Journal of Health Informatics and Statistics

Original Article

J Health Info Stat 2017;42(1):70-76 • https://doi.org/10.21032/jhis.2017.42.1.70

pISSN 2465-8014|eISSN 2465-8022

로지스틱 회귀모형을 사용한 율의 표준화 방법: 국민건강보험공단 건강검진코호트 사용

조상훈[1], 강근석[1], 김현창[2]

[1]숭실대학교 정보통계보험수리학과, [2]연세대학교 의과대학 예방의학교실

Illustration of Calculating Standardized Rates Utilizing Logistic Regression Models: The National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS)

Sang-Hoon Cho[1], Gunseog Kang[1], Hyeon Chang Kim[2]

[1]Department of Statistics and Actuarial Science, Soongsil University, Seoul; [2]Department of Preventive Medicine, Yonsei University College of Medicine, Seoul, Korea

Objectives: To illustrate an approach for standardizing rates utilizing logistic regression models that leads to the enhanced reliability of estimation with reduced calculation cost. **Methods:** For illustrative purposes, data regarding metabolic syndrome patients in 2013 were extracted from the National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS). The detailed step-by-step calculations of age-sex adjusted prevalence rates of metabolic syndrome were demonstrated by both direct and logistic regression standardization approaches whose results were then compared. **Results:** Standardization of rates using logistic regression models facilitated relatively simple calculation that can be easily implemented by using widely employed analytical programs such as R, SPSS, and SAS. Treating age as a continuous variable, the logistic regression approach produced confidence intervals of age-sex adjusted prevalence rates that were much narrower as compared to confidence intervals obtained by the direct standardization. **Conclusions:** Standardization of rates utilizing logistic regression models may be a competitive alternative to the direct standardization in terms of computational efficiency and estimation reliability.

Key words: Standardization, Logistic regression, R, Metabolic syndrome, NHIS-HEALS

서 론

보건·역학 관련 연구에서는 서로 다른 모집단에서 산출한 통계량을 비교하게 된다. 예를 들면, 어떤 한 지역의 유병률이 다른 지역보다 높은지 또는 어떤 지역의 특정 질병의 원인이 무엇인지를 밝히고자 할 때가 있다[1]. 대부분의 선행 연구를 통해 연령은 건강수준에 영향을 미치는 중요한 요인으로 알려져 있으므로[2,3], 지역별 차이를 연구할 때, 지역별 모집단의 상이한 연령 구조를 고려해야 한다[4]. 이와 같이 인구학적(demographic) 구조(성별, 사회·경제적 지위, 인종 등)의 지역별 차이는 건강관련 지표를 산출하는 데 영향을 미친다. 인구학적 구조가 다른 지역 간 또는 기간별 유병률을 비교할 때, 연구의 주요 관심대상이 아닌 교란변수(confounding variable)의 효과를 제거하고 지역

Corresponding author: Sang-Hoon Cho
369 Sangdo-ro, Dongjak-gu, Seoul 06978, Korea
Tel: +82-2-820-0444, E-mail: sanghcho@ssu.ac.kr

Received: January 20, 2017 **Revised:** February 21, 2017 **Accepted:** February 25, 2017

*This study used NHIS-HEALS data (NHIS-2017-2-303) from the National Health Insurance Service (NHIS). The authors declare no conflicts of interest with the NHIS.

No potential conflict of interest relevant to this article was reported.

How to cite this article:
Cho SH, Kang G, Kim HC. Illustration of calculating standardized rates utilizing logistic regression models: The National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS). J Health Info Stat 2017;42(1):70-76. Doi: https://doi.org/10.21032/jhis.2017.42.1.70

간 비교를 단순화시켜주는 지표로 표준화(standardization)를 사용한다. 표준화율(age-adjusted rate)은 1844년 사망률 자료를 분석하는 연구에서 처음으로 사용되었고[5], 이후 표준화율을 산출하는 데 다양한 방법이 적용되고 있다[6].

이 연구에서는 가장 보편적으로 선호되는 방법 중 하나인 직접 표준화(direct standardization) 방법[7,8]과 로지스틱 회귀모형(logistic regression model) [9]을 활용하여 표준화율을 계산하는 방법을 비교하여 소개한다. 로지스틱 회귀모형은 대부분의 학술연구에서 사건 발생 확률과 독립 변수 간의 연관성을 연구하기 위한 목적으로 또는 사건 발생을 예측하기 위한 목적으로 사용되고 있으나[10], 직접 표준화 방법의 효과적이고 합리적인 대안으로 표준화율을 산출하는 데에 활용할 수 있다. 직접 표준화 방법은 범주형 자료에만 사용할 수 있으며, 어떤 특정 연령대에 해당하는 자료수가 적을 때, 연령대별 추정량의 표준오차가 커지므로 추정값의 정확도가 낮아진다[11]. 반면에 로지스틱 회귀모형을 통한 표준화는 범주형뿐만이 아닌 연속형 변수를 사용해서도 추정할 수 있으며, 특정 연령대의 자료 수가 적을 때도 적합된 모형의 함수관계를 사용하여 안정적인 추정이 가능하다[11]. 또한 로지스틱 회귀모형은 변수 간의 상호작용에 관련한 가설 검정을 쉽게 실행할 수 있다. R, SPSS, SAS와 같이 현재 널리 사용되고 있는 통계 분석 프로그램에는 로지스틱 회귀분석을 수행하는 데 필요한 함수들이 잘 구현되어 있어, 표준화에 필요한 계산도 빠르고 손쉽게 수행할 수 있다.

이 연구에서는 로지스틱회귀 모형을 사용해 비편향 추정값인 직접 표준화율을 계산할 수 있음을 예시를 통해 보이고, 오픈 소스인 R을 사용하여 직접 표준화율을 어떻게 실제적으로 계산할 수 있는지 분석 코드를 제시하고자 한다. 예시를 위해, 국민건강보험공단의 건강검진코호트(The National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS)) [12]를 사용하여 대사증후군의 표준화 유병률을 계산해 본다. 먼저, 직접 표준화 방법을 사용하여 대사증후군 관련 연령대별 성별 조유병률(age-sex specific prevalence rate, ASSPR), 표준화 유병률(adjusted prevalence rate, APR), 표준오차 및 신뢰구간을 단계별로 나누어 계산해 보고, 로지스틱 회귀모형을 사용하여 동일한 결과를 도출할 수 있음을 보여준다. 더 나아가 로지스틱 회귀모형에 연령을 연속형 변수로 포함시켰을 때, 표준화 유병률의 표준오차가 줄어들게 되어, 추정량의 신뢰도를 높일 수 있음을 보여준다[13]. 본 연구에서는 일반적인 통계 분석부터 대용량 자료 분석에도 적합한 R 프로그램[14]을 사용하고, 실제적으로 회귀모형을 표준화율을 산출하는 데 적용해 보기 원하는 연구자를 고려하여 분석에 사용한 코드도 함께 제공한다.

연구 방법

대사증후군의 정의

본 연구에서는 National Cholesterol Education Program—Third Adult Treatment Panel (NCEP ATP III)의 진단기준[15]을 수정하여, 2005년 미국심장협회와 국립심폐혈관연구소에서 제시한 modified ATP III [16]에 따라 다음의 5개 대사 위험요인들 중 세 가지 이상에 해당되는 대상을 대사증후군으로 진단하였다.

- 복부비만: 허리둘레 남자 ≥90 cm, 여자 ≥80 cm
- 혈중 고중성지방: ≥150 mg/dL 또는 고중성지질혈증 치료제 복용
- 저하된 고밀도지단백 콜레스테롤: 남자 <40 mg/dL, 여자 <50 mg/dL 또는 저·고밀도지질단백질 콜레스테롤혈증을 치료제 복용
- 상승된 혈압: ≥130/85 mmHg 혹은 고혈압 치료제 복용
- 상승된 혈당: ≥100 mg/dL 혹은 당뇨병 치료제 복용

위의 복부비만 항목의 허리둘레 절단점은 대한비만학회의 1998년 국민건강영양조사 자료를 이용한 연구 결과[17]를 적용하였다.

연구대상

국민건강보험공단은 전 국민의 자격, 건강검진결과, 진료내역, 요양기관 현황 등의 방대한 자료를 보유하고 있으며, 정책수립과 학술연구에 국민건강정보가 활용될 수 있도록 국민건강보험자료공유서비스(National Health Insurance Sharing Service, NHISS) [18]를 통해 다양한 표본연구 DB (sample cohort database) 제공을 계획하고 있으며, 2014년 표본코호트 DB를 시작으로 2016년 건강검진코호트 DB와 노인코호트 DB를 순차적으로 공개해 왔다. 건강검진코호트 DB는 2002년 12월 말 기준 40세에서 79세 사이의 건강보험 자격 유지자 중 약 10%인 51만 명에 대한 2002-2013 (12개년) 동안의 자격 및 소득정보(사회경제적 변수), 병·의원 이용 내역 및 건강검진결과, 요양기관 정보를 포함하고 있는 코호트이다. 코호트와 관련한 자세한 설명은 국민건강보험공단의 빅데이터운영실에서 작성한 건강검진코호트 DB 사용자 매뉴얼[19]을 참조하면 된다.

본 연구에서는 건강검진코호트 DB에서 2013년 건강검진수검자 총 202,064명 중 대사증후군 진단이 가능한 201,845명을 최종 연구 대상으로 하였다. 대사증후군 진단을 위해서는 코호트를 구성하고 있는 아래에 나열된 세부 DB에 포함된 대사 위험요인들과 관련된 변수를 사용하였다.

- 자격 DB: 성(SEX)
- 건강검진 DB: 허리둘레(WAIST), 중성지방(TRIGLYCERIDE), HDL콜레스테롤(HDL_CHOLE), 수축기혈압(BP_HIGH), 이완기혈압(BP_LWST), 공복혈당(BLDS)

• 진료 DB: 진료내역(30t) 중 약제코드(GNL_NM_CD)

표준인구로는 국가통계포털(Korean statistical information service, KOSIS) 사이트(http://kosis.kr)에서 제공하고 있는 2013년 주민등록원앙인구(5세별, 1세별) 자료를 추출하여 사용하였다.

연령별 성별 조유병률

이 논문에서는 (i=1, 2, ..., I)는 총 I개의 연령대를 나타내는데, j (j=1: 남성, j=2: 여성)는 성(gender)을 나타내는 데 각각 사용한다. 연령대는 전체 연구대상자를 연령에 따라 5세별로 범주화하여 정의하였다. 연령대별 성별 조유병률은 각 성별 연령대에 해당하는 대사증후군 환자수를 전체인구 수로 나눈 값으로(일반적으로 인구 100,000명당 비율로 표시하나) 이 연구에서는 인구 100명당 비율로 정의하며 다음의 식을 사용하여 계산한다[7,8].

$$ASSPR \equiv \hat{p}_{ij} \times 100 = \frac{c_{ij}}{n_{ij}} \times 100$$

- c_{ij}: i번째 연령대에 해당하는 j번째 성별 대사증후군 환자 수
- n_{ij}: i번째 연령대에 해당하는 j번째 성별 전체인구 수
- $\hat{p}_{ij}$: i번째 연령대에서 j번째 성별 환자가 대사증후군에 걸릴 확률(p_{ij})의 추정량. 여기서, $\hat{p}_{ij}$은 비편향 추정량. 즉, $E(\hat{p}_{ij})=p_{ij}$

직접 표준화(direct standardization) 방법

표준화 유병률은 각 성별 연령대에 해당하는 조유병률에 표준인구의 비율을 가중치로 주어 산출한 가중평균으로 정의되며, 다음의 식을 사용하여 계산한다[7,8].

$$APR \equiv \sum_{i,j} \frac{N_{ij}}{N} \hat{p}_{ij} = \sum_{i,j} w_{ij} \hat{p}_{ij}$$

- N_{ij}: i번째 연령대에 해당하는 j번째 성별 표준인구 수
- N: 전체표준인구 수. 즉, $N = \sum_{i,j} N_{ij}$
- w_{ij}: i번째 연령대에 해당하는 j번째 성별 표준인구의 비율. 즉, $w_{ij} = \frac{N_{ij}}{N}$

각 성별 연령대에서 환자가 대사증후군에 걸릴 확률분포는 동일한 베르누이분포(Bernoulli distribution)를 따른다는 가정과 환자들 사이의 독립성 가정 아래에서, 표준화 유병률의 분산은 다음과 같이 계산할 수 있으며,

$$Var(APR) = \sum_{i,j} w_{ij}^2 \, Var(\hat{p}_{ij}) = \sum_{i,j} w_{ij}^2 \frac{p_{ij}(1-p_{ij})}{n_{ij}}$$

분산의 추정값 $\widehat{Var}(APR)$은 위의 식에 p_{ij} 대신 $\hat{p}_{ij}$ 추정값을 대입하여 계산하면 된다. 그리고 표준화 유병률의 95% 신뢰구간은 다음과 같이 계산할 수 있다.

$$\left(APR - 1.96 \sqrt{\widehat{Var}(APR)}, APR + 1.96 \sqrt{\widehat{Var}(APR)} \right)$$

로지스틱 회귀모형을 이용한 표준화 방법

대사증후군 환자의 개별 자료를 사용할 수 있을 때, 다음의 로지스틱 회귀모형을 사용하여 표준화 유병률을 계산할 수 있다.

$$\text{logit}(p_{ij}) = \ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 \cdot age\_group_i + \beta_2 \cdot sex_j + \beta_{12} \cdot (age\_group_i \times sex_j)$$

- $\beta_0, \beta_1, \beta_2, \beta_{12}$: 미지(unknown)의 모수

위의 로지스틱 회귀모형을 자료에 적합하여 각 성별 연령별 로짓 추정값($\hat{\pi}_{ij}$=logit($\hat{p}_{ij}$))을 계산할 수 있으며, 로짓함수(logit function)의 역변환(back transformation)을 통해 조유병률을 다음과 같이 계산할 수 있다.

$$ASSPR = \hat{p}_{ij} \times 100 = \frac{\exp(\hat{\pi}_{ij})}{1+\exp(\hat{\pi}_{ij})} \times 100$$

다음으로 조유병률 값과 표준인구 비율을 아래의 식에 대입하여 표준화 유병률을 계산할 수 있다.

$$APR \equiv \sum_{i,j} \frac{N_{ij}}{N} \hat{p}_{ij} = \sum_{i,j} w_{ij} \hat{p}_{ij}$$

또한 로짓 추정량($\hat{\pi}_{ij}$)의 표준오차값을 사용하여 델타 방법(delta method) [20]을 대표본 가정 아래에서 적용하면, 유병률 추정량($\hat{p}_{ij}$)의 표준오차를 다음과 같이 추정할 수 있으며,

$$s.e.(\hat{p}_{ij}) = \hat{p}_{ij}(1-\hat{p}_{ij}) \times s.e.(\hat{\pi}_{ij})$$

- s.e.: standard error.

더 나아가 APR의 표준오차와 95% 신뢰구간을 다음과 같이 계산할 수 있다.

$$s.e.(APR) = \sqrt{\sum_{i,j} w_{ij}^2 \left(s.e.(\hat{p}_{ij})\right)^2} = \sqrt{\sum_{i,j} w_{ij}^2 \left(\hat{p}_{ij}(1-\hat{p}_{ij}) \times s.e.(\hat{\pi}_{ij})\right)^2}$$

(APR-1.96×$s.e.$(APR), APR+1.96×$s.e.$(APR))

Table 1. Characteristics of the study population according to metabolic syndrome (MetSyn) status (n=201,845)

	No MetSyn (n=158,847)		MetSyn (n=42,998)		*p*-value
	Men (n=88,213)	Women (n=70,634)	Men (n=23,931)	Women (n=19,067)	
Age (y)	60.6±8.2	61.4±8.4	61.1±8.3	65.6±9.0	<0.001
Age group					<0.001
50-54	25,614 (29.0)	18,284 (25.9)	6,315 (26.4)	2,544 (13.3)	
55-59	23,184 (26.3)	17,446 (24.7)	6,269 (26.2)	3,297 (17.3)	
60-64	16,787 (19.0)	15,580 (22.1)	4,742 (19.8)	4,185 (21.9)	
65-69	5,973 (6.8)	3,896 (5.5)	1,679 (7.0)	1,470 (7.7)	
70-74	9,998 (11.3)	9,306 (13.2)	2,975 (12.4)	4,264 (22.4)	
75-79	4,161 (4.7)	3,778 (5.3)	1,276 (5.3)	2,041 (10.7)	
80-84	2,083 (2.4)	1,989 (2.8)	573 (2.4)	1,055 (5.5)	
≥85	413 (0.5)	355 (0.5)	102 (0.4)	211 (1.1)	
Waist (cm)	82.8±6.9	77.4±7.5	90.4 ± 7.0	85.7±7.9	<0.001
Triglyceride (mg/dL)	123.3±73.5	108.8±56.5	217.6±116.2	187.9±93.5	<0.001
HDL_cholesterol (mg/dL)	53.4±14.8	58.9±15.4	44.3±15.0	47.2±11.2	<0.001
Systolic blood pressure (mmHg)	123.9±13.6	121.3±14.3	133.2±13.4	133.1±14.2	<0.001
Diastolic blood pressure (mmHg)	77.0±9.2	74.6±9.2	81.8±9.4	80.0±9.3	<0.001
Fasting blood sugar (mg/dL)	100.6±22.3	96.1±17.3	120.9±36.0	113.6±33.7	<0.001

MetSyn, metabolic syndrome.

Roalfe et al. [13]에서도 로지스틱 회귀모형을 사용하여 표준화율을 산출하는 방법을 제안하였다. 하지만 이 연구에서 제안하는 것과 같이 표준인구 비율을 로짓 추정량($\hat{\pi}_{ij}$)에 가중치로 부여하게 되면, APR의 추정량은 편향되게 되며, 직접 표준화 방법을 적용하여 계산한 APR의 추정값과 일치하지 않게 된다. 또한 이 연구에서 제시된 신뢰구간이 APR의 비편향 추정량을 포함한다는 것이 보장될 수 없다.

연구 결과

본 연구에서는 국민건강보험공단 건강검진코호트 DB의 2013년 자격 DB에 포함된 건강검진 수검자 202,064명 중 결측값(missing value)으로 인해 대사증후군을 진단할 수 없는 대상자 219명을 제외한 201,845명을 최종 분석 대상으로 하였다. 전체 연구대상자 중 남성의 비율은 55.6%(112,114명)로 여성에 비해 상대적으로 높았다(Table 1). 평균연령은 여성 대사증후군 환자(65.6±9세)에서 가장 높았으며, 대사증후군 환자 중에서 남자는 50대(52.6%)가, 여자는 70대(33.1%)가 가장 많았다. 정상 군에 비해 대사증후군 환자에서 허리둘레, 중성지방, 수축기 혈압, 이완기 혈압, 공복혈당이 더 높았으며($p<0.001$), HDL 콜레스테롤은 정상 군에서 더 높았다($p<0.001$).

직접 표준화 방법을 사용하여 표준화 유병률을 계산하기 위해 먼저 전체 연구대상자를 성별, 연령대별로 분류하였고, 계산에 필요한 각 단계를 Table 2에 정리하였다. 전체 대사증후군의 조유병률은 21.3명(100명당)이었으며, 성별에 따라 차이를 보이지 않았다(Table 2). 전체 대사증후군 표준화 유병률은 21.9명(100명당)이었으며, 성별에 따른 연령 구조를 고려한 후 여자가 11.9명(100명당)으로 남자의 10명(100명당)보다 더 높았다. 남자의 경우 표준화 유병률이 70대까지 증가하다가 이후 감소하였고, 여자의 경우 표준화 유병률이 연령대가 높아질수록 증가하였다(Table 2). Table 2의 결과를 사용하여, 표준화 유병률, 표준오차, 신뢰구간을 다음과 같이 계산하였다.

$$APR = \sum_{i,j} w_{ij}\hat{p}_{ij} = 21.9\text{명 (100명당)}$$

$$s.e.(APR) = \sqrt{\widehat{Var}(APR)} = \sqrt{\sum_{i,j} w_{ij}^2\,\widehat{Var}(\hat{p}_{ij})} =$$

$$\sqrt{0.010775} = 0.1038\text{명 (100명당)}$$

95% 신뢰구간 = 21.9 ± 1.96 × 0.1038 = (21.66, 22.06)명 (100명당)

먼저 로지스틱 회귀모형을 사용하여 직접 표준화 방법으로 계산한 결과를 손쉽게 도출할 수 있음을 보여주기 위해, 전체 연구대상자를 연령에 따라 50세부터 5세별로 범주화한 연령대(age_gp) 변수(50-54, 55-59, …, 80-84, ≥85)를 정의하였다. 또한 로지스틱 회귀분석을 위해 대사증후군 상태를 나타내는 이분형 종속변수인 ms (=0: 정상, =1: 대사증후군)를 정의하였으며, 연령대, 성(sex), 그리고 두 변수 간의 상호작용항을 독립변수로 모형에 포함시켰다. 분석에는 R 통계 프로그램을 사용하였으며, 분석 코드는 간단한 주석과 함께 Table 3에 기술하였다.

로지스틱 회귀모형을 사용하여 계산한 표준화 유병률(100명당 21.9

Table 2. Standardized rates calculated by the direct standardization method for metabolic syndrome (MetSyn)

Age group	Sex	c_{ij}	n_{ij}	ASSPR	N_{ij}	W_{ij}	$W_{ij}\hat{p}_{ij}\times 100$	$w_{ij}^2\widehat{Var}(\hat{p}_{ij})\times 10{,}000$
50-54	m	6,315	31,929	19.8	2,187,740	0.135	2.7	0.000902
55-59	m	6,269	29,453	21.3	1,721,495	0.106	2.3	0.000639
60-64	m	4,742	21,529	22	1,186,814	0.073	1.6	0.000426
65-69	m	1,679	7,652	21.9	908,015	0.056	1.2	0.0007
70-74	m	2,975	12,973	22.9	779,034	0.048	1.1	0.000313
75-79	m	1,276	5,437	23.5	479,575	0.03	0.7	0.000288
80-84	m	573	2,656	21.6	217,303	0.013	0.3	0.000114
≥85	m	102	515	19.8	110,082	0.007	0.1	0.000142
Subtotal		23,931	112,144	21.3	7,590,057		10	0.003524
50-54	f	2,544	20,828	12.2	2,148,726	0.132	1.6	0.000901
55-59	f	3,297	20,743	15.9	1,728,364	0.106	1.7	0.00073
60-64	f	4,185	19,765	21.2	1,236,338	0.076	1.6	0.000489
65-69	f	1,470	5,366	27.4	1,012,218	0.062	1.7	0.00144
70-74	f	4,264	13,570	31.4	993,280	0.061	1.9	0.000594
75-79	f	2,041	5,819	35.1	742,665	0.046	1.6	0.000818
80-84	f	1,055	3,044	34.7	461,035	0.028	1	0.006
≥85	f	211	566	37.3	327,349	0.02	0.8	0.001678
Subtotal		19,067	89,701	21.3	8,649,974		11.9	0.007251
Total		42,998	201,845	21.3	16,240,030		21.9	0.010775

MetSyn, metabolic syndrome; m, male; f, female.

c_{ij}, number of MetSyn; n_{ij}, population; ASSPR, age-sex specific prevalence rate per 100, i.e., $\hat{p}_{ij}\times 100=\frac{c_{ij}}{n_{ij}}\times 100$; N_{ij}, standard population; N, total standard population, ie., $N=\sum N_{ij}$; $w_{ij}=\frac{N_{ij}}{N}$; APR, adjusted prevalence rate per 100, i.e., $\sum_{i,j}\frac{N_{ij}}{N}\hat{p}_{ij}\times 100=\sum_{i,j}w_{ij}\hat{p}_{ij}\times 100$, $\widehat{Var}(\hat{p}_{ij})=\frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{n_{ij}}$.

Table 3. Illustrative R codes for calculating an adjusted prevalence rate by logistic regression analysis: age group as a categorical variable

```
> str(mydat) #print the data structure of the object "mydat" in a data frame
'data.frame':          201845 obs. of  4 variables:
 $ ms    : num  0 0 0 0 0 0 0 0 0 1 0 ...
 $ age   : int  90 90 90 90 90 90 90 90 90 90 ...
 $ sex   : Factor w/ 2 levels "1","2": 1 2 1 2 2 2 1 1 1 2 1 ...
 $ age_gp: Factor w/ 8 levels "[50,55)","[55,60)",..: 8 8 8 8 8 8 8 8 8 8 ...
> mydat$age_gp <- cut(mydat$age, c(seq(50,85,by=5),200), right = F) #assigns the values of the numeric vector (continuous variable) "age" into one of age groups: [50-54], ..., [80-84], [85 or above], converting "age" into the factor (categorical variable) "age_gp"
> levels(mydat$age_gp) #print the levels (age groups) of the factor "age_gp"
[1] "[50,55)" "[55,60)" "[60,65)" "[65,70)" "[70,75)" "[75,80)" "[80,85)" "[85,200)"
> fit1 <- glm( ms ~ age_gp * sex, family = binomial(), data = mydat ) #fit a logistic regression model to the data and assign the fitted results into the object "fit1"
> mm <- length(levels(mydat$age_gp)) #assign the number of age groups into the object "mm"
> pred.dat1 <- data.frame(sex = as.factor(c(rep(1,mm),rep(2,mm))), age_gp = rep(levels(mydat$age_gp),2))
#construct a data frame "pred.dat1" that contains all combinations of "sex" and "age_gp" values
> pred.res1 <- predict(fit1, pred.dat1, se.fit = T, type = "response")
#calculate the estimates of age-sex specific prevalence rates by using the fitted model "fit1"
and assign into the object "pred.res1"
> as.std.pop1 <- c(2187739.5, 1721494.5, 1186814.0, 908015.0, 779034.0, 479575.0, 217303.0, 110081.5, 2148725.5, 1728363.5, 1236338.0, 1012218.0, 993280.0, 742665.0,  461035.0, 327348.5)
#assign the mid-year population in 2013 for age groups into the object "as.std.pop1"
> as.std.pop.prop1 <- as.std.pop1/sum(as.std.pop1)
#calculate the proportions of the mid-year population for age groups and assign into the object "as.std.pop.prop1"
> round( sum(pred.res1$fit * as.std.pop.prop1) * 100, 2) #calculate the adjusted prevalence rate rounded off to two decimal digits
[1] 21.86
> round( sqrt(sum(pred.res1$se^2 * as.std.pop.prop1^2)) * 100, 4) #calculate the standard error rounded off to four decimal digits
[1] 0.1038
> round( ( sum(pred.res1$fit * as.std.pop.prop1) + c(-1,1) * 1.96 * sqrt(sum(pred.res1$se^2 * as.std.pop.prop1^2)) ) * 100, 2 ) #calculate the 95% confidence interval rounded off to two decimal digits
[1] 21.66 22.06
```

Table 4. Logistic regression model for metabolic syndrome: age as a continuous variable

Parameter	df	Estimate	SE	95% CI	Wald z-statistic	*p*-value=Pr(>|z|)
Intercept	1	-1.7213	0.0536	(-1.8263, -1.6161)	-32.1098	<0.001
Age	1	0.0068	0.0009	(0.0051, 0.0085)	7.8632	<0.001
Sex	1	-2.9197	0.0803	(-3.0771, -2.7624)	-36.3716	<0.001
Age:Sex	1	0.0457	0.0013	(0.0432, 0.0482)	36.1259	<0.001

SE, standard error; CI, confidence interval.

Table 5. Illustrative R codes for calculating an adjusted prevalence rate by logistic regression analysis: age as a continuous variable

```
> fit2 <- glm( ms ~ age * sex, family = binomial(), data = mydat ) #fit a logistic regression model (including "age" as a continuous variable) to the data and assign the fitted results into the object "fit2"
> sort(unique(mydat$age)) #sort unique ages in an increasing order
 [1] 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90
> mm <- length(51:90) #assign the length of the vector (51, ..., 90) into the object "mm"
> pred.dat2 <- data.frame(sex = as.factor(c(rep(1,mm),rep(2,mm))), age = rep(c(51:90),2)) #construct a data frame "pred.dat2" that contains all combinations of "sex" and "age" values
> pred.res2 <- predict(fit2, pred.dat2, se.fit = T, type = "response") #calculate the estimates of age-sex specific prevalence rates by using the fitted model "fit2"and assign into the object "pred.res2"
> round( sum(pred.res2$fit * as.std.pop.prop2) * 100, 2) #calculate the adjusted prevalence rate rounded off to two decimal digits where as.std.pop.prop2 is the mid-year population proportion in 2013
[1] 22.26
> round( sqrt(sum(pred.res2$se^2 * as.std.pop.prop2^2)) * 100, 4) #calculate the standard error rounded off to four decimal digits
[1] 0.0235
> round( ( sum(pred.res2$fit * as.std.pop.prop2) + c(-1,1) * 1.96 * sqrt(sum(pred.res2$se^2 * as.std.pop.prop2^2)) ) * 100, 2 ) #calculate the 95% confidence interval rounded off to two decimal digits
[1] 22.22 22.31
```

명), 표준오차(0.1038), 95% 신뢰구간(21.66, 22.06)은 직접 표준화 방법을 사용하여 계산한 이전 결과와 동일하였다.

다음으로 로지스틱 회귀모형에 범주형 변수인 연령대(age_gp) 대신 연령(age)을 연속형 변수로 포함하여 자료에 적합하였다(Table 4).

적합된 회귀모형을 사용하여 각 성별 연령별 로짓 추정값과 표준화 유병률, 표준오차, 95% 신뢰구간(22.22, 22.31)을 계산하였고, 분석에 사용한 R 코드는 주석과 함께 Table 5에 기술하였다.

연령을 연속형 변수로 모형에 포함하였을 때, 표준화 유병률(100명당 22.3명)은 크게 달라지지 않았지만, 표준오차(0.024)가 확연히 줄어들었으며, 결과적으로 신뢰구간의 길이도 짧아졌다.

고 찰

로지스틱 회귀모형은 대부분 연관성 분석이나 예측 모형 개발에 사용되고 있으나, 본 연구에서는 로지스틱 회귀모형을 사용하여 유병률을 표준화하는 방법을 소개하였다. 직접 표준화 방법은 범주형 자료에만 사용할 수 있으나, 로지스틱 회귀모형을 통한 표준화는 범주형 변수뿐만이 아닌 연속형 변수를 사용해서도 추정할 수 있다. 연령 변수를 범주화하여 로지스틱 회귀모형에 사용하였을 때는 직접 표준화 방법을 사용했을 때와 동일한 결과를 산출할 수 있고, 연령을 연속형 변수로 모형에 포함하였을 때는 추정량의 표준오차가 줄어들어 신뢰구간의 길이가 짧아지며, 결과적으로 더 신뢰할 수 있는 추정값을 얻을 수 있다. 현재 널리 사용되고 있는 통계 분석 프로그램에는 로지스틱 회귀분석을 수행하는 데 필요한 함수들이 잘 구현되어 있어서, 대용량 자료를 분석할 때, 예를 들어, 다양한 질병의 유병률을 여러 지역별로 산출할 때, 로지스틱 모형을 활용한 표준화 방법이 계산상 효율적일 수 있다. 또한 로지스틱 회귀모형을 통한 표준화 방법은 추정된 모형의 함수관계를 사용하여 자료 수가 부족한 특정 연령대의 유병률도 상대적으로 안정적인 추정을 할 수 있다.

로지스틱 회귀모형에 연령을 연속형 변수로 포함하여 표준화율을 산출하는 방법은 Roalfe et al. [13]에서도 제안되었다. 그렇지만 해당 논문에서는 추정된 로짓에 표준화 인구 분포를 가중치로 부여하여, 결과적으로 표준화 유병률의 추정량은 편향되게 되며, 더 나아가 제시된 신뢰구간이 직접 표준화 방법에 의해 계산된 비편향 추정량을 항상 포함한다는 보장이 없게 된다. 이러한 측면에서 본 연구의 방향과 근본적인 차이가 있다.

로지스틱 회귀모형을 사용한 표준화는 직접 표준화 방법과 비교하여 여러 가지 장점을 갖고 있지만, 직접 표준화에 비해 많이 사용되고 있지 않다. 이 논문을 통해 직접 표준화 방법을 적용하기에 적절하지 않은 상황에서 합리적 대안으로 로지스틱 회귀모형이 널리 사용되기

를 기대한다. 또한 본 논문에서 예시로 사용한 자료 분석 결과(Table 2)에서 고연령층에서 더 급격한 유병률의 변화는 관측되지 않았으나, 일반적으로 고연령층에서 유병률의 증가가 더 심할 수 있으므로, 연령층에 따른 유병률의 증가 속도의 변화를 반영할 수 있는 좀 더 유연한 모형에 대한 추후 연구가 필요할 것이다.

REFERENCES

1. Kang HJ, Kwon S. Regional disparity of cardiovascular mortality and its determinants 2016;26(1):12-23.
2. Gribbin B, Pickering TG, Sleight P, Peto R. Effect of age and high blood pressure on baroreflex sensitivity in man. Circ Res 1971;29(4):424-431.
3. Cho KI, Cho SH, Her A, Singh GB, Shin E. Prognostic utility of neutrophil-to-lymphocyte ratio on adverse clinical outcomes in patients with severe calcific aortic stenosis.PLoS ONE 2016;11(8):e0161530.
4. Kwon GY, Lim DS, Park EJ, Jung JS, Kang KW, Kim YA, et al. Assessment of applicability of standardized rates for health state comparison among areas: 2008 community health survey. J Prev Med Public Health 2010;43(2):174-184 (Korean).
5. Neison FGP. On a method recently proposed for conducting inquiries into the comparative sanatory condition of various districts, with illustrations, derived from numerous places in great britain at the period of the last census. J Statistical Society of London 1844;7(1):40-68.
6. Inskip H, Beral V, Fraser P, Haskey J. Methods for age-adjustment of rates. Stat Med 1983;2(4):455-466.
7. Chiang CL. Standard error of the age-adjusted death rate. US Department of Health, Education, and Welfare, Public Health Service, National Vital Statistics Division, 1961.
8. Curtin LR, Klein RJ. Direct standardization (age-adjusted death rates). US Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics, 1995.
9. McCulloch CE, Neuhaus JM. Generalized linear mixed models. Wiley Online Library, 2001.
10. Kim B, Cho S, Cho K, Kim H, Heo J, Cha T. The combined impact of neutrophil-to-lymphocyte ratio and type 2 diabetic mellitus on significant coronary artery disease and carotid artery atherosclerosis. J Cardiovasc Ultrasound 2016;24(2):115-122.
11. Wilcosky TC, Chambless LE. A comparison of direct adjustment and regression adjustment of epidemiologic measures. J Chronic Dis 1985;38(10):849-856.
12. Lee J, Lee JS, Park SH, Shin SA, Kim K. Cohort profile: the national health insurance service-national sample cohort (NHIS-NSC), South Korea. Int J Epidemiol 2015;1-8.
13. Roalfe AK, Holder RL, Wilson S. Standardisation of rates using logistic regression: a comparison with the direct method. BMC Health Serv Res 2008;8(1):275.
14. R Development Core Team. R: A language and environment for statistical computing. R foundation forstatistical computing, Vienna, Austria, 2016. Available at https://www.R-project.org/.
15. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive summary of the third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). JAMA 2001;285(19):2486-2497.
16. Grundy SM, Cleeman JI, Daniels SR, Donato KA, Eckel RH, Franklin BA, et al. Diagnosis and management of the metabolic syndrome: An american heart Association/National heart, lung, and blood institute scientific statement. Circulation 2005;112(17):2735-2752.
17. Lee SY, Park HS, Kim DJ, Han JH, Kim SM, Cho GJ, et al. Appropriate waist circumference cutoff points for central obesity in korean adults. Diabetes Res Clin Pract 2007;75(1):72-80.
18. National Health Insurance Sharing Service (NHISS). Available at https://nhiss.nhis.or.kr [accessed January 20, 2017].
19. National Health Insurance Service-National Medical Examination Sample Cohort (NHIS-NMES) User manual https://nhiss.nhis.or.kr/bd/ab/bdaba006cv.do [accessed January 20, 2017].
20. Casella G, Berger RL. Statistical inference (2nd ed). Pacific Grove, CA: Duxbury/Thomson Learning, 2002.