The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.5.573
31(5), 573–585

Nomogram comparison conducted by logistic regression and naïve Bayesian classifier using type 2 diabetes mellitus (T2D)

Jae-Cheol Park[a] · Min-Ho Kim[a] · Jea-Young Lee[a,1]

[a]Department of Statistics, Yeungnam University

(Received April 9, 2018; Revised May 23, 2018; Accepted June 8, 2018)

Abstract

In this study, we fit the logistic regression model and naïve Bayesian classifier model using 11 risk factors to predict the incidence rate probability for type 2 diabetes mellitus. We then introduce how to construct a nomogram that can help people visually understand it. We use data from the 2013–2015 Korean National Health and Nutrition Examination Survey (KNHANES). We take 3 interactions in the logistic regression model to improve the quality of the analysis and facilitate the application of the left-aligned method to the Bayesian nomogram. Finally, we compare the two nomograms and examine their utility. Then we verify the nomogram using the ROC curve.

Keywords: logistic regression, naïve Bayesian classifier, nomogram, receiver operating characteristic curve, type 2 diabetes mellitus

1. 서론

당뇨병(diabetes mellitus)은 혈액 속 포도당이 몸의 세포에 공급하는 인슐린의 분비가 원활하지 않게 되면서 소변으로 넘쳐 나오게 되는 질병이다. 당뇨병을 초기에 발견하고 치료를 병행하지 않으면 몸 안에 인슐린이 부족해져 급성 합병증이 생길 위험이 매우 커진다. 당뇨병은 제 1형 당뇨병(type 1 diabetes mellitus)과 제 2형 당뇨병(type 2 diabetes mellitus; T2D)으로 나뉜다. 제 1형 당뇨병은 전체 당뇨병의 2% 정도를 차지하고 주로 소아에게 많이 발생한다. 제 1형 당뇨병은 인슐린 분비 기능이 거의 정상적으로 작용하지 않아 혈당 조절을 위해 인슐린 치료가 절대적으로 필요하다. 반면 T2D은 한국인 대부분의 당뇨병을 차지하며 주로 40세 이후 성인에게 발생한다. 본 논문에서는 T2D만 다루기로 한다. 당뇨병 발병률은 계속 증가하는 추세이다. 특히 한국에서는 주요 사망원인 중 6번째 이다 (Statistics Korea, 2014). 그러나 당뇨병을 앓는 사람들 중에 30%는 자신이 당뇨병 임을 알지 못했다 (Korean Diabetes Association, 2017). 따라서 당뇨병 발병에 영향을 미치는 위험요인으로 당뇨병 발병률을 예측하여 대비를 할 필요가 있다.

이에 로지스틱 회귀분석과 순수 베이지안 분류기 방법은 발병률을 예측하는데 효과적이다. 로지스틱 회귀분석에서 종속변수를 범주형 변수로 설정하고, 로짓(logit) 연결 함수를 이용하여 중회귀 분석하였다

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

(Heo와 Lee, 2008; Park, 2018). 순수 베이지안 분류기 모형은 각 위험 요인이 주어질 때 발병이 일어날 조건부 확률 값을 계산하는 베이즈 법칙을 이용한다 (Možina 등, 2004). 로지스틱 회귀분석에서 도출되는 회귀계수 값과 순수 베이지안 분류기 모형에서 도출되는 로그 오즈비 값은 각 위험요인이 발병유무에 영향을 미치는 정도를 나타낸다. 하지만 두 모형을 구축하는 방법이 다르기 때문에 같은 위험요인일지라도 발병유무에 영향을 미치는 정도가 다를 수 있다. 따라서 이 두 모델을 비교하여 각 모델의 장단점을 알아보고자 한다. 이 때, 초기 당뇨병 발병률이 어느 정도인지를 시각화 하여 쉽게 파악할 수 있게 돕는 통계적 도구는 노모그램이다.

노모그램은 복잡한 계산식 없이 발병에 영향을 미치는 위험요인들로 질병의 진단, 재발 또는 생존예측 모델을 만들고 이를 쉽게 보여주는 시각적 도구이다 (Lee 등, 2009; Iasonos 등, 2008). 통계적 지식이 적은 비전문가들도 노모그램을 이용하면 의사결정을 할 때 도움을 받을 수 있다. 본 연구는 발병률 예측을 위한 로지스틱 회귀모형과 순수 베이지안 분류기를 이용해 각 속성값에 대한 영향 정도와 예측 확률을 계산한 다음 이를 점수화 하여 노모그램을 구축하고 최종적으로 두 노모그램을 비교해 보고자 한다. 2절에서는 로지스틱 노모그램과 순수 베이지안 분류기 노모그램 구축 방법을 제시하고, 구축된 모형을 비교하기 쉽도록 하기 위해 베이지안 노모그램을 left-aligned 시키는 방법을 소개한다. 3절에서는 실험 자료에 대한 설명과 노모그램 검증 방법을 소개한다. 4절에서는 로지스틱과 베이지안 노모그램을 구축하고 비교 및 검증한다. 마지막으로 5절에서는 비교한 결과를 살펴보고 의견을 제시한다.

2. 노모그램 구축

의료 분야에서 질병이나 사망에 관련된 위험 요인을 선별하고 발생률을 예측하는 것은 치료 계획을 수립하고, 의학적 의사결정을 내릴 때 매우 중요하다. 위험 인자 선별과 발병률 예측에는 주로 로지스틱 회귀분석과 Cox 비례위험모형을 사용한다. 하지만 이러한 방법들로 얻은 모형에서 위험요인을 선별하고 발병률을 계산하는 것은 비전공자들에게는 어려움이 있다. 따라서 복잡한 모형 대신 질병과 위험요인 사이에 수치적 관계를 계산해 시각화하여 쉽고 직관적으로 이해 할 수 있는 통계적 도구인 노모그램을 소개한다 (Lee 등, 2009; Iasonos 등, 2008). 노모그램은 구축하는데 복잡하지 않으며 간단한 선으로 표현되어서 해석이 용이하다. Figure 2.1은 로지스틱 노모그램의 한 예시이다. 노모그램을 구성하는 요소들에는 위에서부터 Points 선, Risk Factor 선(Symptoms, Sex male), Total Points 선, Probability 선이 있다.

첫 번째 Points 선은 각 위험요인들의 범주가 가지는 점수를 확인하는 선이다. 로지스틱 노모그램의 경우 0∼100점으로 구성되고, 베이지안 노모그램의 경우 −100∼100점으로 구성된다. 두 번째와 세 번째 선인 Risk Factor 선은 각 위험 요인들의 범주가 Target 변수가 yes일 사건에 미치는 영향 정도를 수치적으로 계산한 선이다. Risk Factor 선은 로지스틱 회귀분석의 경우 회귀계수 값을 이용하고, 순수 베이지안 분류기 모형의 경우 로그 오즈비 값을 이용하여 계산한다. 이 예에서는 Symptoms, Sex male 두 가지 변수의 각각의 범주가 Target 변수가 yes일 사건에 미치는 영향 정도를 계산하였다. 다섯 번째 Probability 선은 0에서 1사이의 확률을 적절한 기준으로 나누어 표시한다. 그리고 각 기준에 대응되는 Total Points를 모델에 대입해 계산하여 네 번째 Total Points 선을 구성한다. 이때, Total Points는 각 Points 들의 합으로 구할 수 있다. 구축된 노모그램을 이용하여 Target 변수가 yes일 사건의 확률을 예측할 수 있다. 예를 들어, Symptoms가 0, Sex male이 1인 경우 Total Points는 23.5점이고 Total Points 선에서 수직선을 그어 Probability 선에 대응되는 확률은 0.59임을 알 수 있다.

2장에서는 당뇨병 발병률의 예측을 위해 로지스틱 노모그램과 순수 베이지안 분류기 노모그램 구축 방법을 제시하고 두 노모그램을 비교하기 쉽게 하도록 하기 위해 베이지안 노모그램을 left-aligned하는 방법을 소개한다.

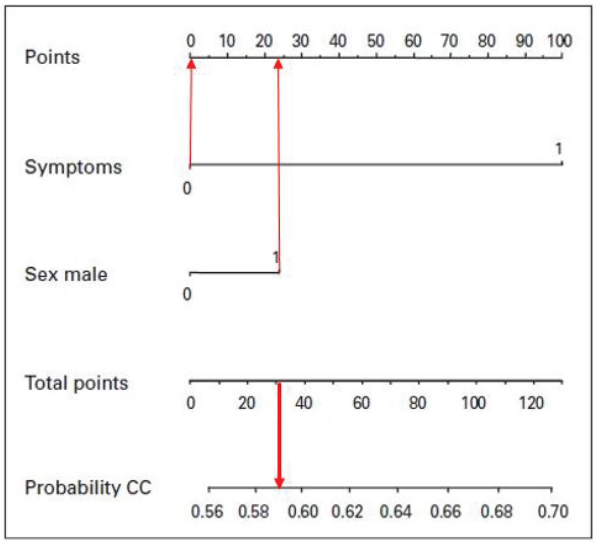


Figure 2.1. Example of nomogram (Iasonos *et al.*, 2008).

2.1. T2D 로지스틱 노모그램 구축

로지스틱 회귀모형은 종속변수가 범주형 변수인 경우 속성값들이 주어질 때 특정 사건이 발생할 확률을 계산할 때 사용된다 (Lee 등, 2005; Heo와 Lee, 2008; Park과 Lee, 2017). 본 연구에서는 종속변수가 이항 변수일 때의 경우만 살펴 보도록 한다. 독립변수 X가 x일 때 종속변수 Y가 성공 확률을 갖는 분포라고 가정하자. 여기서 독립변수 X는 범주형 변수라 가정한다. 이 때 독립변수 X가 x일 때 종속변수 Y가 y일 확률은 아래와 같다.

$$P(Y=y|X=x)=p_x^y(1-p_x)^{1-y}, \quad y=0,1.$$

그리고 이를 로지스틱 회귀모형에 적합시킨 식은 아래와 같다.

$$\log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right)=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k.$$

β값은 X들의 회귀계수 값으로 각 위험요인의 영향도를 나타낸다. 독립변수 X가 x일 때 종속변수 Y의 성공 확률은 로지스틱 회귀모형을 $P(Y=1|X=x)$에 대해 정리하고 독립변수 $X=(X_1,\ldots,X_k)$를 대입하여 구할 수 있다.

$$P(Y=1|X=x)=\frac{\exp\{\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k\}}{1+\exp\{\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k\}}=\frac{\exp\left\{\alpha+\sum_{i=1}^k\beta_iX_i\right\}}{1+\exp\left\{\alpha+\sum_{i=1}^k\beta_iX_i\right\}}$$

이제부터는 노모그램을 이루는 각 선들이 어떻게 구축되는지 설명한다 (Iasonos 등, 2008; Yang, 2014; Park, 2018).

- Points 선

  Points 선은 0점에서 100점으로 구성된다.

- Risk Factor 선

  로지스틱 회귀모형에서 적합시켜 도출된 회귀계수 β값으로 linear predictor (LP_{ij})값을 계산한다. 독립변수 X가 범주형 변수이고 j개의 범주를 가지는 경우, $j-1$개의 가변수(dummy variable)가 생성되는데 이 때 기준 범주의 회귀계수 값을 0으로 둔다.

$$\mathrm{LP}_{ij} = \beta_{ij} \times X_{ij},$$
$$\mathrm{Point}_{ij} = \frac{\mathrm{LP}_{ij} - \min\limits_{j} LP_{ij}}{\max\limits_{j} \mathrm{LP}_{*j} - \min\limits_{j} \mathrm{LP}_{*j}} \times 100,$$

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

- Probability 선

  Probability 선은 0에서 1까지 확률을 적절한 기준으로 분할해 구간을 만든다. 본 논문에서는 0.1씩 10가지 구간으로 나누되, 좌측에 0과 0.05, 우측에 0.95 값을 추가하였다.

- Total Points 선

  Total Points는 각 Points_{ij}값들의 누적 합으로 표시할 수 있다.

$$\text{Total Points} = \sum_{i,j} \mathrm{Points}_{ij} = \frac{100}{\max\limits_{j} \mathrm{LP}_{*j} - \min\limits_{j} \mathrm{LP}_{*j}} \times \sum_{i,j} \left(\mathrm{LP}_{ij} - \min_{j} \mathrm{LP}_{ij} \right).$$

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

$$\text{Total Points} = \frac{100}{\max\limits_{j} \mathrm{LP}_{*j} - \min\limits_{j} \mathrm{LP}_{*j}} \times \left(\log \left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} \right) - \alpha - \sum_{i,j} \min_{j} \mathrm{LP}_{ij} \right).$$

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

2.2. T2D 베이지안 노모그램 구축

순수 베이지안 분류기 모델에서는 속성 값이 서로 독립적으로 발생한다는 가정 하에서 시작한다. 이 가정 하에서, Bayes 정리를 이용하여 속성 값들이 종속변수에 어느 정도 영향을 주는지를 계산하는 방법이다 (Možina 등, 2004). 간단하고 직관적으로 이해하기 쉽고, 강력한 예측력을 가지기 때문에 통계적으로 매우 좋은 도구이다. 종속변수 Y는 로지스틱 회귀모형과 마찬가지로 성공 확률을 갖는 이항분포라고 가정한다. 이때 속성값 $X = (a_1, a_2, \ldots, a_k)$가 주어졌을 때 종속변수 Y가 성공일 사후 확률(posterior probability)은 아래와 같다.

다른수식

$$P(Y=1|X) = \frac{1}{1+\exp\left\{-\mathrm{logit}P(Y=1) - \sum_i \log \mathrm{OR}(a_i)\right\}}$$

여기서,

$$\mathrm{logit}P(Y=1) = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$$

는 종속변수 Y가 성공일 사전확률(prior probability)의 오즈비(odds ratio)이고,

$$\mathrm{OR}(a_i) = \frac{P(a_i|Y=1)}{P(a_i|Y=0)} = \frac{\frac{P(Y=1|a_i)}{P(Y=0|a_i)}}{\frac{P(Y=1)}{P(Y=0)}} = \frac{\text{posterior odds}}{\text{prior odds}}$$

는 사전확률, 사후확률의 오즈비로서 정의한다. 이때 도출되는 각 속성값의 $\log \mathrm{OR}(a_i)$를 이용하여 노모그램을 구축한다 (Možina 등, 2004; Park, 2018).

- Points 선

  Points 선은 -100점에서 100점으로 구성된다.

- Risk Factor 선

  순수 베이지안 분류기 모형에서 적합시켜 도출된 $\log \mathrm{OR}(a_{ij})$값으로 각 위험 요인의 범주 별 Points_{ij}를 계산한 후 Points 선에 맞추어 정렬한다.

$$\mathrm{Points}_{ij} = \frac{\log \mathrm{LR}(a_{ij})}{\max_{ij} |\log \mathrm{LR}(a_{ij})|} \times 100.$$

- Probability 선

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

- Total Points 선

  Total Points는 각 Points_{ij}값들의 누적 합으로 표시할 수 있다.

$$\text{Total Points} = \sum_{i,j} \mathrm{Points}_{ij} = \frac{100}{\max_{ij} |\log \mathrm{LR}(a_{ij})|} \times \sum_{i,j} \log \mathrm{OR}(a_{ij}).$$

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

$$\text{Total Points} = \frac{100}{\max_{ij} |\log \mathrm{LR}(a_{ij})|} \times \left(-\log \left(\frac{1}{P(Y=1|X=x)} - 1 \right) - \mathrm{logit} P(Y=1) \right)$$

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

제 2.2절에서 소개한 베이지안 노모그램 구축 방법을 사용하면, Points의 점수 범위는 -100~100점이다. 하지만 로지스틱 노모그램의 점수 범위는 0~100점으로 차이가 있다. 만약 Left-aligned 방법으로 베이지안 노모그램을 구축하면 로지스틱 노모그램과의 비교가 좀 더 용이하다. 2.3절에서는 합리적인 비교를 위해 베이지안 노모그램에 left-aligned 방법을 적용하여 점수 범위를 0~100점으로 바꾸는 방법을 소개한다.

2.3. Left-aligned 방법을 적용한 T2D 베이지안 노모그램 구축

2.2절에서 소개한 순수 베이지안 분류기 모형에서 도출된 $\log \mathrm{OR}(a_i)$값으로 새로 Points, Risk Factor, Total Points, Probability를 정의한다.

- Points 선

  Points 선은 0점에서 100점으로 구성된다.

- Risk Factor 선

  순수 베이지안 분류기 모형에서 적합시켜 도출된 $\log \mathrm{OR}(a_{ij})$값으로 각 위험 요인의 범주 별 Points_{ij}를 계산한 후 Points 선에 맞추어 정렬한다.

$$\mathrm{Point}_{ij} = \frac{\log \mathrm{OR}(a_{ij}) - \min\limits_{i,j}(\log \mathrm{OR}(a_{ij}))}{\max\limits_{j}(\log \mathrm{OR}(a_{*j})) - \min\limits_{j}(\log \mathrm{OR}(a_{*j}))} \times 100$$

- Probability 선

  Probability 선은 0에서 1까지 확률을 적절한 기준으로 분할해 구간을 만든다. 본 논문에서는 0.1씩 10가지 구간으로 나누되, 좌측에 0.01과 0.05, 우측에 0.99 값을 추가하였다.

- Total Points 선

  Total Points는 각 Points_{ij}값들의 누적 합으로 표시할 수 있다.

$$\begin{aligned}\text{Total Points} &= \sum_{i,j} \mathrm{Points}_{ij} \\ &= \frac{100}{\max\limits_{j}(\log \mathrm{OR}(a_{*j})) - \min\limits_{j}(\log \mathrm{OR}(a_{*j}))} \times \sum_{i,j} \left(\log \mathrm{OR}(a_{ij}) - \min_{i,j}(\log \mathrm{OR}(a_{ij})) \right).\end{aligned}$$

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

$$\begin{aligned}\text{Total Points} = &\frac{100}{\max\limits_{j}(\log \mathrm{OR}(a_{*j})) - \min\limits_{j}(\log \mathrm{OR}(a_{*j}))} \\ &\times \left(-\log\left(\frac{1}{P(Y=1|X=x)} - 1 \right) - \mathrm{logit} P(Y=1) - \sum_{i,j} \min_{i,j}(\log \mathrm{OR}(a_{ij})) \right)\end{aligned}$$

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

3. 실험자료 및 검증 방법

3.1. 실험자료의 특징

본 연구에 사용된 데이터는 2013–2015년도 6기 국민건강영양조사(Korean national health and nutrition examination survey; KNHANES) 데이터이다. 국민건강영양조사는 국민의 건강 및 영양 상태를 파악하여 국가 보건 정책 시행에 근거를 제공하며 그것이 유의미한 결과를 나타내는지 평가할 수 있는 통계를 산출한다. 본 연구는 건강 설문조사, 검진 및 영양조사를 실시한 20–85세 사이 성인을 대상으로 진행하였고 임산부와 결측치는 제외하였다. 최종적으로 13,474명의 데이터가 사용되었고, 그 중 1,543명이 T2D라고 진단하였다. 당뇨병 진단 기준은 조사 당시 당뇨병 약을 복용 중 이거나 적어도 8시간 금식 후 혈당 수치가 126 mg/dl 이상인 사람으로 선정하였다.

사용된 위험요인은 2018년 Park이 보고한 11개의 위험요인(나이(Age), 성별(Sex), 교육수준(Edu), 고용상태(Employment), 수입(Income), 흡연상태(Smoking status), 비만상태(Waist Circumference; WC), 당뇨병 가족력(Family history), 고혈압(Hypertension), 이상지질혈증(Dyslipidemia), 심혈관계 질환(Cardiovascular Disease; CVD))을 사용하였다. Age 변수는 20–39세, 40–59세, 60–85세로 세가

지 범주로 나누었고 Edu 변수는 고등학교 졸업 미만인가 고등학교 졸업 이상인가로 나누었다. Employment 변수는 직업 유무로 나누었고 Income 변수는 가족 수에 비례해 수입을 책정한 뒤, 사분위수로 나누고 1, 2 사분위수에 해당하는 사람들을 low, 3, 4 사분위수에 해당하는 사람들을 high로 범주화 하였다. Smoking status 변수는 과거에 100개피 이상을 피었지만 지금 흡연을 하지 않는 사람들을 Ex-smoker, 현재 흡연을 하는 사람들을 Current-smoker, 과거 100개피 이하를 피었고 지금 흡연을 하지 않는 사람들을 None으로 범주화 하였다. WC 변수는 허리둘레가 남자는 90cm 이상, 여자는 85cm 이상일 때 복부 비만이라고 범주화하였다 (Lee 등, 2006). Family history 변수는 당뇨병 가족력 유무로 범주화 하였고, Hypertension, Dyslipidemia, 변수는 '의사에게 해당 질병의 진단을 받은 적이 있는가?'라는 질문지 문항에 답변으로 예/아니오로 범주화 하였다. 마지막으로 CVD 변수는 심혈관과 관계가 있는 협심증, 심근경색, 뇌졸증 세 가지 질병 중 하나 이상 진단을 받은 경우 심혈관계 질환이 있다고 범주화 하였다. 여기에 로지스틱 회귀모형에서는 변수 간 상호작용을 고려하여 분석의 질을 높이고자 하였다.

3.2. 노모그램 검증 방법

노모그램을 검증하는 방법은 receiver operating characteristic (ROC) curve를 사용하였다 (Cook, 2008). ROC 곡선은 도출된 확률이 실제 발병유무에 얼마나 예측이 되는지 평가하는 도구로서 의학 부문에서 자주 쓰이고 있다. ROC 그래프는 X축에 1 − 특이도 (1 − Specificity), Y축에 민감도(sensitivity)로 하여 그려진다. 이 ROC 곡선의 아래 면적의 넓이를 area under curve (AUC)라고 하며 AUC가 1에 가까울수록 예측력이 좋다고 할 수 있다.

4. 노모그램 구축 및 결과

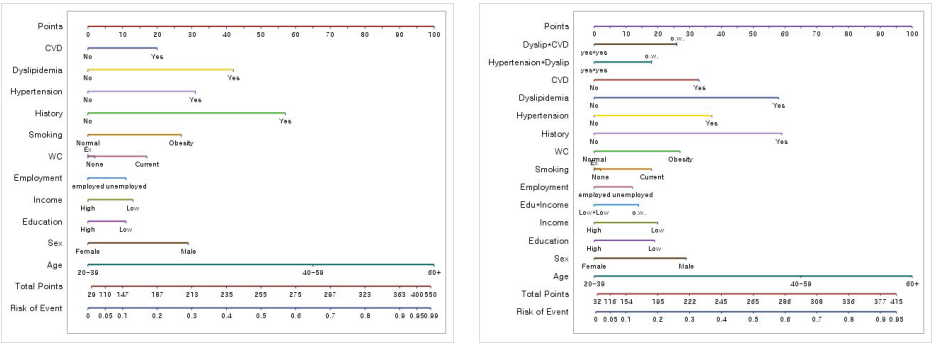
4.1. 상호작용을 고려한 T2D 로지스틱 노모그램 구축과 비교

기존 Park (2018)이 연구한 로지스틱 노모그램에서는 11가지 위험요인들의 당뇨병 발병에 영향을 미치는 독립적인 효과만을 나타냈다. 반면 순수 베이지안 분류기 모형은 조건부 확률 계산과정에서 각 위험요인들의 상호작용 효과가 모두 포함되어 계산된다. 만약 독립적인 효과만으로 로지스틱 회귀모형을 구축한다면 요인 간 상호작용이 전혀 고려되지 않아 예측력이 떨어질 수 있다 (Možina 등, 2004). 따라서 여기서는 위험 요인들간 연관성을 고려하여 예측력을 높이고자 하였다. 먼저 Kim 등 (2015)이 낮은 Socio-Economic Status (SES)일수록 당뇨 발병률이 올라간다고 보고한 사실을 바탕으로 SES와 관련된 Edu 변수, Income 변수, Employment 변수 간 2차 상호작용 항을 잠정적 상호작용 항으로 설정하였다 (Kim 등, 2015). 또 심혈관계 질환이나 이상지질혈증이 당뇨병과 밀접한 관련이 있고 고혈압 또한 심혈관계 질환과 관련성이 높다 (Chung 등, 2018). 따라서 Hypertension 변수, Dyslipidemia 변수 그리고 CVD 변수 간 2차 상호작용 항을 잠정적 상호작용 항으로 설정하였다. 전진선택법으로 로지스틱 회귀 모형을 적합시킨 결과, 11가지 주 효과와 SES 변수에서 한가지 2차 상호작용 효과, 그리고 심혈관 질환 관련 변수에서 두 가지 2차 상호작용 효과까지 총 14가지 효과가 통계적으로 유의하게 나왔다. Table 4.1은 14가지 효과의 로지스틱 회귀 모형식의 오즈비와 Points 값을 계산한 결과이다. 결과를 살펴보면, Edu 변수, Income 변수의 상호작용 항의 유의성 검정 유의확률은 0.025, Hypertension 변수, Dyslipidemia 변수의 상호작용 항의 유의확률은 0.007, Dyslipidemia 변수와 CVD 변수의 상호작용 항의 유의확률은 0.011로 통계적으로 유의하였다. 또한 Hosmer-Lemeshow 검정 결과 유의 확률이 0.0931 이므로 상호작용이 고려된 로지스틱 회귀모형이 적합하고 할 수 있다. 통계 분석은 SPSS 23을 사용하였다 (SPSS Inc., Chicago, IL, USA).

Table 4.1. Multiple logistic regression analysis results considering interactions

Variable	Level	Odds ratio	95% CI	p-value	Points
Age	20–39	1			0
	40–59	3.594	2.757, 4.685	0.000	65
	60–85	7.177	5.426, 9.493	0.000	100
Sex	Female	1			0
	Male	1.761	1.468, 2.113	0.000	29
Edu	High	1			0
	Low	1.467	1.201, 1.792	0.000	19
Employment	Employed	1			0
	Unemployed	1.258	1.104, 1.432	0.001	12
Income	High	1			0
	Low	1.474	1.238, 1.756	0.000	20
Smoking	None	1			2
	Ex-smoker	0.964	0.793, 1.173	0.716	0
	Current-smoker	1.386	1.137, 1.689	0.001	18
WC	Normal	1			0
	Obesity	1.705	1.513, 1.922	0.000	27
History	No	1			0
	Yes	3.216	2.820, 3.668	0.000	59
Hypertension	No	1			0
	Yes	2.073	1.787, 2.404	0.000	37
Dyslipidemia	No	1			0
	Yes	3.107	2.518, 3.834	0.000	58
CVD	No	1			0
	Yes	1.919	1.475, 2.497	0.000	33
Edu∗Income	Low∗Low	0.752	0.585, 0.965	0.025	0
	O.W.	1			26
Hypertension∗Dyslipidemia	Yes∗Yes	0.533	0.533, 0.906	0.007	0
	O.W.	1			14
Dyslipidemia∗CVD	Yes∗Yes	0.603	0.409, 0.889	0.011	0
	O.W.	1			18
Hosmer-Lemeshow goodness-of-fit test : $\chi^2 = 13.59$, df $= 8$, p-value $= 0.0931$.					

Figure 4.1은 주 효과만 포함된 로지스틱 노모그램과 상호작용이 고려된 로지스틱 노모그램이다 (Park, 2018). Figure 4.1(a)는 Park (2018)이 보고한 주 효과만 포함된 로지스틱 노모그램이고 (b)는 본 논문에서 구축한 상호작용이 고려된 로지스틱 노모그램이다. 먼저 나이 변수에서 60세에서 85세 이하 성인의 점수가 100점으로 당뇨 발병에 가장 큰 영향이 있음을 알 수 있다. 두 노모그램을 살펴보면, 상호작용이 고려되지 않은 Age, Sex, Employment, Smoking, WC, History 변수의 Points 값은 두 노모그램이 서로 같다. 하지만 상호작용 이 고려된 Edu, Income, Hypertension, Dyslipidemia, CVD 변수는 그 값이 크게 증가했음을 알 수 있다. Edu 변수의 Points 값은 11에서 19로 증가하였고, Income 변수는 11에서 20으로 증가하였다. 또한 Hypertension 변수는 31에서 37로 증가하였고, Dyslipidemia 변수는 42에서 58로, CVD 변수는 20에서 33으로 증가하였다. 상호작용이 고려된 변수들이 당뇨병 발병 유무에 미치는 효과의 크기가 증가했음을 알 수 있다. 노모그램과 구축에는 SAS 9.4버전이 사용되었다 (SAS Institute Inc., Cary, NC, USA; Yang, 2014).



(a) Logistic nomogram (b) Logistic nomogram with 3 interactions

Figure 4.1. Comparison of the logistic nomogram.

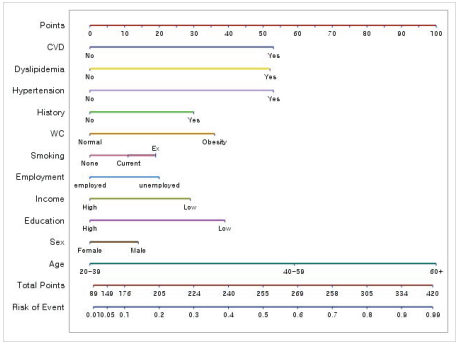


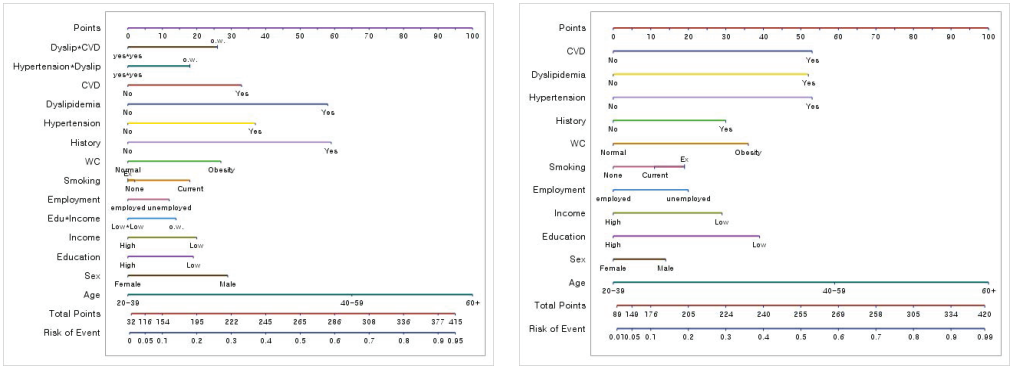
Figure 4.2. Left-aligned Bayesian nomogram.

4.2. Left-aligned 방법을 적용시킨 T2D 베이지안 노모그램 구축

4.1절에서 소개한 로지스틱 노모그램과 Park (2018)이 제안한 베이지안 노모그램의 점수 범위가 다르기 때문에 비교하기에 어려움이 있었다. 따라서 기존 베이지안 노모그램에서 2.3절에서 설명한 left-aligned 방법을 적용하여 우도비를 새로 점수화 한 뒤, 새로운 T2D 베이지안 노모그램을 구축하였다 (Park과 Lee, 2018). Figure 4.2는 새로운 left-aligned 방법을 적용한 베이지안 노모그램이다. 노모그램을 살펴보면 나이 변수가 당뇨 발병에 가장 큰 영향을 미치고 그 다음으로 심혈관계 질환 유무, 이상지질혈증 유무, 고혈압 유무가 당뇨 발병에 큰 영향을 미쳤다. 하지만 상대적으로 성별과 흡연 유무 변수는 영향도가 적었다.

4.3. 두 노모그램의 비교 및 검증

4.3절에서는 4.1절에서 구축한 로지스틱 노모그램과 4.2절에서 구축한 베이지안 노모그램을 비교하고자 한다. Figure 4.3은 상호작용이 고려된 로지스틱 노모그램과 left-aligned 방법을 적용한 베이지안 노모그램이다. 노모그램을 살펴보면, 로지스틱 노모그램 (Figure 4.3(a))과 베이지안 노모그램 (Figure 4.3(b)) 모두 나이 변수가 당뇨병 발병에 가장 큰 영향을 미치는 것을 알 수 있었다. 그리고 left-aligned 방법이 적용된 베이지안 노모그램의 점수가 로지스틱 노모그램 보다 대체로 높음을 알 수 있었다. 이는 순수 베이지안 분류기 방법을 이용하여 조건부 확률을 구할 때 요인 간 상호작용이 포함되어 계산되기



(a) Logistic nomogram with 3 interactions (b) Left-aligned Bayesian nomogram

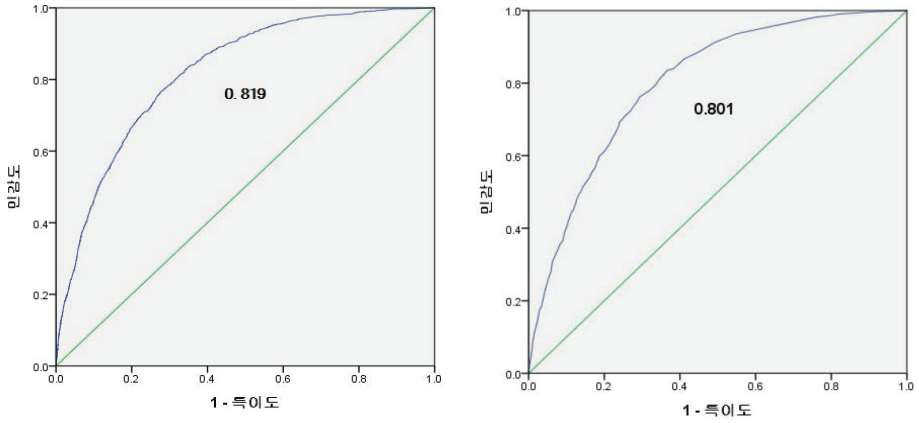
Figure 4.3. Comparison of the T2D nomogram.

때문이다 (Možina 등, 2004). 로지스틱 회귀모형에서는 위험요인 간의 상호작용 항을 넣어 모형을 적합시키면 상호작용을 고려할 수 있다.

먼저 발병에 영향을 주는 개별적인 위험요인을 살펴 보면, 로지스틱 노모그램의 경우 나이(Age)가 40–59세인 경우 65점, 60세에서 이상인 경우 100점으로 가장 크고, 가족력(History)이 있는 경우가 59점, 이상지질혈증(Dyslipidemia)이 있는 경우 58점, 고혈압(Hypertension)이 있는 경우 37점, 심혈관계 질환(CVD)이 있는 경우 33점, 비만(WC)이 있는 경우가 27점 순으로 큰 영향도를 보였다. 그리고 베이지안 노모그램의 경우 나이(Age)가 40–59세인 경우 59점, 60세에서 이상인 경우 100점으로 가장 크고 고혈압(Hypertension)이 있는 경우 53점, 심혈관계 질환(CVD)이 있는 경우 53점, 이상지질혈증(Dyslipidemia)이 있는 경우 52점, 비만(WC)이 있는 경우가 36점 가족력(History)이 있는 경우가 30점 순으로 큰 영향도를 보였다. 두 노모그램 모두 공통적으로 나이 변수가 발병에 가장 큰 영향을 나타내었고, 당뇨병과 연관성이 높은 고혈압, 이상지질혈증, 심혈관계 질환을 가진 사람일수록 당뇨 발병률이 높게 나타났다 (Chung 등, 2018). 또한 가족 중에 당뇨 발병자가 있거나 현재 비만인 사람일수록 당뇨 발병률이 높게 나타났다.

또한 4.1절에서 구축한 로지스틱 노모그램에서 상호작용 항이 있는 변수들의 경우, 두 변수 모두 Points 값이 양수라면 점수 계산 시 상호작용 항의 해당되는 범주의 점수와 합해 준다. 예를 들어 고혈압이 있고 이상지질혈증이 있는 환자의 점수는 두 가지 주 효과 점수 37와 58에서 상호작용항의 점수 0점을 더해 95점이다. 만약 고혈압이 있고 이상지질혈증이 없는 환자의 점수는 37점에서 상호작용항의 점수 14점을 더한 51점이다. 베이지안 노모그램은 각각 105점, 53점이고 상호작용이 없는 기존 로지스틱 노모그램은 각각 73, 31점이다. 기존 로지스틱 노모그램 보다 상호작용을 고려한 경우가 베이지안 노모그램의 점수에 더 가까워 짐을 알 수 있었다.

앞서 설명한 대로 구축한 노모그램을 평가하기 위해 ROC 곡선과 AUC 값을 사용한다. Figure 4.4는 상호작용이 포함된 로지스틱 노모그램과 베이지안 노모그램의 ROC 곡선이다. 상호작용이 포함된 로지스틱 노모그램 (Figure 4.4(a))의 AUC 값이 0.819 (p-value < 0.001)로 1에 가까운 값을 가지고 통계적으로 유의하므로 예측력이 좋은 노모그램이라고 할 수 있었다. 베이지안 노모그램 (Figure 4.4(b))의 경우 Park (2018)이 보고한 ROC 곡선과 같다. AUC 값이 0.801 (p-value < 0.001)로 마찬가지로 1에 가까운 값을 갖고 통계적으로 유의하였다. 따라서 두 노모그램 모두 통계적으로 유의하게 잘 구축되었다고 할 수 있었다. ROC 곡선 구축에는 SPSS 23을 사용하였다 (SPSS Inc., Chicago, IL, USA).



(a) ROC curve of logistic nomogram with 3 interactions

(b) ROC curve of Bayesian nomogram

Figure 4.4. ROC curve of T2D nomograms.

5. 토의 및 결론

본 논문에서는 T2D 발병 여부를 이용하여 로지스틱 회귀모형과 순수 베이지안 분류기 모형을 만들고 이를 시각화 하기 위한 통계적 도구인 노모그램을 구축하였다. 나아가 로지스틱 회귀모형에서는 독립적인 효과뿐만 아니라 상호작용을 고려하여 노모그램을 새로 구축하고, 베이지안 노모그램에서는 left-aligned 방법을 사용해서 비교를 용이하게 하고자 하였다. 분석 자료는 2013–2015년 6기 국민건강영양조사 데이터를 사용하였고 최종적으로 13,474명의 데이터가 사용되었다. 위험요인으로는 11개(Age, Sex, Edu, Employment, Income, Smoking status, WC, Family history, Hypertension, Dyslipidemia, CVD)가 사용되었다. 두 노모그램 모두 나이가 60–85세 사이 범주에서 100점의 점수를 가지며 가장 큰 영향도를 나타내었다. 로지스틱 노모그램의 경우 가족력, 이상지질혈증, 고혈압, 심혈관계 질환, 비만여부 순으로 영향이 크게 나타났다. 그리고 베이지안 노모그램의 경우 고혈압, 심혈관계 질환, 이상지질혈증, 비만여부, 가족력 순으로 영향이 크게 나타났다.

로지스틱과 베이지안의 노모그램은 직관적으로 비교하여도 차이가 보인다. 이는 실제로 베이지안의 경우 조건부 확률 계산 과정에서 요인간 상호작용이 포함되는 반면, 로지스틱 회귀 모형의 경우 요인간의 상호작용은 따로 항에 포함하지 않는 이상 고려되지 않기 때문이다. 로지스틱 회귀모형의 경우 상호작용을 고려하고자 할 때, 의학적인 근거와 통계적 유의성을 모두 고려하여 상호작용 항을 선정한다면, 보다 예측력이 좋은 모형을 구축할 수 있다 (Iasonos 등, 2008). 하지만 항이 많아질수록 모형이 복잡해져 통계적인 의미가 없어 질 수 있으므로 주의가 필요하다. 베이지안 노모그램의 경우 요인 간 상호작용이 각 위험요인 안에 포함되어 있어 사용이 편리하다. 하지만 위험요인의 범주 하나하나에 계산이 필요한 것이 단점이다. 따라서 자료의 특성에 맞게 모형을 적합시키고 노모그램을 구축하는 것이 효율적이다. 이렇게 구축된 연구 모형은 향후 2016년 제 7기 국민건강영양조사 데이터를 활용한 코호트 연구 목적으로 적용될 계획이다.

References

Chung, S. M., Park, J. C., Moon, J. S., and Lee, J. Y. (2018). Novel nomogram for screening the risk of developing diabetes in a Korean population, *Diabetes Research and Clinical Practice*, **142**, 286–293.

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

Heo, M. H. and Lee, Y. G. (2008). *Data Mining Modeling and Example*, Hannarae, Seoul.

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, Y. J., Jeon, J. Y., Han, S. J., Kim, H. J., Lee, K. W., and Kim, D. J. (2015). Effect of socio-economic status on the prevalence of diabetes, *Yonsei Medical Journal*, **56**, 641–647.

Korean Diabetes Association (2017). Korean diabetes fact sheet in Korea 2016. Publish: diabetes fact sheet in Korea, Available from: http://www.diabetes.or.kr/pro/news/admin.php?category=A&code=admin&number=1428&mode=view

Lee, J. W., Park, M. R., and Yu, H. N. (2005). *Statistical Method for Bioscience Research*, Freedom academy, Seoul.

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.

Lee, S. Y., Park, H. S., Kim, S. M., *et al.* (2006). Cut-off points of waist circumference for defining abdominal obesity in the Korean population, *The Korean Journal of Obesity*, **15**, 1–9.

Možina, M., Demšar, J., Kattan, M., and Zupan, B. (2004). Nomogram for Visualization of Naive Bayesian Classifier, *Knowledge Discovery in Databases: PKDD 2004*, **4**, 337–348.

Park, J. C. and Lee, J. Y. (2017). Risk factors for type 2 diabetes among Korean adults in 2014, *Quantitative Bio-Science*, **36**, 15–21.

Park, J. C. (2018). Proposal of nomogram using logistic and Bayesian technique for type 2 diabetes (Master's thesis), Yeungnam University, Gyeongsan.

Park, J. C. and Lee, J. Y. (2018). How to build nomogram for type 2 diabetes using a naïve Bayesian classifier technique, *Journal of Applied Statistics*, **45**, 2999-3011.

Statistics Korea (2014). Causes of death statistics 2014. Policy News, Available from: http://kostat.go.kr/portal/korea/kor_nw/3/index.board?bmode=read&aSeq=348541

Yang, D. (2014). Build prognostic nomograms for risk assessment using SAS. In *Proceedings of SAS Global Forum 2013*, Available from: http://support.sas.com/resources/papers/proceedings13/264-2013.pdf

제 2형 당뇨병을 이용한 로지스틱과 베이지안 노모그램 구축 및 비교

박재철[a] · 김민호[a] · 이제영[a,1]

[a]영남대학교 통계학과

(2018년 4월 9일 접수, 2018년 5월 23일 수정, 2018년 6월 8일 채택)

요 약

본 연구에서는 제 2형 당뇨(type 2 diabetes mellitus)의 발병 확률을 예측하기 위해 11가지 위험요인을 가지고 로지스틱 회귀모형과 순수 베이지안 분류기 모형에 적합시킨다. 그런 다음 이를 시각적으로 쉽게 이해하는데 도움을 주는 노모그램 구축 방법을 소개한다. 분석은 2013–2015년 6기 국민건강영양조사 데이터를 가지고 분석하였다. 또 로지스틱 회귀모형에 세 가지 상호작용 항을 넣어 분석의 질을 높이고자 하였고 베이지안 노모그램에 left-aligned 방법을 사용하여 비교하기 쉽게 만들었다. 최종적으로 두 노모그램을 비교하고 효용성을 알아보았다. 마지막으로 ROC 곡선을 이용하여 노모그램이 적절한지 검증하였다.

주요용어: 제 2형 당뇨(type 2 diabetes mellitus), 로지스틱 회귀모형, 순수 베이지안 분류기 모형, 노모그램, ROC curve

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