0001) | −7.1789 (0.4452) | −16.125 (< 0.0001) |
| $X_1$_2 | 0.1866 (0.0691) | 2.70 (0.0069) | 0.1866 (0.0691) | 2.70 (0.0069) | 0.1867 (0.0690) | 2.704 (0.0068) | 0.1867 (0.0691) | 2.703 (0.0069) |
| $X_4$_2 | 0.3132 (0.1513) | 2.07 (0.0385) | 0.3135 (0.1513) | 2.07 (0.0383) | 0.3115 (0.1512) | 2.061 (0.0394) | 0.3116 (0.1513) | 2.060 (0.0394) |
| $X_4$_3 | 0.4748 (0.1520) | 3.12 (0.0018) | 0.4742 (0.1520) | 3.12 (0.0018) | 0.4725 (0.1519) | 3.110 (0.0019) | 0.4726 (0.1520) | 3.109 (0.0019) |
| $X_4$_4 | 0.6306 (0.1581) | 3.99 (< 0.0001) | 0.6306 (0.1581) | 3.99 (< 0.0001) | 0.6285 (0.1579) | 3.979 (< 0.0001) | 0.6286 (0.1580) | 3.977 (< 0.0001) |
| $X_4$_5 | 0.9643 (0.1524) | 6.33 (< 0.0001) | 0.9645 (0.1524) | 6.33 (< 0.0001) | 0.9622 (0.1522) | 6.321 (< 0.0001) | 0.9624 (0.1523) | 6.318 (< 0.0001) |
| $X_5$_2 | 1.0691 (0.1881) | 5.68 (< 0.0001) | 1.0695 (0.1882) | 5.68 (< 0.0001) | 1.0715 (0.1881) | 5.698 (< 0.0001) | 1.0714 (0.1882) | 5.693 (< 0.0001) |
| $X_5$_3 | 1.3387 (0.1647) | 8.13 (< 0.0001) | 1.3390 (0.1647) | 8.13 (< 0.0001) | 1.3412 (0.1647) | 8.145 (< 0.0001) | 1.3411 (0.1648) | 8.138 (< 0.0001) |
| $X_5$_4 | 1.9358 (0.1662) | 11.65 (< 0.0001) | 1.9360 (0.1662) | 11.65 (< 0.0001) | 1.9383 (0.1662) | 11.664 (< 0.0001) | 1.9383 (0.1663) | 11.656 (< 0.0001) |
| $X_5$_5 | 2.6943 (0.2173) | 12.40 (< 0.0001) | 2.6951 (0.2173) | 12.41 (< 0.0001) | 2.6978 (0.2172) | 12.421 (< 0.0001) | 2.6977 (0.2173) | 12.414 (< 0.0001) |
| $X_6$ | −0.4319 (0.1148) | −3.76 (0.0002) | −0.4317 (0.1148) | −3.76 (0.0002) | −0.4317 (0.1148) | −3.762 (0.0002) | −0.4317 (0.1148) | −3.760 (0.0002) |
| $X_7$ | 0.9208 (0.1121) | 8.22 (< 0.0001) | 0.9207 (0.1121) | 8.21 (< 0.0001) | 0.9207 (0.1120) | 8.219 (< 0.0001) | 0.9207 (0.1121) | 8.214 (< 0.0001) |
| $X_8$ | 0.4190 (0.0861) | 4.87 (< 0.0001) | 0.4190 (0.0861) | 4.87 (< 0.0001) | 0.4191 (0.0860) | 4.872 (< 0.0001) | 0.4191 (0.0861) | 4.868 (< 0.0001) |

SD = standard deviation.

$(-0.4317, 0.9207, 0.4191)$으로 추정된다. 자아존중감이 높고 생활실태에 만족할수록 자원봉사활동을 활발하게 하며, 우울함이 높을수록 자원봉사활동을 비교적 선호하지 않는다고 해석할 수 있다.

지역별 임의효과가 포함된 로지스틱 회귀모형의 유도 가능도 우도와 잔차 유도 가능도 우도의 결과는 Table 4.11과 같으며, 고정효과의 추정된 계수가 대부분 같은 것을 확인할 수 있다. 임의효과의 분산은 유사가능도 우도 방법이 $(0.1693, -0.0332, 0.1145, 0.1069, -0.1796, 0.0839, -0.2618)$, 잔차 유도 가능도 우도 방법이 $(0.1753, -0.0336, 0.1189, 0.113, -0.191, 0.0896, -0.2722)$로 나타났다.

성별 변수의 계수는 유도 가능도 우도, 잔차 유도 가능도 우도 방법 각 $(0, 0.1866)$, $(0, 0.1864)$로 추정되었다. 그리고 소득의 추정 계수는 $(0, 0.3115, 0.4724, 0.6283, 0.9617)$, $(0, 0.3114, 0.472, 0.6282, 0.961)$이다. 성별이 여자이며 소득이 높을수록 자원봉사활동에 긍정적인 영향을 주는 것을 볼 수 있다. 또한 최종학력 변수의 계수도 $(0, 1.0712, 1.3408, 1.9374, 2.696)$, $(0, 1.0706, 1.3407, 1.9369, 2.696)$이다. 최종

**Table 4.11.** Parameter estimation applied to pseudo-likelihood and Residuals pseudo-likelihood by individual in logistic random effect model

| Variable | SAS (PROC GLIMMIX) | | | | R (lme4) | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimation | SD | $t$ statistic | $p$-value | Estimation | SD | $t$ statistic | $p$-value |
| Intercept | −7.1738 | 0.4446 | −16.14 | $< 0.0001$ | −7.1793 | 0.4458 | −16.10 | $< 0.0001$ |
| $X_1$_2 | 0.1866 | 0.06905 | 2.70 | 0.0069 | 0.1864 | 0.06906 | 2.70 | 0.0070 |
| $X_4$_2 | 0.3115 | 0.1513 | 2.06 | 0.0395 | 0.3114 | 0.1513 | 2.06 | 0.0395 |
| $X_4$_3 | 0.4724 | 0.1520 | 3.11 | 0.0019 | 0.4720 | 0.1520 | 3.11 | 0.0019 |
| $X_4$_4 | 0.6283 | 0.1580 | 3.98 | $< 0.0001$ | 0.6282 | 0.1580 | 3.98 | $< 0.0001$ |
| $X_4$_5 | 0.9617 | 0.1523 | 6.32 | $< 0.0001$ | 0.9610 | 0.1523 | 6.31 | $< 0.0001$ |
| $X_5$_2 | 1.0712 | 0.1882 | 5.69 | $< 0.0001$ | 1.0706 | 0.1882 | 5.69 | $< 0.0001$ |
| $X_5$_3 | 1.3408 | 0.1648 | 8.14 | $< 0.0001$ | 1.3407 | 0.1648 | 8.14 | $< 0.0001$ |
| $X_5$_4 | 1.9374 | 0.1663 | 11.65 | $< 0.0001$ | 1.9369 | 0.1663 | 11.65 | $< 0.0001$ |
| $X_5$_5 | 2.6960 | 0.2172 | 12.41 | $< 0.0001$ | 2.6960 | 0.2173 | 12.41 | $< 0.0001$ |
| $X_6$ | −0.4314 | 0.1148 | −3.76 | 0.0002 | −0.4317 | 0.1148 | −3.76 | 0.0002 |
| $X_7$ | 0.9201 | 0.112 | 8.21 | $< 0.0001$ | 0.9198 | 0.1121 | 8.21 | $< 0.0001$ |
| $X_8$ | 0.4188 | 0.0857 | 4.89 | $< 0.0001$ | 0.4210 | 0.08579 | 4.91 | $< 0.0001$ |
| Random effects | Estimation | SD | $t$ statistic | $p$-value | Estimation | SD | $t$ statistic | $p$-value |
| $X_3$_1 | 0.1693 | 0.0947 | 1.79 | 0.0738 | 0.1753 | 0.1006 | 1.74 | 0.0813 |
| $X_3$_2 | −0.0332 | 0.0921 | −0.36 | 0.7187 | −0.0336 | 0.0980 | −0.34 | 0.7320 |
| $X_3$_3 | 0.1145 | 0.0955 | 1.20 | 0.2306 | 0.1189 | 0.1013 | 1.17 | 0.2408 |
| $X_3$_4 | 0.1069 | 0.1052 | 1.02 | 0.3095 | 0.1130 | 0.1113 | 1.02 | 0.3101 |
| $X_3$_5 | −0.1796 | 0.1113 | −1.61 | 0.1067 | −0.1910 | 0.1180 | −1.62 | 0.1056 |
| $X_3$_6 | 0.0839 | 0.1126 | 0.74 | 0.4567 | 0.0896 | 0.1191 | 0.75 | 0.4520 |
| $X_3$_7 | −0.2618 | 0.0996 | −2.63 | 0.0086 | −0.2722 | 0.1057 | −2.58 | 0.0100 |

SD = standard deviation.

학력이 높을수록 자원봉사활동에 긍정적인 영향을 받는 것으로 해석된다. 변수 $(X_6, X_7, X_8)$는 유사가능도 함수를 이용한 방법에서 $(-0.4314, 0.9201, 0.4188)$, 잔차 유사가능도 함수를 이용한 경우는 $(-0.4317, 0.9198, 0.421)$으로 추정되었다. 개체별 임의효과의 결과와 같이 우울함이 적을수록, 자아존중감이 높고 생활실태의 만족도가 클수록 자원봉사활동을 더 많이 하는 것으로 나타났다.

## 5. 결론

일반화 선형 혼합 모형에서는 알지 못하거나 숨겨져 있는 분산의 구조를 임의효과를 통해 구체화할 수 있다. 개체별 분산을 고려하는 방법 외에도 개체 안에서의 상관성을 고려하는 경우에도 사용이 용이하여 관측값을 반복측정하는 경우에도 임의효과 모형을 고려한다. 그 이유로 사회학, 보건학 등 여러 분야에서 활발하게 쓰인다. 본 논문의 분석에 주로 쓰인 로지스틱 임의효과 모형은 이분형의 결과를 가진 일반화 선형 혼합 모형에서도 특별한 경우이다. 데이터의 특성을 고려하여 일반화 선형 혼합 모형의 선택과 임의효과의 존재 여부를 정해야 하며, 임의효과를 포함한 적절한 모형과 모수의 사용은 분석의 정확성을 높일 수 있다.

모수의 추정치와 분산을 구하기 위한 여러 가지 최대우도 추정법이 존재한다. 하지만 모형이 선형이 아닌 경우 우도 방정식이 복잡하여 계산하기 힘들다. 이 경우는 실제 우도와 근사한 우도 함수가 필요하다. 근사 우도 함수는 뉴튼-라프슨 방법을 이용하여 최대우도추정량을 구할 수 있다. 근사 우도를 얻는 방법으로 라플라스 근사법, 가우스-에르미트 구적법, 적응 가우스-에르미트 구적법, 유사 가능도 함수

등이 있으나 제공되는 근사 우도 방법이 통계 패키지마다 차이가 있다.

자원봉사활동의 미치는 영향을 알기 위해서 로지스틱 임의효과 모형으로 다양한 최대우도 추정 방법을 고려하여 분석을 진행하였다. 통계 패키지는 SAS의 PROC GLIMMIX와 R의 lme4를 사용하였다. GLIMMIX와 lme4는 라플라스 근사법, 가우스-에르미트 구적법을 동시에 제공하고 있고, 추가로 GLIMMIX는 유사가능도 함수, 잔차 유사가능도 함수가 제공된다. 자원봉사활동 여부를 반응변수로 지정하였고, 8개의 변수에 대해서 고정효과의 모수 추정치, 임의효과의 분산 추정치를 알아보았다.

프로그램 실행 결과 개체별 임의효과가 적용된 로지스틱 회귀모형인 경우 R에서 lme4의 임의효과 분산 추정 결과가 분산의 왜곡으로 인해 높은 값을 보였다. 개체별 임의효과를 포함한 로지스틱 회귀모형의 경우 라플라스 근사법보다 가우스-에르미트 구적법에서 얻은 추정치의 표준편차가 더 작아서 안정적이다. 그 이유로 개체별 임의효과를 적용하는 경우 가우스-에르미트를 사용하는 것이 좋다. 그리고 프로그램 실행 속도는 지역별 임의효과가 포함된 로지스틱 회귀모형인 경우 개체별 로지스틱 임의효과 모형보다 월등하게 빨랐다. 모수를 추정한 결과 지역별 임의효과 로지스틱 모형은 SAS와 R의 결과가 크게 변하지 않은 것으로 보아 패키지에 크게 영향을 받지 않는다는 것을 알아냈다. 또한 최대우도를 추정하기 위한 근사 방법들도 모수 추정에 크게 영향을 미치지 않는 것을 확인할 수 있었다. 두 모형의 고정효과를 추정한 결과 여자이며 소득이 높을수록, 최종학력이 높을수록, 자아존중감과 생활실태에 대한 만족도가 큰 응답자가 자원봉사활동을 더 많이 참여하였다. 반면 우울함이 큰 응답자는 자원봉사활동을 비교적 적게 참여하는 경향을 보였다.

지역별 임의효과는 추정치가 크지 않지만 어떠한 최대우도 추정법을 적용해도 추정량의 분산을 안정화시키는 것을 볼 수 있었다. 지역별 효과를 주효과로 가정하고 로지스틱 선형 모형을 적용했을 때에도 유의한 차이가 나는 지역들이 존재했기 때문에 지역별 임의효과 모형에서 추정량의 분산 안정화를 발견할 수 있었다. 그러므로 지역별 자원봉사 여부의 차이에 큰 관심이 없다면, 개체별 다른 분산을 가정하는 개체별 임의 효과 모형과는 달리 지역별로 같은 임의 효과를 공유하여 지역별 다른 분산을 가정하는 지역별 임의 효과 모형이 안정적인 모형이라 할 수 있다. 그러나 지역별 효과도 자원봉사 여부의 차이에 영향을 미치는지에 관심이 있다면 개체별 임의효과 모형보다는 단순 로지스틱 선형 모형을 사용하는 것이 더 좋다고 할 수 있다.

시대가 발전하면서 데이터를 얻는 플랫폼이 다양해짐에 따라 데이터의 형태도 복잡하고 다양해진다. 복잡한 데이터에 적절한 모형을 적용하기 위해서는 임의효과를 효과적으로 사용하는 것이 중요하다. 그 때문에 변량 효과를 포함하고 있는 혼합모형의 활발한 연구가 요구될 것이다.

## References


Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions*, Dover, New York.

Agresti, A. (1990). *Categorical Data Analysis* (1st ed), Wiley, New York.

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed), Wiley, New York.

Bates, D. and Maechler, M. (2009). Package 'lme4': linear mixed-effects models using S4 classes (Version 0.999375-32), http://cran.r-project.org/web/packages/lme4/lme4.pdf

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer, New York.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution, *Trends in Ecology & Evolution*, **24**, 127–135.

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

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component

of dispersion, *Biometrika*, **82**, 81–91.

Buonaccorsi, J. P. (1996). Measurement error in the response in the general linear model, *Journal of the American Statistical Association*, **91**, 633–642.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models* (2nd ed), Springer, New York.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*, John Wiley & Sons, New York.

Kim, Y., Choi, Y. K., and Emery, S. (2013). Logistic regression with multiple random effects: a simulation study of estimation methods and statistical packages, *The American Statistician*, **67**, 171–182.

Laplace, P. S. (1986). Memoir on the probability of the causes of events, *Statistical Science*, **1**, 364–378.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: an example, *Journal of the Royal Statistical Society, Series C*, **50**, 325–335.

Li, B., Lingsma, H. F., Steyerberg, E. W., and Lesaffre, E. (2011). Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes, *BMC Medical Research Methodology*, **11**, 77.

Liu, Q. and Pierce, D. A. (1994). A note on Gauss-Hermite quadrature, *Biometrika*, **81**, 624–629.

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

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*, Wiley, New York.

Nelder, J. A. and Wedderburn, R. W. M. (1972). General linearized models, *Journal of the Royal Statistical Society. Series A*, **135**, 370–384.

Pearl, R. and Reed, L. J. (1920). On the rate of growth of the population of the United States since 1790 and its mathematical representation. In *Proceedings of the National Academy of Sciences*, **6**, 275–288.

Pearl, R., Reed, L. J., and Kish, J. F. (1940). The logistic curve and the census count of 1940, *Science*, **92**, 486–488.

SAS Institute (2008). *Sas/Stat 9.2 User's Guide: The Glimmix Procedure*, SAS Pub.

SAS Institute (2009). *Sas/Stat 9.2 User's Guide: The Glimmix Procedure*, SAS Pub.

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

Schultz, H. (1930). The standard error of a forecast from a curve, *Journal of the American Statistical Association*, **25**, 139–185.

Solomon, P. J. and Cox, D. R. (1992). Nonlinear component of variance models, *Biometrika*, **79**, 1–11.

StataCorp, L. P. (2013). *Stata Multilevel Mixed-Effects Reference Manual*, StataCorp LP, Texas.

Verhulst, P. F. (1838). Notice sur la loi que la population suit dans son accroissement, *correspondance Mathematique et Physique Publiee Par a. Quetelet*, **10**, 113–121.

Verhulst, P. F. (1845). Recherches Mathйmatiques sur La Loi D'Accroissement de la Population, *Nouveaux Mйmoires de l'Acadйmie Royale des Sciences et Belles-Lettres de Bruxelles*, **18**, 1–45.

Wang, N., Lin, X., Gutierrez, R. G., and Carroll, R. J. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models, *Journal of the American Statistical Association*, **93**, 249–261.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models, *Biometrika*, **80**, 791–795.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach, *Journal of statistical Computation and Simulation*, **48**, 233–243.

# 로지스틱 임의선형 혼합모형의 최대우도 추정법

김민아[a] · 경민정[a,1]

[a]덕성여자대학교 정보통계학과

(2017년 10월 10일 접수, 2017년 11월 29일 수정, 2017년 11월 30일 채택)

**요 약**

관측되지 않는 효과 또는 고정효과로 설명할 수 없는 분산 구조가 포함되어 정확한 모수 추정이 어려운 경우 체계적인 분석을 위해 일반화 선형 모형은 임의효과가 포함된 일반화 선형 혼합 모형으로 확장되었다. 본 연구에서는 일반화 선형 모형 중에서도 이분적인 반응변수를 다루는 로지스틱 회귀모형에 임의효과를 포함한 최대 우도 추정 방법을 설명한다. 그중에서도 라플라스 근사법, 가우스-에르미트 구적법, 적응 가우스-에르미트 구적법 그리고 유사가능도 우도에 대한 최대우도 추정법을 자세히 알아본다. 또한 제안한 방법을 사용하여 한국 복지 패널 데이터에서 정신건강과 생활만족도가 자원봉사활동에 미치는 영향에 대해 분석한다.

주요용어: 로지스틱 회귀모형, 일반화 선형 혼합 모형, 임의효과, 최대우도 추정법

이 논문은 덕성여자대학교 교내연구비 3000002744 지원을 받아 수행되었습니다.

[1]교신저자: (01369) 서울시 도봉구 삼양로 144길 33, 덕성여자대학교 정보통계학과.
E-mail: mkyung@duksung.ac.kr