The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.1.025
33(1), 25–46

Bayesian ordinal probit semiparametric regression models: KNHANES 2016 data analysis of the relationship between smoking behavior and coffee intake

Dasom Lee[a] · Eunji Lee[b] · Seogil Jo[c] · Taeryeon Choi[b,1]

[a]Department of Statistics, North Carolina State University;
[b]Department of Statistics, Korea University;
[c]Department of Statistics (Institute of Applied Statistics), Jeonbuk National University

(Received October 22, 2019; Revised December 2, 2019; Accepted December 10, 2019)

Abstract

This paper presents ordinal probit semiparametric regression models using Bayesian Spectral Analysis Regression (BSAR) method. Ordinal probit regression is a way of modeling ordinal responses - usually more than two categories - by connecting the probability of falling into each category explained by a combination of available covariates using a probit (an inverse function of normal cumulative distribution function) link. The Bayesian probit model facilitates posterior sampling by bringing a latent variable following normal distribution, therefore, the responses are categorized by the cut-off points according to values of latent variables. In this paper, we extend the latent variable approach to a semiparametric model for the Bayesian ordinal probit regression with nonparametric functions using a spectral representation of Gaussian processes based BSAR method. The latent variable is decomposed into a parametric component and a nonparametric component with or without a shape constraint for modeling ordinal responses and predicting outcomes more flexibly. We illustrate the proposed methods with simulation studies in comparison with existing methods and real data analysis applied to a Korean National Health and Nutrition Examination Survey (KNHANES) 2016 for investigating nonparametric relationship between smoking behavior and coffee intake.

Keywords: BSAR, Gaussian process, KNHANES data, Markov chain Monte Carlo, Ordinal probit, Semiparametric regression

1. 서론

범주형 자료는 자료의 값이 범주(category)로 주어지는 경우로, 일반적으로, 범주의 개수에 따라 범주가 2개일 경우 이항형(binary), 3개 이상일 경우 다항형(polychotomous)으로 나뉜다. 또한 범주의 순서의 여부에 따라, 순서가 없을 때 명목형(nominal), 순서가 있을 때 순서형(ordinal)으로 분류할 수 있다 (Agresti, 2013). 예를 들어, 질병에 걸렸는 지, 질병에 걸리지 않았는 지의 여부는 이항형이면서 명목

This research was supported by a Korea University Grant (K1910951).
[1]Corresponding author: Department of Statistics, Korea University, 145 Anam Rd, Seongbuk Gu, Seoul 02841, Korea. E-mail: trchoi@korea.ac.kr

형에 해당하고, 소비자가 물건 A, B, C 중 하나를 선택하는 경우는 다항형이면서 명목형이다. 또한, 성적, 신체검사 결과, 소득 수준, 행복의 정도를 여러 등급으로 분류하는 경우 각 범주의 순서가 정해져 있기 때문에 다항형이면서 순서형에 해당한다. 이러한 범주형 자료 분석은 통계학 뿐만 아니라 사회과학, 경제학, 의학, 심리학, 보건학 등 다양한 분야에서 활용되고 있으며, 최근 더욱 많은 활용과 사례연구들이 진행되고 있다 (Kim, 2015; Park 등, 2018; Lee와 Heo, 2014; Kockelman과 Kweon, 2002; Clark 등, 2001; Seok 등, 2017). 이러한 범주형 자료 분석을 위해서는, 연속형 자료 분석에 적합한 선형 회귀 모형 대신 일반화 선형 모형(generalized linear model)을 사용한다. 일반화 선형 모형은 Nelder과 Wedderburn (1972)가 처음 고안한 모형으로, 반응변수의 분포가 지수분포족(exponential family)일 때 설명변수와 반응변수를 특정한 연결함수(link function) $g(\cdot)$를 통하여 유연하게 모형을 적합하는 방식이다. 예를 들어, 이항형 반응변수를 위한 일반화 선형모형의 경우, 설명변수들의 선형결합과 반응변수의 평균을 연결하는 함수가 로짓(logit)함수일 때는 로짓 또는 로지스틱 모형, 프로빗(probit)함수일 때는 프로빗 모형이라 부른다. 로짓 함수는 $\text{logit}(p) = \log(p/(1-p))$를, 프로빗 함수는 표준정규분포의 누적분포함수의 역함수 $\Phi^{-1}(p)$를 각각 의미한다. 또한, 설명변수에 대한 선형모형 뿐 아니라, 특정한 비선형 모형를 가정하거나, 보다 일반적으로 선형모형과 불특정한 비선형모형을 가정하는 비모수 함수 모형이 결합된 준모수 모형을 적합하려는 연구도 활발히 진행되어왔다. 대표적인 예는 Hastie와 Tibshirani (1990)의 generalized additive model (GAM)으로서, 이항자료에 대하여 연결 함수를 로짓 함수로 사용하고, 설명변수간의 선형결합 대신, 알려지지 않은 비모수 함수의 합으로 대체하고, 각각의 비모수 함수를 스플라인 모형을 통해 추정하는 방식이다. GAM은 다양한 방식으로 확장되어 왔으며, `R`의 `mgcv`패키지 (Wood 2017)를 활용하여 회귀모형 뿐 아니라 이항형 로지스틱 및 프로빗 모형, 다항형 로지스틱 및 프로빗 모형 등을 적합할 수 있다.

명목형 자료분석을 위한 베이지안 방법의 기본적인 모형은 Albert와 Chib (1993)의 잠재변수를 이용한 이항형 프로빗 회귀모형이라 할 수 있다. Albert와 Chib (1993)은 이항확률에 프로빗 연결함수를 직접 적용하는 방식 대신 선형회귀모형을 통해 설명된는 잠재변수 값에 따라 이항형 반응변수를 분류하는 방식을 채택하였다. 이러한 접근방식은 사전 분포와 사후 분포 간의 켤레성(conjugacy)으로 인해 사후분포 도출에 용이하기 때문에 베이지안 범주형 자료 분석의 토대가 되었다. 명목형 변수를 위한 비모수적 베이지안 방법에 있어서는 Chipman 등 (2010)의 Bayesian additive regression tree (BART)가 대표적인 방법이라고 할 수 있다. BART는 정규화 사전 분포(regularization prior)를 사용하고, 반응변수의 확률을 프로빗 연결함수를 통해 m개의 설명력이 약한 나무들의 합으로 표현하는 방식으로, `R`의 `BART` 패키지의 `pbart`함수를 통해 적합할 수 있다. BART를 통한 비모수적 접근 방법 외에, Wood (2017)에서는 `mgcv`패키지의 `jagam`함수를 통해 GAM 모형을 베이지안 방식으로 적합할 수 있도록 하였다. 가장 최근에 이루어진 비모수적 베이지안 접근 방법에는 Bayesian Spectral Analysis Regression (BSAR) (Lenk와 Choi, 2017)로 알려진 가우지안 확률과정을 통한 비모수 함수 적합 방법이 있으며, Jo 등 (2019)에서 제공하는 `bsamGP`패키지의 `gbsar`함수를 통해 선형모형과 비모수모형의 합으로 표현되는 준모수(semiparametric) 일반화 모형을 적합할 수 있다.

순서형 자료는 명목형 자료와 달리 반응변수의 범주 간 순서를 고려하는 자료로서, 순서형 로지스틱 회귀 분석 혹은 누적 로짓 모형(cumulative logit model)을 사용하여 설명변수가 주어졌을 때, Q개의 순서형 범주를 가진 반응변수의 누적 확률을 모형화 할 수 있다. 베이지안 방법론은 Albert와 Chib (1993)이 제안한 순서형 프로빗 회귀가 대표적으로, 자료 확대 방법(data augmentation)과 프로빗 연결함수를 사용하고, 선형회귀모형을 통해 설명되는 잠재변수가 속하는 구간에 따라 y를 분류한다. 이항형 프로빗 회귀모형과 마찬가지로 켤레성을 이용하여 깁스 표집(Gibbs sampling)을 사용할 수 있기 때문에 선호되는 베이지안 접근방법이라 하겠다. 이러한 Albert와 Chib (1993)은 여러 방식으로

확대되어, 다양한 실제 자료분석에 활용되었다 (Cowles 등, 1996; Chen과 Dey, 2000; Xie 등, 2009; Hasegawa, 2010). 예를 들어, Cowles 등 (1996)은 Albert와 Chib (1993)이 적용한 깁스 표집방식을 확장한 혼합 깁스/메트로폴리스-해스팅스 알고리즘(hybrid Gibbs/Metropolis-Hastings algorithm)을 사용하였고, Chen과 Dey (2000)는 프로빗 연결함수 대신 혼합 다변량 정규분포 연결함수를 이용해 상관관계가 있는 순서형 자료를 분석하였다. 또한 Harris와 Zhao (2007)은 영과잉 순서형 프로빗 모형(zero inflated ordered probit; ZIOP)을 발전시켰고, 담배 소비 설문조사 자료에 모형을 적합시켰으며, Sha와 Dechi (2019)에서는 모수의 사후 분포 표본 추출의 비효율성을 해결하고자 무정보 사전분포(non-informative prior) 대신 디리슈레(Dirichlet) 사전 분포를 사용하는 베이지안 순서형 분류기법 활용을 연구하였다. 이러한 베이지안 접근방법들은 주로 모수적 선형모형에 국한되고, 비모수 함수를 고려하는 모형들은 많은 연구가 이루어지지 않았으나, 최근 이루어진 연구의 예로는 Wood (2017)과 Jara 등 (2009)가 있다. Wood (2017)은 비모수 모형 GAM을 순서형 자료 분석에 적용시키기 위해서 y의 범주를 잠재변수 μ에 따라 분류하는 누적 분계점 모형(cumulative threshold model)을 바탕으로 로짓 연결함수를 사용하였고, Jara 등 (2009)는 Albert와 Chib (1993)의 방식에 따라 고정 효과를 모형화하고, 변량 효과에 다변량 폴랴 나무 혼합 분포(mixtures of multivariate Polya trees)를 사용하는 베이지안 준모수 일반화 선형 혼합 모형(Bayesian semiparametric linear mixed model)으로 발전시켰다.

이에, 본 논문에서는 순서형 자료 분석을 위한 보다 일반적인 베이지안 준모수 모형을 제안하고자 한다. 구체적으로는 순서형 프로빗 회귀 모형에 Lenk와 Choi (2017)에서 연구된 BSAR 방법을 접목하여, 잠재변수를 선형모형과 비선형 모형의 합으로 표현하는 순서형 준모수 프로빗 회귀 모형을 제안하고자 한다. BSAR 방법론은 알려지지 않은 비선형 함수를 추정하기 위해 함수에 가우시안 과정(Gaussian process; GP) 사전 분포를 가정하고, 단조증가 또는 감소, 오목 또는 볼록 등과 같은 함수의 형태 제약을 줄 수 있는 방법이다. 이를 바탕으로, 본 논문에서 제안하는 순서형 준모수 프로빗 회귀 모형에서는 잠재변수를 설명하는 비모수 함수에 단조증가와 같은 형태제약을 주어, 설명변수와 반응변수 간의 관계에 대한 사전 지식을 활용하여 보다 나은 자료 분석을 할 수 있다는 점에서 기존의 연구와 차별되는 장점을 갖는다.

이를 위한, 본 논문의 구성은 다음과 같다. 2절에서는 본 논문이 제안하는 베이지안 순서형 프로빗 준모수 회귀모형에 대하여 설명한다. 이를 위하여, 먼저 잠재변수를 이용한 기본적인 베이지안 순서형 프로빗 회귀모형을 설명하고, 비모수 함수를 추정하기 위해 필요한 BSAR 의 개념을 설명한다. 이를 바탕으로 제안하는 순서형 프로빗 준모수 회귀 모형 구조에 대하여 논의하고, 베이지안 순서형 프로빗 준모수 회귀 모형 하에서의 사후 분포와 이를 통한 베이지안 사후 추론 과정을 설명한다. 3절에서는 모의실험을 통하여 이항형 프로빗 준모수 회귀모형과 기존의 다른 모형들 간의 적합결과를 비교하고, 형태 제약에 따른 순서형 프로빗 준모수 회귀모형의 적합결과를 비교 분석하도록 한다. 4절에서는 순서형 프로빗 준모수 회귀 모형에 대한 실제 자료 분석에 대해 고찰한다. 구체적으로, 국민건강영양조사 제7기 1차년도 (2016) 자료(Korean National Health and Nutrition Examination Survey (KNHANES), 2016)를 바탕으로, 프로빗 준모수 회귀모형을 활용하여 흡연과 커피의 섭취량 간의 관계에 대한 실증적 분석을 수행한다. 5절에서는 본 논문을 정리하고 추가 연구 방향과 향후 과제에 대해 논의하도록 한다.

2. 베이지안 순서형 프로빗 준모수 회귀모형

2.1. 잠재변수를 이용한 베이지안 순서형 프로빗 회귀모형

베이지안 순서형 프로빗 회귀모형의 토대가 되는 Albert와 Chib (1993)의 방법은 이항형 프로빗 회귀를 정규분포를 따르는 잠재 변수로 표현한 것으로부터 시작한다. 즉, 반응변수 y_i가 확률 p_i를 가진 베

르누이 분포를 따를 때, Albert와 Chib (1993)은 확률 p_i에 식 (2.1)과 같이 정규분포의 누적분포함수의 역함수를 직접 적용하는 방식 대신 잠재변수 z_i의 값에 따라 y_i를 분류하는 식 (2.2)의 방식을 채택한다.

$$y_i \sim \text{Bernoulli}(p_i), \quad \Phi^{-1}(p_i) = \boldsymbol{w}_i^\top \boldsymbol{\beta}, \tag{2.1}$$

$$y_i = \begin{cases} 0, & z_i \le 0, \\ 1, & z_i > 0, \end{cases} \tag{2.2}$$

$$z_i = \boldsymbol{w}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

이러한 잠재변수를 이용한 이항형 프로빗 방식을 확장하여 Albert와 Chib (1993)은 Q개의 순서화된 범주형 변수 $y_i \in \{1, 2, \ldots, Q\}$, $i = 1, \ldots, n$를 모형화하는 순서형 프로빗 회귀 방법을 제안하였다. 이항형 프로빗 회귀는 식 (2.2)와 같이, 0을 기준으로 한 z_i의 값에 따라 y_i의 이항 범주가 0과 1로 나눠지는 반면, 순서형 프로빗 회귀는 $Q-1$개의 지점 $(a_1, \ldots, a_{Q-1})$에 따라 y가 Q개의 범주로 구분된다.

$$y_i = \begin{cases} 1, & -\infty < z_i \le 0, \\ 2, & 0 < z_i \le a_2, \\ \vdots & \vdots \\ Q, & a_{Q-1} < z_i < \infty, \end{cases}$$

$$z_i = \boldsymbol{w}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1),\ i = 1, \ldots, n,$$

여기서 $a_0 = -\infty$, $a_1 = 0$, $a_Q = \infty$으로 정의된다. 순서형 프로빗 모형에서는, 절단점 $a_2, \ldots, a_{Q-1}$은 모수로 취급되기 때문에 사후분포를 통해 추정되며, 절단점을 바탕으로 명목형 변수 y_i가 q번째 순서에 속할 확률은 다음과 같이 표준정규분포의 누적분포함수 $\Phi(\cdot)$로 표현됨을 알 수 있다.

$$\text{P}\left(y_i = q | \boldsymbol{\beta}, \{a_q\}\right) = \text{P}\left(a_{q-1} < z_i \le a_q\right) = \Phi\left(a_q - \boldsymbol{w}_i^\top \boldsymbol{\beta}\right) - \Phi\left(a_{q-1} - \boldsymbol{w}_i^\top \boldsymbol{\beta}\right).$$

프로빗 회귀 모수 $\boldsymbol{\beta}$, 절단점 모수 벡터 $\boldsymbol{a} = [a_2, a_3, \ldots, a_{Q-1}]$, 그리고 잠재변수 벡터 $\boldsymbol{z}$의 결합 사후 분포는 다음과 같이 표현된다.

$$p(\boldsymbol{a}, \boldsymbol{\beta}, \boldsymbol{z} | y) \propto p(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{\beta}, \boldsymbol{a}) p(\boldsymbol{\beta}, \boldsymbol{a}), \tag{2.3}$$

식 (2.3)에서 $p(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{\beta}, \boldsymbol{a}) = p(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{a}) p(\boldsymbol{z} | \boldsymbol{\beta}, \boldsymbol{a})$는 완전 가능도(full likelihood) 또는 확대 가능도(augmented likelihood) 함수를 의미한다. 이 경우 $\boldsymbol{a}$와 $\boldsymbol{\beta}$에 무정보 사전 분포를 가정하면, 식 (2.3)의 결합 사후 분포를 통하여, $\boldsymbol{\beta}$의 완전 조건부 사후 분포(full conditional posterior distribution)는 정규분포를, z_i의 완전 조건부 사후 분포는 $(a_{y_i} - 1, a_{y_i})$에서 절단된 정규분포를 따름을 유도할 수 있다.

$$\boldsymbol{\beta} | \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{y} \sim N\left(\left(X'X\right)^{-1} X'\boldsymbol{z}, \left(X'X\right)^{-1}\right)$$

$$z_i | \boldsymbol{a}, \boldsymbol{\beta}, z_{-i}, y \overset{\text{ind}}{\sim} TN_{(a_{y_i - 1}, a_{y_i}]}\left(\boldsymbol{w}_i^\top \boldsymbol{\beta}, 1\right), \quad i = 1, 2, \ldots, n.$$

또한, 이를 바탕으로, 절단점 모수 벡터 $\boldsymbol{a}$의 완전 조건부 사후 분포가 다음과 같이 균일분포를 따르게 됨을 확인할 수 있다.

$$a_q | a_{-q}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{y} \propto \prod_{i: y_i = q} I\left(a_{q-1} < z_i \le a_q\right) \prod_{i: y_i = q+1} I\left(a_q < z_i \le a_{q+1}\right)$$

$$\Leftrightarrow a_q | a_{-q}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{y} \overset{\text{ind}}{\sim} \text{Unif}\left[l_q, u_q\right], \ \ l_q = \max\left\{a_{q-1}, \max_{i: y_i = q} z_i\right\}, \ u_q = \min\left\{a_{q+1}, \min_{i: y_i = q+1} z_i\right\}.$$

이러한 각각의 완전 조건부 사후분포를 이용하여 Albert와 Chib (1993)은 깁스표집방법을 통한 마르코프 연쇄 몬테칼로(Markov chain Monte Carlo; MCMC) 방법을 적용하는 베이지안 순서형 프로빗 모형을 연구하였다.

2.2. 베이지안 순서형 프로빗 준모수 회귀모형

본 논문에서 제안하는 베이지안 순서형 프로빗 준모수 회귀모형은 2.1절에서 설명한 Albert와 Chib (1993)의 순서형 프로빗 선형모형을 확장하여, 잠재변수 z_i가 설명변수 $\boldsymbol{w}_i$에 대한 선형모형 $\boldsymbol{w}_i^\top \boldsymbol{\beta}$와 설명변수 x_i의 비모수 함수 $f(x_i)$의 합으로 표현되는 준모수 프로빗 회귀모형으로서, 식 (2.4)의 모형으로 표현된다.

$$y_i = \begin{cases} 1, & -\infty < z_i \le 0, \\ 2, & 0 < z_i \le a_2, \\ \vdots & \vdots \\ Q, & a_{Q-1} < z_i < \infty, \end{cases} \tag{2.4}$$

$$z_i = \boldsymbol{w}_i^\top \boldsymbol{\beta} + f(x_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1),\ i = 1, \ldots, n.$$

특히, 본 논문이 제안하는 베이지안 순서형 프로빗 준모수 회귀모형에서는 식 (2.4)의 비모수 함수 $f(x_i)$를 Lenk와 Choi (2017)에서 연구한 BSAR 방법을 통해 추정하고자 한다.

BSAR 방법은 Lenk와 Choi (2017)에서 제안된 비모수 베이지안 함수 추정방법으로서, 알려지지 않은 비모수 함수에 GP 사전 분포를 가정하고, 이를 푸리에 급수(Fourier series)의 무한 합으로 근사하여 함수를 적합하게 되는 접근방식이며, 비모수 함수에 대한 적합 뿐 아니라, 함수가 단조증가, 단조감소, 오목, 볼록 등의 형태 제약을 갖는 경우에도 함수적합이 용이하다. 구체적으로, BSAR 방법에서는 비모수 함수 $f(x)$를 가우시안 확률과정 $Z(x)$라고 가정하고, Karhunen-Loéve 전개(expansion)를 통하여 다음과 같이 무한개의 직교기저함수(orthogonal basis function)의 선형 결합으로 표현할 수 있다.

$$f(x) = Z(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x).$$

Lenk와 Choi (2017)는 구체적으로 cos 함수를 직교기저함수$\{\varphi_j\}$ ($\varphi_0(x) = 1$, $\varphi_j(x) = \sqrt{2}\cos(\pi j x)$, $j = 1, 2, \ldots$, $0 \le x \le 1$)로 사용하였고, 실제 적합을 위하여, J개의 코사인 함수들을 사용하는 절단형태의 표현을 통해 주어진 함수 $f(x)$를 근사하는 방식을 적용하였다.

$$f(x) \approx f_J(x) = \sum_{j=1}^{J} \theta_j \varphi_j(x).$$

BSAR 방법은 베이지안 비모수 회귀 함수 추정에 있어서 유연하게 적합할 뿐 아니라, 회귀 함수에 특정한 제약조건을 쉽게 반영할 수 있다는 장점을 제공한다. 회귀 함수 $f(x)$가 단조증가/감소 또는 오목/볼록과 같은 형태적 제약(shape restriction)이 필요한 경우, BSAR 방법은 도함수(derivative)를 모형화함으로써, 회귀함수의 형태 제약을 회귀 모형에 반영할 수 있게 된다. 구체적으로, 회귀함수 f가 단조증가/감소한다고 가정하면, 1차 도함수를 다음과 같이 가우시안 확률과정 $Z(x)$의 제곱으로 모형화시킨다.

$$f'(x) = \delta Z^2(x), \quad Z(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x),\ \delta \in \{-1, 1\}, \tag{2.5}$$

여기서 δ는 회귀함수가 증가 또는 감소라는 제약에 따라 각각 1과 -1의 값을 갖게된다. 이를 바탕으로, 비모수 함수 $f(x)$는 단조증가 및 감소의 형태를 가정한 $f'(x) = \delta Z^2(x)$의 적분을 통해, 다음과 같은 형태로 표현됨을 알 수 있다.

$$f(x) = \delta \left[\int_0^x Z^2(s)ds - \int_0^1 \int_0^x Z^2(s)dsdx \right] = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \theta_j \theta_k \varphi_{j,k}^a(x),$$
$$\varphi_{j,k}^a(x) = \int_0^x \varphi_j(s)\varphi_k(s)ds - \int_0^1 \int_0^x \varphi_j(s)\varphi_k(s)dsdx.$$

BSAR를 결합한 순서형 프로빗 준모수 회귀 모형을 위하여, 다음과 같은 사전분포를 고려하도록 한다. 먼저, BSAR 방식으로 표현되는 비모수 함수 $f(x)$에 대하여, cos 기저함수인 $\varphi_j(x)$의 계수 θ_j에 대하여 다음과 같은 계층적 정규분포를 따르는 사전 분포를 설정한다.

$$\theta_j | \tau^2, \gamma \sim \mathcal{N}\left(0, \tau^2 \exp[-j\gamma]\right), \quad j \geq 1. \tag{2.6}$$

형태제약이 없는 경우에는, 비모수 함수의 절편(상수항)에 해당하는 θ_0는 모수항의 절편 β_0와 합쳐지므로 $\theta_0 = 0$으로 설정하고, 형태제약이 있는 경우에는 θ_0에 대하여 $\theta_0 \sim \mathcal{N}(0, \sigma_{\theta_0}^2) I(\theta_0 \geq 0)$의 절단(truncated) 정규분포를 설정한다. 아울러, 계층적 정규분포의 초모수 τ^2와 γ에 대해서도 각각 역감마분포(inverse-gamma; IG)와 지수분포의 추가적인 사전 분포를 설정하고, 모수적 모형의 선형회귀계수벡터 $\boldsymbol{\beta}$에 대해서는 다차원 정규분포를 다음과 같이 설정하도록 하도록 한다.

$$\tau^2 \sim \text{IG}\left(\frac{r_0}{2}, \frac{s_0}{2}\right), \quad \gamma \sim \text{Exp}(w_0), \quad \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{m}_\beta, \boldsymbol{V}_\beta).$$

이러한 사전분포와 식 (2.4)의 잠재변수를 통한 순서형 프로빗 준모수 회귀모형의 가능도를 바탕으로, 알려지지 않은 모수 $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, τ^2, γ와 잠재변수벡터 $\boldsymbol{z}$의 결합사후분포는 다음과 같이 표현된다.

$$\begin{aligned}
&\pi\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2, \gamma, \psi, \boldsymbol{z} | \boldsymbol{y}\right) \\
&\propto \prod_{i=1}^{n} \Phi\left(a_{y_i} - \boldsymbol{w}_i^\top \boldsymbol{\beta} - f(x_i)\right) - \Phi\left(a_{y_i - 1} - \boldsymbol{w}_i^\top \boldsymbol{\beta} - f(x_i)\right) \times \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left(z_i - \boldsymbol{w}_i^\top \boldsymbol{\beta} - f(x_i)\right)^2\right) \\
&\quad \times \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{m}_\beta\right)^\top \boldsymbol{V}_\beta^{-1}\left(\boldsymbol{\beta} - \boldsymbol{m}_\beta\right)\right) \times \exp\left(-\frac{\theta_0^2}{2\sigma_{\theta_0}}\right) I(\theta_0 \geq 0) \\
&\quad \times \prod_{j=1}^{J} \left\{\left(\tau^2 \exp(-j\gamma)\right)^{-\frac{1}{2}} \exp\left(-\frac{\theta_j^2}{2\tau^2 \exp(-j\gamma)}\right)\right\} \times \exp(-w_0\gamma) \times (\tau^2)^{-\frac{r_0}{2}-1} \exp\left(-\frac{s_0}{2\tau^2}\right).
\end{aligned} \tag{2.7}$$

이 경우, θ_0에 대한 사전 분포는 형태제약이 있는 경우에만 곱해진다.

식 (2.7)과 같이 표현된 결합사후분포를 바탕으로, MCMC 방법을 통하여 각 모수의 사후표본을 추출하여 사후추론을 실시하도록 한다. 구체적으로, $\boldsymbol{\beta}$|Rest는 $\boldsymbol{\beta}$의 완전조건부 분포로서, Rest는 완전조건부 분포가 계산되는 해당 모수($\boldsymbol{\beta}$)를 제외한 나머지 모수 전체와 데이터를 의미하며, 정규 분포의 켤레성으로 인하여 다음과 같은 다차원 정규분포가 됨을 쉽게 확인할 수 있으며, 깁스 표집(Gibbs sampling)을 통해 사후 분포를 쉽게 계산할 수 있다.

$$f(\boldsymbol{\beta}|\text{Rest}) \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left(z_i - \boldsymbol{w}_i^\top \boldsymbol{\beta} - f\left(x_i\right)\right)^2\right) \times \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{m}_{\boldsymbol{\beta}}\right)^\top \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta} - \boldsymbol{m}_{\boldsymbol{\beta}}\right)\right),$$
$$\boldsymbol{\beta}|\text{Rest} \sim N\left(\boldsymbol{m}_{\boldsymbol{\beta}_n}, \boldsymbol{V}_{\boldsymbol{\beta}_n}\right), \quad \boldsymbol{V}_{\boldsymbol{\beta}_n} = \left(\boldsymbol{V}_{\boldsymbol{\beta}}^{-1} + \boldsymbol{W}^\top \boldsymbol{W}\right)^{-1}, \; \boldsymbol{m}_{\boldsymbol{\beta}_n} = \boldsymbol{V}_{\boldsymbol{\beta}_n}\left[\boldsymbol{W}^\top(\boldsymbol{z} - \boldsymbol{f}) + \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{m}_{\boldsymbol{\beta}}\right].$$

비모수 $f(x)$의 cos기저함수 φ_j의 계수 θ_j에 대한 사후분포를 도출하는 방법은 형태 제약 유무에 따라서 달라진다. 형태제약이 없을 때, $\theta_0 = 0$이고, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$의 사전 분포는 식 (2.6)과 같이, τ^2와 γ가 주어진 경우에는, 정규분포로서, 완전조건부 분포역시 다음과 같이 정규분포로 계산됨을 알 수 있다.

$$f(\theta_1, \ldots, \theta_J|\text{Rest}) \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left(z_i - \boldsymbol{w}_i^\top \boldsymbol{\beta} - \sum_{j=1}^{J} \theta_j \varphi_j(x_i)\right)^2\right) \times \exp\left(-\frac{\theta_j^2}{2\tau^2 \exp(-j\gamma)}\right),$$

$$\theta_j|\text{Rest} \sim N\left(m_{j_n}, \sigma_{j_n}^2\right), \quad \sigma_{j_n}^2 = \left(\varphi_j^2 + \frac{1}{\tau^2 \exp(-j\gamma)}\right)^{-1}, \; m_{j_n} = \sigma_{j_n}^2 \left(z_i - \boldsymbol{w}_i^\top \boldsymbol{\beta}\right) \varphi_j.$$

이에 반해, 형태 제약을 갖는 경우에는 $f(x)$는 식 (2.5)와 같이 가우지안 확률과정의 제곱함수의 적분형태로 표현되기 때문에, θ_j의 사후 분포계산을 위해서 정규분포의 켤레성을 이용할 수 없게 된다. 이러한 문제를 해결하기 위해서, Lenk와 Choi (2017)에서는 θ_j의 사후표본 추출을 위하여 적응 마르코프 연쇄 몬테칼로(adaptive MCMC)방법을 적용하였으며, 본 논문에서도 같은 방식의 사후표본추출 방식을 사용하도록 한다. 또한 역감마 사전 분포를 사용하는 τ^2의 완전조건부분포는 켤레성을 바탕으로 역감마 분포로 쉽게 유도되며,

$$f(\tau^2|\text{Rest}) \propto \prod_{j=1}^{J} \left\{ \left(\tau^2\right)^{-\frac{1}{2}} \exp\left(-\frac{\theta_j^2}{2\tau^2 \exp(-j\gamma)}\right)\right\} \times (\tau^2)^{-\frac{r_0}{2}-1} \exp\left(-\frac{s_0}{2\tau^2}\right),$$

$$\tau^2|\text{Rest} \sim \text{IG}\left(\frac{r_n}{2}, \frac{s_n}{2}\right), \quad r_n = r_0 + J, \; s_n = s_0 + \sum_{j=1}^{J} \frac{\theta_j^2}{\exp(-j\gamma)}.$$

지수분포를 사전분포로 사용하는 γ의 완전조건부분포는 알려진 형태가 아니기 때문에 slice 표집을 이용해 사후표본을 추출한다. 적응 마르코프 연쇄 몬테칼로방법과 slice 표집을 이용한 사후표본추출방법에 대한 보다 자세한 설명은 Lenk와 Choi (2017)을 참조하도록 한다. 마지막으로, 절단점 $\boldsymbol{a}$와 잠재변수 $\boldsymbol{z}$에 대한 완전조건부포는 2.1절에 설명된 바와 같은 방식으로 유도되며, 각각 다음과 같이 주어지고,

$$a_q|\text{Rest} \sim \text{Unif}[l_q, u_q], \quad l_q = \max\left\{a_{q-1}, \max_{i:y_i=q} z_i\right\}, \; u_q = \min\left\{a_{q+1}, \min_{i:y_i=q+1} z_i\right\},$$

$$z_i|\text{Rest} \sim \mathcal{N}\left(\boldsymbol{w}_i^\top \boldsymbol{\beta} + f(x_i), 1\right) I(a_{y_i-1} < z_i \le a_{y_i}).$$

이를 통한 사후표본추출이 이루어지게 된다.

3. 실증분석 I : 모의 실험

3절에서는 2절에서 제안한 순서형 프로빗 준모수 회귀모형에 대하여, 모의실험자료를 바탕으로 실증분석을 수행한다. 이를 위하여, 3.1절에서는 반응변수가 이항형이고 비모수 함수에 형태제약을 고려하지 않았을 때, BSAR에 기반한 이항형 프로빗 준모수 모형을 적합해보고, 기존에 제안된 프로빗 BART 모형과 베이지안 GAM (`jagam`) 모형과 비교해보도록 한다. 3.2절에서는 이항형 반응변수이지만 비모수 함수에 특정한 형태를 주었을 때 형태 제약 여부에 따른 BSAR 이항형 프로빗 회귀모형의 결과를 비교하고, 3개 이상의 순서화된 범주를 가진 반응변수를 적합하기 위하여 순서형 프로빗 준모수 회귀모형을 사용하고, 형태 제약 유무에 따른 적합결과를 비교하기 위하여 모의실험자료를 통하여 살펴본다.

3.1. 이항형 프로빗 준모수 회귀 모형 적합

비모수 함수 $f(x)$의 특정한 형태적 제약을 고려하지 않는 경우, BSAR에 기반한 이항형 프로빗 준모수

Table 3.1. Numerical summary of model comparison for three binary probit regression models

	BART	JAGAM	BSAR
LPML	−291.661	−283.675	−289.902
WAIC	583.169	578.270	579.809
RMSE	1.006	1.031	1.060

BART = Bayesian additive regression tree; JAGAM = just another Gibbs additive modeller; BSAR = Bayesian spectral analysis regression; LPML = log pseudo marginal likelihood; WAIC = Watanabe-Akaike information; RMSE = root mean squared error.

회귀 모형 적합을 위하여, 다음과 같이 모의실험자료를 생성하도록 한다.

$$
\begin{aligned}
y_i &= \begin{cases} 0, & z_i \le 0, \\ 1, & z_i > 0. \end{cases} \\
z_i &= f(x_i) + \epsilon_i, \quad f(x_i) = 30(x_i - 0.2)(x_i - 0.4)(x_i - 0.7), \\
x_i &\sim \text{Unif}(0,1),\ \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad i = 1, \ldots, 500.
\end{aligned}
$$

즉, 비모수 함수를 위한 설명변수 x_i는 균등분포 Unif$(0,1)$에서 표본을 추출하고, 평균이 $f(x_i)$이고 분산이 1인 정규분포를 따르는 잠재변수 z_i의 값에 따라 0과 1의 값을 갖는 이항형 반응변수 y_i를 생성하여 500개의 자료를 구성한다. 모의실험에 사용한 비모수 함수 $f(x)$는 단조 형태가 아니기 때문에, BSAR 방법적합에 있어서 형태제약이 없는 경우를 사용한다. 또한, BSAR 방법에 기반한 이항형 프로빗 준모수 회귀 모형의 성능을 비교하기 위해 이항자료 적합을 위한 BART와 베이지안 GAM 비모수 모형도 함께 적합해보도록 한다. 베이지안 GAM 모형에서는 비모수 함수적합을 위하여 B-스플라인 기저함수를 사용하였고, 매듭(knots)은 등간격으로 20개로 두어 `mgcv::jagam`함수를 사용한다. 또한, 세 가지 모형 모두에서 MCMC를 통한 사후표본 추출에서는 10,000개의 소각(burn-in)표본을 제외하고 10번째 표본마다 저장하여 총 1,000개의 사후표본을 얻도록 한다. BART의 초모수 $(\theta, \omega, a, b, \rho, \texttt{power}, \texttt{base}, \texttt{ntree}, \texttt{numcut})$는 $(0, 1, 0.5, 1, 2, 0.95, 50, 100)$로 사용하는 `BART::pbart`의 기본값으로 정하고, 베이지안 GAM과 마찬가지로 BSAR의 기저함수의 개수를 $J = 20$개로 정하여 비교해보도록 한다.

Table 3.1은 이러한 모의실험자료에 대하여 BART, 베이지안 GAM(just another Gibbs additive modeller; JAGAM), BSAR를 각각의 이항형 프로빗 회귀 모형을 통해 적합한 결과를 수치적으로 요약한 값을 나타낸다. 구체적으로, 세 모형간의 적합 성능을 비교하는 모형평가에 있어서, 모의 실험에서 생성한 $f(x_i)$와 추정된 $\hat{f}(x_i)$를 바탕으로 root mean squared error (RMSE)를 계산하고, 두 가지 모형 선택(model selection) 기준 log pseudo marginal likelihood (LPML)과 Watanabe-Akaike information (WAIC)를 계산하여 비교하도록 한다 (Geisser과 Eddy, 1979; Watanabe, 2010). 이러한 세 가지 수치값들은 T개의 사후표본을 바탕으로 계산되며, $\boldsymbol{\vartheta}^{(t)}$, $t = 1, \ldots, T$를 모수의 사후 표본이라 할 때, RMSE, LPML, WAIC는 다음과 같이 정의된다.

$$
\begin{aligned}
\text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - \hat{f}(x_i)\right)^2}, \\
\text{LPML} &= \sum_{i}^{n} \log(\text{CPO}_i), \quad \text{CPO}_i \approx \left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{p(y_i \mid \boldsymbol{\vartheta}^{(t)})}\right],
\end{aligned}
$$

$$\begin{aligned}\text{WAIC} &= -2\sum_{i}^{n}\left[\log \mathbb{E}_{\boldsymbol{\vartheta}|y}\left\{p(y_i \mid \boldsymbol{\vartheta})\right\} - \text{Var}_{\boldsymbol{\vartheta}|y}\left\{\log p(y_i \mid \boldsymbol{\vartheta})\right\}\right] \\ &\approx -2\sum_{i}^{n}\log\left[\frac{1}{T}\sum_{t=1}^{T} p\left(y_i|\boldsymbol{\vartheta}^{(t)}\right)\right] + 2\sum_{i}^{n}\frac{1}{T}\sum_{t=1}^{T}\left[\log p\left(y_i|\boldsymbol{\vartheta}^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T}\log p\left(y_i|\boldsymbol{\vartheta}^{(t)}\right)\right]^2.\end{aligned}$$

모형 간의 성능 비교시, LPML이 클수록, WAIC와 RMSE가 작을수록 해당모형이 자료를 더 잘 설명함을 의미하며, Table 3.1의 결과를 바탕으로 살펴볼 때, BSAR를 이용한 이항형 프로빗 준모수 회귀모형은 기존의 다른 비모수 모형과 비슷한 수준으로 모형을 적합하고 있음을 알 수 있다.

3.2. 형태제약을 반영하는 프로빗 준모수 회귀 모형 적합

3.2절에서는 비모수 함수가 단조증가 또는 감소하는 경우, 본 논문에서 제안하는 바와 같이 BSAR 방법을 통하여 증가 또는 감소의 형태제약을 반영하는 순서형 프로빗 회귀 모형의 성능을 고찰하고자 한다. 이를 위하여, 이항형 반응변수를 BSAR 이항형 프로빗 모형을 통해 적합하고, 3.1절에서 비교한 GAM, JAGAM 모형과 형태제약의 유무에 따른 BSAR 모형 간의 적합 결과를 비교한다. 또한 3개 이상의 순서화된 범주를 가진 반응변수를 적합하기 위하여 순서형 프로빗 준모수 회귀모형을 사용하고, 형태제약의 유무에 따른 적합결과 간의 비교를 모의실험자료 바탕으로 살펴보고자 한다.

먼저, 이항형 자료분석에 있어서 함수의 형태제약에 따른 BSAR 방법의 적합 결과를 비교하기 위해, 3.1절과 같은 방식으로 이항형 프로빗 회귀 모형에 따른 모의실험자료를 생성한다. 구체적으로, 다음과 같이 형태 제약에 따른 두 종류(단조감소, 단조증가)의 비모수 함수를 각각 고려하고, 3.1절의 이항형 프로빗 준모수 회귀 모형 구조로부터 표본크기가 500인 데이터를 생성하도록 한다.

- 함수 1 (단조감소): $f(x) = -\pi(x-0.5) - 3\sin(\pi(x-0.5)) + 3\sin(2\pi(x-0.5)) - 2\sin(3\pi(x-0.5)) + \sin(4\pi(x-0.5))$
- 함수 2 (단조증가): $f(x) = 0.8\log(0.2x+3) - 0.9$
- 함수 3 (단조증가): $f(x) = 2\sinh(4x-2) + 1$

함수 1은 단조감소, 함수 2와 함수 3은 모두 단조증가 함수이나, 함수 2는 오목한 형태의 단조증가 함수이고, 함수 3은 약간의 S곡선을 가진 단조증가 함수라는 차이가 있으며, 형태제약을 고려한다면 보다 나은 함수적합이 될 것으로 예상된다. 이러한 3개의 함수로부터 표현되는 이항 프로빗 회귀 모형으로부터 생성한 데이터에 형태 제약이 없는 BSAR 방법과 단조증가 또는 감소에 대한 형태 제약을 준 BSAR 방법을 각각 적합하여 적합결과를 비교하도록 한다. 이항형 프로빗 모형에 BSAR 방법을 적용하기 위해서, R의 `bsamGP` 패키지의 `gbsar` 함수를 통해 모형을 적합하고, 형태제약 유무에 따른 모형 간의 평가 기준은 RMSE, LPML, WAIC값을 사용하기로 한다. Table 3.2는 모의실험자료를 바탕으로 한 RMSE, LPML, WAIC의 값을 각 함수별로 형태제약 유무에 따른 요약값을 제공하며, 이 경우, RMSE가 더 작을 수록, LPML이 더 클 수록, WAIC가 더 작을 수록 더 선호되는 모형으로 판단한다. Table 3.2에서 형태제약을 반영하는 모형은 '단조', 반영하지 않는 모형은 '비단조'로 표시되며, 세가지 함수는 단조감소와 단조증가하는 함수이기 때문에 단조증가 또는 감소의 형태 제약을 주어 적합한 모형(단조)이 형태 제약을 주지 않은 모형(비단조)보다 좋은 결과를 나타냄을 알 수 있다. 따라서, 만약 잠재변수 z의 비모수 함수 $f(x)$가 어느 정도 일정한 단조성의 함수 형태를 지니고 있다면, 적절한 형태제약을 반영하는 적합방법을 사용하는 것이 적합면에서 더 좋은 결과를 낼 수 있음을 보여준다.

추가적으로, BART와 GAM 방법에서는 형태제약을 반영할 수 없으나, 형태제약 BSAR 방법의 성능을 확인해보기 위하여, 함수 2와 3을 바탕으로 생성된 동일한 모의실험자료를 적합해보고, 잠재변수 z를

Table 3.2. Numerical summary of LPML, WAIC, and RMSE for BSAR with/without shape restrictions

모형평가기준	함수 1 (단조감소)		함수 2 (단조증가)		함수 3 (단조증가)	
	비단조	단조	비단조	단조	비단조	단조
LPML	−183.848	−185.327	−347.821	−347.925	−88.369	−87.176
WAIC	367.476	370.660	695.647	695.853	176.666	174.343
RMSE	0.298	0.241	0.052	0.076	1.257	0.657

BSAR = Bayesian spectral analysis regression; LPML = log pseudo marginal likelihood; WAIC = Watanabe-Akaike information; RMSE = root mean squared error.

Table 3.3. Numerical summary of model assessment for JAGAM and BSAR with/without shape restrictions under functions (2) and (3)

평가기준	함수 2 (단조증가)			함수 3 (단조증가)		
	BART	JAGAM	BSAR (단조)	BART	JAGAM	BSAR (단조)
LPML	−350.973	−348.969	−347.925	−89.39	−87.314	−87.176
WAIC	701.863	697.943	695.853	178.643	174.638	174.343
RMSE	0.169	0.107	0.076	1.439	1.151	0.657
Accuracy	0.612	0.554	0.540	0.926	0.916	0.916
Sensitivity	0.609	0.551	0.541	0.920	0.932	0.932
Specificity	0.617	0.562	0.537	0.937	0.889	0.889

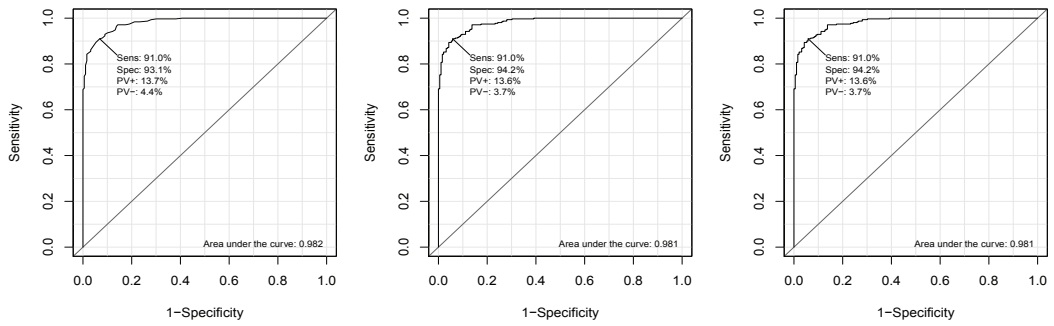
BART = Bayesian additive regression tree; JAGAM = just another Gibbs additive modeller; BSAR = Bayesian spectral analysis regression; LPML = log pseudo marginal likelihood; WAIC = Watanabe-Akaike information; RMSE = root mean squared error.

설명하는 비모수 함수 추정 값에 대한 RMSE를 계산해보도록 한다. 아울러, 잠재변수 추정 뿐 만 아니라, 반응변수 y를 분류하는 성능 평가를 위하여, 다음과 같은 세가지 평가기준, 정확도(accuracy), 민감도(sensitivity), 특이도(specificity) 값을 계산하여 비교하고자 하였다. 본 논문이 고려하는 프로빗 모형의 경우, 추정된 $\hat{f}(x_i)$를 바탕으로 $\hat{y}_i$를 분류하고, 이 추정된 $\hat{y}_i$와 실제 y_i를 비교를 통해, 다음과 같이 계산하도록 한다.

$$\begin{aligned}
\text{True Positive (TP)} &: (\text{실제 } y_i\text{가 1일 때 추정된 } \hat{y}_i\text{가 1인 수}),\\
\text{False Positive (FP)} &: (\text{실제 } y_i\text{가 0일 때 추정된 } \hat{y}_i\text{가 1인 수}),\\
\text{True Negative (TN)} &: (\text{실제 } y_i\text{가 0일 때 추정된 } \hat{y}_i\text{가 0인 수}),\\
\text{False Negative (FN)} &: (\text{실제 } y_i\text{가 0일 때 추정된 } \hat{y}_i\text{가 1인 수}),
\end{aligned}$$

$$\begin{aligned}
\text{정확도(accuracy)} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \quad \text{민감도(sensitivity)} = \frac{\text{TP}}{\text{TP}+\text{FN}},\\
\text{특이도(specificity)} &= \frac{\text{TN}}{\text{TN}+\text{FP}}.
\end{aligned} \tag{3.1}$$

식 (3.1)에서 알 수 있듯이, 정확도는 y_i의 값을 제대로 분류한 비율, 민감도는 $y_i = 1$일 때 이를 실제로 1로 분류한 비율, 특이도는 $y_i = 0$일 때 이를 0으로 분류한 경우를 의미하며, 이러한 세가지 평가기준의 값이 모두 높을수록 분류 성능이 좋은 모형임을 나타낸다. Table 3.3은 세 가지 분류 성능 평가기준(정확도, 민감도, 특이도)과 세 가지 모형 선택 기준(LPML, WAIC, RMSE)에 대한 수치적 요약을 나타내며, Figure 3.1은 함수 2에서의 BART, JAGAM, 형태제약 BSAR 세 가지 방법을 적합했을 때, y_i를 1로 분류하는 확률 값의 절단점에 따른 (1 − 특이도)와 (민감도)의 그래프를 나타낸 ROC 곡선(receiver operating characteristic curve)을 나타낸다.



(a) BART (b) JAGAM (c) BSAR

Figure 3.1. Receiver operating characteristic curves of three binary probit regression models under function (3). BART = Bayesian additive regression tree; JAGAM = just another Gibbs additive modeller; BSAR = Bayesian spectral analysis regression.

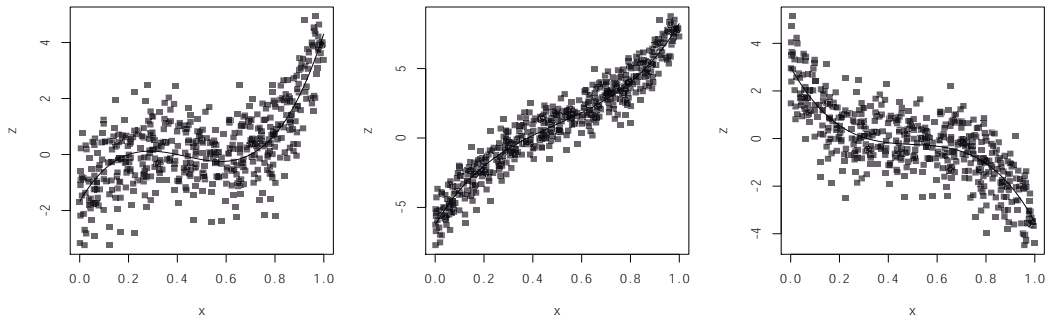
Table 3.3의 세가지 모형선택 평가기준 요약값을 통해, 단조증가 형태제약을 준 BSAR 모형이 LPML, WAIC, RMSE 모두 형태제약을 반영할 수 없는 BART, JAGAM 보다 우수함을 알 수 있다. 세가지 분류성능 평가기준이나 Figure 3.1의 ROC 곡선을 통한 예측력 측면에서는, BSAR 모형은 JAGAM과 비슷한 결과를 내었고, BART 모형이 좀 더 잘 y_i의 값을 예측함을 알 수 있었다. 이러한 결과들은, 여러가지 베이지안 비모수 함수 추정 방법들에 있어서, 형태 제약 함수 적합에서는 BSAR 모형이, 예측에 있어서는 BART 모형이 뛰어나다는 기존의 결과 (Chipman 등, 2010; Lenk와 Choi, 2017; Tan과 Roy, 2019)들과 유사하다고 할 수 있다.

추가적으로, 이항 프로빗 회귀 모형 뿐 아니라, 반응변수가 순서형일 때 BSAR 방법을 이용한 베이지안 순서형 프로빗 회귀 모형에 대해서도 형태제약 유무에 따른 적합결과를 비교하는 모의실험자료분석을 수행한다. 이를 위하여, 이항형 프로빗 회귀 모형에서의 모의실험과 유사한 방식으로, 형태 제약에 따른 세 가지(비단조, 단조증가, 단조감소) 비모수 함수를 각각 고려하도록 한다. 구체적으로는, 비모수 함수를 위한 설명변수 x_i는 균등분포 Unif$(0,1)$에서 표본을 추출하고, 순서형 반응변수 $y_i \in \{1,2,3,4\}$는 평균이 $f(x_i)$이고 분산이 1인 정규분포를 따르는 z_i의 값에 따라 500개의 데이터를 생성하였다. 또한 고정된 절단값 $a_0 = -\infty$, $a_1 = 0$, $a_4 = \infty$를 제외한 절단값은 총 두 개로, z_i의 60% 분위수 $z_{0.6}$과 90% 분위수 $z_{0.9}$를 이용하였다. 즉, $\boldsymbol{a} = (-\infty, 0, z_{0.6}, z_{0.9}, \infty)$이며, 이를 바탕으로 최종 생성된 데이터 구조는 다음과 같다:

$$y_i = \begin{cases} 1, & -\infty < z_i \le 0, \\ 2, & 0 < z_i \le z_{0.6}, \\ 3, & z_{0.6} < z_i \le z_{0.9}, \\ 4, & z_{0.9} < z_i < \infty, \end{cases}$$

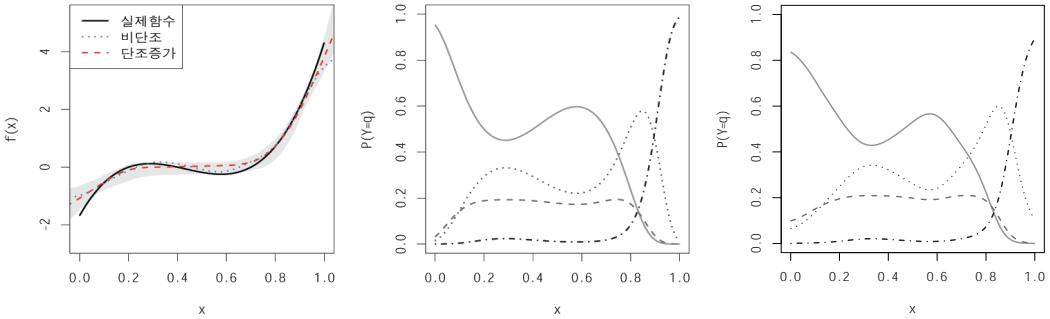
$$z_i = f(x_i) + \epsilon_i, \quad x_i \sim \text{Unif}(0,1),\ \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1),\ i = 1, \ldots, 500.$$

- 함수 1 (비단조) : $f(x) = 30(x-0.2)(x-0.4)(x-0.7)$,
- 함수 2 (단조증가) : $f(x) = 2\sinh(4x-2) + 1$,
- 함수 3 (단조감소) : $f(x) = -3(2x-1)^3 - \frac{1}{2}x$.



(a) 함수 1 (b) 함수 2 (c) 함수 3

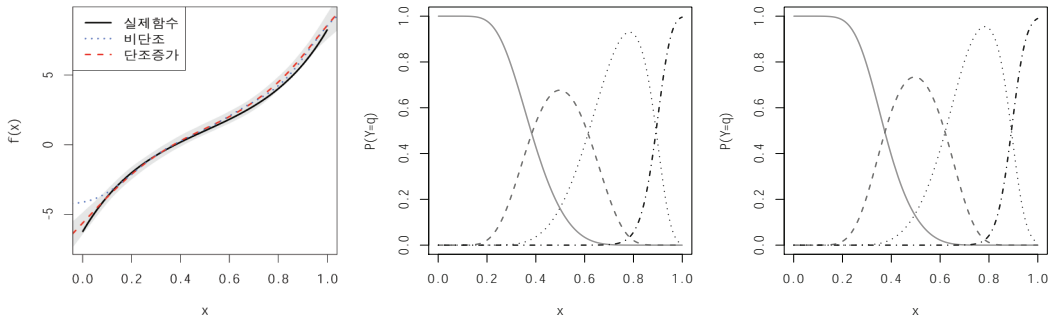
Figure 3.2. Simulation data (x, z) with $n = 500$ for ordinal probit regression model with shape restrictions.



(a) 함수 1의 적합 결과 (b) 반응변수 y의 실제 확률 (c) 반응변수에 대한 추정 확률

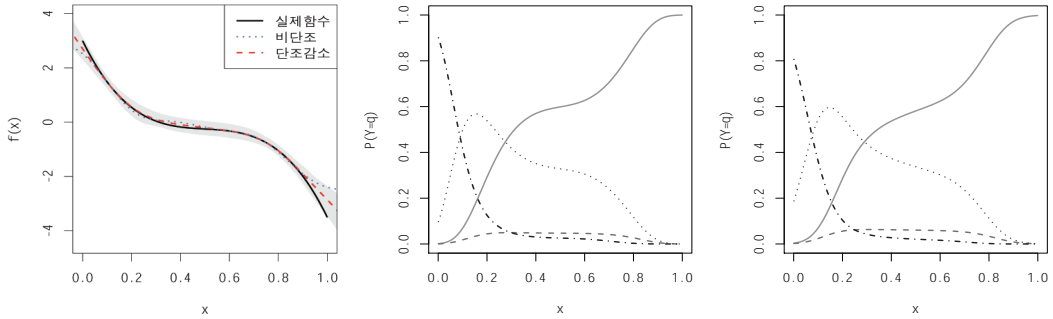
Figure 3.3. Model fitting of simulation data with function (1) based on ordinal probit Bayesian spectral analysis regression.

Figure 3.2는 함수 1–3으로부터 생성한 500개의 데이터와 실제 함수를 각각 보여준다. 함수 1은 이항 형 프로빗 회귀 모형의 모의 실험에서 사용한 것과 동일한 비단조 함수로, 형태제약을 고려하지 않는 적합이, 함수 2와 함수 3은 각각 단조증가와 단조감소하는 함수이기 때문에 단조성 형태제약을 고려하는 BSAR 순서형 프로빗 회귀 모형이 더 나은 적합을 할 것으로 예상되며, 이를 모의실험을 통해서 확인해보도록 한다. 사후추론을 위한 MCMC 방법에서는 이전의 모의실험들과 마찬가지로 첫 10,000개의 사후표본을 소각하고, 5번째 표본마다 저장하여 총 1,000개의 표본을 얻었고, 추출된 사후표본의 trace plot과 R의 `CODA` 패키지에서 제공하는 값들을 바탕으로 MCMC 알고리즘이 수렴성이 보장된다. 이를 바탕으로, 각 모의실험자료에 대한 순서형 프로빗 모형의 적합결과는 Figures 3.3–3.5에서 확인할 수 있으며, (a)의 결과는 실제 함수 $f(x)$에 대하여, 형태제약 유무에 따라 적합된 함수 $\hat{f}(x)$와 95% 신용 구간을, (b)의 결과는 모의실험에서 생성된 실제 반응변수 y의 확률, $P(Y = q|\boldsymbol{x}, \boldsymbol{z})$, $q = 1, 2, 3, 4$를, 그리고 (c)의 결과는 형태제약 유무에 따른 순서형 프로빗 모형을 통해 추정한 반응변수 y의 확률, $\hat{P}(Y = q|\boldsymbol{x}, \boldsymbol{z})$를 나타낸다. 또한 Figures 3.3–3.5의 (b)와 (c)에서 실선(solid line), 파선(dashed line), 점선(dotted line), 쇄선(dash-dotted line)은 각각 $P(Y = 1|x), P(Y = 2|x), P(Y = 3|x), P(Y = 4|x)$에 해당한다. 이를 통해, BSAR 순서형 프로빗 모형에서 단조성 제약을 고려하게 되면, 비모수 함수에 대한 적합에 있어서 더 나은 결과를 보여주며, 이에 대응하는 반응변수의 확률도 형태제약을 반영함으로써 더욱 잘 추정함을 알 수 있다. 예를 들어, $P(y_i = 1|x_i, \boldsymbol{w}_i) = \Phi(-f(x_i) - \boldsymbol{w}_i^\top \boldsymbol{\beta})$로 표현되



(a) 함수 2의 적합결과 (b) 반응변수 y의 실제 확률 (c) 반응변수 y의 추정 확률

Figure 3.4. Model fitting of simulation data with function (2) based on ordinal probit Bayesian spectral analysis regression.



(a) 함수 3의 적합결과 (b) 반응변수 y의 실제 확률 (c) 반응변수 y의 추정 확률

Figure 3.5. Model fitting of simulation data with function (3) based on ordinal probit Bayesian spectral analysis regression.

Table 3.4. Summary of LPML and WAIC for ordinal probit BSAR with/without shape restrictions

모형평가기준	함수 1(비단조)		함수 2(단조증가)		함수 3(단조감소)	
	비단조	단조증가	비단조	단조증가	비단조	단조감소
LPML	−503.057	−501.039	−221.936	−218.084	−388.302	−385.553
WAIC	1006.115	1002.088	442.355	436.156	776.283	771.109

BSAR= Bayesian spectral analysis regression; LPML = log pseudo marginal likelihood; WAIC = Watanabe-Akaike information.

기 때문에, $f(x_i)$에 단조증가 제약을 주면 x_i가 증가할수록 확률은 감소하게 되고, 단조감소 제약을 주면 x_i가 증가할수록 확률은 증가하는 모습을 보인다. 이와 반대로 $P(y_i = Q|x_i, \boldsymbol{w}_i) = 1 - \Phi(a_{q-1} - f(x_i) - \boldsymbol{w}_i^\top \boldsymbol{\beta})$에서는 $f(x_i)$에 단조증가 제약을 주면 x_i가 증가할수록 확률은 증가하고, 단조감소 제약을 주면 x_i가 증가할수록 확률은 감소하게 된다. 따라서 BSAR 방식을 사용하는 순서형 프로빗 회귀 모형에서는 함수의 형태제약 뿐 아니라, 이에 대응하는 설명변수 x_i에 따른 처음 범주와 마지막 범주에 대한 반응변수의 확률의 형태제약을 고려함을 알 수 있다.

이러한 결과들은 Figures 3.4–3.5의 (c)의 $P(Y = 1)$로 표시된 실선을 통해 알 수 있으며, 이 경우, Figure 3.4에서는 형태제약이 없는 경우, Figure 3.4와 Figure 3.5에서는 형태제약을 고려하는 모형으

로부터 추정한 확률값을 표시한다. 추가적으로, 형태제약의 유무에 따른 두 가지 모형간의 비교를 위하여, Table 3.4에서 요약된 바와 같이 LPML과 WAIC의 두가지 모형선택기준을 사용하고, 형태제약이 있는 경우에 형태제약을 고려하는 BSAR 프로빗 회귀 모형이 더 선호됨을 알 수 있다.

4. 실증분석 II : 국민건강영양조사자료분석

4.1. 국민건강영양조사 자료: 흡연 강도와 커피 섭취빈도 간의 관계

흡연은 고혈압, 당뇨, 뇌졸중, 폐질환 등의 수많은 질병과 깊은 관련이 있는 것으로 알려져 있기에 (Jung 등, 2018), 정부에서는 다각도의 정책을 통해 흡연 규제를 적극적으로 하고 있는 실정이다. 그러나, 이러한 노력에도 불구하고, 흡연율의 감소는 미미하며, 흡연 규제 정책의 실효성에 대한 문제제기와 함께, 보다 효과적인 규제 정책을 펼치기 위해서 흡연과 관련한 여러 요인들을 분석하는 연구들이 수행되고 있다 (Cho, 2013; Moon, 2016; Ahn 등, 2017). 4절에서는, 이러한 선행연구에서 이루어진 연구가설과 결과들을 바탕으로, 본 논문이 제안하는 베이지안 프로빗 준모수 회귀 모형을 국민건강영양조사 자료분석에 활용하여, 커피 섭취량과 흡연 간의 연관성을 실증적으로 고찰하고자 한다.

국민건강영양조사는 보건복지부 산하 질병관리본부의 주관으로 실시되는 전국 규모의 건강 및 영양 조사로, 본 절의 실제자료분석에서 커피 섭취량을 포함하고 있는 가장 최근 조사 년도의 자료인 국민건강영양조사 제 7기 1차년도 (2016년) 자료(Korean National Health and Nutrition Examination Survey, (KNHANES) 2016)의 건강설문조사 및 식품섭취빈도 조사결과(https://knhanes.cdc.go.kr/knhanes/main.do)를 이용하도록 한다. 구체적으로는, 해당 자료를 바탕으로, 흡연강도와 커피 섭취빈도간의 관계를 파악하기 위하여, 반응변수를 나누는 기준에 따라 이항형 프로빗 준모수 회귀 모형과 순서형 프로빗 준모수 회귀 모형을 각각 적합하도록 한다. 반응 변수를 “비흡연”, “흡연”의 두가지로 구분할 경우, 0과 1의 값을 갖는 이항형 범주이기 때문에 이항형 프로빗 준모수 회귀 모형을 사용하고, 반응변수를 흡연의 강도에 따라 세분화하여 “비흡연”, “가끔 흡연”, “매일 흡연”, “중독 흡연”의 4가지 흡연 강도로 구분할 경우, 순서형 프로빗 준모수 회귀 모형을 사용하도록 한다. 흡연 강도를 나누는 방식은 Moon (2016)과 Cho (2013)가 제시한 기준을 따르도록 한다. Moon (2016)에서는 “현재 담배를 피우십니까?”라는 국민건강영양조사 설문지에 대한 응답을 기준으로 ‘과거에는 피웠으나 현재는 피우지 않음’, ‘가끔 피움’, ‘매일 피움’에 따라 응답자를 비흡연, 가끔 흡연, 매일 흡연으로 분류하고, 또한, 평생 담배를 피운 적이 없는 응답자도 비흡연에 포함시킨다. 중독 흡연(heavy smoking)은 Cho (2013)의 하드코어흡연자(hardcore smoker) 정의에 따른 분류기준을 사용한다. 중독 흡연의 기준은 현재 흡연자 중 매일 흡연자, 흡연기간 5년 이상, 일 15개비 이상 흡연하고 지난 1년 간 금연을 시도한 경험이 없으며 향후 6개월 내 금연계획이 없고 만 26세인 자를 의미한다. 이와 같은 방식으로 분류된 흡연여부 및 강도에 따른 응답자 수(비율)를 주 단위 커피섭취빈도에 따라 계산한 결과는 Table 4.1과 같이 요약된다.

Table 4.1의 커피섭취빈도는 주 단위 섭취빈도로서, 연속형 설명변수로 간주되며, 국민건강영양조사자료의 지침서(Korean Centers for Disease Control and Prevention, 2016) (https://knhanes.cdc.go.kr/knhanes/sub03/sub03_06_02.do)의 계산방식을 따른다. 구체적으로, 설문조사 항목에서 “커피의 최근 1년간 평균섭취빈도” 및 “하루 3회 초과 섭취 시 하루 평균 섭취 횟수”를 기준으로 계산되며, “커피의 최근 1년간 평균섭취빈도”는 월 단위, 주 단위, 일 단위의 9개 항목으로 나누어져 있고(무응답 항목은 조사 대상에서 제외), 주 단위 섭취빈도는 그대로, 월 단위 평균섭취빈도와 일 단위 섭취빈도는 각각 주 단위 섭취빈도로 다시 계산된다. Table 4.1를 통해, 커피의 주당 섭취빈도가 증가할수록 비흡연율은 감소하고 매일 흡연율과 중독 흡연율이 높아지는 경향을 보이고 있음을 알 수 있다. 따라서, 이러한

Table 4.1. Frequency table of coffee intake and smoking behavior

커피섭취빈도	비흡연	흡연	가끔 흡연	매일 흡연	중독 흡연
0	119 (79.9)	30 (20.1)	7 (4.7)	18 (12.1)	5 (3.4)
0.23	24 (82.8)	5 (17.2)	2 (6.9)	3 (10.3)	0 (0.0)
0.58	39 (86.7)	6 (13.3)	1 (2.2)	4 (8.9)	1 (2.2)
1	45 (84.9)	8 (15.1)	3 (5.7)	5 (9.4)	0 (0.0)
3	114 (85.7)	19 (14.3)	6 (4.5)	12 (9.0)	1 (0.8)
5.5	57 (79.2)	15 (20.8)	1 (1.4)	13 (18.1)	1 (1.4)
7	263 (81.7)	59 (18.3)	11 (3.4)	42 (13.0)	6 (1.9)
14	375 (79.3)	98 (20.7)	22 (4.7)	60 (12.7)	16 (3.4)
21	159 (64.1)	89 (35.9)	13 (5.2)	61 (24.6)	15 (6.0)
28	49 (54.4)	41 (45.6)	3 (3.3)	29 (32.2)	9 (10.0)
35	28 (41.8)	39 (58.2)	4 (6.0)	26 (38.8)	9 (13.4)
42	6 (46.2)	7 (53.8)	0 (0.0)	5 (38.5)	2 (15.4)
49	6 (50.0)	6 (50.0)	0 (0.0)	3 (25.0)	3 (25.0)
합계	1284	422	73	281	68

흡연강도와 커피섭취량간의 함수적 관계를 설명하기 위하여, 먼저 흡연/비흡연으로 구분된 이항자료를 BSAR 방법을 통한 이항형 프로빗 준모수 모형을 적합하고, 커피의 주당 섭취빈도에 따른 응답자가 비흡연에 속하는 확률과 흡연자에 속하는 확률을 모형화하여 흡연과 커피섭취량과의 관계를 고찰한다. 아울러, 흡연유무에 따른 분류 뿐 아니라, 흡연강도에 따라 네 개의 범주로 구분된 순서형 자료에 대하여, 형태제약을 반영하는 BSAR 방법을 통한 순서형 프로빗 준모수 회귀 모형을 적합하고, 각 범주별 확률을 추정하고, 이에 따른 흡연강도와 커피섭취량간의 관계를 보다 세분화하여 설명하도록 한다.

커피섭취빈도 외에 추가적으로 흡연 여부 및 강도와 연관이 있는 명목형 설명 변수로는 연령, 교육수준, 폭음 여부, 직업군 등으로서, 기존의 선행연구 (Kang 등, 2017; Moon, 2016)에서도 고려된 흡연과 관련된 사회, 경제적 요인들을 설명한다. 구체적으로, Table 4.2에 요약된 바와같이, 연령은 50세를 기준으로 50 이상과 19세 이상 50세 미만의 두 범주로 구분하고, 직업군은 표준직업분류 대분류 코드 중 관리자, 전문가 및 관련종사자, 그리고 사무 종사자는 “비육체 노동자”로, 서비스종사자와 판매 종사자는 “서비스 종사자”로, 농림어업숙련종사자, 기능원 및 관련기능 종사자, 장치·기계조작 및 조립 종사자, 그리고 단순노무종사자는 “육체노동자”로 분류하며, 폭음 여부는 일주일에 한 번 이상 폭음하는 지의 여부에 따라 0과 1의 값을 부여하도록 한다.

4.2. 베이지안 프로빗 준모수 회귀모형 : 흡연 강도와 커피 섭취빈도 간의 관계

본 절에서는, 4.1절에서 설명된 국민건강영양조사 제 7기 1차년도 (2016년)의 건강설문조사 및 식품섭취빈도조사 자료를 바탕으로, 커피 섭취량과 흡연과의 관계를 설명하는 이항형 및 순서형 프로빗 준모수 회귀 모형 적합을 통한 실증적 분석을 실시하도록 한다. 반응 변수를 “비흡연”, “흡연”의 두가지로 구분하여, 이항형 프로빗 준모수 회귀 모형을 사용하고, 반응변수를 흡연의 강도에 따라 세부화해하여 “비흡연”, “가끔 흡연”, “매일 흡연”, “중독 흡연”의 4가지 흡연 강도로 구분하여, 순서형 프로빗 준모수 회귀 모형을 적합한다. 잠재변수를 통한 준모수 회귀모형의 구조는, 선형모형과 비모수 모형의 합으로 표현되며, 명목형 설명변수인 성별(`sex`), 연령(`isOld`), 교육수준(`edu2`,`edu3`,`edu4`), 폭음여부(`isBinge`), 직업군(`worker_m`, `worker_s`)은 선형모형 $\boldsymbol{w}_i^\top \boldsymbol{\beta}$에 가변수(dummy variable)형태로 포함되고, 연속형 설명변수인 커피섭취빈도(x)는 비모수 함수에 사용된다. 이를 바탕으로 한, 베이지안 이

Table 4.2. Frequency table of smoking behavior with ordinal covariates

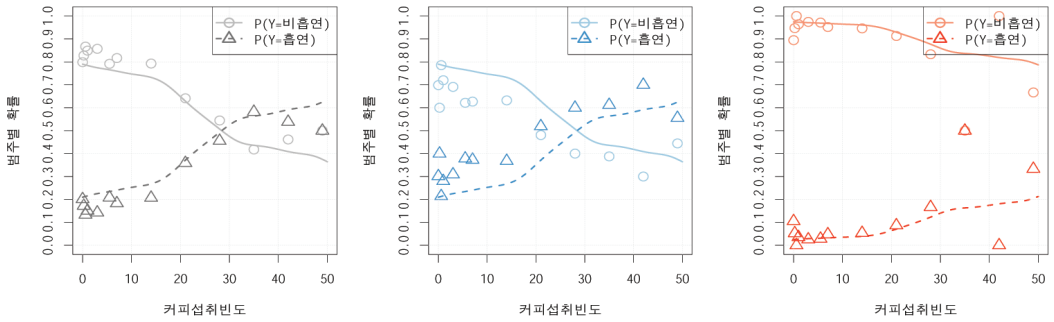
설명 변수		응답자 수				
		비흡연	흡연	가끔 흡연	매일 흡연	중독 흡연
성별	여성	771	59	26	31	2
(sex)	남성	513	364	47	251	66
연령	< 50	902	328	59	221	48
(isOld)	≥ 50	382	94	14	60	20
교육수준	초졸이하	92	25	2	15	8
(edu2)	중졸	92	32	5	20	7
(edu3)	고졸	438	150	23	108	19
(edu4)	대졸이상	662	215	43	138	34
폭음빈도	주 1회 이하	1024	201	41	130	30
(isBinge)	주 1회 이상	260	221	32	151	38
직업군	비육체노동자	647	182	34	120	28
(worker_m)	육체노동자	331	156	22	103	31
(worker_s)	서비스종사자	306	84	17	58	9
합계		1284	422	73	281	68

항형 프로빗 준모수 회귀 모형은 다음과 같다:

$$y_i = \begin{cases} 0, & z_i \leq 0, \\ 1, & z_i > 0, \end{cases}$$

$$z_i = \boldsymbol{w}_i^\top \boldsymbol{\beta} + f(x_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1),\ i = 1, \ldots, 1706. \tag{4.1}$$

식 (4.1)의 비모수함수 $f(x_i)$를 추정하기 위해 BSAR 방법을 적용하고, $f(x_i)$에 대한 단조증가의 형태 제약을 반영하도록 한다. Table 4.1에서 볼 수 있듯이, 커피의 주당 섭취빈도가 증가할수록 비흡연자에 속할 확률은 줄어들고, 흡연자일 확률은 증가하는 경향을 보이며, 흡연과 커피 섭취 간의 선행 연구 (Kang 등, 2017; Moon, 2016)들을 종합해볼 때, 커피의 주당 섭취빈도가 증가할수록 응답자가 흡연자에 속할 확률이 증가함을 예상해볼 수 있다. 식 (4.1)에서 표현된 바와 같이, 본 논문이 고려하는 이항형 프로빗 준모수 회귀모형에서는 다른 설명변수 $\boldsymbol{w}_i$가 고정된 잠재변수 또는 해당 설명변수의 효과를 제외한 $z_i - \boldsymbol{w}_i^\top \hat{\beta}$변수가 x_i와의 관계에 단조증가 또는 단조감소의 특성이 있을 때, $f(x_i)$에 단조증가 형태 제약을 반영하게 되면, 잠재변수 z_i에 대응하는 이항반응변수 y_i의 확률이 x_i와의 단조증가 관계를 잘 설명할 수 있게 된다. 구체적으로, 커피의 주당 섭취빈도와 흡연확률과의 단조관계를 설명하기 위해서, 응답자가 흡연자일 확률 $P(y_i = 1|x_i, \boldsymbol{w}_i)$는 $P(z_i > 0) = 1 - \Phi(-f(x_i) - \boldsymbol{w}_i^\top \boldsymbol{\beta})$, $z_i \sim \mathcal{N}(f(x_i) + \boldsymbol{w}_i^\top \boldsymbol{\beta}, 1)$로 표현되기 때문에 $f(x_i)$가 단조증가하면, 대응하는 $P(z_i > 0)$가 증가하게 됨을 알 수 있다. 이항형 프로빗 준모수 회귀모형을 통해 적합한 결과는 Figure 4.1과 Table 4.3에 요약되어 있으며, Figure 4.1의 그래프에는 커피의 주당 섭취빈도 흡연자와 비흡연자의 예측 확률과 관측값을, Table 4.3은 선형모형 $\boldsymbol{w}_i^\top \boldsymbol{\beta}$에 포함된 설명변수들의 계수 추정값들을 각각 나타낸다. Table 4.3에 요약된 값들은 MCMC 방법을 통해 추출된 $\boldsymbol{\beta}$의 사후표본들에 대한 평균 및 분위수들을 요약한 값을 의미한다. Table 4.3을 살펴보면 성별, 연령, 폭음 여부가 흡연 확률에 유의하게 영향을 미침을 확인할 수 있다. 예를 들어, 성별에 대한 회귀 계수가 양수이고, 연령에 대한 회귀 계수가 음수이므로, 주당 커피섭취빈도가 동일할 경우, 남성일 때 흡연자에 속할 확률이 더 높아지고, 연령에 50세 이상일 때 흡연자에 속할 확률이 더 낮아짐을 알 수 있다. 또한, 폭음 여부에 대해서도, 성별과 마찬가지로 폭음 빈도가 주 1회 이상인 사람은 1회 이하인 사람보다 흡연자에 속할 확률이 높아짐을 알 수 있다.



(a) 단조증가 (b) 남성 (c) 여성

Figure 4.1. Model fitting with binary probit BSAR with shape restrictions : Coffee intake and smoking behavior with gender.

Table 4.3. Posterior estimates of β with binary probit BSAR (monotone increasing)

요약값	변수명							
	`sex`	`isOld`	`edu2`	`edu3`	`edu4`	`isBinge`	`worker_m`	`worker_s`
사후평균	1.142	−0.508	−0.109	−0.216	−0.389	0.585	−0.114	0.095
2.5%	0.968	−0.701	−0.481	−0.556	−0.735	0.433	−0.310	−0.118
97.5%	1.319	−0.317	0.276	0.108	−0.015	0.738	0.086	0.312

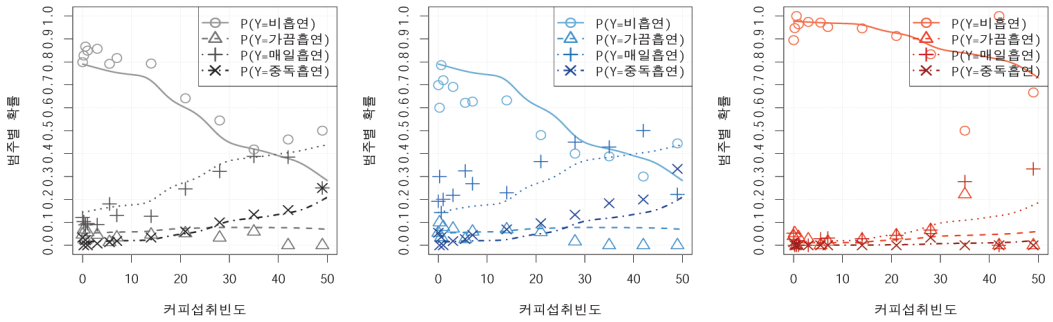
Figure 4.1에서 볼 수 있듯이, 비모수 함수 $f(x)$에 단조증가 형태제약조건을 반영함으로써, 흡연자의 확률이 커피의 섭취빈도에 따라 증가하게 되는 패턴을 더욱 잘 설명함을 알 수 있다. 이를 통하여, 성별, 연령, 직업군, 교육수준에 따른 커피의 주당 섭취빈도에 따른 흡연자와 비흡연자의 확률을 추정하고 예측할 수 있다. 예를 들어, 50세 미만의 대졸 이상의 학력, 폭음을 하지 않고, 서비스 직업군을 가진 남성의 경우, 이에 대응하는 설명변수 값들은 $\texttt{sex}_i, \texttt{isOld}_i, \texttt{edu2}_i, \texttt{edu3}_i, \texttt{edu4}_i, \texttt{isBinge}_i, \texttt{worker_m}_i, \texttt{worker_s}_i) = (1, 0, 0,\ 0, 1, 0, 1, 0)$에 대응하며, 이를 바탕으로 Table 4.3에 요약된 선형회귀계수 추정값과 추정된 비모수 함수를 결합하여, 잠재변수 추정값 $\hat{z}$를 계산하고, 이에 대응하는 확률을 추정할 수 있다. 이러한 방식으로 추정된 확률값들은 Figure 4.1의 (b)와 (c)와 같이 나타낼 수 있으며, 이 경우, 남성이 흡연자에 속할 확률은 여성이 흡연자에 속할 확률보다 더 높고, 커피의 주당 섭취빈도가 증가할수록 더 빠르게 증가하는 경향이 있음을 알 수 있다.

흡연여부에 따른 이항형 반응변수에 대한 모형화 뿐 아니라, 흡연의 강도에 따라 구분한 순서형 반응변수에 대한 확률을 프로빗 준모수 회귀모형을 통해 적합하도록 한다. 이를 위하여, 반응 변수 y_i를 비흡연(none) “1”, 가끔 흡연(sometimes) “2”, 매일 흡연(daily) “3”, 중독 흡연(heavy) “4”의 4개의 값을 갖는 순서형 자료로 구분하여, 다음과 같은 순서형 프로빗 준모수 회귀모형을 고려한다.

$$y_i = \begin{cases} 1, & -\infty < z_i \le 0, \\ 2, & 0 < z_i \le a_2, \\ 3, & a_2 < z_i \le a_3, \\ 4, & a_3 < z_i < \infty, \end{cases}$$

$$z_i = \boldsymbol{w}_i^\top \boldsymbol{\beta} + f(x_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1),\ i = 1, \ldots, 1706. \tag{4.2}$$

식 (4.2)와 같이, 4가지 범주로 구분에 필요한 잠재변수 z_i는 이항형 프로빗 준모수 회귀 모형과 마찬가



(a) 단조증가 (b) 남성 (c) 여성

Figure 4.2. Model fitting with ordinal probit BSAR with shape restrictions: coffee intake and smoking behavior with gender.

지로, 동일한 설명변수들을 사용하고, $f(x_i)$에 형태제약을 반영하여, 커피의 주당 섭취빈도 증가에 따른 각 범주에 속하는 확률을 보다 효과적으로 추정하고 예측하고자 한다. 예를 들어, 다른 설명변수 $\boldsymbol{w}_i$를 고정시키고 $f(x_i)$에 단조증가 형태제약을 주게 되면, 커피의 주당 섭취빈도가 증가할수록, 반응변수가 범주 1(비흡연자)에 속할 확률 $P(y_i = 1|x_i, \boldsymbol{w}_i) = P(z_i < 0) = \Phi(-f(x_i) - \boldsymbol{w}_i^\top \boldsymbol{\beta})$은 단조감소하게 되고, 범주 4(중독 흡연자)에 속할 확률 $P(y_i = 4|x_i, \boldsymbol{w}_i) = P(a_3 < z_i) = 1 - \Phi(a_3 - f(x_i) - \boldsymbol{w}_i^\top \boldsymbol{\beta})$은 단조증가하는 제약을 줄 수 있기 때문에, 해당 범주의 확률들을 보다 잘 설명할 수 있고, 이를 통해 보다 정확한 확률예측이 가능할 것이다.

Figure 4.2는 이러한 순서형 프로빗 준모수 회귀모형을 비모수 함수 $f(x)$에 대한 형태제약을 반영하여 적합한 결과로서, 흡연강도에 따른 4가지 범주에서의 추정확률을 나타낸다. 비모수 함수 $f(x)$에 대하여 단조증가의 형태제약을 반영하게 되면, Figure 4.2(a)와 같이, 이에 대응하는 순서형 범주의 확률의 단조증가/감소의 패턴을 더욱 명확하게 추정할 수 있음을 알 수 있다. 특히 커피의 주당 섭취빈도가 14회 이상(하루 평균 2회이상)을 기점으로 비흡연자일 확률은 현저하게 감소하고 매일 흡연자의 확률은 급격히 증가함을 확인할 수 있으며, 이를 통해, 형태제약을 반영하는 모형이 흡연과 커피 주당 섭취빈도 간의 관계를 보다 명확하게 설명함을 알 수 있다. 또한, 중독 흡연일 확률도, 비록 관측값이 다른 흡연강도의 범주에 비해 많지 않더라도, 형태제약을 반영함으로써, 대응하는 확률을 보다 잘 추정할 수 있음을 알 수 있다.

Table 4.4는 단조증가 순서형 프로빗 준모수 회귀모형의 회귀계수 β의 사후표본 요약값을 보여준다. 이항형과 마찬가지로, 성별, 연령, 폭음여부가 유의한 변수로 파악되고, 추가적으로 교육수준도 유의한 변수로 파악된다. 또한 특정한 설명변수의 값이 한 다누이 증가할 때의 반응변수의 확률의 변화를 설명하는 주변효과(marginal effect)는 첫번째 범주와 마지막 범주의 경우 회귀계수의 부호에 따라서 쉽게 확인될 수 있기 때문에(Agresti, 2013; Koop 등, 2007 등 참조), 유의한 회귀계수를 바탕으로, 비흡연과 중독흡연에 속할 확률의 증감을 설명할 수 있다. 예를 들어 성별의 회귀계수가 1.191로 양수이기 때문에, $P(y_i = 1|x_i, \boldsymbol{w}_i) = \Phi(-f(x_i) - \boldsymbol{w}_i^\top \boldsymbol{\beta})$이므로, 다른 명목형 설명변수 값을 고정시키고 주당 커피 섭취 빈도가 같다고 가정할 때, 남성($w_{i1} = 1$)이 여성($w_{i1} = 0$)보다 비흡연자에 속할 확률이 낮게 됨을 알 수 있다. 이와 반대로, $P(y_i = 4|x_i, \boldsymbol{w}_i) = 1 - \Phi(a_3 - f(x_i) - \boldsymbol{w}_i^\top \boldsymbol{\beta})$이므로, 남성이 여성보다 중독흡연일 확률이 높게 됨을 알 수 있다. 폭음여부의 경우 회귀계수가 양수이기 때문에, 다른 설명변수들이 고정되고, 동일한 커피 섭취빈도를 갖는 경우, 폭음빈도가 주 1회 이상일 때 주 1회 이하일 때보다 중독 흡연일 확률은 높아지고, 비흡연일 확률은 낮아짐을 알 수 있다.

Table 4.4. Posterior estimates of β with ordinal probit BSAR (monotone increasing)

요약값	변수명							
	sex	isOld	edu2	edu3	edu4	isBinge	worker_m	worker_s
사후평균	1.191	−0.463	−0.165	−0.324	−0.491	0.537	−0.099	0.071
2.5%	1.030	−0.661	−0.605	−0.655	−0.809	0.392	−0.288	−0.131
97.5%	1.358	−0.278	0.201	−0.020	−0.151	0.679	0.083	0.253

이러한 결과들을 바탕으로, 이항형 프로빗 준모수 모형에서 추정하였던, 50세 미만의 대졸 이상의 학력, 폭음을 하지 않는 서비스 직업군에 대한 성별간의 확률을, 형태제약을 반영하는 순서형 프로빗 준모수 모형을 통해 4가지 흡연 강도에 따라 세분화하여 살펴보도록 한다. Figure 4.2 (b)와 (c)는 이러한 4가지 흡연 강도에 따른 추정확률을 나타내고 있으며, 커피의 주당 섭취빈도가 증가할수록 비흡연자에 속할 확률이 감소하고, 이와 반대로 매일 흡연, 중독 흡연에 속할 확률은 증가하게 됨을 알 수 있다. 또한, 남성이 비흡연자에 속할 확률은 여성이 비흡연자에 속할 확률보다 매우 낮고, 커피의 주당 섭취빈도가 증가할수록 더 빠른 속도로 비흡연자에 속할 확률이 감소하는 모습을 보이며, 형태제약을 반영함으로써, 커피의 주당 섭취빈도가 높아지게 될 때, 비흡연자에 속할 확률과 중독흡연에 속할 확률이 더 명확하게 드러남을 알 수 있다. 다만, 여성일 때 흡연자에 속할 확률은 관측된 비율을 잘 추정하지 못하는 것으로 보인다. 이는 성별에 따른 확률을 추정하는 데 있어서, 잠재변수 z의 평균이 성별의 회귀계수 β만큼만 값이 다르기 때문에, 회귀계수에 의존하는 주변효과만으로 설명하기에는 불충분함으로 기인한다고 판단된다. 또한, 선형구조로 설명되는 성별과 다른 설명변수, 그리고 비모수 함수를 가정하는 커피섭취빈도와의 복합적인 관계로 인하여 확률 추정이 어렵기 때문이라고 판단되며, 보다 정확한 확률추정을 위해서는 계층적 베이지안 접근방법과 같은 대안적인 방법을 통해 성별과 다른 설명변수간의 교호작용 등을 고려하는 것이 필요해보인다.

5. 결론

본 논문에서는 잠재변수를 이용한 베이지안 이항형 및 순서형 프로빗 준모수 회귀 모형에 대해서 고찰하였다. 이를 위하여, Albert와 Chib (1993)에서 제안된 잠재변수를 통한 프로빗 선형 회귀 모형을 선형모형과 비모수(비선형)모형의 표현되는 준모수 회귀모형으로 확장하고, 특히, 비선형 관계를 갖는 함수를 적합하기 위하여 Lenk와 Choi (2017)에서 제안한 BSAR 방법을 통한 베이지안 비모수 함수 추정기법을 적용하였다. 이러한 준모수 회귀모형을 통해, 잠재 변수와 설명변수간의 관계를 보다 유연하게 설명하고, 특히, BSAR 방법을 바탕으로, 단조증가/감소와 같은 형태 제약을 비모수 회귀함수에 반영함으로써, 이에 대응하는 반응변수의 확률을 보다 효과적으로 추정하고자 하였다.

이항형과 순서형 자료를 따르는 모의실험자료를 생성하여, BSAR 방법을 통한 베이지안 이항형 및 순서형 프로빗 준모수 회귀모형에 대한 적합 결과를 살펴보고, 기존에 연구된 유사모형과의 성능을 비교하였다. 형태제약 유무에 따른 모형 적합 결과를 살펴보고, 적절한 모형선택 기준에 따라 형태제약 유무를 결정하는 방식을 적용하여 모형 간의 비교 분석을 수행하였다. 아울러, 본 논문에서 고찰한 베이지안 이항형 및 순서형 프로빗 준모수 회귀모형을 국민건강영양조사 제 7기 1차년도 (2016) 자료 (Korean National Health and Nutrition Examination Survey, (KNHANES) 2016)에 적용하여, 흡연과 커피의 섭취량 간의 관계에 대한 실증적 분석을 수행하였다. 이를 위하여 성별, 연령, 교육 수준, 폭음 여부, 직업군을 흡연에 영향을 주는 선형관계에 있는 설명변수로, 커피 섭취량의 지표인 커피의 주당 섭취빈도를 비선형(비모수) 관계를 설명하는 변수를 사용하는 준모수 모형을 적합하였다. 흡연과 비흡연의 이항자료 뿐 아니라, 흡연강도에 따라 범주를 세분화한 순서형 자료에 대하여, 각각 이항형 프로빗 준

모수 모형과 순서형 프로빗 준모수 모형을 적합하였다. BSAR 방법을 통하여 비모수 함수에 대하여 단조증가의 형태제약을 반영하여, 이에 대응하는 범주별 확률을 보다 명확하게 추정할 수 있었다. 이러한 분석 결과들은 기존의 연구(Carmody 등, 1985; Moon, 2016; Ahn 등, 2017)에서 이루어진 커피섭취와 흡연의 관계를 보다 실증적이고 구체적으로 규명하고, 흡연유무 뿐 아니라, 흡연강도에 영향을 미치는 요인을 다각적이고 심층적으로 분석했다는 점에서 유의미한 결과라고 할 것이다. 본 연구에서는 커피섭취빈도 x를 설명하는 비선형 함수 적합을 위하여 BSAR 방법을 사용하였으나, 대안적으로는 다항식 회귀(polynomial regression)나 회귀 스플라인(regression spline)을 활용하는 것을 고려해 볼 수 있을 것이다. 이 경우, 다항식의 차수나 스플라인 항의 갯수를 설정하는 문제를 고려야할 것이며, 단조증가 또는 감소와 같은 형태제약을 갖는 함수를 적합하기 위해서는 추가적인 연구가 필요할 것으로 예상된다. 또한, 순서형 범주를 적합하기 위한 스플라인 회귀 모형은 대해서는 Wood (2017)의 `gam`의 `ocat` family를 참조해 모형을 적합할 수 있지만, 현재 로짓 링크만 제한적으로 사용할 수 있기 때문에, 실용적인 면에 있어서는 다소 간의 한계를 보인다. 이에 반해, BSAR 방법은 가우지안 확률과정과 코사인 기저를 바탕으로 한 다소 복잡한 적합방식이지만, 단조증가, 감소 뿐 아니라 오목,볼록, U자형, S자형 등의 다양한 형태제약을 반영할 수 있으며 (Lenk와 Choi, 2017), R 패키지 `bsamGP` (Jo 등, 2019)에서 다양한 모형적합을 제공하기 때문에, 본 논문에서의 연구와 같은 실제 자료분석에 있어서 매우 유용할 것이라고 판단된다.

본 연구에서 고찰한 이항 및 순서형 반응변수를 위한 베이지안 프로빗 준모수 회귀모형은, 아직 많은 연구가 되지 않은 실정이며, 특히 형태제약을 반영하는 모형과 이를 통한 범주별 확률 추정과 관련한 응용문제 해결에 있어서 다양하게 활용될 것으로 예상한다. 아울러, 본 논문에서 고찰한 BSAR 방법을 더욱 확장하여, 이산형 반응변수에 대한 계층적 베이지안 프로빗 준모수 회귀모형이나 이산형 반응변수의 빈도(frequency)를 0과 1사이의 연속형 반응 변수로 모형화하는 베이지안 베타 준모수 회귀모형(beta semiparametric regression model)을 향후 과제로 고려하고자 한다.

References

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed), John Wiley & Sons, NJ.

Ahn, H. J., Gwak, J. I., Yun, S. J., Choi, H. J., Nam, J. W., and Shin, J. S. (2017). The influence of coffee consumption for smoking behavior, *Korean Journal of Family Practice*, **7**, 218–222.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, **88**, 669–679.

Carmody, T. P., Brischetto, C. S., Matarazzo, J. D., O'Donnell, R. P., and Connor, W. E. (1985). Co-occurrent use of cigarettes, alcohol, and coffee in healthy, community-living men and women. *Health Psychology*, **4**, 323.

Chen, M. H. and Dey, D. K. (2000). Bayesian analysis for correlated ordinal data models. In *Generalized Linear Models: A Bayesian Perspective* (volume 5, pages 133–157), Dekker, New York.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees, *The Annals of Applied Statistics*, **4**, 266–298.

Cho, K. S. (2013). Prevalence of hardcore smoking and its associated factors in Korea, *Health and Social Welfare Review*, **33**, 603–628.

Clark, A., Georgellis, Y., and Sanfey, P. (2001). Scarring: The psychological impact of past unemployment, *Economica*, **68**, 221–241.

Cowles, M. K., Carlin, B. P., and Connett, J. E. (1996). Bayesian tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness, *Journal of the American Statistical Association*, **91**, 86–98.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection, *Journal of the American Statistical Association*, **74**, 153–160.

Harris, M. N. and Zhao, X. (2007). A zero-inflated ordered probit model, with an application to modelling tobacco consumption, *Journal of Econometrics*, **141**, 1073–1099.

Hasegawa, H. (2010). Analyzing tourists' satisfaction: a multivariate ordered probit approach, *Tourism Management*, **31**, 86–97.

Hastie, T. J. and Tibshirani, R. J. (1990). Generalized additive models, *Monographs on Statistics and Applied Probability* (Vol 43), Chapman and Hall, London.

Jara, A., Hanson, T. E., and Lesaffre, E. (2009). Robustifying generalized linear mixed models using a new class of mixtures of multivariate Polya trees, *Journal of Computational and Graphical Statistics*, **18**, 838–860.

Jo, S., Choi, T., Park, B., and Lenk, P. (2019). bsamGP: An R package for Bayesian spectral analysis models using Gaussian process priors, *Journal of Statistical Software*, **90**, 1–41.

Jung, K. W., Won, Y. J., Kong, H. J., Lee, E. S., and Community of Population-Based Regional Cancer Registries (2018). Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2015, *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, **50**, 303–316.

Kang, E., Lee, J. A., and Cho, H. J. (2017). Characteristics of hardcore smokers in South Korea from 2007 to 2013, *BMC Public Health*, **17**, 521.

Kim, M. (2015). Semiparametric approach to logistic model with random intercept, *Korean Journal of Applied Statistics*, **28**, 1121–1131.

Kockelman, K. M. and Kweon, Y. J. (2002). Driver injury severity: an application of ordered probit models, *Accident Analysis & Prevention*, **34**, 313–321.

Koop, G., Poirier, D. J., and Tobias, J. L. (2007). *Bayesian Econometric Methods (Econometric Exercises)*, Cambridge University Press, Cambridge.

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

Lee, J. H. and Heo, T. Y. (2014). A study of effect on the smoking status using multilevel logistic model, *Korean Journal of Applied Statistics*, **27**, 89–102.

Lenk, P. J. and Choi, T. (2017). Bayesian analysis of shape-restricted functions using Gaussian process priors, *Statistica Sinica*, **27**, 43–69.

Moon, S. (2016). Types of smoking statuses and associated factors among Korean wageworkers, *Journal of Korean Public Health Nursing*, **30**, 495–511.

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

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 (T2D), *Korean Journal of Applied Statistics*, **31**, 573–585.

Seok, H. E., Bang, H. J., and Kim, S. Y. (2017). Bayesian analysis of KBSID-III adaptive behavior data using a zero-inflated ordered probit model, *Korean Journal of Psychology: General*, **36**, 215–239.

Sha, N. and Dechi, B. O. (2019). A Bayes inference for ordinal response with latent variable approach, *Stats*, **2**, 321–331.

Tan, Y. V. and Roy, J. (2019). Bayesian additive regression trees and the general BART model, *Statistics in Medicine*, **38**, 5048–5069.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research*, **11**, 3571–3594.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed), CRC Press, Florida.

Xie, Y., Zhang, Y., and Liang, F. (2009). Crash injury severity analysis using Bayesian ordered probit models, *Journal of Transportation Engineering*, **135**, 18–25.

베이지안 순서형 프로빗 준모수 회귀 모형 : 국민건강영양조사 2016 자료를 통한 흡연양태와 커피섭취 간의 관계 분석

이다솜[a] · 이은지[b] · 조성일[c] · 최태련[b,1]

[a]Department of Statistics, North Carolina State University; [b]고려대학교 통계학과;
[c]전북대학교 통계학과 (응용통계연구소)

(2019년 10월 22일 접수, 2019년 12월 2일 수정, 2019년 12월 10일 채택)

요 약

본 논문에서는 Bayesian spectral analysis regression (BSAR) 방법론을 이용한 베이지안 순서형 프로빗 준모수 회귀모형에 대해서 고찰한다. 순서형 프로빗 회귀모형은 순서가 있는 범주형 자료를 모형화하는 방법으로, 정규 분포의 분포함수의 역함수인 프로빗 연결함수를 이용해 각 범주의 확률과 설명변수을 연결함으로써 반응변수의 확률을 모형화한다. 베이지안 프로빗 회귀 모형은 정규 분포를 따르는 잠재변수를 도입함으로써 사후 분포 도출을 용이하게 하고, 절단점에 따라 나뉘어지는 잠재변수들의 값에 따라서 반응 변수들이 범주화된다. 본 논문에서는 이러한 잠재 변수 방법을 확장해 BSAR 방법론에 기반하여 단조증가/감소와 같은 형태제약을 반영할 수 있는 베이지안 이항형 및 순서형 프로빗 준모수 회귀모형에 대해 연구한다. 모의실험을 통하여 이항형 프로빗 준모수 회귀모형과 기존의 다른 모형들 간의 적합결과를 비교하고, 형태 제약에 따른 순서형 프로빗 준모수 회귀모형의 적합결과를 비교 분석하도록 한다. 아울러, 국민건강영양조사 제 7기 1차년도 (2016) 자료(Korean National Health and Nutrition Examination Survey (KNHANES), 2016)를 바탕으로, 본 논문에서 고찰한 이항형 및 순서형 프로빗 준모수 회귀모형을 적용하여, 흡연양태와 커피섭취 간의 관계에 대한 실증적 분석을 수행한다.

주요용어: BSAR, 가우지안 과정, 국민건강영양조사 자료, 마르코프 연쇄 몬테칼로, 순서형 프로빗, 준모수 회귀

본 연구는 고려대학교 연구비에 의하여 수행되었음 (K1910951).
[1]교신저자: (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. E-mail: trchoi@korea.ac.kr