The Korean Journal of Applied Statistics
2024, Vol. 37, No. 3, 339–353

DOI: https://doi.org/10.5351/KJAS.2024.37.3.339
Print ISSN1225-066X / Online ISSN 2383-5818

A study on Bayesian beta regressions for modelling rates and proportions

Jeongin Lee[a], Jaeoh Kim[1,a], Seongil Jo[2,a]

[a]Department of Statistics and Data Science, Inha University

Abstract

In cases where the response variable in proportional data is confined to a limited interval, a regression model based on the assumption of normality can yield inaccurate results due to issues such as asymmetry and heteroscedasticity. In such cases, the beta regression model can be considered as an alternative. This model reparametrizes the beta distribution in terms of mean and precision parameters, assuming that the response variable follows a beta distribution. This allows for easy consideration of heteroscedasticity in the data. In this paper, we therefore aim to analyze proportional data using the beta regression model in two empirical analyses. Specifically, we investigate the relationship between smoking rates and coffee consumption using data from the 6th National Health Survey, and examine the association between regional characteristics in the U.S. and cumulative mortality rates based on COVID-19 data. In each analysis, we apply the ordinary least squares regression model, the beta regression model, and the extended beta regression model to analyze the data and interpret the results with the selected optimal model. The results demonstrate the appropriateness of applying the beta regression model and its extended version in proportional data.

Keywords: beta regression, coffee consumption, COVID 19, proportion data, smoking rate

1. 서론

흡연은 수많은 질병과의 연관성이 높아 Kang 등 (2010)에서 언급한 바와 같이 건강 위험행위로 여겨지며 한국보건사회연구원의 발표에서 우리나라 성인의 통상 금연의 1년 성공률은 18.4%에 불과하고 우리나라 정부도 금연의 성공률을 올리기 위해 다양한 정책들을 시도하고 있다 (Ahn 등, 2017). 이러한 정책적인 노력에 더해 흡연자의 생활습관의 변화도 금연에 도움이 되는것으로 알려져있는데, 커피섭취와 관련한 생활 습관이 그 중 하나이다 (Ahn 등, 2017). 따라서 본 논문에서는 흡연율과 커피 섭취 행태를 비롯한 연령, 소득, 교육수준과 같은 연구 대상자의 특성이 흡연율에 미치는 영향을 살펴보고 흡연에 유의한 연관을 미치는 요인들을 분석해 보고자 한다. 또한, 흡연율과 더불어 세계적으로 큰 혼란을 야기했던 코로나 19의 사망률과 지역별 특성간에 유의한 관련성을 갖는지에 대한 분석을 수행한다. 코로나 사망율에 대한 지역적 특성 연구는 Li 등 (2021)

This study was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022R1A5A7033499, RS-2023-00209229), Jaeoh Kim's research was supported by the MSIT(Ministry of Science, ICT), Korea, under the National Program for Excellence in SW), supervised by the IITP(Institute of Information & communications Technology Planing & Evaluation) in 2024"(2022-0-01127).

[1]Corresponding author: Department of Statistics and Data Science, Inha University, Incheon 22212, Korea. E-mail: jaeoh.k@inha.ac.kr

[2]Corresponding author: Department of Statistics and Data Science, Inha University, Incheon 22212, Korea. E-mail: bstatsjo@inha.ac.kr

Published 30 June 2024 / journal homepage: http://kjas.or.kr

에서와 같이 개인 특성으로는 포착되지 않는 지역적 추세를 설명할 수 있으며, 이후 추가적인 감염병 도래 예방에 있어 대응 정책 결정을 위해 해당 연구의 결과가 시사하는 바를 반영할 수 있다는데에 이점이 있다. 따라서 이를 위해 미국 플로리다주의 코로나 19 사망률 자료와 해당 주의 카운티별 특성간의 유의한 관계를 파악해보고자 한다.

연속형 반응변수가 단위 구간 혹은 특정 구간내에 존재하는 경우, 연속형 반응변수의 회귀모형 적합값이 상한과 하한이 존재하고 오차항이 정규분포를 따르지 않기 때문에, 보통최소제곱(ordinary least squares; OLS) 기반의 선형회귀모형을 통한 추정과 검정은 적절하지 않을 수 있다. 이러한 경우 대안으로 반응변수를 로짓(logit)변환한 뒤 정규성 가정을 기초로 하는 선형회귀모형을 통해 자료를 분석할 수 있으나, 해당 방법은 반응변수의 비대칭도(skewness)를 높여 잔차를 비대칭적으로 만드는 문제를 야기하는 것이 알려져 있다 (Cribari-Neto와 Zeileis, 2010). 이에 본 연구에서는 0과 1사이의 구간에서 정의되는 비율자료(proportion/rate data) 분석을 위해 Ferrari와 Cribari-Neto (2004)에서 소개한 재모수화(reparametrization)된 베타분포(beta distribution) 기반의 베타회귀모형(beta regression model)을 고려하고자 한다.

Ferrari와 Cribari-Neto (2004)에 의해 제안된 재모수화 베타분포는 기존 베타분포의 형태 모수(shape parameter)들을 평균과 정밀도 모수(precision parameter)로 재모수화한 분포로, 표준 단위구간 내에 속하는 연속형 변수를 모형화 할 때 비대칭성과 이분산성(heteroskedasticity)을 쉽게 고려할 수 있는 장점을 가지고 있다. 게다가, 설명변수를 고려할 수 있는 베타회귀모형으로의 확장시 추정된 회귀계수에 대해 평균 반응변수의 관점에서 해석이 용이하고, 많은 면에서 일반화선형모형과 유사하다는 것이 잘 알려져 있다 (Bayes 등, 2012).

비율자료 분석을 위한 베타회귀모형은 다양한 연구에서 제안되어왔다. 먼저 Ferrari와 Cribari-Neto (2004)에서는 본인들이 제안한 재모수화된 베타분포를 기초로 빈도론적 관점에서의 베타회귀모형을 제안하였다. 이때, 모형은 정밀도 모수를 고정된 상수로 가정하고, 평균 모수에 대해 설명변수와의 관계를 고려하였다. 이에 반해, Smithson와 Verkuilen (2006)에서는 Ferrari와 Cribari-Neto (2004)의 모형에서 고정된 정밀도 모수의 역수인 산포 모수(dispersion parameter)를 모형화하는 것의 유용성을 확인하고, 평균에 대한 위치 하위모형과 정밀도에 대한 산포하위모형을 가지는 베타회귀모형을 제안하였으며 이는 Simas 등 (2010)과 Ferrari 등 (2011)에서 변수산포 베타회귀모형이라는 용어로 명명되었다.

베이지안 관점에서의 베타회귀모형도 많은 연구에서 제안되었다. 베이지안 관점의 모형 추정은 모수에 대한 사전 분포와 관측된 자료에 대한 가능도를 함께 고려함으로써 사전 지식과 자료의 영향력을 통합적으로 반영하여 모형을 추정할 수 있으며, 모수에 대한 사후분포를 추론하기 때문에 모수 추정에 대한 불확실성을 정량적으로 나타낼 수 있다는 데에 이점이 있다. 관련 선행연구로 Branscum 등 (2007)은 평균 모수를 모형화하는 베이지안 베타회귀모형을 제안하고 마코프체인 몬테 카를로(Markov chain Monte Carlo; MCMC) 방법으로 회귀계수를 추정하는 알고리즘을 제시하였으며, Bayes 등 (2012)는 이상치(outlier)의 영향을 덜 받게 하기 위해 두 개의 베타분포를 혼합한 베타직사각형분포(beta rectangular distribution) 기반의 베타직사각형 회귀모형(beta rectangular regression)을 제안하였다. 또한, 최근 Zhou와 Huang (2022)는 네 개의 모수를 가지는 베타분포를 기초로 하여 반응변수의 받침(support)에 대한 정보가 없을 때 적용이 가능한 베타회귀모형을 제안하였다.

국내의 경우에는 Jang 등 (2018)이 베이지안 혼합 베타회귀모형에서 랜덤효과(random effects)를 고려한 베이지안 계층 모형(Bayesian hierarchical model)에 대한 연구와 Jang (2017)의 베타회귀모형을 이용한 건강 관련 삶의 질 자료 분석 연구, Han 등 (2020)의 이분산성이 있는 데이터에 대해 베타회귀모형을 활용한 동원훈련 분석 연구 등이 있다. 이 외에도, 베타회귀모형은 경제학, 유전학, 의학 등 다양한 분야에서 비율자료에 적용되고 있으며 관련 참고문헌으로는 Branscum 등 (2007), Peplonska 등 (2012), 그리고 Buntaine (2011) 등이 있다.

본 연구의 구성은 다음과 같다. 2절에서는 베타분포, 베타회귀모형에 대하여 설명한다. 이를 위하여, 먼저 베타분포와 베타분포의 재모수화 과정을 설명하고 베타회귀모형과 확장된 베타회귀모형에 대한 개념을 설

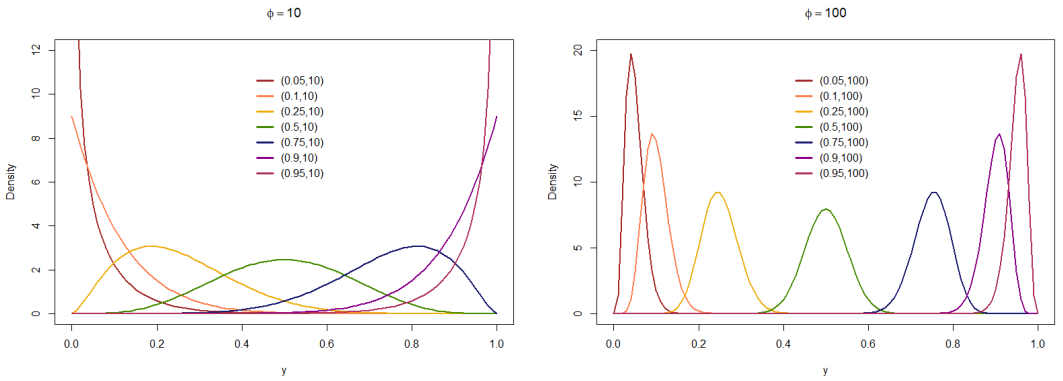

Figure 1: *Examples of Beta pdf for* $\phi = 10$ *(light panels) and* $\phi = 100$ *(right panels) with different values of* μ.

명한 뒤, 베이지안 방식으로 모수를 추정하는 베이지안 베타회귀의 내용을 다룬다. 이어서, 3절에서 모형비교 기준을 설명한 후 4절에서는 2절에서 설명한 베타회귀모형을 바탕으로 제 6기 국민건강영양조사 원시자료 (Korea national health and nutrition examination survey; KHANES)와 미국의 카운티별 COVID 19 (Coronavirus Disease 2019) 사망률에 대한 자료를 실증분석한다. 먼저, 보건복지부 산하 질병관리본부의 주관으로 실시되는 제 6기 KHANES의 자료를 바탕으로 흡연율과 커피 섭취 빈도간의 관계를 베타회귀를 이용하여 분석한 결과를 제시하고, 다음으로 R의 'betaBayes' 패키지에 포함된 미국의 카운티별 특성 및 COVID 19 확진자 및 사망자 수 자료를 통하여 카운티 별 특성과 COVID 19 누적 사망률간의 관계를 베타회귀모형을 적합하여 고찰한다. 또한, 각 자료에 대해 OLS 기반의 선형회귀모형, 베타회귀모형, 확장된 베타회귀모형, 베이지안 베타회귀모형, 확장된 베이지안 베타회귀모형을 적합하고, 결정계수(coefficient of determination) 및 베이지안 정보 기준(Bayesian information criterion; BIC) 비교를 통해 최적의 모형을 선택한다. 마지막으로, 5절에서는 실증 분석의 결과에 대한 해석을 정리하고 본 연구에 대한 전반적인 내용을 요약한다.

2. 베타회귀모형의 소개

베타분포는 그 형태 모수에 따라 확률밀도함수(probability density function)가 다양한 형태를 취할 수 있는 매우 유연한 분포로, 자료분석을 위한 통계적 모델링에 있어 상당히 용이 하다는 것이 잘 알려져 있다 (Cribari-Neto와 Zeileis, 2010). 이에 본 절에서는 베타분포에 대해 간략하게 설명하고, 베타회귀모형과 확장된 베타회귀모형, 그리고 각 회귀모형에 대한 베이지안 추론 방법에 대해 설명한다.

2.1. 베타분포

0과 1사이의 구간에 받침을 갖는 연속형 확률변수 Y가 베타분포를 따른다고 가정하자. 그때, Y에 대한 확률밀도함수는 아래 식 (2.1)과 같이 정의된다.

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1, \tag{2.1}$$

이때 p, q 는 형태 모수로 $p > 0, q > 0$ 이고 $\Gamma(\cdot)$ 는 감마 함수이며, 평균과 분산은 각각 아래와 같이 주어진다.

$$E(Y \mid p, q) = \frac{p}{(p+q)}, \quad \mathrm{Var}(Y \mid p, q) = \frac{pq}{(p+q)^2(p+q+1)}. \tag{2.2}$$

베타분포를 활용한 회귀모형에서는 전형적으로 반응치의 평균에 대하여 모형화하는 것이 유용하므로, Ferrari와 Cribari-Neto (2004)에서는 식 (2.1)의 형태 모수 p, q 에 대하여 $\mu = p/(p+q)$, $\phi = p+q$로 재모수화를 하여 베타분포의 확률밀도함수를 아래와 같이 정의하였다.

$$f(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \tag{2.3}$$

여기서 $0 < \mu < 1, \phi < 0$ 이고, 평균과 분산은 식 (2.4)와 같다.

$$E(Y \mid \mu, \phi) = \mu, \quad \mathrm{Var}(Y \mid \mu, \phi) = \frac{\mu(1-\mu)}{1+\phi}. \tag{2.4}$$

본 논문에서는 위의 식 (2.3)에서 새롭게 정의된 베타분포를 $Y \sim \mathrm{Beta}(\mu, \phi)$로 나타낸다. 참고로, 재모수화된 베타분포의 평균과 분산에서 ϕ는 μ가 고정되었을 때, ϕ가 큰 값을 가질수록 $\mathrm{var}(Y)$가 작은 값을 가지기 때문에 정밀도 모수로써 해석되고, 반대로 ϕ^{-1}는 산포 모수로써 해석될 수 있다.

Figure 1은 평균과 정밀도 모수에 따라 재모수화된 베타분포의 다양한 확률밀도함수 형태를 보여준다. 각 그림을 통해 정밀도 모수 ϕ가 일정할 때 평균과 정밀도 모수의 조합에 따라 유연하게 변화하는 베타분포의 형태를 확인할 수 있다. 또한 정밀도 모수가 클수록 평균을 중심으로 확률 분포가 밀집되어 있어 작은 산포를 갖고, 정밀도 모수가 작을수록 평균으로부터 확률 분포가 널리 퍼져있으며 큰 산포를 갖는 형태를 보임을 확인할 수 있다.

2.2. 베타회귀모형

2.2.1. 베타회귀모형

반응변수가 비율과 같이 연속형이고, $(0, 1)$ 구간 내에 속하는 경우 반응변수에 대한 회귀모형의 적합값은 상한과 하한이 존재하여 일반적으로 정규분포 가정하의 선형회귀모형 사용이 적절치 않다. 반응변수를 실수 전체로 확장될 수 있도록 변환하여 선형회귀모형을 적용할 수도 있지만 이러한 접근법 역시 모형의 모수들이 반응변수 Y에 대한 평균으로 직접적으로 해석될 수 없고, 특히 비율 데이터의 경우 일반적으로 갖고 있는 비대칭성이나 이분산성의 문제들로 정규성 기반 구간추정이나 가설검정에서 표본이 적은 경우 부정확한 결과를 도출하는 문제를 가져올 수 있다 (Simas 등, 2010). 따라서 이를 보완하기 위해 Ferrari와 Cribari-Neto (2004)가 제안한 표준 단위구간 내에 속하는 연속 변수에 대한 베타회귀모형을 통해 회귀분석을 실시한다. 해당 모형의 가장 큰 장점은 반응변수가 베타분포를 따른다는 가정하에 기존의 모형 해석적 단점과 반응 변수에 대한 비대칭성, 이분산성을 모두 고려하여 회귀모형을 적용할 수 있다는 것이다.

$i = 1, \ldots, n$에 대하여 y_i가 평균을 μ_i, 정밀도를 ϕ로 하는 베타분포를 따른다고 하자. 즉, $y_i \sim \mathrm{Beta}(\mu_i, \phi)$. 그때, 베타회귀모형은 다음과 같이 정의된다.

$$\begin{aligned} Y_i &\sim \mathrm{Beta}(\mu_i, \phi) \\ g(\mu_i) &= x_i^\top \beta = \eta_i, \end{aligned} \tag{2.5}$$

여기서 $g(\cdot) : (0,1) \mapsto \mathbb{R}$는 순증가(strictly increasing)하는 함수로, 두번 미분가능(differentiable)한 특성을 갖는 연결함수(link function)이다. 그리고 $\beta = (\beta_1, \ldots, \beta_k)^\top$는 $k \times 1$ 인 회귀계수벡터(regression coefficients) (단, $k < n$), $x_i = (x_{i1}, \ldots, x_{ik})^\top$는 k차원의 설명변수벡터를 나타낸다. 본 모형에서 정밀도 모수 ϕ는 하나의 상수로서 취급되며 평균에 대한 하위모형만 존재하는 형태이다. 참고로, 위의 모형에서 회귀계수와 정밀도 모수는 Ferrari와 Cribari-Neto (2004)에서 처럼 최대가능도법(maximum likelihood method; MLE) 혹은 Kelley 등 (2007)에서 소개된 베이지안 계산 방법에 의해 추정될 수 있다.

2.2.2. 확장된 베타회귀모형

기존의 베타회귀모형은 정밀도 모수를 고정된 상수로 가정하여 현실의 문제를 다루기에 적절하지 않을 수 있다. 이에 Smithson와 Verkuilen (2006)은 베타회귀모형의 확장을 제안하였고, Simas 등 (2010)은 좀 더 확장된 베타회귀모형을 제안하여 이를 변수산포 베타회귀모형(beta regression with dispersion covariates)이라고 명명하였다.

변수산포 베타회귀모형은 정밀도 모수를 상수로 취급하지 않고 평균과 같이 설명변수들에 대한 선형 예측자를 두어 평균과 정밀도 모수에 대한 두개의 하위모형을 갖는 모형으로 아래의 식 (2.6)과 같이 정의된다.

$$\begin{aligned} Y_i &\sim \text{Beta}\left(\mu_i, \phi_i\right) \\ \mu_i &= g_1^{-1}\left(x_i^\top \beta\right), \\ \phi_i &= g_2^{-1}\left(z_i^\top \gamma\right), \end{aligned} \tag{2.6}$$

위 식에서, $g_1^{-1}(\cdot)$와 $g_2^{-1}(\cdot)$는 연결함수의 역함수를 나타내고, β와 γ는 각 $k \times 1$, $h \times 1$인 회귀계수 벡터, x_i는 k차원, z_i는 h차원의 설명변수 벡터이며 각 설명변수 벡터 간 변수가 서로 중복되거나 완전히 동일한 경우도 허용한다 (단, $k + h < n$).

2.3. 빈도론적 추론

본 절에서는 2.2절에서 설명한 베타회귀모형과 확장된 베타회귀모형의 모수에 대해 빈도론적 추정 방법을 설명한다.

2.3.1. 가능도 함수

베타회귀모형과 확장된 베타회귀모형의 회귀 모수는 빈도론적 관점에서 가능도가 최대가 될 때의 모수인 최대가능도추정량(maximum likelihood estimator; MLE)으로 추정될 수 있다. 이때 베타회귀모형 및 확장된 베타회귀모형이 최대화하여야 하는 가능도함수는 각각 식 (2.7), 식 (2.8)과 같다.

$$\begin{aligned} L\left(y_1, \ldots, y_n \mid \beta, \phi\right) &= \prod_{i=1}^{n} f\left(y_i \mid g^{-1}\left(x_i^\top \beta\right), \phi\right) \\ &= \prod_{i=1}^{n} \frac{\Gamma(\phi)}{\Gamma(g^{-1}(x_i^\top \beta)\phi)\Gamma((1 - g^{-1}(x_i^\top \beta))\phi)} y^{g^{-1}(x_i^\top \beta)\phi - 1}(1-y)^{(1-g^{-1}(x_i^\top \beta))\phi - 1}. \end{aligned} \tag{2.7}$$

$$\begin{aligned} L(y_1, \ldots, y_n \mid \beta, \gamma) &= \prod_{i=1}^{n} f\left(y_i \mid g_1^{-1}(x_i^\top \beta), g_2^{-1}(z_i^\top \gamma)\right) \\ &= \prod_{i=1}^{n} \frac{\Gamma(g_2^{-1}(z_i^\top \gamma))}{\Gamma(g_1^{-1}(x_i^\top \beta) \cdot g_2^{-1}(z_i^\top \gamma))\Gamma((1 - g_1^{-1}(x_i^\top \beta)) \cdot g_2^{-1}(z_i^\top \gamma))} \\ &\quad \times y^{g_1^{-1}(x_i^\top \beta) \cdot g_2^{-1}(z_i^\top \gamma) - 1}(1-y)^{(1-g_1^{-1}(x_i^\top \beta)) \cdot g_2^{-1}(z_i^\top \gamma) - 1}. \end{aligned} \tag{2.8}$$

2.3.2. 최대가능도추정량의 계산

식 (2.7)과 식 (2.8)에서 주어진 가능도함수를 최대화 하는 모수를 구하기 위한 점수방정식의 계산을 통해 Ferrari와 Cribari-Neto (2004) 와 Simas 등 (2010)에서 제안된 방식으로 최대가능도추정량이 구해질 수 있다. 그러나 실제 모형 적합시에는 보통 BFGS와 같은 비선형적 최적화 알고리즘을 사용하여 로그가능도함수에

대한 수치적 최적화를 거쳐 최대가능도 추정량을 추정한다 (Simas 등, 2010). 본 논문에서는 베타회귀모형 및 확장된 베타회귀모형의 모수의 최대가능도를 추정하기 위해 Cribari-Neto와 Zeileis (2010)이 제안한 R의 "betareg" 패키지를 사용한다.

2.4. 베이지안 추론

본 절에서는 2.2절에서 설명한 베타회귀모형과 확장된 베타회귀모형의 모수에 대해 베이지안 추정 방법을 설명한다.

2.4.1. 사전분포와 사후분포

식 (2.5)에서 정의된 베타회귀모형의 모수에 대한 사후분포를 추정하기 위해, 본 논문에서는 아래와 같이 회귀계수 β의 사전분포로 평균을 a, 공분산을 B로 가지는 k차원의 다변량 정규분포를 고려하고, 정밀도 모수의 사전분포로 형태모수와 비율모수(rate parameter)가 각각 $c > 0$와 $d > 0$인 감마분포(gamma distribution)를 가정한다.

$$\beta \sim N_k(a, B), \quad \phi \sim \text{Gamma}(c, d).$$

이때 사후분포는 아래 식 (2.9)와 같이 주어진다.

$$\begin{aligned} p(\beta, \phi \mid y_1, \ldots, y_n) &\propto \prod_{i=1}^{n} f\left(y_i \mid g^{-1}(x_i^\top \beta), \phi\right) \times N_k(\beta \mid a, B) \times \text{Gamma}(\phi \mid c, d) \\ &\propto \prod_{i=1}^{n} \frac{\Gamma(\phi)}{\Gamma(g^{-1}(x_i^\top\beta)\phi)\Gamma((1-g^{-1}(x_i^\top\beta))\phi)} y^{g^{-1}(x_i^\top\beta)\phi-1}(1-y)^{(1-g^{-1}(x_i^\top\beta))\phi-1} \\ &\quad \times N_k(\beta \mid a, B) \times \text{Gamma}(\phi \mid c, d), \end{aligned} \tag{2.9}$$

여기서 $g(\cdot)$는 순증가 (strictly increasing)하며 두번 미분가능(differentiable)한 특성을 갖는 연결함수이다.

확장된 베타회귀모형의 경우 식 (2.6)에서 정의된 베타회귀모형의 모수에 대한 사후분포를 추정하기 위해, 아래와 같이 회귀계수 β의 사전분포로 평균을 a, 공분산을 B로 가지는 k차원의 다변량 정규분포를 고려하고, γ의 사전분포로 평균을 m, 공분산을 N로 가지는 h차원의 다변량 정규분포를 고려한다.

$$\beta \sim N_k(a, B), \quad \gamma \sim N_h(m, N).$$

이때 사후분포는 아래 식 (2.10)과 같이 주어진다.

$$\begin{aligned} p(\beta, \gamma \mid y_1, \ldots, y_n) &\propto \prod_{i=1}^{n} f\left(y_i \mid g_1^{-1}(x_i^\top \beta), g_2^{-1}(z_i^\top \gamma)\right) \times N_k(\beta \mid a, B) \times N_h(\gamma \mid m, N) \\ &\propto \prod_{i=1}^{n} \frac{\Gamma(g_2^{-1}(z_i^\top\gamma))}{\Gamma(g_1^{-1}(x_i^\top\beta)\cdot g_2^{-1}(z_i^\top\gamma))\Gamma((1-g_1^{-1}(x_i^\top\beta))\cdot g_2^{-1}(z_i^\top\gamma))} \\ &\quad \times y^{g_1^{-1}(x_i^\top\beta)\cdot g_2^{-1}(z_i^\top\gamma)-1}(1-y)^{(1-g_1^{-1}(x_i^\top\beta))\cdot g_2^{-1}(z_i^\top\gamma)-1} \\ &\quad \times N_k(\beta \mid a, B) \times N_h(\gamma \mid m, N), \end{aligned} \tag{2.10}$$

식 (2.9)와 동일하게 위의 식에서 $g_1(\cdot), g_2(\cdot)$는 순증가 (strictly increasing)하며 두번 미분가능(differentiable)한 특성을 갖는 연결함수이다.

2.4.2. 사후분포의 계산

본 논문에서는 베타회귀모형 및 확장된 베타회귀모형의 모수의 사후확률분포를 추정하기 위해 R에서 Joao와 Vinicius (2022)이 제안한 Carpenter (2017)의 "rstan" 패키지 기반인 "bayesbr" 패키지를 사용한다. 해당 패키지는 해밀토니안 몬테카를로(Hamiltonian Monte Carlo) 방법 중 Hoffman과 Gelman (2014)에 의해 제안된 No-U-Turn Sampler (NUTS)를 사용하여 모수의 사후분포로 부터 사후표본을 추출한다.

3. 모형비교

보통최소제곱 기반의 선형회귀모형과 최대가능도법을 활용하여 추정한 베타회귀모형과 확장된 베타회귀모형, 베이지안 베타회귀모형과 확장된 베타회귀모형의 적합 결과를 비교하기 위해 본 논문에서는 결정계수(R^2)와 BIC를 사용한다. 두 평가 기준은 아래 식(3.1)과 같다.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad \text{BIC} = -2\ln(\mathcal{L}) + K\log(n), \tag{3.1}$$

여기서 y_i는 반응변수의 i번째 관측값이며, $\hat{y}_i$는 모형에 의해 추정된 반응변수의 값, $\bar{y}$는 반응변수의 평균 값을 의미한다. 그리고 BIC의 식에서 $\mathcal{L}$은 추정된 모수 하에서의 가능도함수(likelihood function)를 의미하며, 모형의 복잡성을 나타내는 측도인 K는 주어진 모형의 모든 모수의 개수, n은 관측치의 개수를 나타낸다. 참고로, 모수의 값이 아닌 분포를 추정하는 베이지안 방법의 경우에는 모수에 대한 사후분포에서의 평균(즉, 사후평균(posterior mean))을 해당 모수의 추정값으로 두고 Pseudo R^2와 Psuedo BIC를 구하여 모형 비교를 수행한다. 평가 기준에 대해 결정계수는 값이 큰 모형이 최적의 모형을 나타내고, BIC의 경우에는 값이 작은 모형이 최적의 모형을 나타낸다.

4. 실증자료분석

4.1. 실증자료분석 1 : KNHANES

4.1.1. 연구자료 및 연구대상

본 절에서는 보건복지부 산하 질병관리본부의 주관으로 실시되는 국민건강영양조사 원시자료를 바탕으로 흡연율과 커피 섭취 빈도간의 관계를 베타회귀를 이용하여 실증적으로 고찰하고자 한다. 국민건강영양조사는 1995년 제정된 국민건강증진법 제16조에 근거하여 시행하는 전국 규모의 건강 및 영양조사로 가구원확인조사, 건강설문조사, 검진조사, 영양조사를 통해 조사자료를 수집하고 있다. 본 논문에서는 그 중 연구 대상자의 일반적 특성, 흡연 행태, 커피 섭취빈도의 조사 결과가 반영되어 있는 건강 설문조사와 식품섭취빈도 조사 설문 결과를 이용하여 연구를 진행하였다. 분석에 사용한 자료는 커피 섭취빈도자료를 포함하고 있는 제 6기(2013년–2015년)의 국민영양조사(KNHANES)의 원시 자료(http://knhanes.cdc.go.kr/)이고, 연구 대상자는 제6기 중 음료섭취빈도가 조사된 응답자 중 결측치가 없는 만 19세 이상 성인 12,314명으로 설정하였다.

4.1.2. 자료 전처리 및 탐색

커피섭취빈도와 흡연율에 대한 베타회귀모형 적합을 위해 흡연 여부와 흡연 상태에 대해 정의하고 데이터 전처리 및 탐색을 실시하였으며, 구체적인 방법은 다음과 같다.

첫째, 반응변수인 흡연율 계산을 위해 흡연 여부에 대한 기준을 질병관리청 및 NCHS (National Center for Health Statistics)의 기준을 반영하여, 평생 100개비 이상의 담배를 피웠는지에 대한 건강 설문조사 문항 응답 결과로 평생 100개비 이상의 담배를 피운 경우를 흡연자 집단, 평생 100개비 미만의 담배를 피운 경우를

Table 1: KHANES : dataset

	Smoking prop	ncof	incm	Edu	Age	BMI
1	0.4254	0	2.3509	3.1404	29	23.7417
2	0.3582	0.2326	2.3333	3.0833	27	23.2588
3	0.4571	0.5814	2.0625	3.1563	29	25.1747
4	0.3889	1	2.4857	3.2	31	24.3544
5	0.4576	3	2.3306	3.2016	31	23.2973
6	0.5196	5.5	2.3208	3.0755	36	24.2219
7	0.4192	7	2.4541	3.1239	39.5	23.6318
8	0.4889	14	2.3977	3.1590	42	23.8319
9	0.5990	21	2.3361	3.1372	45	24.0770
10	0.7227	28.0875	2.5157	3.1509	47	24.3215
11	0.7783	35	2.3314	3.0814	45	24.6585
12	0.8229	55.3163	2.2986	3.0694	41.5	23.9393

Table 2: KHANES data set description

Variable	Type	Min	Median	Mean	Max
Smoking prop	continuous	0.36	0.47	0.54	0.82
ncof	Continuous	0.00	6.25	14.23	55.32
incm	Continuous	2.06	2.33	2.35	2.52
Edu	Continuous	3.07	3.14	3.13	3.20
Age	Continuous	27.00	37.25	36.92	47.00
Bmi	Continuous	23.26	24.01	24.04	25.17

비흡연자 집단으로 정의하였다. 또한, 흡연자 집단 중 현재도 매일 혹은 가끔 담배를 피우고 있다고 응답한 대상자를 현재 흡연상태, 현재 담배를 피지 않는다고 응답한 대상자를 현재 금연상태로 정의하였다.

둘째, 설명변수는 커피섭취빈도와 연구 대상자의 특성들을 반영할 수 있는 변수들을 선택하여 사용하였다. 커피섭취빈도의 경우 연구 대상자들의 최근 1년간 평균 커피섭취빈도에 대한 문항 응답 결과를 주간 단위로 환산하였으며, 흡연율에 영향을 주거나 연구 대상자의 특성이 반영될 수 있는 나이, 교육수준, 소득수준, BMI 를 추가적인 설명변수들로 설정하였다. 이때, 나이의 경우 10 단위로 범주화를 거쳤으며, 교육과 소득수준의 변수는 각 수준을 반영한 수치형 변수들로 사용하였다. 교육과 소득수준의 구체적인 구분은 국민건강영양조사의 답변을 그대로 사용하여 교육의 경우 초등학교 졸업이하, 중학교 졸업, 고등학교 졸업, 학사 이상의 4단계로 구분하였으며, 소득의 경우에는 소득액에 따라 하, 중하, 중상, 상의 4단계로 구분하여 1,2,3,4 의 정수형 값을 부여하여 사용하였다.

마지막으로, 주간 커피섭취빈도를 기준으로 흡연군과 금연군을 12개의 군집으로 분리하였으며, 각 군의 소득수준, 교육수준, 나이, BMI의 평균을 계산하여 변수 값으로 구성하였다. 반응변수의 경우 해당하는 커피섭취 빈도 내에서의 전체 인원대비 흡연군의 비율을 흡연율로 정의하여 계산하였다. 이때 주간 커피섭취빈도 기준은 $(0, 0.2326, 0.5814, 1, 3, 5.5, 7, 14, 21, 28.0875, 35, 55.3163)$으로 지정하였으며, 본 기준에 따른 사용자료의 형태는 Table 1과 같다.

Table 2는 분석에서 활용하는 변수들의 요약통계량을 제시한 것이다. 반응변수인 흡연율(smoking prop)의 경우 $[0.36, 0.82]$의 범위내에 존재하는 특정 범위내에 존재하는 연속형 비율 자료이다. 해당 자료를 자세히 살펴보면 보편적인 비율데이터가 갖는 비대칭, 이분산성의 특징을 갖고 있지는 않아 보통최소제곱법 기반의 회귀모형을 적합하여도 합당하다고 할 수 있으나, 베타회귀모형과의 비교를 통해 커피섭취빈도와 흡연율의

Table 3: Analysis of the smoking rate using linear regression and beta regression

	Dependent variable:Smoking rate							
	Linear Regression			Beta Regression			Bayesian Beta Regression	
	Estimate	SE	*P*-value	Estimate	SE	*P*-value	Posterior mean	95%CrI
Intercept	−0.467	1.485	0.764	0.204	0.043	<0.001	0.193	(0.062, 0.326)
ncof	0.007	0.001	0.004	0.041	0.006	<0.001	0.037	(0.020, 0.054)
incm	−0.048	0.196	0.813	−0.126	0.424	0.766	−0.224	(−2.169, 1.789)
Edu	−0.059	0.422	0.893	−0.364	0.906	0.688	−0.387	(−4.840, 2.601)
Age	−0.003	0.004	0.458	−0.011	0.011	0.315	0.005	(−0.037, 0.039)
Bmi	0.045	0.036	0.260	0.132	216.499	0.080	0.150	(−0.212, 0.512)
Intercept (ϕ)				216.5	88.2	0.014	91.523	(22.474, 179.741)
R^2	0.938			0.956			0.951	
BIC	−27.231 (df=7)			−29.018 (df=7)			−22.032 (df=7)	

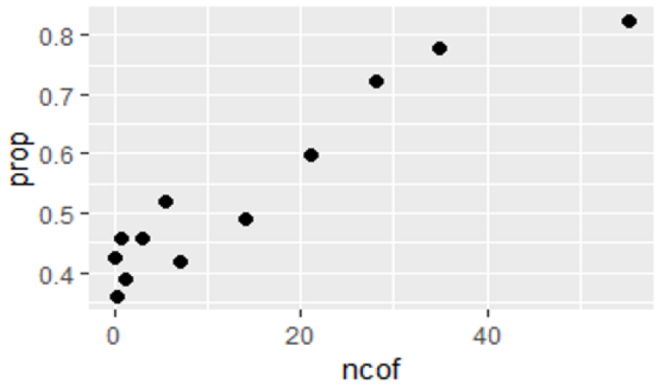

Figure 2: *Scatter plot of smoking rates according to coffee consumption for smoking group.*

관계를 더 잘 설명할 수 있는 모형이 무엇인지 살펴본다. 설명변수에 해당하는 주당커피섭취빈도(ncof)와 소득수준(incm), 교육수준(edu), 나이(age), BMI(bmi) 변수들도 연속형 변수들로 구성되어 있다.

Figure 2는 커피섭취빈도와 흡연율간의 관계를 탐색해보기위해 산점도를 통해 확인해본 결과이다. 그림으로부터 커피섭취빈도가 증가할수록 해당 커피 섭취빈도에서 흡연율이 전반적으로 증가하는 경향을 보이는 것을 알 수 있다.

4.1.3. 모형 비교 및 분석 결과

이 절에서는 흡연율과 커피 섭취빈도간의 관계를 보기 위해서 적합한 모형의 분석 결과를 제시한다. 반응변수는 흡연율, 설명변수는 커피섭취빈도, 교육수준, 소득수준, 나이, BMI로 하였고, 보통최소제곱법 기반 선형회귀모형과 베타회귀모형을 적합하였다. 베타회귀모형에 대한 추정은 최대가능도법과 베이지안 추론 두 가지 방식을 모두 이용하였으며, 연결함수는 logit, probit, cloglog, cauchit, log, loglog 중 관측치에 대한 가장 높은 로그 가능도 값과 가장 낮은 BIC 값을 보이는 cauchit 연결함수를 선택하였다. 베이지안 추론의 경우, 총 100,000개의 마코프 체인(Markov chain)을 추출하여 그 중 초기의 60,000개는 버리고 남은 40,000개를 사

Table 4: Breusch-Pegan test result in linear regression with KHANES data

Statistic	BP	df	p-value
OLS	5.155	4	0.2718

후표본으로 하여 추론에 이용하였으며, 결정계수와 BIC는 추정량으로 사후평균(posterior mean)을 사용하여 Pseudo-R^2, Pseudo-BIC로 값을 도출하였다.

Table 3은 각 모형에 대한 적합결과를 나타낸다. 표에서 선형회귀모형 적합 결과를 살펴보면 커피 섭취빈도 변수만이 회귀 계수에 대한 검정에서 유의수준 0.05하에 p-value가 0.004로 유일하게 흡연율의 변화를 설명하는데 유의한 변수로 확인되었다. 빈도론 관점의 베타회귀모형의 경우에도 커피 섭취빈도 변수의 p-value만 0.001 보다 작아 유일하게 유의한 변수임이 확인 되었으며, 해당 변수는 베이지안 추론 방법에서도 95% 신용구간(credible interval, 95%CrI)에 0이 포함되지 않은 것으로 나타났다. 이러한 결과는 커피섭취빈도가 증가할수록 흡연율의 증가에 유의한 영향을 주는 것으로 판단할 충분한 근거가 있음을 의미한다고 할 수 있다. 각 모형간 적합 결과의 비교에서는 커피 섭취빈도 변수를 제외한 타 변수들과 흡연율간의 관계가 대체로 유사하였으나 age 변수의 경우, 빈도론 관점의 선형회귀모형 및 베타회귀모형에서 흡연율과 음의 상관관계를 갖고, 베이지안 베타회귀모형에서는 양의 상관관계를 갖는 것으로 상이한 결과가 확인 되었다. 그러나 해당 변수의 회귀계수 추정치를 살펴보면 0에 가깝고 유의하지 않은 변수이기에 방향성의 차이는 의미가 없는 것으로 판단된다.

그 다음으로, 본 실증분석에서의 최적모형을 결정하기 위해 정밀도 모수인 ϕ의 추정값과 결정계수 및 BIC의 값을 살펴보면, 보통최소제곱법 기반 선형회귀모형과 베이지안 방식으로 추정한 베타회귀모형보다는 정밀도 모수가 더 큰 값을 갖고 결정계수가 더 높으며, BIC가 더 작은 값을 갖는 최대가능도를 이용한 베타회귀모형의 적합이 더 적절할 것으로 판단된다. 따라서 흡연율과 커피섭취간의 관계를 파악하기 위한 흡연율 자료의 회귀모형 적합에서 최적모형은 cauchit 연결함수를 사용하는 빈도론 방법의 베타회귀모형이라 판단된다.

비율자료의 특성상 이분산성(heteroscedasticity)이나 비대칭(assymetric)을 보이는 경향이 있으므로, 추가적으로 보통최소제곱법 기반 선형회귀모형의 이분산성 여부를 확인하기 위하여 Breusch와 Pagan (1979)이 제안한 Breusch–Pagan test를 실시하고 해당 결과를 Table 4에 제시하였다. 유의수준 0.05하에 해당 검정은 p-value가 0.2809로 이분산성을 갖는다고 보기 어렵다. 따라서 선형회귀모형을 적합하여도 모형 가정에 큰 문제를 갖지는 않으나, 결정계수와 BIC를 모두 고려하였을 때 보통최소제곱법 기반 선형회귀모형보다 베타회귀모형을 적합한 경우 반응변수의 평균에 대하여 더 유의하게 적합될 수 있음을 확인할 수 있었다.

4.2. 실증자료분석 2 : COVID 19

4.2.1. 연구자료 및 연구대상

본 절에서는 R의"betaBayes" 패키지에 내재되어 있는 미국의 카운티별 특성 및 COVID 19 확진자와 사망자 수 자료를 통하여, 카운티 별 특성과 COVID 19 누적 사망률간의 관계를 베타회귀모형을 이용하여 실증적으로 고찰하고자 한다. 모형 적합시 설명변수로는 미국의 인구 조사국(United States Census Bureau)의 2018년 ACS (American Community Survey) 5년 추정치 중 남성비율(MaleP), 65세 이상 인구 비율(Age65plusP), 빈곤층 비율(PovertyP)을 사용하였으며, 추가적으로 USDA (U.S. Department Of Agriculture) 산하 ERS(Economic Research Service)에서 10년 단위로 대도시 지역의 인구 규모에 따라 대도시와 소도시 및 비수도권의 카운티를 구분하는 분류체계인 2013년 기준 RUCC (Rural-Urban Continuum Code)를 함께 고려하였다. 해당 변수명은 RUCC_2013으로 0.1–0.3의 값을 갖는 경우, 값이 작을수록 대도시에 가까우며, 0.4–0.9값을 갖는 경우 값이 클수록 시골 지역으로 간주된다. 반응 변수로는 비영리 단체 USA Facts의 2020년 10월 13일 기준 카운티

Table 5: Summary statistics of covid data without outliers

Variable	Type	Min	Median	Mean	Max
Deathrate	Continuous	0.00	0.02	0.02	0.06
MaleP	Continuous	0.41	0.50	0.50	0.79
Age65plusP	Continuous	0.06	0.18	0.18	0.56
PovertyP	Continuous	0.02	0.15	0.15	0.55
RUCC_2013	Discrete	0.10	0.60	0.50	0.90

Table 6: Breusch-Pegan test result in linear regression with covid data without outliers

Statistic	BP	df	p-value
OLS	57.217	5	0.000^{***}

Table 7: Analysis of the deathrate using linear regression, beta regression and extended beta regression

	Dependent variable: Death rate								
	Linear regression			Beta regression			Extended beta regression		
	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
Intercept (μ)	0.018	0.000	<0.001	−4.102	0.019	<0.001	−4.058	0.020	<0.001
MaleP (μ)	−0.031	0.012	0.006	−4.695	0.742	<0.001	−5.098	0.732	<0.001
Age65plusP (μ)	0.030	0.006	<0.001	−0.214	0.372	0.565	−0.170	0.355	0.632
PovertyP (μ)	0.051	0.004	<0.001	2.756	0.241	<0.001	2.469	0.235	<0.001
RUCC_2013 (μ)	−0.015	0.001	<0.001	−1.268	0.066	<0.001	−0.633	0.078	<0.001
Intercept (ϕ)	-	-	-	59.701	1.755	<0.001	4.113	0.030	<0.001
RUCC_2013 (ϕ)	-	-	-	-	-	-	−1.482	0.108	<0.001
R^2	0.105			0.197			0.178		
BIC	−16752.71 (df = 6)			−18395.83 (df = 6)			−18574.47 (df = 7)		

별 COVID-19 누적 사망자 수를 카운티별 누적 확진자 수로 나누어 누적 사망률로 사용하였다. 원 자료는 총 3,142개의 카운티별 특성과 코로나 누적 확진 및 사망 수를 담고 있지만 1.5×IQR (interquatile range) 기준으로 이상치와 결측치 변수를 제거하여 결과적으로 총 49개 주의 2,887개의 카운티의 자료를 분석에 활용하였다.

4.2.2. 자료 전처리 및 탐색

모형을 적합하기 위해, 자료의 전처리를 실시하였다. 구체적으로, 누적 확진자가 없어 누적 사망률을 계산할 수 없는 경우 누적 사망률을 0으로 대체하여 사용하였으며, 해당 자료들의 범위 변환을 위해 Liu와 Eugenio (2016)에서 제안한 아래의 식 (4.1)를 이용하였다. 즉, 사망률의 범위를 $[0, 1)$에서 $(0, 1)$로 변환하였다.

$$y^{**} = (y(n-1) + 0.5)/n. \tag{4.1}$$

변환을 거친 자료에 대한 요약 통계량은 Table 5에 제시하였다. 이때 반응변수인 deathrate의 최솟값은 0.000159 로 기존의 0의 값을 가진 사망률 자료의 변환 결과값이며, 최댓값은 0.059696으로 $(0.00, 0.06)$ 범위 내에 속하는 연속형 비율 자료이다. 설명변수의 경우 연속형 변수 MaleP, Age65plusP, PovertyP 와 이산형 변수 RUCC_2013 으로 구성되어 있다.

Table 8: Analysis of the deathrate using Bayesian beta regression and Bayesian extended beta regression

	Bayesian beta regression		Bayesian extended beta regression	
	Posterior mean	95% CrI	Posterior mean	95% CrI
Intercept	−4.102	(−4.138, −4.064)	−4.100	(−4.137, −4.064)
MaleP	−4.682	(−6.239, −3.137)	−4.742	(−6.245, −3.150)
Age65plusP	−0.220	(−0.914, 0.460)	−0.225	(−0.920, 0.471)
PovertyP	2.275	(2.303, 3.185)	2.742	(2.296, 3.216)
RUCC_2013	−1.268	(−1.390, −1.143)	−1.258	(−1.378, −1.138)
Intercept(ϕ)	59.675	(56.145, 63.037)	4.085	(4.029, 4.142)
RUCC_2013(ϕ)	-	-	−0.012	(−0.031, 0.008)
R^2	0.162		0.162	
BIC	−18398(df=6)		−18567(df=7)	

4.2.3. 모형 비교 및 분석 결과

이 절에서는 COVID 19 사망률과 미국의 카운티 별 특성간의 연관성에 대해 OLS기반 선형회귀모형과 베타회귀모형, 확장된 베타회귀모형 적합결과를 제시한다. 또한 각 모형의 반응변수는 deathrate, 설명변수는 MaleP, Age65plusP, PovertyP, RUCC_2013 이며, 확장된 베타회귀모형의 경우 정밀도에 대한 하위모형의 설명변수로 RUCC_2013 만을 두어 각 설명변수들의 유의성을 확인하였다. 모형 적합시, 베타회귀모형의 경우 logit 연결함수를 사용하였으며, 확장된 베타회귀모형의 경우 평균에 대한 하위모형의 연결함수를 logit, 정밀도에 대한 하위모형의 연결함수를 log로 사용하였다. 베이지안 추론 방법에서는 100,000개의 마코프 체인을 추출하여 그 중 초기의 60,000개는 버리고 남은 40,000개를 사후표본으로하여 추론을 실시하였다.

Table 6은 이분산성 여부를 확인하기 위해 OLS기반 선형회귀모형을 적합한 후, 실시한 Breusch-Pegan test의 결과를 보여준다. 해당 결과를 통해 알 수 있듯이, 반응변수인 Covid 19 사망률에 대하여 이분산성이 존재한다. 이러한 경우 반응변수를 로짓(logit) 변환한 뒤 OLS 방법으로 분석할 수 있으나, 해당 방법은 반응변수의 비대칭도를 높여 잔차를 비대칭적으로 만드는 문제를 야기한다는 것이 잘 알려져 있다 (Cribari-Neto와 Zeileis, 2010).

다음 Table 7은 OLS 기반의 선형회귀모형 및 최대 가능도 기반의 베타회귀 및 확장된 베타회귀 모형의 결과이다. 우선 Table 7에 제시된 선형회귀 모형의 결과를 보면, 남성의 비율 및 고령 인구의 비율이 높고 대도시의 기준에 멀수록 코로나 발병률이 유의하게 낮아지는 경향을 보였으며, 빈곤층의 비율이 높을수록 코로나의 발병률이 유의하게 높아지는 경향을 보였다. 그 다음으로 베타회귀모형의 결과를 살펴보면, 남성의 비율이 높고 대도시의 기준에서 멀수록 Covid 19 사망률이 유의하게 낮아지는 경향을 보였으며 빈곤층의 비율이 높을수록 Covid 19 사망률이 유의하게 높아지는 경향을 보였다. 이는 OLS 기반의 선형회귀와 유사한 결과이나 OLS 기반 모형과 달리 고령인구비율과 코로나 사망률간의 유의한 관계성은 확인할 수 없었다. 마지막으로, 확장된 베타회귀모형의 결과를 보면, 평균에 대한 하위 모형은 베타회귀와 동일한 결과를 보였고 정밀도에 대한 하위 모형의 경우에는 RUCC_2013 변수가 Covid 19 사망률에 유의한 영향을 미침을 확인할 수 있었다. 추가적으로 모형별 결정계수를 확인한 결과, OLS기반 선형회귀모형의 경우 0.105, 베타회귀모형의 경우 0.197, 확장된 베타회귀모형의 경우 0.178로 베타회귀모형, 확장된 베타회귀모형, OLS 기반의 선형회귀모형의 순으로 높은 결과를 보였다. 또한 BIC의 값을 살펴보면 확장된 베타회귀모형에서 BIC의 값이 −18574.47로 가장 작은 값을 가졌다.

Table 8은 Covid 19 사망률에 대한 베이지안 베타회귀 및 확장된 베타회귀모형의 모수 추정 결과로 모수의 사후평균과 95% 신용 구간을 제시하였다. 먼저 좌측의 베이지안 베타회귀모형의 적합 결과를 살펴보면,

남성의 비율이 작을수록, 빈곤층의 비율이 높을수록, 대도시의 기준에 멀수록 COVID 19 사망률이 유의하게 높았으며 고령인구비율의 경우 COVID 19 사망률과 유의한 관계임을 확인할 수 없었다. 이는 최대 가능도 추정법으로 추정한 베타회귀모형과 유사한 결과임을 확인할 수 있으며, 확장된 베이지안 베타회귀모형 역시 최대가능도 추정을 통한 베타회귀 모형과 동일한 결론이 도출되었다. 결정계수는 베타회귀모형이 0.16165, 확장된 베타회귀모형이 0.16182로 근소한 차이를 보였으며 BIC의 경우 확장된 베타회귀모형에서 −18567로 더 작은 값을 가졌다.

지금까지의 자료분석결과를 종합해 보면, 모형 비교에서는 최대 가능도 추정법을 통한 베타회귀모형의 결정계수 값이 가장 높았으며, 최대 가능도 추정법을 통한 확장된 베타회귀모형에서 가장 낮은 BIC 결과 값을 얻을 수 있었다. 따라서 OLS 기반의 선형회귀모형보다 베타회귀모형 및 확장된 베타회귀모형이 Covid 19 사망률에 대한 모형 적합에서 더 우수함을 확인할 수 있었으며, 모형의 해석에서는 남성의 비율이 낮고 빈곤층의 비율이 높고, 소도시 혹은 도시화가 진행되지 않은 시골 지역에 가까울수록 Covid 19 사망률이 유의하게 낮음을 확인할 수 있었다.

5. 결론

연속형 반응변수가 단위 구간 혹은 특정 구간내에 존재하는 경우 모형 적합값에 상한과 하한이 존재하고 오차항이 정규분포를 따르지 않기 때문에, OLS 기반의 선형회귀모형을 통한 추정과 검정은 적절치 않을 수 있다. 특히 전형적으로 이분산성과 비대칭적인 특징을 고려해야 하는 비율 자료의 경우 이를 고려할 수 있는 대안이 되는 모형 적합이 필요하다. 이에 본 논문에서는 평균에 대한 하위 모형을 갖는 베타회귀모형과 평균 및 정밀도 모수에 대한 하위 모형을 갖는 확장된 베타회귀모형을 고려하였다. 또한 흡연율에 대한 실증 자료와 코로나 사망률에 대한 실증 자료를 OLS 기반의 선형회귀모형을 적합하여 베타회귀모형과의 적합 결과를 비교하는 실증분석을 수행하였다. 모형 적합 결과의 비교에서 흡연율과 사망률에 대한 두 실증분석 결과 모두 OLS 기반의 선형회귀모형보다 베타회귀모형이 더 유연한 적합을 통해 높은 결정계수 값과 낮은 BIC 값을 갖는 것을 확인하였다. 이는 베타회귀모형의 유연한 적합을 통한 유용성을 확인한 것으로, 반응변수가 특정 구간 내에 존재하는 경우 베타회귀모형 적합의 적절성을 확인한 것으로 생각할 수 있다.

적합 결과의 해석에서는 베타회귀모형을 적합한 흡연율 자료의 경우, 커피 섭취빈도가 높을수록 흡연율이 높아지는 유의한 양의 상관관계가 있음을 확인하였다. 또한 베타회귀모형을 적합한 Covid 19 의 사망률 자료의 경우, 남성의 비율이 낮고, 빈곤층의 비율이 높으며, 도시화가 진행되지 않은 시골 지역에 가까울수록 Covid 19 사망률이 유의하게 낮은 관계를 가짐을 확인할 수 있었다. 따라서, 흡연율과 사망률에 대한 본 분석 결과를 통해 베타회귀모형 적합의 적절성을 확인하였다는 점과 흡연에 유의한 영향을 미치는 요인을 확인하였다는 점, 코로나 사망률에 대한 지역적 특성을 연구함으로써 개인 특성으로는 포착되지 않는 지역적 추세를 설명하고 추가적인 감염병 도래 예방에 있어 대응 정책 결정을 위해 해당 연구의 결과가 시사하는 바를 반영할 수 있다는데에 본 연구의 의의를 둔다.

References

Ahn HJ, Gwak JI, Yun SJ, Choi HJ, Nam JW, and Shin JS (2017). The influence of coffee consumption for smoking behavior, *Korean Journal of Family Practice*, **7**, 218–222.

Bayes CL, Bazán JL, and García CB (2012). A new robust regression model for proportions, *Bayesian Analysis*, **21**, 841–866.

Branscum AJ, Johnson WO, and Thurmond MC (2007). Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses, *Australian and New Zealand*

Journal of Statistics, **49**, 287–301.

Buntaine MT (2011). Does the Asian development bank respond to past environmental performance when allocating environmentally risky financing?, *World Development*, **39**, 336–350.

Breusch TS and Pagan AR (1979). A simple test for heteroscedasticity and random coefficient variation, *Econometrica*, **47**, 1287–1294.

Carpenter B, Gelman A, Hoffman MD *et al* (2017). Stan: A probabilistic programming language, *Journal of Statistical Software*, **76**, 1–32.

Cribari-Neto F and Zeileis A (2010). Beta regression in R, *Journal of Statistical Software*, **34**, 1–24.

Ferrari SL and Cribari-Neto F (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics*, **31**, 799–815.

Ferrari SL, Espinheira PL, and Cribari-Neto F (2011). Diagnostic tools in beta regression with varying dispersion, *Statistica Neerlandica*, **65**, 337–351.

Han B, Yun W, and Kim J (2020). Analysis of mobilization training data using beta regression, *Journal of the Korean Data and Information Science Society*, **31**, 611–620.

Homan MD and Gelman A (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, *Journal of Machine Learning Research*, **15**, 1593–1623.

Jang E, Choi S, and Kin D (2018). Robust Bayesian beta regression analysis, *Journal of the Korean Data and Information Science Society*, **29**, 27–36.

Jang E (2017). Analysis of health-related quality of life using beta regression, *Journal of the Korean Data and Information Science Society*, **28**, 547–557.

Joao M and Vinicius M (2022). bayesbr: Beta regression on a Bayesian model, Retrieved Oct. 12, 2022, Available from: https://cran.r-project.org/web/packages/bayesbr/bayesbr.pdf

Kang K, Sung J, and Kim CY (2010). High risk groups in health behavior defined by clustering of smoking, alcohol, and exercise habits, *National Heath and Nutrition Examination Survey*, **43**, 73–83.

Kelley GO, Garabed R, Branscum A, Perez A, and Thurmond M (2007). Prediction model for sequence variation in the glycoprotein gene of infectious hematopoietic necrosis virus in California, U.S.A, *Diseases of Aquatic Organisms*, **78**, 97–104.

Li D, Gaynor SM, Quick C, Chen JT, Stephenson BJK, Coull BA, and Lin X (2021). Identifying US County-level characteristics associated with high COVID-19 burden, *BMC Public Health*, **21**, 1–10.

Liu F and Eugenio EC (2016). A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression, *Statistical Methods in Medical Research*, **27**, 1024–1044.

Peplonska B, Bukowska A, Sobala W *et al* (2012). Rotating night shift work and mammographic density, *Cancer Epidemiology, Biomarkers and Prevention*, **21**, 1028–1037.

Smithson M and Verkuilen J (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distribut ed dependent variables, *Psychological Methods*, **11**, 54–71.

Simas AB, Barreto-Souza W, and Rocha AV (2010). Improved estimators for a general class of beta regression models, *Computational Statistics and Data Analysis*, **54**, 348–366.

Zhou H and Huang X (2022). Bayesian beta regression for bounded responses with unknown supports, *Computational Statistics and Data Anslysis*, **167**, 107345.

Received December 21, 2023; Revised January 18, 2024; Accepted January 23, 2024

비율자료 모델링을 위한 베이지안 베타회귀모형의 비교 연구

이정인[a], 김재오[1,a], 조성일[2,a]

[a]인하대학교 통계·데이터사이언스학과

요 약

비율자료와 같이 반응변수가 제한된 구간에 속하는 경우 비대칭성이나 이분산성의 문제들로 인해 정규성 가정을 기반으로 하는 회귀모형의 적용은 부정확한 결과가 도출될 수 있다. 이러한 경우 대안으로 베타회귀모형이 고려된다. 베타회귀모형은 베타분포를 평균과 정밀도 모수로 재모수화 하였을 때, 반응변수가 베타분포를 따른다는 가정하에 평균과 정밀도에 대한 하위모형을 갖는 회귀모형으로 자료의 이분산성을 쉽게 고려할 수 있다. 본 연구에서는 두 가지 실증분석에서 비율자료에 베타회귀모형을 적합하여 분석하고자 한다. 특히, 제6기 국민 건강조사자료를 통해 흡연율과 커피 섭취와의 연관성을, COVID-19 자료를 기반으로 미국의 지역 특성들과 누적 사망률의 연관성을 고찰한다. 각 분석에서는 보통최소제곱 회귀모형과 베타회귀모형 및 확장된 베타회귀모형을 적용하여 최적의 모형을 선택하고 결과를 해석한다. 분석의 결과는 비율자료에서 베타회귀모형 및 확장된 베타회귀모형 적용의 적절성을 입증한다.

주요용어: 비율자료, 베타회귀모형, 흡연율, 커피섭취, COVID 19

이 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구 사업임 (NRF-2022R1A5A7033499, RS-2023-00209229), 김재오의 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음(2022-0-01127).

[1]교신저자: (22212) 인천시 미추홀구 인하로 100, 인하대학교 통계·데이터사이언스학과. E-mail: jaeoh.k @inha.ac.kr

[2]교신저자: (22212) 인천시 미추홀구 인하로 100, 인하대학교 통계·데이터사이언스학과. E-mail:bstatsjo @inha.ac.kr