The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.1.089
27(1), 89–102

A Study of Effect on the Smoking Status using Multilevel Logistic Model

Ji Hye Lee[a] · Tae-Young Heo[a,1]

[a]Department of Information and Statistics, Chungbuk National University

(Received November 15, 2013; Revised January 1, 2014; Accepted January 6, 2014)

Abstract

In this study, we analyze the effect on the smoking status in the Seoul Metropolitan area using a multilevel logistic model with Community Health Survey data from the Korea Centers for Disease Control and Prevention. Intraclass correlation coefficient (ICC), profiling analysis and two types of predicted value were used to determine the appropriate multilevel analysis level. Sensitivity, specificity, percentage of correctly classified observations (PCC) and ROC curve evaluated model performance. We showed the applicability for multilevel analysis allowed for the possibility that different factors contribute to within group and between group variability using survey data.

Keywords: Multilevel analysis, multilevel logistic regression model, hierarchical data, community health survey data.

1. 서론

흡연은 담배 연기에 포함된 4,000여종의 화학물질 중 건강에 해로운 250여종 이상의 물질로 인하여 건강에 유해하다는 것은 분명한 사실이다. 흡연자의 조기 사망률은 비흡연자에 비해 매우 높으며 중대 질병의 발병률 또한 매우 높은 것으로 익히 알려져 있다 (WHO, 2013).

2002년 서울 시민의 보건의식 행태 조사 결과 매일 흡연자와 가끔 흡연자를 합한 흡연율이 전체 인구 중 27.68%로 매우 높은 수준으로 나타났다. 또한 여성, 청소년의 흡연율이 지속적으로 증가하며 흡연 시작 연령이 감소하는 추세를 보였다. 이에 서울시는 각종 금연 정책을 추진하여 2010년까지 연령별, 성별 흡연율을 감소시켜 "담배연기 없는 건강한 서울"을 만들고자 하였다 (Seoul Metropolitan Government, 2013).

본 연구에서는 서울시에서 추진하는 각종 금연 정책 효과를 알아보기 위하여 25개 각 행정구간 흡연율의 차이를 다수준 분석을 통하여 알아보고자 하였다. 이는 다수준 분석을 통해 각 행정구간 흡연율의 차이가 존재하는지에 대한 여부를 판단하여 서울시 금연 정책의 지역적 파급효과의 실효성을 확인할 수 있다는 장점이 있다.

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(No. NRF-2012R1A1A1040358).

[1]Corresponding author: Associate professor, Department of Information and Statistics, Chungbuk National University, Cheong-Ju, Chungbuk 361-763, Korea. E-mail: theo@cbnu.ac.kr

본 연구의 목적을 위해 질병관리본부의 2010년도 지역사회 건강조사 자료를 바탕으로 서울시 25개 행정구역에서 관할하고 있는 보건소의 자료를 추출하여 분석하였다. 지역사회 건강조사는 지역 주민의 건강수준, 흡연율, 음주율 등 건강 통계를 매년 생산하고 조사내용 및 수행체계 표준화로 주민의 건강수준을 지역 간 비교 가능케 하는 것을 목표로 하고 있다. 개인의 행태는 단순한 개인의 특성만으로 결정되는 것이 아니라 개인이 속한 지역이나 집단의 특성이 공통적으로 결정에 영향을 미치기된다. 개인의 행태에 영향을 미치는 요인이 개인 수준이 아닌 집단 또는 지역과 같은 상위 수준에 존재할 경우, 일반적인 회귀 분석 방법으로는 정확한 연관성을 파악할 수 없으며 지역 특성의 영향력을 분석하고자 할 경우에는 다수준 분석을 사용하는 것이 타당하다.

다수준 분석은 교육학 둥 사회학적 연구에서 개발되어 활용되었으며 최근에는 보건학, 의학, 역학 자료분석에 많은 연구가 진행되고 있다. 다양한 분야에서의 다수준 분석을 이용한 연구를 살펴보면 다음과 같다.

Kim 등 (2012)은 주사제에 대한 환자의 인식수준을 살펴보고 환자의 주사제 요청에 영향을 미치는 영향을 파악하고자 다수준 분석에서 일반적인 회귀 분석과 로지스틱 회귀분석을 이용하였으며 Lee 등 (2012)은 연구대상 남녀 노인의 문제음주에 영향을 미치는 요인을 확인하기 위해 위계적 방법의 로지스틱 회귀분석을 실시하였다.

Lee 등 (2012)은 위계적 로지스틱 회귀분석을 통해 인구 사회학적 특성, 심리적 특성, 가족 및 사회적 관계 특성이 노인 자살 생각에 미치는 영향력을 살펴보고자 하였다. Jung 등 (2010)은 다수준 분석을 이용하여 의료기관 혹은 의사특성으로 인한 군집 효과를 보정하여 건강보험 환자와 의료급여 환자 간 제공되는 혈액투석 서비스(과정, 결과 측면)의 차이와 의료보장 형태가 혈액투석 서비스 제공에 미치는 영향을 살펴보고자 하였다.

Khan과 Shaw (2011)는 피임을 통한 생식의 제한에 영향을 미치는 요인을 알아보기 위하여 다수준 로지스틱 회귀분석을 실시하였으며 Schabenberger (2005), Dai 등 (2006), Li 등 (2006), Flom 등 (2006)은 SAS의 glimmix 프로시저를 이용한 다수준 로지스틱 회귀분석의 활용성을 입증하였다.

다수준 자료를 분석하는 통계적 방법은 두 가지로 분류해 볼 수 있다. 첫 번째는 개인 수준의 자료를 합산하여서 상위 수준 단위의 구성 개념으로 사용가능한지의 여부를 파악하는 것이며 두 번째는 다수준 모형 자체가 적합한지를 검정해보는 것이다. 다수준 모형 자체가 적합한지를 검증하는 데에는 애초에 다수준 모형을 세워서 이후에 통계적으로 검증하는 경우와 처음에는 단일 수준 혹은 교차 수준의 모형을 세웠는데 이러한 모형이 적절한가의 여부를 파악하기 위한 두 가지 유형이 존재한다 (Park과 Ko, 2005).

이에 본 연구에서는 다수준 모형을 먼저 세운 후 다수준 모형의 타당성 여부를 검정하는 방법을 이용하여 서울시의 금연 정책에 따른 각 행정구간 흡연율의 차이를 알아보기 위해 개인 수준과 지역 수준의 두 가지 수준의 변수를 동시에 고려할 수 있는 다수준 로지스틱 회귀분석을 실시하였다. 다수준 로지스틱 회귀분석에는 SAS 9.3의 glimmix 프로시저를 이용하였다.

본 논문의 구성은 다음과 같이 2장에서 다수준 로지스틱 모형을 간략히 소개하였으며, 3장에서는 분석에 이용된 프로그램을 설명하고 다수준 분석 결과를 제시하였으며 4장에서는 결론으로 맺음하였다.

2. 적용 모형 및 자료

2.1. 다수준 로지스틱 모형

2.1.1. 로지스틱 모형 로지스틱 회귀 모형은 종속변수인 Y_i의 로짓을 설명변수 x_k를 통하여 예측하

는 모형으로 식 (2.1)과 식 (2.2)와 같다.

$$y_i \sim \text{Ber}(p_i), \tag{2.1}$$

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}, \tag{2.2}$$

$$k = 1, 2, \ldots, K,\ i = 1, 2, \ldots, n,$$

여기서 p_i는 종속변수 Y_i의 관심 있는 사건이 일어날 확률이며 $\beta_1, \beta_2, \ldots, \beta_k$는 회귀계수, $x_{1i}, x_{2i}, \ldots, x_{ki}$는 연속 또는 범주형인 설명변수이다.

식 (2.2)의 모형은 확률 p에 관한 함수로 쉽게 전환할 수 있으며 식 (2.2)에 역함수를 취하면 식 (2.3)과 같이 기술된다.

$$p_i = \frac{e^{(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki})}}{1+e^{(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki})}}. \tag{2.3}$$

식 (2.2)와 식 (2.3)은 설명변수 $x_{1i}, x_{2i}, \ldots, x_{ki}$로서 종속변수를 예측하는 동일한 로지스틱 회귀모형이다. 식 (2.2)는 예측되는 결과 값이 로짓의 단위이므로 회귀계수 해석에 어려움이 발생하여 식 (2.3)을 통해 설명변수가 특정 값을 가질 때의 확률 p_i를 예측함으로써 일반적인 해석이 가능해져 보편적으로 많이 사용되고 있다.

2.1.2. 다수준 로지스틱 모형 다수준 로지스틱 모형은 위계적 자료 분석을 위해 일반적인 로지스틱 회귀모형을 확장시킨 형태로서 단일 수준만 고려하는 일반적인 로지스틱 회귀모형과는 달리 개인 수준과 집단 수준, 즉, 집단 내 변동과 더불어 집단 간 변동까지 고려하는 모형이다.

위계적 자료는 개인 수준과 집단 수준의 변수가 동시에 존재하는 형태이므로 이를 단일 수준 모형을 이용하여 분석하는 경우 자기상관성으로 인해 일반적인 회귀모형을 적용시킬 수 없기 때문에 대안으로 다수준 모형을 이용해야 한다 (Park과 Ko, 2005).

다수준 로지스틱 모형에 사용되는 로짓 연결 함수 p_{ij}는 식 (2.4)와 같이 설명할 수 있다.

$$p_{ij} = P(y_{ij} = 1) \tag{2.4}$$

여기서, y_{ij}는 집단 j에 속한 개인 i의 종속변수라 하고 이항 분포를 따른다고 가정한다.

다수준 로지스틱 모형의 가장 단순한 형태인 설명변수가 하나인 2-수준 로지스틱 모형은 수준-1(개인 수준)과 수준-2(집단 수준)로 나누어 표현할 수 있으며 수준-1은 식 (2.5)와 같다

$$\text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j} x_{ij}, \quad \text{(개인 수준)}. \tag{2.5}$$

여기서, x_{ij}는 개인 수준에서의 설명변수이며 β_{0j}는 절편, β_{1j}은 개인 수준에서 종속변수에 미치는 설명변수의 효과를 나타낸다.

수준-2는 모형 전체에서 관측된 분산 중 집단 수준을 설명하는 분산을 확인할 수 있는 모형으로 식 (2.6)에서와 같이 수준-1의 절편 β_{0j}를 분해한 형태로 절편 γ_{00}와 수준-2 공변량(covariate)인 z_j가 포함되며 집단 j의 오차항 δ_{0j}로 이루어져 있다. 이 때, 개인 수준의 변수들의 효과는 고정효과로 다시 말해서, 집단 수준에 의한 변동만이 흡연 여부에 영향을 미치는 것으로 가정하여 모형에 적용하였다.

$$\beta_{0j} = \gamma_{00} + \gamma_{01} z_j + \delta_{0j}, \qquad \beta_{1j} = \gamma_{10}, \quad \delta_{0j} \sim N(0, \sigma_\delta^2), \quad \text{(집단 수준)}. \tag{2.6}$$

또한 수준-1과 수준-2를 결합하여 최종적으로 2-수준 로지스틱 모형을 식 (2.7)과 같이 나타낼 수 있다.

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_{00} + \gamma_{01}z_j + \gamma_{10}x_{ij} + \delta_{0j}, \quad \text{여기서 } \delta_{0j} \sim N\left(0, \sigma_\delta^2\right), \quad \text{(최종 모형)}. \tag{2.7}$$

따라서 2-수준 로지스틱 모형은 고정 효과와 임의 효과가 동시에 존재하며 로짓 연결 함수가 있으므로 로지스틱 혼합 효과 모형(logistic mixed effect model)이라고 할 수 있다.

2.2. 분석 방법 및 자료

본 연구에서 사용된 지역사회 건강조사 자료는 서울시에서 25개의 각 구에 속하는 총 20330명의 개인으로 이루어진 계층적 자료이므로 같은 구에 속하는 개인은 상관성이 존재한다. 또한 종속변수는 개인의 흡연여부로 현재 흡연을 하고 있음과 현재 흡연을 하고 있지 않음 두 가지로 나뉘어지는 이분형 구조이며 설명변수는 Kim (1999)의 연구에 근거하여 범주형 변수의 형태인 성별, 학력과 연령, 연속형 변수의 형태인 연간 소득을 이용하였다. 이 때, 지역사회 건강조사 원자료에서는 연령이 연속형 변수의 형태이나 본 연구에서는 연령대별 흡연율의 차이를 알아보고자 범주형 변수의 형태로 변환하였다.

본 연구에서는 다수준 로지스틱 회귀분석을 위하여 SAS 9.3에서 제공하고 있는 PROC GLIMMIX 프로시저를 사용하였다. 이는 GLMM(generalized linear mixed model)을 다룰 수 있는 구문으로 GLMM은 GLM(generalized linear model)과 LMM(linear mixed model)을 연결시킨 모형으로써 범주형 자료의 분석에서 자료가 집단의 형태로 얻어진 계층적 자료인 경우 사용할 수 있다.

3. 결과

3.1. 로지스틱 회귀분석

다수준 로지스틱 모형에 대한 분석에 앞서 일반적인 로지스틱 회귀 모형을 자료에 적합시켰다. 종속변수는 개인의 흡연여부이며 설명변수로는 개인의 성별, 연령, 연간 소득, 학력이 사용되었으며 모형은 식 (3.1)과 같다.

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \beta_4 X_{4j} \tag{3.1}$$

여기서, X_{1j}는 성별, X_{2j}는 연령, X_{3j}는 연간 소득, X_{4j}는 학력 변수를 의미한다. 분석 결과는 각각의 설명변수와 종속변수와의 관계를 알 수 있으며 어느 설명변수가 종속변수인 흡연여부와 어떠한 영향이 있는지를 확인할 수 있다. Table 3.1은 다중 로지스틱 회귀 분석의 결과이다.

Table 3.1에서와 같이 먼저 성별의 경우 여성을 기준으로 남성과 오즈비(odds ratio)를 비교한 결과 남성이 여성에 비해 더 많이 흡연을 하는 것으로 나타났다. 연령별로는 70대 이상의 경우가 다른 연령대보다 흡연율이 낮음을 알 수 있으며 70대를 기준으로 60대, 10대, 50대 20대, 40대, 30대 순으로 흡연율이 높아지는 것을 알 수 있다.

연간 소득의 경우에는 소득이 증가할수록 흡연을 적게 하는 것으로 나타났다. 학력의 경우에는 대학원 이상을 기준으로 4년제 대학, 2년/3년제 대학, 중학교, 고등학교, 초등학교, 서당, 무학 순으로 흡연을 많이 하는 것으로 나타나 학력이 낮아질수록 흡연율이 높은 것을 알 수 있다.

3.2. 다수준 로지스틱 회귀 분석

개인의 관측치를 1-수준 변수로 하고 지역을 2-수준 변수로 투입한 다수준 로지스틱 회귀모형을 적합시켰다. 본 연구에서는 다수준 로지스틱 분석을 위해 세가지 모형을 고려하였다. 첫째로, 자료의 기초정

Table 3.1. Results of traditional logistic regression model

Parameter		Estimate	Standard Error	P-value	Odds ratio (95% C.I.)
Intercept		−5.2228	0.1411	<.0001	· ·
Sex	Male	3.1334	0.0558	<.0001	22.952 (20.598, 25.635)
	Female	·	·	·	· ·
age	10	0.7074	0.2206	0.0013	2.029 (1.303, 3.101)
	20	1.5078	0.1061	<.0001	4.517 (3.675, 5.571)
	30	1.7992	0.1013	<.0001	6.045 (4.965, 7.387)
	40	1.5171	0.1009	<.0001	4.559 (3.748, 5.566)
	50	1.0481	0.0998	<.0001	2.852 (2.349, 3.474)
	60	0.3529	0.1029	0.0006	1.423 (1.165, 1.743)
	>70	·	·	·	· ·
income		−0.00348	0.000766	<.0001	0.997 (0.995, 0.998)
academic ability	illiteracy	1.7262	0.1957	<.0001	5.619 (3.813, 8.217)
	village school	1.6297	0.7486	0.0295	5.102 (1.004, 20.359)
	elementary school	1.0959	0.1256	<.0001	2.992 (2.339, 3.829)
	middle school	0.9813	0.1150	<.0001	2.668 (2.131, 3.345)
	high school	1.0477	0.0909	<.0001	2.851 (2.389, 3.412)
	college	0.8017	0.1005	<.0001	2.229 (1.832, 2.718)
	university	0.4270	0.0869	<.0001	1.553 (1.294, 1.820)
	Graduate or higher	·	·	·	· ·

Table 3.2. Estimates of variance and ICC for random effect

Parameters		Subject	Estimate	Standard Error	P-value	ICC
null model	Intercept($\hat{\sigma}_\delta^2$)	health center	0.0118	0.0056	0.0336	0.0035
model1	Intercept($\hat{\sigma}_\delta^2$)	health center	0.0054	0.0046	0.2425	0.0016
model2	Intercept($\hat{\sigma}_\delta^2$)	health center	0.0074	0.0057	0.1937	0.0022

보를 확인하기 위하여 1-수준과 2-수준의 변수를 투입하지 않은 절편만을 포함하는 기초모형을 살펴볼 것이다. 둘째, 기초모형을 바탕으로 식 (2.5)와 같이 지역의 고정 효과를 제외한 임의 효과만을 포함하는 (1-수준 변수만을 포함시킨 모형) 연구모형1에 대해서 살펴볼 것이다.

연구모형1에 사용된 변수로 종속변수는 개인의 흡연여부이며 설명변수는 1-수준 변수인 개인 수준에 해당하는 성별, 연령, 연간 소득, 학력을 이용하였다. 마지막으로 연구모형1에 지역 수준의 고정 효과를 포함시키기 위하여 2-수준 변수를 추가하여 식 (2.7)과 같이 연구모형2를 구축하였으며 2-수준 변수로는 지역별 남성의 비율, 평균 연령, 평균 연간 소득, 평균 학력 변수가 이용되었다. 이 때, 2-수준 변수는 1-수준 변수로 사용된 성별, 연령, 연간 소득, 학력 변수를 이용하여 각 지역별 비율 및 평균을 계산한 값이다.

3.2.1. 모형 결과 본 연구에서 제시한 세가지 모형에 대한 타당성을 검토하기 위해 임의 효과에 대한 분산을 추정하고 급내 상관 계수(ICC; intraclass correlation coefficient)를 추정하였으며 Table 3.2와 같다.

Table 3.2의 기초 모형의 모수 추정 결과를 통하여 다른 설명변수를 고려하지 않았을 때 흡연여부에 대한 지역별 분산을 분석함으로써, 연구모형1과 연구모형2에서 다른 설명변수들의 설명력을 살펴보게 된다.

본 논문은 설명변수가 범주형 자료와 연속형 자료가 뒤섞여 있는 형태로 1-수준의 분산은 이분산적이기 때문에 각 1-수준 단위마다 다르게 된다. 따라서 총 분산에 대한 각 수준의 분산비율 정도를 살펴보는 것은 의미가 없게 되지만 이 때 1-수준 분산의 척도인수를 1로 고정할 수 있기 때문에 편의상 1-수준의 분산을 1로 가정하였고 본 논문에서는 2-수준 분산만을 설명하였다.

흡연여부와 관련하여 지역 수준의 분산은 0.0118로 나타났고 표준 오차는 0.0056으로 나타났다. 이 때, 이를 이용하여 p-값을 계산하여 추정된 값이 0보다 유의하게 크면 흡연여부에 행정구별 보건소의 효과가 존재한다고 볼 수 있으며 다시 말해 특정 행정구의 보건소는 다른 행정구의 보건소와 흡연율의 차이를 보인다는 것을 의미한다. 위의 결과에서는 p-값이 0.0336으로 0보다 유의하게 크다고 할 수 있으므로 각 행정구 간 흡연율의 차이를 보인다고 할 수 있다.

그리고 제 2-수준의 집단 수준에서의 변이(집단 간의 분산)와 제 1-수준에서의 개인 수준에서의 변이(집단 내의 분산)를 가지고 집단 간 분산 비율(ICC)를 구할 수 있는데 이는 종속변수의 전체 분산 중에서 집단 수준의 분산이 차지하는 비율이라고 할 수 있다. 집단 수준의 분산을 이용하여 ICC를 계산하는 방법은 식 (3.2)와 같다.

$$\text{ICC} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \pi^2/3}. \tag{3.2}$$

본 자료의 기초모형에서는 지역 수준의 분산이 0.0118로 ICC 값이 약 0.0035로 나타났으며 이를 통해 전체 분산 중에서 지역 수준의 분산이 차지하는 비율이 0.35%정도임을 확인할 수 있다.

연구모형1에서의 각 행정구 간의 변동성을 측정하는 보건소의 모수($\hat{\sigma}_\delta^2$) 추정 결과 Table 3.2와 같이 보건소 절편의 추정된 분산은 0.0054이며 표준 오차는 0.0046으로 나타났다. 이 때, p-값이 0.2425로 추정된 값이 0보다 유의하게 크다고 할 수 없으므로 각 행정구의 간 흡연율의 차이가 존재하지 않는다고 할 수 있다.

연구모형2에서의 각 행정구 간의 변동성을 측정하는 보건소의 모수($\hat{\sigma}_\delta^2$) 추정 결과 Table 3.2와 같이 보건소 절편의 추정된 분산은 0.0074이며 표준 오차는 0.0057로 나타났다. 이를 이용하여 p-값을 계산한 결과 그 값이 0.1937로 추정된 값이 0보다 유의하게 크다고 할 수 없으므로 각 행정구 간 흡연율의 차이가 존재하지 않는다고 할 수 있다.

위의 결과를 종합해보면 기초 모형에서는 각 행정구 간 흡연율의 차이가 존재한다고 나타났으나 개인 수준과 지역 수준의 변수가 추가된 모형에서는 각 행정구 간 흡연율의 차이가 존재하지 않는 것으로 나타났다. 이는 개인 수준과 지역 수준의 변수가 흡연여부에 영향을 미치고 있는 유의한 변수라는 사실을 유추할 수 있다.

따라서 개인 수준과 지역 수준의 변수가 흡연여부에 어떠한 영향을 미치고 있는지 알아보기 위하여 지역의 고정 효과를 제외한 모형인 연구모형1의 예측된 고정 효과와 지역의 고정 효과를 포함시킨 모형인 연구모형2의 예측된 고정 효과를 연구하였다. 적용된 연구모형1과 연구모형2는 식 (3.3)과 식 (3.4)와 같다.

$$\text{logit}(p_{ij}) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij}, \tag{3.3}$$

$$\begin{aligned}\text{logit}(p_{ij}) = {} & \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{30}X_{3ij} + \gamma_{40}X_{4ij} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{03}Z_{3j} \\ & + \gamma_{04}Z_{4j} + \delta_{0j}\end{aligned} \tag{3.4}$$

여기서, X_{1ij}는 성별, X_{2ij}는 연령, X_{3ij}는 연간 소득, X_{4ij}는 학력으로 개인 수준의 변수를 의미하며 Z_{1j}는 남성의 비율, Z_{2j}는 평균 연령, Z_{3j}는 평균 연간 소득, Z_{4j}는 평균 학력으로 지역 수준의 변수를

Table 3.3. Parameter estimates for fixed effects for two types of model

Parameter		Model1				Model2			
		Estimate	Standard Error	P-value	Odds ratio	Estimate	Standard Error	P-value	Odds ratio
Intercept		−5.220	0.142	<.0001	·	−4.998	2.434	0.0538	·
Sex	Male	3.133	0.056	<.0001	22.936	3.131	0.056	<.0001	22.907
	Female	·	·	·	·	·	·	·	·
age	10	0.707	0.221	0.0014	2.028	0.707	0.221	0.0014	2.027
	20	1.505	0.106	<.0001	4.503	1.502	0.106	<.0001	4.490
	30	1.797	0.101	<.0001	6.029	1.794	0.102	<.0001	6.013
	40	1.515	0.101	<.0001	4.550	1.512	0.101	<.0001	4.536
	50	1.050	0.100	<.0001	2.856	1.049	0.100	<.0001	2.855
	60	0.353	0.103	0.0006	1.423	0.353	0.103	0.0006	1.423
	>70	·	·	·	·	·	·	·	·
income		−0.003	0.001	<.0001	0.997	−0.003	0.001	<.0001	0.997
academic ability	illiteracy	1.721	0.196	<.0001	5.592	1.709	0.196	<.0001	5.521
	village school	1.600	0.749	0.0327	4.952	1.582	0.749	0.0346	4.865
	elementary school	1.086	0.126	<.0001	2.962	1.070	0.127	<.0001	2.916
	middle school	0.976	0.115	<.0001	2.654	0.963	0.116	<.0001	2.620
	high school	1.042	0.091	<.0001	2.834	1.030	0.092	<.0001	2.801
	college	0.800	0.101	<.0001	2.225	0.789	0.101	<.0001	2.202
	university	0.426	0.087	<.0001	1.531	0.422	0.087	<.0001	1.524
	Graduate or higher	·	·	·	·	·	·	·	·
ave(age)		·	·	·	·	−0.002	0.029	0.9566	0.998
ave(income)		·	·	·	·	−0.003	0.010	0.7815	0.997
ave(academic ability)		·	·	·	·	−0.039	0.263	0.8851	0.962
prop(male)		·	·	·	·	0.427	1.394	0.7629	1.532

의미한다. 이 때, 학력 변수는 범주형 변수로 평균을 계산하는 것이 불가능하지만 지역 수준의 변수로 변형하기 위하여 학력을 1부터 7까지 저학력에서 고학력 순으로 나열하여 연속형 변수로 취급하여 평균을 계산하였다. 연구모형1과 연구모형2의 예측된 고정 효과 결과는 Table 3.3과 같다.

여기서 성별, 연령, 연간 소득, 학력은 개인 수준의 고정 효과 추정량이며 평균 연령, 평균 연간 소득, 평균 학력, 남성 비율은 지역 수준의 고정 효과 추정량이다.

이를 이용하여 연구모형1을 개인 수준에서 살펴볼 때 성별의 경우 여성을 기준으로 남성과 오즈비(odds ratio)를 비교한 결과 남성이 여성에 비해 더 많이 흡연을 하는 것으로 나타났다. 연령별로는 70대 이상의 경우가 다른 연령대보다 흡연율이 낮음을 알 수 있으며 70대를 기준으로 60대, 10대, 50대 20대, 40대, 30대 순으로 흡연율이 높아지는 것을 알 수 있다.

연간 소득의 경우에는 소득이 증가할수록 흡연을 적게 하는 것으로 나타났다. 학력의 경우에는 대학원 이상을 기준으로 4년제 대학, 2년/3년제 대학, 중학교, 고등학교, 초등학교, 서당, 무학 순으로 흡연을 많이 하는 것으로 나타나 학력이 낮아질수록 흡연율이 높은 것을 알 수 있다.

이는 다수준 분석에 앞서 실시한 일반적인 로지스틱 분석의 결과와 같게 나온 것을 확인할 수 있다. 여기서 모든 값들이 유의한 것으로 나타났으므로 따라서 개인의 특성이 흡연여부의 유의한 설명변수가 될 수 있다고 할 수 있다.

연구모형2의 결과를 살펴보면 개인 수준에서 고려할 때 성별의 경우 여성을 기준으로 남성과 오즈비(odds ratio)를 비교한 결과 남성이 여성에 비해 더 많이 흡연을 하는 것으로 나타났다. 연령별로는 70대 이상의 경우가 다른 연령대보다 흡연율이 낮음을 알 수 있으며 70대를 기준으로 60대, 10대, 50대 20대, 40대, 30대 순으로 흡연율이 높아지는 것을 알 수 있다.

연간 소득의 경우에는 소득이 증가할수록 흡연을 적게 하는 것으로 나타났다. 학력의 경우에는 대학원 이상을 기준으로 4년제 대학, 2년/3년제 대학, 중학교, 고등학교, 초등학교, 서당, 무학 순으로 흡연을 많이 하는 것으로 나타나 학력이 낮아질수록 흡연율이 높은 것을 알 수 있다. 이를 통해 다수준 분석에 앞서 실시한 일반적인 로지스틱 분석과 연구모형1의 분석 결과와 같게 나온 것을 확인할 수 있다. 여기서 모든 값들이 유의한 것으로 나타났으므로 따라서 개인의 특성이 흡연여부의 유의한 설명변수가 될 수 있다고 할 수 있다.

지역 수준에서 살펴볼 때에는 지역 수준의 변수들의 유의확률이 모두 0.05보다 크게 나타나므로 유의 수준 0.05에서 유의하지 않아 흡연여부와 평균 연령, 평균 연간 소득, 평균 학력, 남성 비율과는 유의한 관계가 없음을 알 수 있다. 즉, 지역 보건소의 속성이 흡연의 유의한 설명변수가 될 수 없다고 할 수 있다.

연구모형1과 연구모형2의 결과를 종합해 볼 때 개인의 특성은 흡연여부의 유의한 설명변수가 될 수 있는 것으로 나타났으나 지역 보건소의 속성은 흡연의 유의한 설명변수가 될 수 없는 것으로 나타났다.

3.2.2. 예측 흡연 비율에 관하여 각 행정구의 보건소별로 차이가 존재하는지 확인하기 위해 보건소 프로파일링을 적용시켰다. 프로파일링은 특정 행정구의 보건소의 흡연율과 특정 행정구를 제외한 나머지 보건소의 흡연율의 평균을 비교하는 방법이며 귀무가설은 식 (3.5)와 같으며 보건소 프로파일링 결과는 Table 3.4와 같다.

$$H_0 : \mu_i - \frac{1}{J}\sum_{j=1}^{J}\mu_j = 0, \tag{3.5}$$

여기서 추정값의 지수 변화값은 특정 행정구의 보건소와 나머지 행정구 보건소의 흡연율 평균을 비교할 수 있는 척도이다. 만약 특정 행정구 보건소의 추정값의 지수 변화값이 다른 행정구 보건소의 값보다 유의하게 크다면 특정 행정구 보건소는 나머지 행정구 보건소들의 평균 흡연율보다 흡연율이 높다는 것을 의미하며, 특정 행정구 보건소의 추정값의 지수 변화값이 다른 행정구 보건소의 값보다 유의하게 작다면 특정 행정구 보건소는 나머지 행정구 보건소들의 평균 흡연율에 비해 흡연율이 낮다는 것을 의미한다.

따라서 연구모형1의 프로파일링 결과 강동구와 용산구가 다른 구에 비해 흡연율이 높으며 양천구와 중랑구가 다른 구에 비해 흡연율이 낮은 것처럼 보이나 통계적으로 유의한 의미는 갖지 않으므로 서울시 보건소 간에 흡연율의 차이가 없는 것을 알 수 있다. 연구모형2에서도 통계적으로 유의한 값이 존재하지 않아 보건소 간에 흡연율의 차이가 없는 것으로 나타나 연구모형1과 동일한 결과를 보이는 것을 알 수 있다.

앞서 각 행정구의 보건소에 대하여 흡연율을 예측한 것과는 달리 20330명의 각 개인에 대하여 흡연여부를 예측하였다. 두 가지 모형의 예측값이 예측에 사용되었으며 그 값이 0.5보다 작은 경우 흡연을 하지 않는 것으로 판단하고 0.5보다 큰 경우에는 흡연을 하는 것으로 판단하였다.

첫 번째 모형은 임의 효과의 결과에서 사용되며 두 번째 모형은 고정 효과에 기초한 예측에서 사용되는 모집단 평균 예측이다. 이를 계산하는 방법은 다음과 같다.

다수준 로지스틱 모형이 식 (3.6)과 같다고 가정할 때,

$$Y = X\beta + Z\gamma + e. \tag{3.6}$$

Table 3.4. Health center profiling

Label	Model1				Model2			
	Estimate	Standard Error	$\Pr > \|t\|$	exp (estimate)	Estimate	Standard Error	$\Pr > \|t\|$	exp (estimate)
Gangnam-gu	−0.026	0.060	0.662	0.9739	0.006	0.073	0.932	1.0063
Gangdong-gu	0.084	0.059	0.168	1.0874	0.106	0.067	0.126	1.1115
Gangbuk-gu	0.007	0.058	0.901	1.0073	−0.007	0.068	0.919	0.9931
Gangseo-gu	−0.010	0.059	0.873	0.9905	−0.007	0.067	0.918	0.9931
Gwanak-gu	−0.011	0.057	0.854	0.9894	−0.025	0.074	0.740	0.9755
Gwangjin-gu	0.050	0.058	0.393	1.0514	0.062	0.070	0.389	1.0636
Guro-gu	−0.041	0.058	0.492	0.9602	−0.051	0.067	0.449	0.9501
Geumcheon-gu	0.025	0.056	0.658	1.0255	0.012	0.071	0.871	1.0117
Nowon-gu	0.020	0.058	0.735	1.0201	0.027	0.068	0.699	1.0271
Dobong-gu	0.022	0.058	0.709	1.0222	0.019	0.067	0.782	1.0189
Dongdaemun-gu	0.005	0.058	0.936	1.0047	−0.015	0.069	0.828	0.9850
Dongjak-gu	−0.011	0.058	0.847	0.9888	−0.012	0.065	0.859	0.9885
Mapo-gu	0.030	0.059	0.620	1.0301	0.041	0.067	0.548	1.0414
Seodaemun-gu	−0.008	0.058	0.888	0.9917	−0.014	0.066	0.832	0.9860
Seocho-gu	−0.017	0.059	0.775	0.9832	0.011	0.071	0.878	1.0110
Seongdong-gu	−0.003	0.058	0.958	0.9969	−0.009	0.064	0.895	0.9915
Seongbuk-gu	0.016	0.058	0.787	1.0160	0.010	0.065	0.880	1.0100
Songpa-gu	−0.074	0.059	0.219	0.9285	−0.078	0.071	0.286	0.9250
Yangcheon-gu	−0.085	0.059	0.160	0.9186	−0.097	0.066	0.156	0.9073
Yeongdeungpo-gu	−0.042	0.058	0.480	0.9591	−0.042	0.066	0.531	0.9587
Yongsan-gu	0.100	0.059	0.102	1.1053	0.113	0.069	0.110	1.1204
Eunpyeong-gu	−0.002	0.058	0.978	0.9984	−0.004	0.065	0.946	0.9955
Jongno-gu	0.036	0.058	0.546	1.0363	0.050	0.068	0.466	1.0515
Jung-gu	−0.002	0.058	0.971	0.9979	−0.012	0.066	0.862	0.9884
Jungnang-gu	−0.063	0.058	0.291	0.9391	−0.083	0.067	0.225	0.9201

Table 3.5. Example of comparison of observed value and two types of predicted values

Observed value	Predictive probability by random effect	Predictive probability by fixed effect
1	0.57733 (1)	0.58376 (1)
0	0.03808 (0)	0.03906 (0)
1	0.49643 (0)	0.50303 (1)
0	0.48990 (0)	0.50560 (1)

첫 번째 모형의 예측값은 식 (3.7)과 같이 나타내며 두 번째 모형의 예측값은 식 (3.8)과 같이 나타낸다.

$$E[Y|\gamma] = g^{-1}\left(X'\hat{\beta} + Z'\hat{\gamma}\right), \tag{3.7}$$

$$E[Y] = g^{-1}\left(X'\hat{\beta}\right). \tag{3.8}$$

이를 이용하여 예측한 전체 관측치의 흡연여부에 관한 두 가지 모형의 예측값 결과 중 일부를 Table 3.5로 나타냈다.

Table 3.5를 살펴보면 첫 번째 줄의 흡연을 하는 개인에 대하여 예측값을 계산한 결과 두 가지 모형의 예측값 모두 0.5 이상으로 나타나 흡연을 하는 것으로 예측이 적절히 되었으며 두 번째 줄에 서술된 흡

Table 3.6. Results of prediction performance for various suggested models

Observed value	Logistic model		Total
	1	0	
1	1624	2690	4314
0	1123	14893	16016
Total	2747	17583	20330

Observed value	Random effect (model1)		Total
	1	0	
1	1643	2671	4314
0	1111	14905	16016
Total	2754	17576	20330

Observed value	Fixed effect (model1)		Total
	1	0	
1	1621	2693	4314
0	1123	14893	16016
Total	2744	17586	20330

Observed value	Random effect (model2)		Total
	1	0	
1	1641	2673	4314
0	1115	14901	16016
Total	2756	17574	20330

Observed value	Fixed effect (model2)		Total
	1	0	
1	1641	2673	4314
0	1133	14883	16016
Total	2774	17556	20330

연을 하지 않는 개인에 대하여 예측값을 계산한 결과 두 가지 모형의 예측값 모두 0.5 이하로 나타나 흡연을 하지 않는 것으로 예측이 적절히 되었다.

그러나 세 번째 줄에 서술된 흡연을 하는 개인에 대하여 예측값을 계산한 결과 첫 번째 모형의 예측값은 0.5이하로 흡연을 하지 않는다고 예측한 반면 두 번째 예측값은 0.5 이상으로 흡연을 한다고 예측하였다. 또한 마지막 줄에 서술된 흡연을 하지 않는 개인에 대하여 예측값을 계산한 결과 첫 번째 예측값은 0.5이하로 흡연을 하지 않는다고 적절히 예측한 반면 두 번째 예측값은 0.5 이상으로 흡연을 한다고 예측하였다. 이처럼 두 가지 모형 모두에서 정확한 예측 결과를 보이는 경우도 있었으나 서로 상반된 예측 결과를 보이는 경우도 있는 것으로 나타났다.

따라서 다수준 로지스틱 모형에서 두 가지 예측값 중 어느 모형의 예측값이 더 정확성을 나타내는지 알아보기 위하여 전체 관측치에 대한 예측값 결과를 이용하여 만든 실제 관측치와 예측값의 교차표를 작성하였으며 비교를 위하여 일반적인 로지스틱 모형에서의 실제 관측치와 예측값의 교차표를 작성하였다. 그 결과는 Table 3.6과 같다.

Table 3.6을 토대로 다른 모형에 비해 더 큰 정확성을 나타내는 모형을 알아보기 위하여 민감도(sensitivity)와 특이도(specificity), 정확도(PCC; percentage of correctly classified observations)를 계산하였다. 여기서 민감도는 실제 흡연을 하고 있음(1)으로 관측되었을 때 예측값 또한 흡연을 하고 있음(1)으로 나타날 확률을 뜻하며 특이도는 실제 흡연을 하고 있지 않음(0)으로 관측되었을 때 예측값 또한 실제 흡연을 하고 있지 않음(0)으로 나타날 확률을 뜻한다. 정확도는 측정 결과가 진정한 참값을 반영하는 정도로 전체 관측치 중 예측값 적중 비율을 나타낸다.

민감도와 특이도, 정확도를 계산한 결과는 Table 3.7과 같으며 계산 결과 민감도, 특이도, 정확도 모두에서 임의 효과를 포함하는 연구모형1이 가장 높게 나타나는 것을 알 수 있다.

앞서 예측한 결과 Table 3.6을 이용하여 ROC curve(Receiver Operating Characteristic curve)를 작성하였다. ROC curve는 민감도(sensitivity)와 특이도(specificity)가 어떤 관계를 갖고 변하는지를 이차원 평면상에 표현한 것으로 AUC(area under curve), 즉, ROC curve 아래의 면적이 넓을수록 좋은 진단 방법이라 할 수 있다.

Table 3.7. Result of sensitivity, specificity and PCC

	Logistic model	Random effect (model1)	Fixed effect (model1)	Random effect (model2)	Fixed effect (model2)
Sensitivity	0.3764	0.3808	0.3757	0.3803	0.3803
Specificity	0.9298	0.9306	0.9298	0.9303	0.9292
PCC	0.8124	0.8139	0.8122	0.8136	0.8127

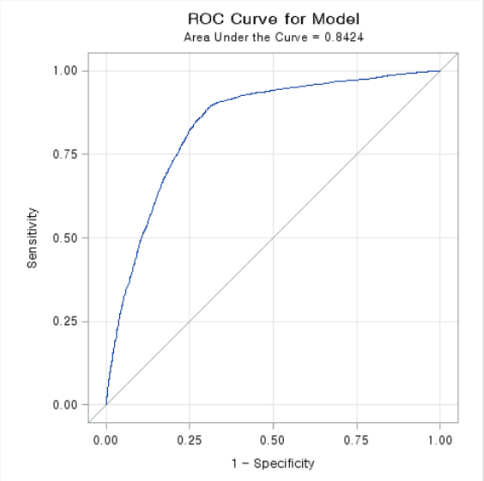

Figure 3.1. ROC curve for logistic model

일반적인 로지스틱 모형의 예측 결과를 이용한 ROC curve는 Figure 3.1로 나타낼 수 있다.

연구모형1과 연구모형2 각각의 임의 효과를 포함하는 예측 결과를 이용한 ROC curve는 Figure 3.2로 나타낼 수 있다.

연구모형1과 연구모형2 각각의 고정 효과를 포함하는 예측 결과를 이용한 ROC curve는 Figure 3.3으로 나타낼 수 있다.

위의 결과를 AUC를 이용하여 비교한 결과, 일반적인 로지스틱 모형의 AUC는 0.8424로 계산되고 임의 효과를 포함하는 경우에서 연구모형1의 AUC는 0.8432로 나타났으며 연구모형2의 AUC는 0.8410으로 나타났다. 고정 효과를 포함하는 경우에서는 연구모형1의 AUC는 0.8425로 나타났으며 연구모형2의 AUC는 0.8424로 나타났다. 이 때, AUC가 클수록 더 좋은 모형으로 판단하므로 AUC가 가장 크게 나타난 임의 효과를 포함하는 연구모형1이 다른 모형에 비해 더 좋은 모형이라 할 수 있다.

4. 결론 및 제언

이 연구는 지역 주민의 성별, 연령, 연간 소득, 학력이 흡연여부에 미치는 영향과 적합한 모형의 수준을 확인하기 위하여 다수준 로지스틱 모형을 이용하여 모형의 적절성 여부를 통해 살펴보았다.

다수준 로지스틱 모형 검정 결과 연구모형1에서는 개인 수준에서 살펴볼 때 남성이 여성에 비해 흡연율이 높으며 연령에서는 70대 이상의 경우가 다른 연령대에 비해 흡연율이 낮게 나타나며 70대를 기준으로 60대, 10대, 50대 20대, 40대, 30대 순으로 흡연율이 높아지는 것을 알 수 있다. 또한 연간 소득이 증가할수록 흡연을 적게 하며 학력의 경우에는 대학원 이상을 기준으로 4년제 대학, 2년/3년제 대학, 중학

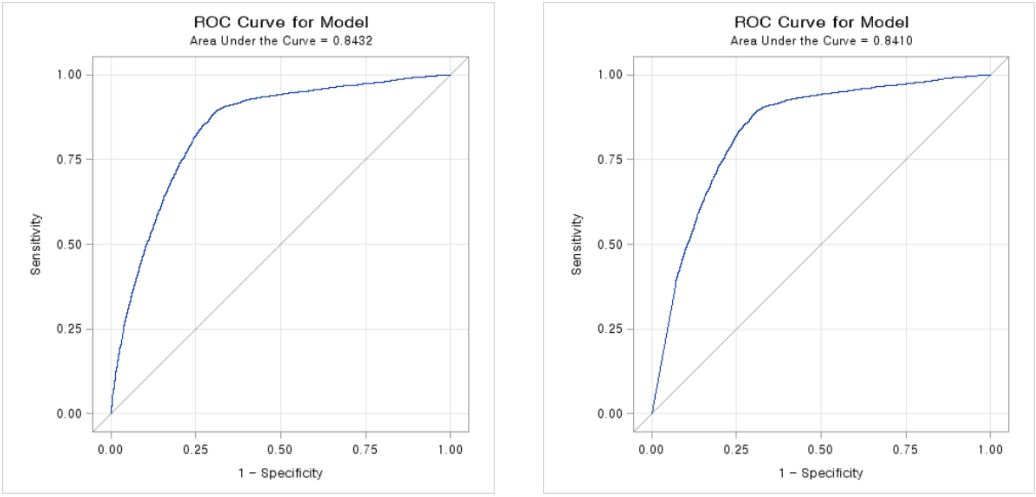

Figure 3.2. ROC curve for model1

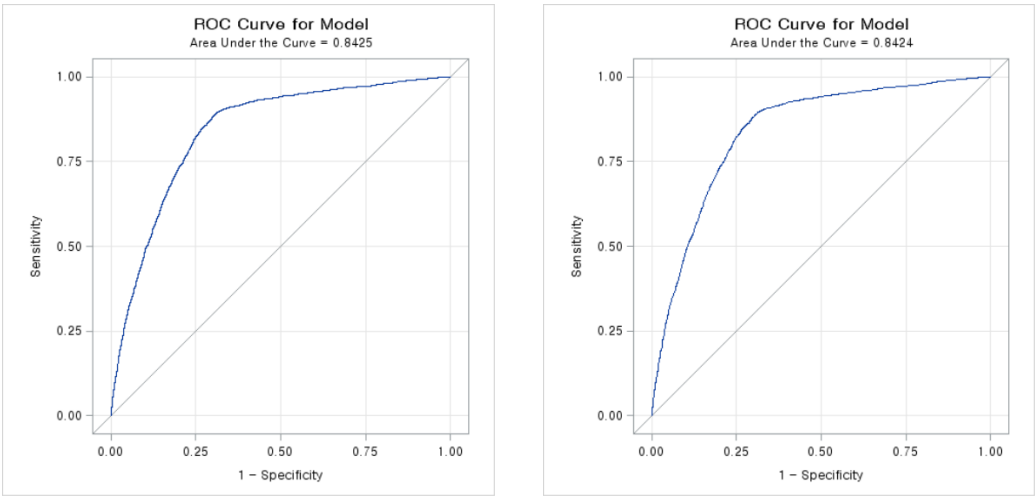

Figure 3.3. ROC curve for model2

교, 고등학교, 초등학교, 서당, 무학 순으로 흡연을 많이 하는 것으로 나타나 학력이 낮아질수록 흡연율이 높은 것을 알 수 있다.

연구모형2에서는 개인 수준에서 살펴볼 때에는 연구모형1의 개인 수준 결과와 동일하게 나타났으며 지역 수준에서 살펴볼 때에는 지역 수준의 변수들의 유의확률이 모두 0.05보다 크게 나타나므로 유의 수준 0.05에서 유의하지 않아 흡연여부와 평균 연령, 평균 연간 소득, 평균 학력, 남성 비율과는 유의한 관계가 없음을 알 수 있다. 즉, 지역 수준의 변수는 흡연여부에 영향을 미치지 않으므로 각 행정구의 보건소는 개인의 흡연에 영향을 미치지 않으며 개인 수준의 변수만 흡연에 영향을 미치므로 다수준이 아닌 단일 수준의 로지스틱 모형이 적절하다고 할 수 있다.

모형을 살펴본 후 각 수준의 모형을 이용하여 행정구별로 보건소의 흡연율을 비교하고 개인의 흡연여부

를 예측한 결과는 다음과 같이 나타났다. 보건소 프로파일링을 이용하여 각 행정구 보건소의 흡연율을 비교한 결과 연구모형1과 연구모형2에서 모두 통계적으로 유의한 차이를 보이는 행정구 보건소가 존재하지 않는 것으로 나타나 서울시 행정구 보건소 간 흡연율의 차이가 존재하지 않는 것을 알 수 있다.

개인의 흡연여부를 두 가지 모형의 예측값을 이용하여 예측한 결과에서는 ROC curve를 작성해본 결과 다른 모형에 비해 AUC가 더 크게 나타난 임의 효과를 포함하는 연구모형1이 다른 모형에 비해 더 좋은 모형이라 할 수 있다.

위와 같은 결과를 통해 다수준 모형의 적용을 통하여 모형의 적합한 수준을 검정하는 방법의 활용성을 입증하였으며 서울시의 금연 정책 시행 결과 각 구의 보건소들의 활발한 정책 추진으로 25개 구의 보건소 모두에서 비슷한 결과를 보이는 것으로 나타났다.

본 연구에서는 일반적인 다수준 모형과 다수준 로지스틱 모형을 적용하였으나 이를 토대로 하여 매년 조사가 시행되는 지역사회 건강조사의 특성을 이용한 패널 다수준 모형의 개발과 공간적 상관성을 반영한 공간 다수준 모형의 개발의 필요성이 제기되었으며 향후 연구 계획 중에 있다.

References

Dai, J., Li, Z. and Rocke, D. (2006). *Hierarchical Logistic Regression Modeling with SAS GLIMMIX*, Western Users of SAS Software 2006.

Flom, P. L., McMahon, J. M. and Pouget, E. R. (2006). *Using PROC NLMIXED and PROC GLIMMIX to Analyze Dyadic Data with Binary Outcomes*, Northeast SAS Users Group 2006.

Guo, G. and Zhao, H. (2000). Multilevel modeling for binary data, *Annual Review of Sociology*, **26**, 441–462.

Jung, J.-H., Kwon, S.-M., Kim, K.-H., Lee, S.-K. and Kim, D.-S. (2010). Impact of health insurance type on the quality of hemodialysis services: A multilevel analysis, *Journal of Preventive Medicine and Public Health*, **43**, 245–256.

Khan, H. and Shaw, J. (2011). Multilevel logistic regression analysis applied to binary contraceptive prevalence data, *Journal of Data Science*, **9**, 93–110.

Kim, C.-H. (1999). Factors related to smoking among male smokers in Seoul, *Inje Medical Journal*, **20**, 699–704.

Kim, D.-S., Hwang, J.-H. and Hwang, J.-I. (2012). A multi-level analysis of injection requests and associated patient characteristics in the Korean acute-care outpatient setting, *Korean College of Clinical Pharmacy*, **22**, 13–20.

Lee, H., Lee, S. and Lee, E. (2012). Characteristics and factors related to problem drinking of the elderly in Korea, *Journal of The Korea Society of Health Informatics and Statistics*, **37**, 34–75.

Li, J., Alterman, T. and Deddens, J. A. (2006). *Analysis of Large Hierarchical Data with Multilevel Logistic Modeling Using PROC GLIMMIX*, SAS Users Group International 31.

Park, W.-W. and Ko, S. (2005). Procedures and methods of multilevel analysis: With a focus on WABA, *Seoul Journal of Business*, **39**, 59–90.

Schabenberger, O. (2005). *Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models*, SUGI 30.

Seoul Metropolitan Government (2013). http://health.seoul.go.kr/archives/825

WHO (2013). http://www.who.int/mediacentre/factsheets/fs339/en/index.html

다수준 로지스틱 모형을 이용한 흡연 여부에 미치는 영향 분석

이지혜[a] · 허태영[a,1]

[a]충북대학교 정보통계학과

(2013년 11월 15일 접수, 2014년 1월 1일 수정, 2014년 1월 6일 채택)

요 약

본 연구에서는 질병관리본부에서 매년 조사하고 있는 지역사회 건강조사 자료를 이용하여 서울시 지역을 대상으로 개인의 흡연 여부에 대한 영향 요인을 확인하고 지역간 차이를 모형에 반영시키는 다수준 로지스틱 모형을 이용하여 분석하였다. 다수준 모형에서의 적합한 분석 모형의 수준을 결정하기 위해 ICC(intraclass correlation coefficient)와 프로파일링 분석, 수준별 모형의 예측정확도를 이용하였다. 제안된 모형들의 성능을 평가하기 위해 민감도, 특이도, 정확도를 구하고 ROC curve를 작성하였다. 결과적으로 지역사회 건강조사 자료와 같이 개인과 집단 변수를 동시에 고려할 수 있다면 다양한 다수준 모형의 적용이 가능하며 활용성이 높다는 것을 알 수 있었다.

주요용어: 다수준 분석, 다수준 로지스틱 회귀 모형, 위계적 자료, 지역사회 건강조사.

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2012R1A1A1040358).

[1]교신저자: (361-763) 충북 청주시 흥덕구 개신동, 충북대학교 정보통계학과, 교수. E-mail: theo@cbnu.ac.kr