The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.5.555
33(5), 555–567

Comparison of nomograms designed to predict hypertension with a complex sample

Min Ho Kim[a] · Min Seok Shin[a] · Jea Young Lee[a,1]

[a]Department of Statistics, Yeungnam University

(Received June 3, 2020; Revised July 9, 2020; Accepted August 6, 2020)

Abstract

Hypertension has a steadily increasing incidence rate as well as represents a risk factors for secondary diseases such as cardiovascular disease. Therefore, it is important to predict the incidence rate of the disease. In this study, we constructed nomograms that can predict the incidence rate of hypertension. We use data from the Korean National Health and Nutrition Examination Survey (KNHANES) for 2013–2016. The complex sampling data required the use of a Rao-Scott chi-squared test to identify 10 risk factors for hypertension. Smoking and exercise variables were not statistically significant in the Logistic regression; therefore, eight effects were selected as risk factors for hypertension. Logistic and Bayesian nomograms constructed from the selected risk factors were proposed and compared. The constructed nomograms were then verified using a receiver operating characteristics curve and calibration plot.

Keywords: hypertension, logistic regression, naïve Bayesian classifier, nomogram, risk factor

1. 서론

고혈압은 수축기 혈압 또는 이완기 혈압이 비정상적으로 높은 질병으로 한국에서는 2011년 30세 이상 성인 중 약 28.5% 발병하였고, 65세 이상 성인 중에서는 남성 58.4%, 여성 61.8%가 발병하였다 (Shin 등, 2015). 고혈압은 관상동맥질환, 심부전증, 뇌졸중, 혈관성 치매와 같은 심혈관계 질환으로 발전할 수 있으며 (Van den Berg 등, 2009), 이런 심혈관계질환은 2017년 한국 인구 10만 명당 119.6명으로 한국인 사망 원인 중 2위를 차지하였다 (Nam 등, 2018; Statistics Korea, 2018). 본인이 고혈압임을 인지하지 못하는 환자들이 많고 고혈압이 다른 합병증을 유발하는 질병이므로 인식과 예방이 중요시 된다.

따라서 위험 요인의 인식과 질병 예측에 도움을 줄 수 있는 도구나 방법의 발전이 중요하다. 이를 도울 수 있는 통계적 도구 중 하나가 바로 노모그램(nomogram)이다. 노모그램은 분석을 통해 예측한 확률을 시각적으로 설명하기 위해 만들어진 그래프이다. 노모그램의 가장 큰 장점은 위험 요인을 한 눈에 확인할 수 있고, 개개인의 특징을 바탕으로 질병이 발생할 확률을 점수를 통해 바로 예측할 수 있다는 점이다 (Mozina 등 2004). 이러한 노모그램은 raw data를 이용하여 당뇨와 이상지질혈증에 대해 구축된 바 있다 (Park 등, 2018; Kim 등, 2019). 질병의 위험 요인들을 규명하기 위해 진행된 연구들은 대부분 질병의 발병률을 예측하는 통계적 모형으로 로지스틱 회귀모형이나 Cox 비례위험모형을 많이 사용해왔다.

[1]Corresponding author: Department of Statistics, Yeungnam University, 280 Daehak-Ro, Gyeongsan, Gyeongbuk 38541, Korea. E-mail: Jlee@yu.ac.kr

이처럼 고혈압에서도 위험 요인을 선별하는데 다양한 선행연구가 진행 되었지만 실제로 통계학적 지식이 부족한 비전공자들은 분석 결과만으로 실제 고혈압의 위험 정도를 인지하는데 어려움이 있다. 그래서 본 연구에서는 고혈압의 노모그램을 로지스틱 회귀모형과 순수 베이지안 분류기 모형으로 각각 구축하였고 (Kim, 2020; Kim과 Lee, 2020), 비교를 통해 두 노모그램의 유용성을 분석하였다. 본 논문의 2절에서는 고혈압의 위험요인을 선별하고 위험도를 추정하는 방법인 Rao-Scott chi-squared test와 복합 표본 하에서 로지스틱 회귀모형과 순수 베이지안 분류기 모형을 이용한 노모그램 구축과 검증 방법을 소개한다. 3절에서는 국민건강영양조사 데이터에 대한 설명과 2절에서 설명한 방법을 적용한 고혈압의 발병률을 예측하는 노모그램을 구축하고 검증한다. 마지막 4절에서는 구축한 노모그램에 대한 의견 및 결론을 제시한다.

2. Methodology

2.1. Rao-Scott chi-squared test

고혈압의 위험요인을 선정할 때 실제로 고혈압 유무에 따라 위험요인의 영향이 있는지 밝히는 과정은 필수적이다. 일반적으로 Pearson chi-squared test를 사용하여 고혈압의 위험요인을 선별한다.

Pearson chi-squared test는 고혈압과 위험 요소로 구성된 분할표에서 각 셀에 들어갈 빈도가 독립이라는 가정이 필요하다. 그러나 본 연구에서 사용된 데이터는 2단계 층화집락추출법을 사용했으므로, 분할표의 각 셀이 독립이라는 가정을 만족하지 못한다. 이 때 Pearson chi-squared test를 사용하면 검정통계량 값이 과도하게 커지므로 귀무가설을 쉽게 기각한다고 알려져 있다 (Rao와 Scott, 1981; Sung, 2012). 따라서 층, 집락, 가중치 등 설계 효과를 고려한 Rao-Scott chi-squared 통계량을 사용한다. Rao-Scott chi-squared 통계량은 다음과 같다.

$$\chi^2_{\text{Rao-Scott}} = \frac{\chi^2}{\hat{\delta}},$$

여기서 분자 χ^2은 Pearson chi-squared 통계량이고,

$$\hat{\delta} = \left[\sum_i \sum_j \left(1 - \hat{\pi}_{i+}\hat{\pi}_{+j}\right) \hat{d}_{ij} - \sum_i \left(1 - \hat{\pi}_{i+}\right) \hat{d}_{i+} - \sum_j \left(1 - \hat{\pi}_{+j}\right) \hat{d}_{+j} \right] \Big/ \left(I-1\right)\left(J-1\right),$$

$$\hat{d}_{ij} = \frac{\widehat{\text{Var}}\left(\hat{\pi}_{ij}\right)}{\hat{\pi}_{ij}\left(1-\hat{\pi}_{ij}\right)/n}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,$$

여기서 $\hat{\pi}_{ij}$는 추정된 i번째 행, j번째 셀의 확률이고, $\widehat{\text{Var}}(\hat{\pi}_{ij})$는 $\hat{\pi}_{ij}$의 추정된 분산, n은 표본의 수이다. 그리고 $\hat{d}_{ij}$는 $\hat{\pi}_{ij}$의 설계 효과 추정치이다.

2.2. Nomogram construction method

의료 분야에서 질병이나 사망과 관련된 위험 요인을 선별하고, 질병 발생률을 예측하는 연구들이 활발하게 진행되고 있다. 위험을 예측하기 위해 질병 또는 사망에 영향을 주는 위험인자를 선별하고, 어느 정도 영향을 주는지 계산을 하는 통계적 기법들을 사용한다. 하지만 비 전공자들이 통계적 결과 만으로 위험률을 계산한다는 것은 다소 어려움이 있다. 따라서 복잡한 계산 없이 한 눈에 여러 위험요인들에 의한 질병이나 사망의 위험률을 알 수 있는 노모그램을 제시한다 (Lee 등, 2009; Iasonos 등, 2008). 노모그램의 구성요소로는 4가지가 있다. 위에서부터 Point 선, Risk Factor 선, Probability 선, Total Point 선이 있다. 질병의 발생 확률은 질병과 관련된 요인들의 Risk Factor 선으로부터 얻은 점수의 합으로 Total Point를 구하고 이에 대응하는 확률을 Probability 선에서 도출함으로써 예측할 수 있다.

2.2.1. Nomogram construction of logistic regression model 로지스틱 회귀모형의 결과를 이용해 노모그램을 구축하는 방법은 다음과 같다 (Iasonos 등, 2008; Park 등, 2018).

- point 선

  Point 선은 0점에서 100점으로 구성된다.

- Risk factor 선

  로지스틱 회귀모형으로부터 도출된 회귀계수 β_{ij}값으로 LP_{ij}값을 계산한다. 독립변수 X가 범주형 변수이고 j개의 범주를 가지는 경우 $j-1$개의 가변수(dummy variable)를 갖는다. 이 때 기준 범주의 회귀계수는 0이다.

$$\mathrm{LP}_{ij} = \beta_{ij} \times X_{ij},$$
$$\mathrm{Point}_{ij} = \frac{\mathrm{LP}_{ij} - \min_j \mathrm{LP}_{ij}}{\max_j \mathrm{LP}_{*j} - \min_j \mathrm{LP}_{*j}} \times 100,$$

  여기서 β_{ij}는 i번째 위험요인의 j번째 범주의 회귀계수 값, X_{ij}는 i번째 위험요인의 j번째 범주의 속성값을 나타낸다. LP_{*j}는 추정된 회귀계수의 편차가 가장 큰 위험요인의 LP값을 나타낸다.

- Probability 선

  Probability 선은 0에서 1까지 확률을 적절한 기준으로 분할해 구간을 만든다.

- Total point 선

  Total point는 각 위험요소의 Point_{ij}들의 총합이다.

$$\text{Total Point} = \frac{100}{\max_j \mathrm{LP}_{*j} - \min_j \mathrm{LP}_{*j}} \sum_i \sum_j \left(\mathrm{LP}_{ij} - \min_j \mathrm{LP}_{ij} \right).$$

  이제 위 Probability 선의 각 값에 대응하는 Total point 값을 구하기 위해 로지스틱 회귀모형을 $\sum_{i,j} \mathrm{LP}_{ij}$에 대해 정리한 뒤, 위 식에 대입하면 다음과 같은 식이 도출된다.

$$\text{Total Point} = \frac{100}{\max_j \mathrm{LP}_{*j} - \min_j \mathrm{LP}_{*j}} \left(\log \frac{P(Y=1 \mid X=x)}{1-P(Y=1 \mid X=x)} - \beta_0 - \sum_i \sum_j \min_j \mathrm{LP}_{ij} \right).$$

  그 뒤 $P(Y=1|X=x)$에 Probability 선의 값을 대입하여 Total point 선을 구축한다.

2.2.2. Nomogram construction of naïve Bayesian classifier model

- point 선

  Point 선은 -100점에서 100점으로 구성된다.

- Risk factor 선

  순수 베이지안 분류기 모형으로 얻어진 $\log \mathrm{OR}(a_i = j)$를 이용해 Point_{ij}를 계산하면 다음과 같다.

$$\mathrm{Point}_{ij} = \frac{\log \mathrm{OR}(a_i = j)}{\max_{i,j} |\log \mathrm{OR}(a_i = j)|} \times 100.$$

- Probability 선

  Probability 선은 0에서 1까지 확률을 적절한 기준으로 분할해 구간을 만든다.

- Total point 선

  Total point는 각 위험요소의 Point_{ij}들의 총합이다.

$$\begin{aligned}\text{Total Point} &= \frac{100}{\max_{i,j}|\log \text{OR}(a_i = j)|}\sum_i\sum_j(\log \text{OR}(a_i = j)) \\ &= \frac{100}{\max_{i,j}|\log \text{OR}(a_i = j)|} \times \left(-\log\left(\frac{1}{P(Y=1|X=x)}-1\right) - \log\frac{P(Y=1)}{1-P(Y=1)}\right).\end{aligned}$$

  그 뒤 $P(Y=1|X=x)$에 Probability 선의 값을 대입하여 Total point 선을 구축한다.

2.2.3. Left-aligned method of nomogram for naïve Bayesian classifier model 순수 베이지안 분류기 모형은 점수가 −100∼100점이므로 left-aligned 방법을 적용한다. 이는 점수가 0∼100점이므로 로지스틱 노모그램과 비교하기 쉽다.

- point 선

  Point 선은 0점에서 100점으로 구성된다.

- Risk factor 선

  순수 베이지안 분류기 모형에서 적합시켜 도출된 $\log \text{OR}(a_i = j)$값으로 각 위험요인의 범주 별 Point_{ij}를 계산한 후 Point 선에 맞추어 정렬한다.

$$\text{Point}_{ij} = \frac{\log \text{OR}(a_i = j) - \min_{i,j}\log \text{OR}(a_i = j)}{\max_j \log \text{OR}(a_* = j) - \min_j \log \text{OR}(a_* = j)} \times 100.$$

- Probability 선

  Probability 선은 0에서 1까지 확률을 적절한 기준으로 분할해 구간을 만든다.

- Total point 선

  Total point는 각 위험요소의 Point_{ij}들의 총합이다.

$$\begin{aligned}\text{Total Point} = &\frac{100}{\max_j \log \text{OR}(a_* = j) - \min_j \log \text{OR}(a_* = j)} \\ &\times \sum_{i,j}\left(\log \text{OR}(a_i = j) - \min_{i,j}\log \text{OR}(a_i = j)\right).\end{aligned}$$

  이제 위 Probability 선의 각 값에 대응되는 Total point 값을 구하기 위해 순수 베이지안 분류기 모형을 $\sum_{i,j}\log \text{OR}(a_i = j)$에 대해 정리한 뒤, 위 식에 대입하면 아래와 같은 식이 도출된다.

$$\begin{aligned}\text{Total Point} = &\frac{100}{\max_j \log \text{OR}(a_* = j) - \min_j \log \text{OR}(a_* = j)} \\ &\times \left(-\log\left(\frac{1}{P(Y=1|X=x)}-1\right) - \text{logit}P(Y=1) - \sum_{i,j}\min_{i,j}\log \text{OR}(a_i = j)\right).\end{aligned}$$

  그 뒤 $P(Y=1|X=x)$에 probability 선의 값을 대입하여 Total point 선을 구축한다.

2.3. Nomogram validation method

노모그램을 구축한 뒤, 노모그램의 정확성을 검증하기 위해 receiver operating characteristics (ROC) curve와 calibration plot을 사용하였다 (Akobeng, 2007; Cook, 2008). ROC curve는 X축은 1 − Specificity, Y축은 Sensitivity로 구성되어 있으며, curve 아래 면적의 넓이(area under curve; AUC)는 예측 정확도의 지표로 사용된다. 다른 도구로써 Calibration plot을 이용하는데, X축은 예측확률, Y축은 실제확률로 이루어져 있다. 만약 예측확률이 실제 확률과 가깝다면, Calibration plot은 45° 각도의 선에 가깝게 그려진다 (D'Agostino 등, 2001). 그리고 예측확률과 실제 확률간의 회귀직선의 적합도 지표인 R^2으로 노모그램을 검증한다. 모든 분석은 R software 3.6.1을 사용했고, 노모그램을 구축하는 툴은 SAS 9.4를 사용하였다 (SAS Institute Inc., Cary, NC, USA).

3. Applications

3.1. Complex sample materials

본 연구에 사용된 데이터는 국민건강영양조사(Korean National Health and Nutrition Examination Survey; KNHANES) 2013∼2016년도 자료이다. 국민건강영양조사는 표본의 대표성 및 추정의 정확성 향상을 위해 복합 표본 설계 방식인 2-staged stratified cluster sampling method을 사용하였다 (Korea Centers for Disease Control and Prevention, 2016). 고혈압의 진단 기준은 수축기 혈압이 140mmHg 이상이거나 이완기 혈압이 90mmHg 이상이거나 고혈압 약을 복용 중이거나 의사의 진단을 받은 사람으로 선정하였다.

사용된 위험요인은 총 10개로 고혈압 발병에 중요한 영향을 미치는 여러 선행 연구에서 선정하였다 (Kshirsagar 등, 2010). 위험요인은 나이, 성별, 흡연 상태, BMI, 고혈압 가족력, 당뇨병, 음주 유무, 운동 유무, 뇌졸중, 이상지질혈증이다. 나이는 20세 이상 44세 이하, 45세 이상 64세 이하, 65세 이상으로 범주화 하였다. 흡연 상태는 현재 흡연을 하는 그룹(present), 과거에 흡연을 했던 그룹(past), 그리고 흡연을 한적이 없는 그룹(no)으로 나누었다. BMI는 25 미만을 정상(normal), 25 이상 30 미만을 과체중(overweight), 그리고 30 이상은 비만(obese)으로 범주화 하였다. 고혈압 가족력 유무는 부모 및 형제 중 누구 한 명이라도 고혈압이 있을 경우를 yes로, 그렇지 않으면 no로 범주화 시켰다. 당뇨병 유무는 공복 혈당이 126mg/dL 이상이거나 의사 진단을 받았거나 혈당 강하제를 복용 중이거나 인슐린 주사를 투여 받은 사람을 yes로, 그렇지 않으면 no로 범주화 하였다. 음주 유무는 음주 경험이 없으면 no, 그렇지 않으면 yes로 범주화 하였다. 운동 유무는 일주일에 적어도 한번은 걷기 또는 근력 운동을 하면 yes로, 그렇지 않으면 no로 범주화 하였다. 뇌졸중 유무는 의사 진단 여부에 따라 범주화 하였다. 이상지질혈증 유무는 의사 진단을 받았거나, 총 콜레스테롤이 240mg/dL 이상이거나 콜레스테롤 강하제를 복용하거나 HDL콜레스테롤이 40mg/dL 미만이거나 고중성지방이 200mg/dL 이상인 경우 중 하나라도 해당되면 yes로, 그렇지 않으면 no로 범주화 하였다.

조사대상자는 20세 이상 성인 총 24,095명 중 건강설문조사에 참여하지 않은 1,727명을 제외한 22,368명이였다. 한 개인이 가지고 있는 결측치의 경우, 수치형 자료는 평균으로, 범주형 자료는 결측치 처리하려는 변수와 가장 관련이 깊은 2개의 변수를 chi-squared test를 통해 선정하고, 그 2개의 변수를 기준으로 그룹화 했을 때 결측치 변수의 최빈값을 구해 값을 대체하였습니다. 그 후 모형의 예측력을 판단하기 위해 데이터를 무작위로 7 : 3의 비율로 나누어 Training data ($n = 15659$)는 모형을 만들어 노모그램을 구축하는데 사용하였고 Test data ($n = 6709$)는 검증하는데 사용하였다.

Table 3.1. Logistic regression analysis results considering interactions

Variable	Level	Odds ratio	p-value	Points
Age	20–44	1.00		0
	45–64	4.29	0.000	54
	65–	15.18	0.000	100
Sex	Man	1.00	0.000	41
	Woman	0.33		0
BMI	Obese	6.42	0.000	68
	Overweight	2.72	0.000	37
	Normal	1.00		0
FH_HTN	Yes	2.17	0.000	29
	No	1.00		0
DM	DM	1.00	0.000	26
	O.W.	0.50		0
ALCOHOL	Yes	1.17	0.030	6
	No	1.00		0
STROKE	Yes	1.85	0.000	21
	No	1.00		0
DYSLIPIDEMIA	Yes	1.77	0.000	11
	No	1.00		0
AGE∗SEX	45–64 & Woman	2.07	0.000	27
	65– & Woman	3.45	0.000	45
	O.W.	1.00		0
AGE∗BMI	45–64 & Overweight	0.82	0.1466	14
	65– & Overweight	0.56	0.000	0
	45–64 & Obese	0.74	0.1797	10
	65– & Obese	0.64	0.1740	5
	O.W.	1.00		21
SEX∗DYSLIPIDEMIA	Woman & Yes	1.37	0.027	12
	O.W.	1.00		0
SEX∗ALCOHOL	Woman & Yes	0.76	0.099	0
	O.W.	1.00		10

3.2. Rao-Scott chi-squared test results

고혈압 유무에 따라 10개의 위험요인의 Rao-Scott chi-squared test 결과, 나이가 많을수록 고혈압 발병률이 증가함을 알 수 있었다. 성별에 따라 남자가 여자보다 발병률이 더 높고, 흡연 경험이 있거나 BMI가 높거나 가족력, 당뇨, 뇌졸중, 이상지질혈증이 있을 경우 고혈압 발병률이 증가한다는 것을 알 수 있었다. Rao-Scott chi-squared test 결과 고혈압 발병 유무 변수와 10개의 위험 요인 모두 p-value가 0.001보다 작았다. 따라서 10개의 위험 요인들은 고혈압 발병 유무와 관련 있다고 할 수 있다.

3.3. Nomogram for hypertension using logistic regression model with complex sample

우리는 10개의 위험요인을 바탕으로 표본 가중치를 고려한 로지스틱 회귀분석을 적합시켰다. 먼저 10개의 위험요인의 주 효과만 고려한 모형을 분석한 결과, 흡연 상태, 운동 유무 두 개의 위험요인은 회귀계수가 유의하지 않게 나타났다. 따라서 두 개의 위험요인을 제외한 8개 위험요인을 최종적인 고혈압

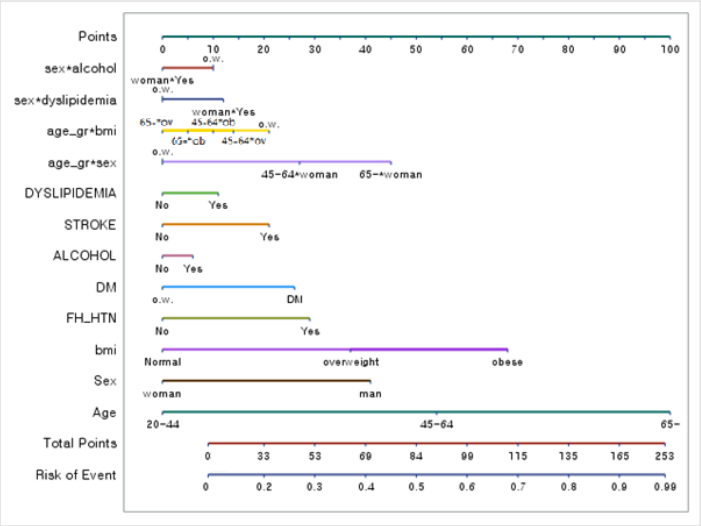


Figure 3.1. Nomogram for hypertension using logistic regression with complex sample.

발병 예측 변수로 설정하였다. 추가로 각 모든 2차 상호작용을 포함한 모형을 분석한 결과 나이와 성별, 나이와 BMI, 성별과 이상지질혈증, 성별과 음주 유무 4가지 항이 유의수준 5%에서 유의하였다. 따라서 8개의 위험요인과 4개의 상호작용을 모두 고려하여 총 12개의 위험요인이 최종적인 회귀 계수로 사용되었다. 8가지 위험요인과 4가지 상호작용을 고려한 로지스틱 회귀분석 결과는 Table 3.1과 같다.

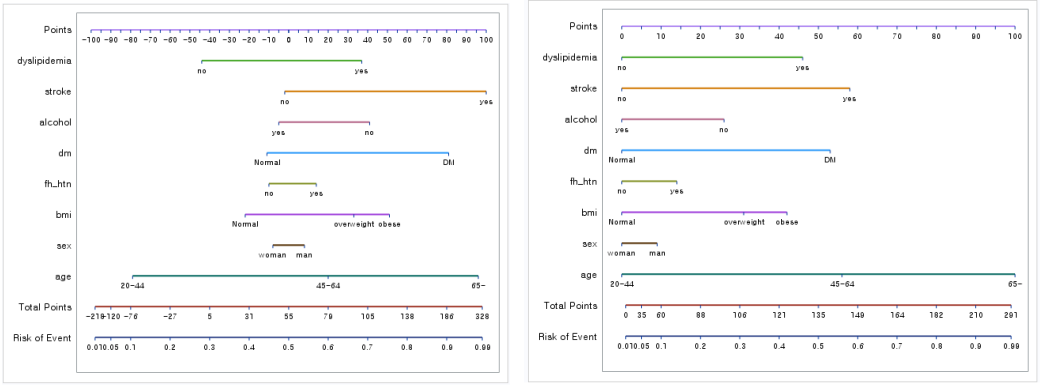
Table 3.1을 바탕으로 고혈압 발병률을 예측하는 로지스틱 노모그램을 구축하였다. 그 결과는 Figure 3.1과 같다. Figure 3.1에서 교호작용 선의 경우 BMI 범주에 해당하는 overweight는 ov로, obese는 ob로 표기하였다. Figure 3.1을 보면 AGE와 BMI가 고혈압에 가장 큰 영향을 주는 것을 알 수 있다. 나이가 많을수록 더 높은 점수를 받아 고혈압 발병률이 더 높아지고, BMI 지수가 높을수록 역시 고혈압 발병률이 더 높아진다는 것을 알 수 있다. 그리고 주 효과 중에서 성별이 나이와 BMI 뒤를 이어 가장 큰 영향을 미치는 것으로 나타났으며, 여자보다 남자가 41점 더 높은 점수를 받아 남자가 고혈압 발병률이 더 높다는 것을 알 수 있다. 그 다음으로 고혈압 발병률에 큰 영향을 미치는 요인은 가족력(FH_HTN)으로 부모나 형제 중 한 사람이라도 고혈압이 있는 사람이 그렇지 않은 사람보다 29점 더 높은 점수를 얻는 것으로 나타났다. 그 다음으로 DM, STROKE, DYSLIPIDEMIA, ALCOHOL 순으로 고혈압 발병률에 큰 영향을 미친다는 것을 알 수 있다. 상호작용 중에서는 AGE∗SEX가 고혈압 발병률에 가장 큰 영향을 주는 것을 알 수 있었고 65세 이상 여성일 경우 45점으로 가장 높은 점수를 받아 고혈압 발병률이 더 높아진다는 것을 알 수 있다. 다른 상호작용 효과 변수들은 10점에서 20점 내외의 점수를 가지는 것을 알 수 있다. 상호작용의 경우 해당하는 두 개의 주 효과의 점수를 먼저 계산한 뒤, 각 주 효과의 범주에 맞는 교호작용의 점수를 마지막으로 구해 합산하여 최종적인 점수를 구하는 방식이다.

3.4. Nomogram for hypertension using naïve Bayesian classifier model with complex sample

순수 베이지안 분류기 모델을 이용해 계산된 속성 값 $\log \mathrm{OR}(a_i = j)$의 j번째 범주에 대한 $\log \mathrm{OR}$ $(a_i = j)$와 노모그램 점수는 Table 3.2에 나와 있다. Figure 3.2(b)는 left-aligned 방법이 적용

Table 3.2. The results of naïve Bayesian classifier model for hypertension and nomogram points

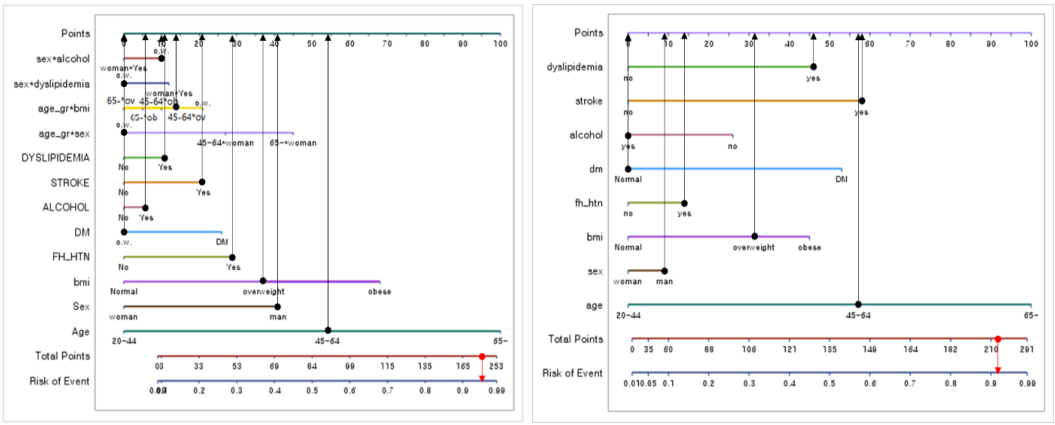
Variable	Category	$P(a_i = j|$ hypertension)	$P(a_i = j|$Non-hypertension)	$\log \mathrm{OR}(a_i = j)$	Points	Left-aligned points
AGE	20–44	0.15	0.59	−1.33	−80	0
	45–64	0.47	0.34	0.33	20	57
	65–	0.37	0.08	1.61	95	100
SEX	Man	0.54	0.47	0.14	8	9
	Woman	0.46	0.53	−0.14	−8	0
BMI	Obese	0.08	0.03	0.86	57	45
	Overweight	0.41	0.23	0.56	33	31
	Normal	0.51	0.73	−0.37	−22	0
FH_HTN	Yes	0.45	0.36	0.24	14	14
	No	0.55	0.64	−0.16	−10	0
DM	DM	0.22	0.06	1.36	82	53
	Normal	0.78	0.94	−0.19	−11	0
ALCOHOL	Yes	0.85	0.92	−0.08	−5	0
	No	0.15	0.08	0.68	41	26
STROKE	Yes	0.04	0.01	1.68	100	58
	No	0.96	0.99	−0.04	−2	0
DYSLIPIDEMIA	Yes	0.70	0.38	0.63	37	46
	No	0.30	0.62	−0.73	−44	0



(a) Bayesian nomogram for hypertension (b) Left-aligned Bayesian nomogram

Figure 3.2. Nomogram for hypertension using naive Bayesian classifier model with complex sample.

된 순수 베이지안 분류기 모형 노모그램이다. Figure 3.2(b)를 보면 나이의 노모그램 점수 범위가 0∼100점으로 고혈압 발병 유무에 가장 큰 영향을 미치는 위험요인이라는 것을 알 수 있다. 그리고 뇌졸중(STROKE)이 나이의 뒤를 이어 가장 큰 영향을 미치는 것으로 나타났으며, 뇌졸중을 앓는 사람이 그렇지 않은 사람보다 58점 더 높은 점수를 얻는 것으로 나타났다. 그 다음으로 고혈압 발병률에 큰 영향을 미치는 요인은 당뇨(DM)로 당뇨가 있는 사람이 그렇지 않은 사람보다 53점 더 높은 점수를 얻는 것으로 나타났다. 다음으로 DYSLIPIDEMIA, BMI, ALCOHOL, FH_HTN, SEX 순으로 고혈압 발병률에 큰 영향을 미치는 것으로 나타났다. Figure 3.2(a)의 노모그램 점수 범위가 −100∼100점인 반면, Figure 3.2(b)는 점수 범위가 0∼100점으로 로지스틱 회귀모형의 노모그램의 점수 범위가 같으므로 서



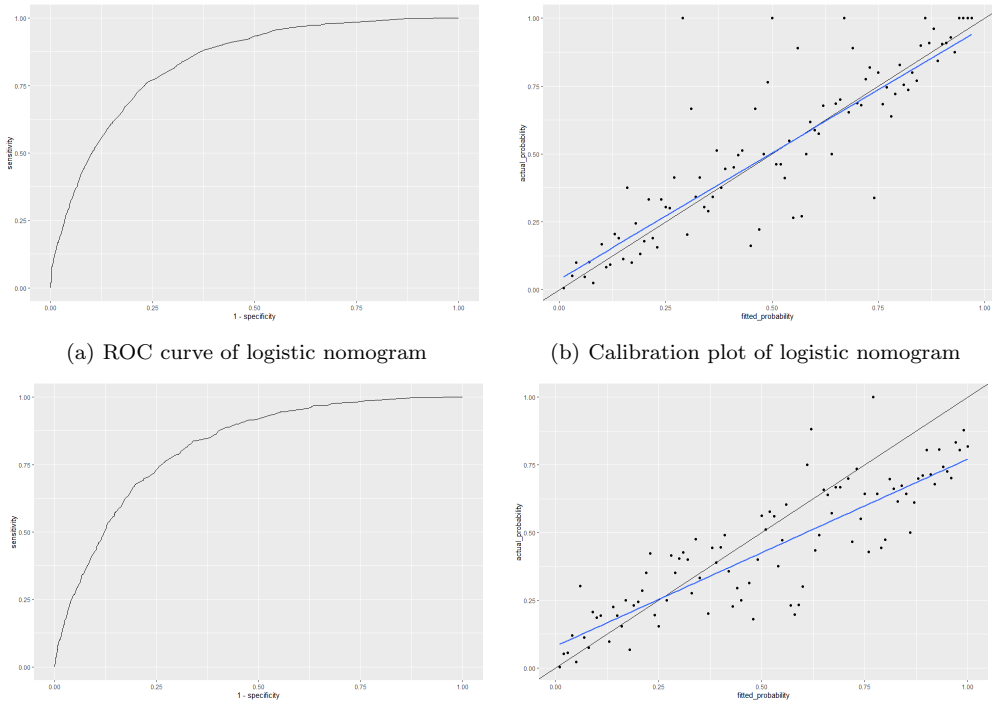
(a) Logistic nomogram for hypertension (b) Left-aligned Bayesian nomogram

Figure 3.3. Comparison of logistic nomogram and left-aligned Bayesian nomogram.

로 비교하기 쉽다.

3.5. Comparison of logistic nomogram and left-aligned Bayesian nomogram

3.5절에서는 로지스틱 노모그램과 베이지안 노모그램의 점수 범위가 다르기 때문에 직접적으로 비교를 하기엔 무리가 있어 로지스틱 노모그램과 점수 범위를 0∼100점으로 수정한 왼쪽 정렬 베이지안 노모그램과 비교한다. Figure 3.3(a)와 (b)는 각각 고혈압을 예측하는 로지스틱 노모그램과 left-aligned 방법을 사용한 베이지안 노모그램이다. 먼저 발병에 영향을 주는 개별적인 위험요인을 살펴 보면, 로지스틱 노모그램의 경우 나이(AGE)가 45–64세인 경우 54점, 65세 이상인 경우 100점으로 가장 크게 나타났다. 그리고 BMI가 비만(obese)인 경우가 68점, 과체중(overweight)인 경우 37점으로 고혈압에 두 번째로 가장 큰 영향을 미치는 것으로 나타났다. 그리고 다음으로 성별(SEX), 가족력(FH_HTN), 당뇨(DM), 뇌졸중(STROKE), 이상지질혈증(DYSLIPIDEMIA), 음주 유무(ALCOHOL) 순으로 큰 영향도를 보였다. 그리고 베이지안 노모그램의 경우 나이(AGE)가 45–64세인 경우 57점, 65세 이상인 경우 100점으로 가장 크게 나타났다. 그리고 뇌졸중(STROKE) 진단을 받은 경우가 58점으로 고혈압에 두 번째로 가장 큰 영향을 미치는 것으로 나타났다. 다음으로 당뇨(DM), 이상지질혈증(DYSLIPIDEMIA), BMI, 음주 유무(ALCOHOL), 가족력(FH_HTN), 성별(SEX) 순으로 큰 영향도를 보였다. 두 노모그램 모두 나이 변수가 발병에 가장 큰 영향을 나타내었다. 로지스틱 노모그램의 경우 예측력을 높이기 위해 위험요인 간 상호작용을 고려하여 노모그램을 구축할 수 있었다. 하지만 너무 많은 상호작용의 고려는 추정해야 할 모수가 많아져 추정의 정확도가 떨어질 수 있고, 노모그램의 해석이 어려울 수 있다. 반면 순수 베이지안 분류기 모형은 위험요인만 사용해 노모그램의 해석적 측면에서 유리하다고 할 수 있다. 실제 예시로, 50세 남성이고 과체중(overweight), 고혈압 가족력이 있고, 당뇨가 없으며 음주 경험이 있고 이상지질혈증과 뇌졸중을 앓는 사람을 두 노모그램에 적용하였다. Figure 3.3(a)의 로지스틱 노모그램의 경우 AGE선에서 45–64세에 해당하는 점수인 54점과, SEX선에서 남자에 해당하는 점수 41점과 두 개의 교호작용선인 AGE∗SEX선에서 45–64세이면서 남자일 때의 점수 0점, BMI선에서 overweight에 해당하는 점수인 37점, AGE∗BMI선에서 45–64세이면서 BMI가 overweight일 때의 점수 14점, FH_HTN선에서 Yes점수 29점, DM선에서 o.w점수 0점, ALCOHOL선에서 Yes점수 6점, SEX∗ALCOHOL선에서 o.w점수 10점, STROKE선에서 Yes점수 21점, DYSLIPI-



(a) ROC curve of logistic nomogram (b) Calibration plot of logistic nomogram

(c) ROC curve of Bayesian nomogram (d) Calibration plot of Bayesian nomogram

Figure 3.4. ROC curve and calibration plot of the nomograms.

DEMIA선에서 Yes점수 11점, SEX∗DYSLIPIDEMIA선에서 o.w점수 0점을 합한 점수 223점이 최종 점수가 된다. 이 점수에 대응되는 확률을 구하면 이 사람의 고혈압 발병 확률은 96%라고 할 수 있다. Figure 3.3(b)의 left-aligned 베이지안 노모그램의 경우 AGE선에서 45–64세에 해당하는 점수인 57점과 SEX선에서 남자에 해당하는 점수 9점, BMI선에서 overweight선에서 overweight에 해당하는 점수 31점, FH_HTN선에서 Yes점수 14점, DM선에서 Normal점수 0점, ALCOHOL선에서 Yes점수 0점, STROKE선에서 Yes점수 58점, DYSLIPIDEMIA선에서 Yes점수 46점을 합한 점수 215점이 최종 점수가 된다. 이 점수에 대응되는 확률을 구하면 이 사람의 고혈압 발병 확률은 91%라고 할 수 있다. 최종적으로 두 노모그램이 비슷한 예측 확률을 도출한다는 것을 확인하였다.

3.6. Validation for nomograms

구축한 노모그램을 검증하기 위해 ROC curve와 calibration plot을 사용하였다. 복합표본 하에서의 로지스틱 회귀모형과 순수 베이지안 모형의 각 ROC curve와 calibration plot은 Figure 3.4와 같다. 복합표본 하에서 로지스틱 회귀모형과 순수 베이지안 분류기 모형의 적합도를 비교하였다. 그 결과는 Table 3.4와 같다.

복합표본 하에서 로지스틱 회귀분석은 ROC curve의 AUC가 train, test 각각 0.8301, 0.8236이고 calibration plot의 R^2가 train, test 각각 0.9146, 0.9002이었다. 복합표본 하에서 순수 베이지안 분류기는 ROC curve의 AUC가 train, test 각각 0.8169, 0.8168이고 calibration plot의 R^2은 train, test 각각 0.9340, 0.9235이었다.

Table 3.3. AUC of ROC and R^2 of calibration plot for each model

	Logistic regression		Naïve Bayesian classifier	
	Train	Test	Train	Test
AUC of ROC	0.8301	0.8236	0.8169	0.8168
R^2 of cali-plot	0.9146	0.9002	0.9340	0.9235

AUC = area under curve; ROC = receiver operating characteristics.

4. 결론 및 토의

본 논문에서는 고혈압의 위험요인을 이용하여 로지스틱 회귀모형과 베이지안 분류기 모형을 만들고 이를 시각화 하기 위한 통계적 도구인 노모그램을 구축하였다. 사용된 데이터는 한국 국민의 건강 행태를 파악할 수 있는 국민건강영양조사 2013–2016년 자료이며, 사용한 데이터는 22,368명이었다. 하지만 각 층, 집락 별로 부여된 개인별 가중치를 적용했을 때, 실질적으로 반영된 인구 수는 39,235,320명으로 실제 2015년 인구 총 조사에서 조사된 내국인 인구수와 0.99% 밖에 차이가 나지 않았다 (Korean Statistical Information Service, 2018). 또한 고혈압의 발병률은 raw data에서는 33.9%, 복합표본을 사용할 때는 28.4%였다. 한국과 다른 나라들의 선행 연구에서의 고혈압 발병률을 보면 약 27.2∼28.5%였고 이는 복합 표본을 사용했을 때 발병률이 현실적인 수치임을 알 수 있었다. 따라서 복합 표본조사 방법이 더 유용하고 합리적이라고 할 수 있었다. 복합 표본의 특성에 맞는 고혈압의 위험요인을 선별하기 위한 방법으로 Rao-Scott chi-squared test를 시행한 결과, 모든 위험요인이 통계적으로 유의하였다. 하지만 로지스틱 회귀분석시, 흡연 상태와, 운동 유무는 유의하지 않았다. 따라서 8개의 주 효과를 고혈압의 위험요인으로 최종 선별하였다. 로지스틱 회귀모형의 경우 최종적으로 선정된 위험요인은 8개의 주 효과 (AGE, SEX, BMI, FH_HTN, DM, ALCOHOL, STROKE, DYSLIPIDEMIA)와 4개의 상호작용(AGE∗SEX, AGE∗BMI, SEX∗DYSLIPIDEMIA, SEX∗ALCOHOL)으로 선정되었다. 두 노모그램에서 공통적으로 나이가 고혈압에 가장 큰 영향을 미치는 위험요인 이였다. 그리고 로지스틱 노모그램의 경우, 음주 유무가 가장 작은 영향을 미치는 위험요인 이였다. 순수 베이지안 분류기 노모그램의 경우, 가장 작은 영향을 미치는 위험요인은 성별 이였다. 로지스틱 회귀 모형의 경우 요인 간의 상호작용 항을 모형에 포함하지 않는 이상 고려되지 않지만, 베이지안 모형의 경우에는 조건부 확률 계산 과정에서 요인 간 상호작용이 포함된다. 로지스틱 회귀모형의 경우 의학적인 근거와 통계적 유의성을 모두 고려하여 상호작용 항을 선정한다면 보다 좋은 예측 모형을 구축할 수 있다. 하지만 너무 많은 요인들을 포함하면 모형이 복잡해져 통계적인 의미가 없어질 수 있으므로 주의할 필요가 있다. 베이지안 노모그램의 경우 요인 간 상호작용이 각 위험요인 안에 포함되어 있어 사용이 편리하나, 위험요인내 범주 하나 하나 계산이 필요하다는 것이 단점이다.

고혈압은 발병률이 꾸준히 증가하고 있을 뿐 아니라 심혈관 질환의 큰 주요원인이 되었다. 그래서 고혈압의 발병을 예측하는 도구로서 고혈압 노모그램을 구축하였다. 노모그램을 통해 의료 분야에서 통계적 지식이 없는 의료진들이나 일반인들이 간편하게 질병을 진단할 수 있다. 이에 본 연구에서 구축한 노모그램이 데이터를 이용한 증거 기반 의료의 요구가 갈수록 매우 커지고 있는 의학분야에서 향후 치료계획을 수립하는데 도움이 될 수 있다고 할 수 있다.

References

Akobeng, A. K. (2007). Understanding diagnostic tests 3: receiver operating characteristic curves, *Acta Paediatrica*, **96**, 644–647.

Cook, N. R. (2008). Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve, *Clinical Chemistry*, **54**, 17–23.

D'Agostino Sr, R. B., Grundy, S., Sullivan, L. M., Wilson, P., and CHD Risk Prediction Group (2001). Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation, *Jama*, **286**, 180–187.

Iasonos, A., Schrag, D., Raj, G. V., and Panageas, K. S. (2008). How to build and interpret a nomogram for cancer prognosis, *Journal of Clinical Oncology*, **26**, 1364–1370.

Kim, M. H. (2020). *Nomogram model for predicting the incidence of hypertension with complex sample* (Master's thesis), Yeungnam University, Gyeongsan.

Kim, M. H. and Lee, J. Y. (2020). How to construct a nomogram for hypertension using complex sampling data from Korean adults, *Communications in Statistics-Theory and Methods*, Published online: 07 June 2020.

Kim, M. H., Seo, J. H., and Lee, J. Y. (2019). Nomogram building to predict dyslipidemia using a naïve Bayesian classifier model, *The Korean Journal of Applied Statistics*, **32**, 619–630.

Korea Centers for Disease Control and Prevention (2016). *The Seventh Korea National Health and Nutrition Examination Survey (KNHANES VII-1).*

Korean Statistical Information Service (KOSIS). Census, Statistic Korea, Republic of Korea. Accessed December 2018. Available from: http://kosis.kr/statHtml3/statHtml.do?orgId=101&tblId=DT_1IN1503 &vw_cd=MT _ZTITLE&list_id=A11_2015_1_10_10&seqNo=&lang_mode=ko&language=kor &obj_var _id=&itm_id=&conn _path=MT_ZTITLE

Kshirsagar, A. V., Chiu, Y. L., Bomback, A. S., August, P. A., Viera, A. J., Colindres, R. E., and Bang, H. (2010). A hypertension risk score for middle-aged and older adults, *The Journal of Clinical Hypertension*, **12**, 800–808.

Lee, K. M., Kim, W. J., and Yun, S. J. (2009). A clinical nomogram construction method using genetic algorithm and naïve Bayesian technique, *Journal of Korean Institute of Intelligent Systems*, **19**, 796–801.

Mozina, M., Demšar, J., Kattan, M., and Zupan, B. (2004). Nomogram for visualization of Naïve Bayesian classifier, *Knowledge Discovery in Databases: PKDD 2004*, 337–348.

Nam, H. R., Pak, S. B., Jung, S. J., Choi, I. Y., and Kim, Y. (2018). Interdependency of Risk Factors for Hypertension: the 2010–2015 Korean National Health and Nutrition Examination Survey, *Korean Journal of Family Practice*, **8**, 372–379.

Park, J. C., Kim, M. H., and Lee, J. Y. (2018). Nomogram comparison conducted by logistic regression and naïve Bayesian classifier using type 2 diabetes mellitus, *The Korean Journal of Applied Statistics*, **31**, 573–585.

Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit the independence in two-way tables, *Journal of the American Statistical Association*, **76**, 221–230.

Shin, J., Park, J. B., Kim, K. I., Kim, J. H., Yang, D. H., Pyun, W. B., Kim, Y. G., Kim, G. H., and Chae, S. C. (2015). 2013 Korean Society of Hypertension guidelines for the management of hypertension: part I–epidemiology and diagnosis of hypertension, *Clinical Hypertension*, **21**, 1.

Statistics Korea (2018). Causes of death statistics 2017. Policy News. Available from: http://kostat.go.kr/ portal/korea/ kor_nw/3/index.board?bmode=read&bSeq=&aSeq= 370711&pageNo=1&rowNum=10& navCount=10&currP g = &sTarget=title&sTxt=2017

Sung, N. K. (2012). *Sampling Methodologies* (2nd ed), Freedom academy, Seoul.

Van den Berg, E., Kloppenborg, R. P., Kessels, R. P., Kappelle, L. J., and Biessels, G. J. (2009). Type 2 diabetes mellitus, hypertension, dyslipidemia and obesity: a systematic comparison of their impact on cognition, *Biochimica et Biophysica Acta -Molecular Basis of Disease*, **1792**, 470–481.

고혈압 예측을 위한 노모그램 구축 및 비교

김민호[a] · 신민석[a] · 이제영[a,1]

[a]영남대학교 통계학과

(2020년 6월 3일 접수, 2020년 7월 9일 수정, 2020년 8월 6일 채택)

요 약

고혈압은 발병률이 꾸준히 증가하고 있을 뿐 아니라, 심혈관 질환과 같은 2차 질병의 주된 위험 요인이 되었다. 게다가 고혈압은 뇌졸중, 혈관성 치매와 같은 다른 합병증을 유발하는 질병이다. 따라서 고혈압 발병률을 예측하는 것은 중요한 일이다. 본 연구에서, 고혈압 발병률을 예측할 수 있는 노모그램을 구축하였다. 데이터는 2013년부터 2016년까지의 국민건강영양조사로부터 얻어졌다. 복합 표본의 특성을 고려하여 Rao-Scott chi-squared test를 통해 고혈압에 영향을 미치는 10가지 요인을 규명하였다. 하지만 로지스틱 회귀분석 시, 흡연 상태와, 운동 유무는 유의하지 않았다. 따라서 8개의 주 효과를 고혈압의 위험요인으로 최종 선별하였다. 그리고 최종 선별된 위험 요인들로 로지스틱 노모그램과 베이지안 노모그램을 제시 및 비교하였다. 마지막으로 ROC curve 그래프와 calibration plot을 통해 노모그램을 검증하였다.

주요용어: 고혈압, 로지스틱 회귀분석, 순수 베이지안 분류기, 노모그램, 위험 요인

[1]교신저자: (38541) 경상북도 경산시 대학로 280, 영남대학교 통계학과. E-mail: Jlee@yu.ac.kr