응용통계연구 (2012)
25(2), 363–376
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.2.363

영 과잉 순서적 프로빗 모형을 이용한 한국인의 음주자료에 대한 베이지안 분석

오만숙[1] · 오현탁[2] · 박세미[3]

[1]이화여자대학교 통계학과, [2]전북대학교 경영학부, [3]이화여자대학교 통계학과

(2012년 1월 13일 접수, 2012년 2월 21일 수정, 2012년 4월 12일 채택)

요 약

순서적 다항 반응변수의 경우 종종 과도하게 많은 수의 관측치가 0 범주에서 발생하는 영 과잉 특성을 지닌다. 이러한 영 과잉 자료에서 0범주를 발생시키는 요인이 여러 개 존재할 때 일반적인 순서적 프로빗 모형은 자료를 설명함에 있어서 한계를 지닌다. 본 논문에서는 영 과잉 특성을 반영한 이 단계 영 과잉 순서적 프로빗 모형의 베이지안 분석 기법을 제시하고 이를 2008년도 통계청에서 조사한 한국인의 음주소비 자료에 적용시킨다. 첫 번째 단계에서는 음주소비가 하나도 없다고 답한 0 범주에 속하는 비음주자들을 신념 또는 영구적 건강상의 문제 등으로 상황에 관계없이 음주를 하지 않는 절대적 비음주자(genuine non-drinker, non-participant)와 현재 소비가 없지만 상황에 따라 음주자가 될 가능성이 있는 잠재적 음주자(zero consumption potential drinker)로 구분하는 프로빗 모형을 적용시켜 분석한다. 두 번째 단계에서는 잠재적 음주자와 1 이상의 범주에 속하는 실제적 음주자를 합하여 음주자 집단으로 보고 이에 대하여 순서적 프로빗 모형을 적용하여 분석한다. 분석결과, 비음주자 중 약 30%가 절대적 비음주자로 음주자료가 일반적 순서적 자료에 비하여 뚜렷한 영 과잉 특성을 가짐을 알 수 있었다. 각 변수의 한계효과를 분석함으로써 같은 설명변수가 절대적 비음주자와 잠재적 음주자에 미치는 영향이 서로 반대로 나타날 수 있음을 발견하였고, 따라서 한국인의 음주자료에 대하여 제안된 영 과잉 순서적 프로빗 모형이 유용함을 보여주었다.

주요용어: 영 과잉, 마코브 체인 몬테칼로, 사후분포, 순서적 범주형 자료.

1. 서론

우리나라는 세계에서 술을 많이 소비하는 국가 중 하나이며 현재도 음주로 인하여 초래되는 사회경제적 비용은 꾸준히 증가하여 연 20조원에 달하고 있다. 이와 같은 음주는 질병과 사망뿐 아니라 교통사고, 산업재해, 가정폭력, 폭행, 자살, 자해 등 심각한 사회적 문제를 발생시키고 있다 (박재홍 등, 2010).

우리나라 성인의 월간 음주율은 98년 이후부터 현재까지 꾸준하게 증가하고 있는 추세이다. 그림 1.1에서 보이는 것처럼 2008년 월간 음주율은 전체 59.5%에 이르며, 특히 남자는 74.6%로 높은 수치를 보인다. 여자의 음주율 역시 2008년에 44.9%로 꾸준히 증가함을 보이고 있다(보건복지부 국민건강영양조사). 술에 관대한 우리사회의 문화는 성인뿐만 아니라 청소년의 음주까지 부추기고 있는 실정으로 청소년 음주율이 전체의 30%에 육박할 정도로 학생의 신분임에도 불구하고 이른 나이에 청소년들이 음주를 경험하고 있다(질병관리본부, 청소년건강행태 온라인조사 통계).

이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행한 기초연구사업임 (No. 2011-0004589).

[1]교신저자: (120-750) 서울시 서대문구 이화여대길 52, 이화여자대학교 통계학과, 교수.
E-mail: msoh@ewha.ac.kr

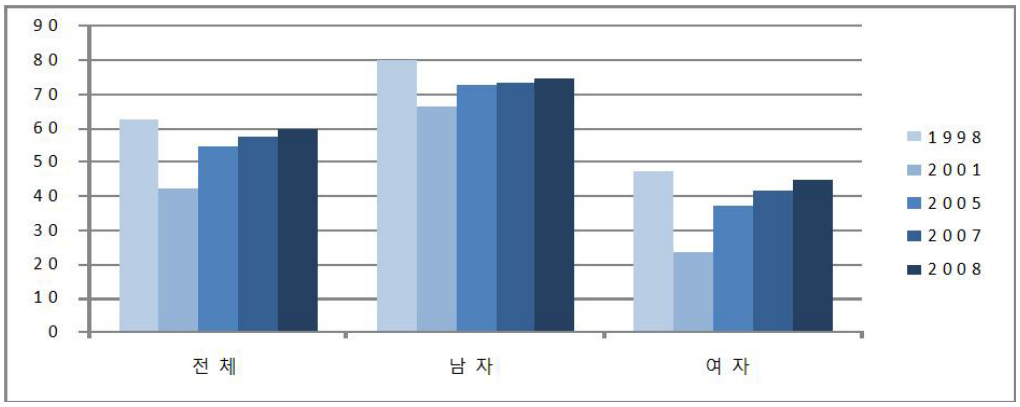

그림 1.1. 19세 이상 성인 월간 음주율(단위: %)
출처: 국민건강영양조사, 건강행태조사; ※ 월간 음주율: 최근 1년 동안 한 달에 1회 이상 음주자 비율

높은 음주율은 자연스럽게 우리사회에 부정적 요인으로 작용하게 되어 개인의 신체적, 정신적, 재정적으로 심각한 문제를 야기하고 있다. 따라서 이를 예방하고 적절한 해결책을 모색하기 위해 음주와 관련된 연구와 조사가 요구되고 있으나 아직까지도 담배에 비해 술에 대한 경각심은 상대적으로 미흡하며 사회적인 관대함으로 인해 음주문제에 대한 대비책에는 여전히 소홀한 실정이다.

음주와 특정 요인간의 관계 및 연관성을 규명하여 음주문제에 대한 해결책의 방향을 제시하기 위해 수행된 최근의 연구를 보면, '중학생의 음주 지식 및 태도와 음주 행위 요인간의 관계' (서경현과 양승애, 2010), '노인의 음주 및 정신건강 특성이 자살생각에 미치는 영향' (윤명숙 등, 2010) '한국인의 사인별 알코올 기여도 산출' (천성수와 손애리, 2008), '대학생 음주실태와 문제음주 변화 추이' (김승수와 정슬기, 2009) 등이 있다.

그러나 대부분 기존의 연구들은 음주자료의 특성인, 음주를 하지 않는다고 답한 비음주자들이 과도하게 많은, 영 과잉(zero-inflated) 자료의 특성을 반영하지 않고 있다. 영 과잉 모형은 최근 과도한 영을 포함하는 이산자료에 적절한 모형으로 관심을 끌고 있으며, 음주자료와 같은 순서적 다항 자료에서 0 반응 값이 과도하게 많은 영 과잉 자료는 일반적인 순서적 프로빗 모형 보다는 영 과잉 특성을 반영한 영 과잉 순서적 프로빗 모형(Zero-Inflated Ordered Probit model; ZIOP)이 더 적절함이 알려지고 있다. (Gurmu와 Dagne, 2009; Harris와 Zhao, 2007; Rodrigues, 2003)

또한 음주에 관한 기존의 연구들은 조사대상 기간 동안 음주를 하지 않는다고 대답한 비음주자들을 모두 동일한 집단으로 간주하고 있다. Harris와 Zhao (2007)는 흡연 자료에서 비흡연 응답자들을 동일한 하나의 집단으로 보지 않고 두 그룹으로 나누어 분석하였는데 음주자료에서도 마찬가지의 접근 방법이 유용할 것이라 생각된다. 구체적으로, 조사대상 기간 동안 음주를 하지 않는다고 대답한 비음주자 중에는 만성적 혹은 영구적인 건강상의 문제, 종교나 신념 등의 이유로 인하여 음주와 관련 있는 요인들(수입, 건강 등)의 값에 관계없이 음주를 하지 않는 절대적 비음주자(genuine non-drinker)와 일시적인 건강문제 또는 경제적인 이유로 조사 대상 기간에는 음주를 하지 않았지만 음주와 관련 있는 요인들의 값이 변하면 앞으로 음주를 할 가능성이 있는 잠재적 음주자(potential drinker)의 두 그룹으로 나눌 수 있을 것이다. 즉, 표면적으로 두 그룹 모두 조사대상 기간에 음주를 하지 않은 비음주자들로 구성되어 있지만 절대적 비음주자는 일종의 비참여자(non participant)이고 잠재적 음주자는 음주자이지만 제로 소비자(zero consumption)라고 볼 수 있다. 이 두 그룹은 알코올에 대한 태도 차이로 인하여 요인별 효과가 다를 수 있는데, 이 두 그룹을 하나로 합하여 분석한다면 Harris와 Zhao (2007)가 흡연 자료에서 보

인 바와 같이 서로의 효과가 상쇄되어 잘못된 결론에 이를 수 있다.

통계청에서 주관한 2008년 사회통계 조사 중 보건, 가족 부문 조사에서는 가구주가 다양한 설문에 대하여 응답하도록 되어 있다. 우리는 그 중 "지난 1년(2007.6.24~2008.6.23)동안 술을 한 잔 이상 마신 적이 있습니까? 있다면, 얼마나 자주 드셨습니까?"라는 질문에 대한 응답을 관심 있는 반응변수로 두고 영 과잉 특성을 반영한 ZIOP 모형을 사용하여 한국인의 음주실태에 대한 베이지안 분석을 수행하고자 한다.

음주자료에 대한 ZIOP 모형 적용의 첫 단계에서는 절대적 비음주자와 그 외, 즉, 절대적 비음주자 그룹과 잠재적 음주자와 실제로 음주를 하는 실제적 음주자로 구성된 음주자 그룹으로 이분하는 프로빗 모형을 도입하고, 두 번째 단계에서 잠재적 음주자와 실제적 음주자에 대하여 순서적 프로빗 모형을 도입한다. 깁스표본기법을 이용한 베이지안 기법을 적용하여 모형의 복잡성으로 인하여 수리적 추정이 어려운 점을 해결한다. 또한 각 설명변수가 절대적 비음주자 확률에 미치는 한계효과와 잠재적 음주자에 미치는 효과를 계산하여 설명변수의 효과가 두 그룹에서 상반되게 나타나는지 알아본다.

2장에서는 ZIOP 모형과 적절한 잠재변수를 소개하고 3장에서는 모수의 사전분포와 깁스표본기법을 적용하기 위한 각 모수의 조건부 사후분포를 유도한다. 4장에서는 우리나라 통계청에서 주관한 2008년 보건, 가족부문조사 자료 중 음주관련 변수들을 선별하여 베이지안 ZIOP 모형 분석을 수행하고 결과를 해석한다. 5장에서는 요약과 결론을 제시한다.

2. 영 과잉 순서형 프로빗 모형(ZIOP 모형)

2.1. ZIOP 모형

관측치가 $0, 1, \ldots, J$의 순서적 범주 중 하나로 얻어지는 순서적 다항변수에 대한 모형으로 순서형 프로빗(Ordered Probit; OP) 모형이 주로 사용된다. 그러나 자료가 과도한 영 값(zero)을 포함하고 있을 때, OP 모형은 과도한 영 값을 설명함에 있어서 제한된 능력을 가진다. 특히 영(zero) 범주를 선택한 대상자들이, 예를 들어 절대적 비음주자와 잠재적 음주자와 같은 두 개의 구별된 그룹으로 구성되어 있을 때 OP 모형은 설명력에 한계를 지닌다 (Harris와 Zhao, 2007). 반면, OP 모형을 확장한 ZIOP 모형은 OP 모형이 가지는 한계를 극복하는 장점이 있어 최근 많은 관심을 끌고 있다.

본 논문에서는 ZIOP 모형에 두 단계의 일반선형모형을 적용한다. 0 범주에 속하는 비음주 응답자를 절대적 비음주자와 잠재적 음주자로 나누고, 1 이상의 범주에 속하는 응답자를 실제적 음주자라고 하자. 첫 번째 단계에서는, 각 대상을 절대적 비음주자와 (잠재적 또는 실제적) 음주자로 구분하는 프로빗 모형을 가정한다.

즉, 이항변수 S에 대하여 다음과 같은 프로빗 모형을 가정한다.

$$S = \begin{cases} 1, & \text{잠재적 또는 실제적 음주자,} \\ 0, & \text{절대적 비음주자} \end{cases}$$

이며 $P(S = 1) = \Phi(\mathbf{x}'\boldsymbol{\beta})$이다. 여기에서 Φ는 표준정규분포의 누적 확률분포함수이고 $\mathbf{x}$는 S와 관련있는 설명변수들의 벡터이이다. 프로빗 모형에서 이항변수 S는 $N(\mathbf{x}'\boldsymbol{\beta}, 1)$을 따르는 연속변수 S^*와 다음과 같이 연결시킬 수 있다.

$$S = \begin{cases} 1, & S^* > 0, \\ 0, & S^* \leq 0 \end{cases}$$

S와 S^* 관측되지 않는 잠재변수들이다.

두 번째 잠재변수 Y^*는 정규분포를 따르는 연속변수로서, 잠재적 음주와 실제적 음주의 다항반응 변수 저변에 존재하는 숨겨진 정규분포를 나타내는 변수이다. 즉, Y^*는 잠재성을 가지고 있는 대상까지 포함된 음주자의 음주 소비정도에 대한 변수라 볼 수 있다.

두 번째 단계에서는 음주자의 음주소비 정도에 관한 모형으로, 순서적 프로빗 모형을 가정한다. $S=1$, 즉, 음주자라는 조건하에서 음주 소비의 정도를 표현하는 이산형 변수를 $\widetilde{Y}$라 하면 $\widetilde{Y}$는 절대적 비음주자를 제외한 음주자의 반응변수이므로 일반적 OP 모형에 의해 설명될 수 있다. 연속 잠재변수 Y^*가 $Y^* \sim N(\mathbf{z}'\boldsymbol{\gamma}, 1)$ 분포를 따를 때,

$$\widetilde{Y} = j, \quad \text{if } \ \delta_{j-1} < Y^* \leq \delta_j, \ \ j = 0, \ldots, J$$

이라 정의하면

$$P\left(\widetilde{Y} = j | \mathbf{z}, S=1\right) = \Phi\left(\delta_j - \mathbf{z}'\boldsymbol{\gamma}\right) - \Phi\left(\delta_{j-1} - \mathbf{z}'\boldsymbol{\gamma}\right), \quad j = 0, \ldots, J$$

이다. 여기서 $\mathbf{z}$는 $\widetilde{Y}$와 관련 있는 설명변수들의 벡터이고 δ_j는 $\widetilde{Y}$값을 결정짓는 경계 값(threshold)이며 $\delta_{-1} = -\infty$, $\delta_0 = 0$, $\delta_J = \infty$이다. 여기에서 $\delta_0 = 0$으로 고정하는 것은 $\mathbf{z}$에 상수항이 포함될 경우 계수의 identifiability 때문에 필요한 조건이다. 만약 $\mathbf{z}$가 상수항을 갖지 않으면 δ_0를 고정할 필요가 없다.

이상의 두 단계를 종합하면, ZIOP 모형에서 우리가 관측할 수 있는 반응변수 Y는

$$Y = S\widetilde{Y}$$

로 표현될 수 있고, Y의 확률밀도함수는 다음과 같다.

$$\begin{aligned} p(y|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) &= \begin{cases} P(Y=0|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = P(S=0|\boldsymbol{\beta}) + P(S=1|\boldsymbol{\beta}) P\left(\widetilde{Y}=0|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, S=1\right), \\ P(Y=j|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = P(S=1|\boldsymbol{\beta}) P\left(\widetilde{Y}=j|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, S=1\right), \quad j=1, \ldots, J \end{cases} \\ &= \begin{cases} [1 - \Phi(\mathbf{x}'\boldsymbol{\beta})] + \Phi(\mathbf{x}'\boldsymbol{\beta})\Phi(-\mathbf{z}'\boldsymbol{\gamma}), \\ \Phi(\mathbf{x}'\boldsymbol{\beta}) \left[\Phi(\delta_j - \mathbf{z}'\boldsymbol{\gamma}) - \Phi(\delta_{j-1} - \mathbf{z}'\boldsymbol{\gamma})\right], \quad j = 1, \ldots, J. \end{cases} \end{aligned} \tag{2.1}$$

즉, $Y=0$일 확률은 프로빗 모형에서 $S=0$(절대적 비음주자)의 확률과 OP과정으로부터의 $\widetilde{Y}=0$(잠재적 음주자)의 확률을 더한 것이다. OP 모형과 비교하면 ZIOP 모형에서는 프로빗 모형에서의 영 값이 더해지면서 일반적인 OP 모형에 비하여 과도한 영값이 관측되었다고 보는 것이다. 즉, ZIOP 모형에서는 일반적인 OP 모형에 비하여 절대적인 비음주자의 추가로 인하여 영 과잉이 발생하였기 때문에, 비음주자 중 절대적 비음주자의 비율인 $P(S=0)$이 중요한 의미를 지니며 이를 Zero-inflation 값이라 한다. 만약 Zero-inflation 값이 0에 가깝다면 ZIOP 모형을 적합 시키는 의미가 없다.

2.2. 한계효과(Marginal effect)

각 설명변수가 어떤 확률에 미치는 영향의 정도를 한계효과(Marginal effect)로 측정할 수 있다. 한계효과란 다른 설명변수를 고정한 상태에서 설명변수 x_k가 변함에 따라 확률이 어떻게 변하는지를 측정하며 일반적으로 x_k에 대한 편미분 값으로 나타낼 수 있다. 만약 x_k가 이항변수라면 x_k가 1인 경우와 0인 경우의 확률의 차이를 의미한다. 예를 들어, $P(S=0)$에 대한 x_k의 한계효과는 절대적 비음주자 비율이 x_k가 변화함에 따라 어떻게 움직이는지를 나타낸다. 또한 $P(S=1, \widetilde{Y}=j)$에 대한 x_k의 한계효과는 음주소비정도가 x_k에 따라 어떻게 변화하는지를 나타낸다 (Harris와 Zhao, 2007).

음주자료에서 특히 관심 있는 것은 어떤 설명변수에서 절대적 비음주자와 잠재적 음주자에 대한 한계효과가 서로 상반되게 나타나는가 하는 것이다. 다음 식으로부터

$$P(Y=0) = P(S=0) + P\left(S=1, \widetilde{Y}=0\right).$$

비음주자 $P(Y=0)$에 대한 한계효과는 절대적 비음주자 $P(S=0)$에 대한 한계효과와 잠재적 음주자 $P(S=1, \widetilde{Y}=0)$에 대한 한계효과의 합으로 표현된다. 상황에 관계없는 절대적 비음주자와 조사대상 기간 동안에는 음주를 하지 않았지만 상황이 바뀌면 음주를 할 가능성이 있는 잠재적 음주자는 서로 다른 특성을 가질 것이며, 이런 특성들로 인하여 설명변수의 변화율에 대하여 다른 양상을 보일 수 있다. 그런데 비음주자에 대한 한계효과는 두 한계효과를 합하여 나타나기 때문에, 만약 두 변화율이 양의 효과와 음의 효과로 방향이 다르다면 서로 상쇄되어 중요한 정보를 놓치는 결과를 초래할 수 있다.

따라서 절대적 비음주자에 대한 한계효과와 잠재적 음주자에 대한 한계효과, 그리고 전체 비음주자에 대한 한계효과를 각각 살펴보는 것이 매우 중요하다. 설명변수의 절대적 비음주자에 대한 한계효과와 잠재적 음주자에 대한 한계효과가 다르다면, 음주 정책 설립에서 이 점을 고려하여 절대적 비음주자에 대한 접근방법과 잠재적 음주자에 대한 접근방법을 달리 해야 할 것이다.

또한 음주자의 음주소비 정도에 대한 한계효과도 중요한 의미를 지닌다. 절대적 비음주자가 아니고 음주를 한다면 어떤 변수들이 음주 소비를 늘이거나 줄이는데 영향을 미치는지 파악하는 것이 중요하기 때문이다.

본 논문에서는 $\mathbf{x}=\mathbf{z}$를 사용하기 때문에 i번째 대상의 관심 있는 확률에 대한 x_k변수의 한계효과 식을 살펴보면 다음과 같이 주어진다.

$$\begin{aligned}
M_k(S=1) &= \frac{\partial P(S=1)}{\partial \mathrm{x}_k} = \phi(\mathbf{x}'\boldsymbol{\beta})\cdot\beta_k \\
M_k(S=0) &= \frac{\partial P(S=0)}{\partial \mathrm{x}_k} = -\phi(\mathbf{x}'\boldsymbol{\beta})\cdot\beta_k \\
M_k(S=1,\widetilde{Y}=0) &= \frac{\partial P(S=1,\widetilde{Y}=0)}{\partial \mathrm{x}_k} \\
&= \Phi(-\mathbf{z}'\boldsymbol{\gamma})\phi(\mathbf{x}'\boldsymbol{\beta})\beta_k - \Phi(\mathbf{x}'\boldsymbol{\beta})\phi(-\mathbf{z}'\boldsymbol{\gamma})\gamma_k \\
M_k(Y=0) &= M_k(S=0) + M_k(S=1,\widetilde{Y}=0) \\
M_k(Y=j) &= \left[\Phi\left(\delta_j - \mathbf{z}'\boldsymbol{\gamma}\right) - \Phi\left(\delta_{j-1} - \mathbf{z}'\boldsymbol{\gamma}\right)\right]\cdot\phi(\mathbf{x}'\boldsymbol{\beta})\beta_k \\
&\quad - \left[\Phi\left(\mathbf{x}'\boldsymbol{\beta}\right)\phi\left(\delta_j - \mathbf{z}'\boldsymbol{\gamma}\right) - \Phi\left(\mathbf{x}'\boldsymbol{\beta}\right)\phi\left(\delta_{j-1} - \mathbf{z}'\boldsymbol{\gamma}\right)\right]\cdot\gamma_k, \quad j=1,\ldots,J-1 \\
M_k(Y=J) &= \left[1 - \Phi\left(\delta_{J-1} - \mathbf{z}'\boldsymbol{\gamma}\right)\right]\cdot\phi\left(\mathbf{x}'\boldsymbol{\beta}\right)\beta_k + \Phi\left(\mathbf{x}'\boldsymbol{\beta}\right)\phi\left(\delta_{J-1} - \mathbf{z}'\boldsymbol{\gamma}\right)\cdot\gamma_k.
\end{aligned}$$

위 식에서 $\phi(\cdot)$은 표준정규분포의 확률밀도함수를 나타낸다.

한계효과들이 모수 $\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta}$의 함수로 주어지므로 우리는 위 한계효과식의 사후기대치로 한계효과를 추정하고 추정오차를 계산한다. 참고로, Harris와 Zhao (2007)에서는 고전적 빈도론자 추정방법으로 한계효과 식에 $\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta}$의 최우추정치를 대입하여 한계효과를 추정하고 델타방법을 이용해 분산을 추정하였다.

3. 사전분포와 사후분포

모수 $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta})$의 사전분포로 다음과 같은 공액사전분포를 가정한다.

$$\pi(\boldsymbol{\beta}) \sim N\left(\boldsymbol{\beta}_0, \Sigma_\beta\right),$$

$$\pi(\boldsymbol{\gamma}) \sim N\left(\boldsymbol{\gamma}_0, \Sigma_\gamma\right),$$
$$\pi(\delta_j) \sim N\left(\delta_{0j}, \sigma_j^2\right) I\left(\delta_{j-1} \le \delta_j < \delta_{j+1}\right), \quad j = 1, \ldots, J-1.$$

ZIOP 모형의 우도함수 식 (2.1)과 사전분포로부터 모수의 결합사후분포는 다음과 같이 주어진다.

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}|\boldsymbol{y}) \propto \prod_{i=1}^{n} \left[1 - \Phi(\mathbf{x_i}'\boldsymbol{\beta}) + \Phi(\mathbf{x_i}'\boldsymbol{\beta})\Phi\left(-\mathbf{z_i}'\boldsymbol{\gamma}\right)\right]^{I(y_i=0)}$$
$$\cdot \prod_{i=1}^{n}\prod_{j=1}^{J} \left[\Phi(\mathbf{x_i}'\boldsymbol{\beta}) \cdot \left(\Phi(\delta_j - \mathbf{z_i}'\boldsymbol{\gamma}) - \Phi(\delta_{j-1} - \mathbf{z_i}'\boldsymbol{\gamma})\right)\right]^{I(y_i=j)} \cdot \pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})\pi(\boldsymbol{\delta}).$$

위의 결합사후분포로부터 모수의 조건부 사후분포를 유도하기 어려우므로 잠재변수 S^*, Y^*를 포함하여 결합사후분포를 구하면

$$\pi(s^*, y^*, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}|\boldsymbol{y})$$
$$\propto \prod_{i=1}^{n} \left[I(s_i^* \le 0)\phi(s_i^* - \mathbf{x_i}'\boldsymbol{\beta}) + I(s_i^* > 0)\phi(s_i^* - \mathbf{x_i}'\boldsymbol{\beta})I(y_i^* \le 0)\phi(y_i^* - \mathbf{z_i}'\boldsymbol{\gamma})\right]^{I(y_i=0)}$$
$$\cdot \prod_{i=1}^{n} \left[I(s_i^* > 0)\phi(s_i^* - \mathbf{x_i}'\boldsymbol{\beta})\right]^{I(y_i\neq 0)} \cdot \prod_{j=1}^{J} \left[I(\delta_{j-1} \le y_i^* < \delta_j)\phi\left(y_i^* - \mathbf{z_i}'\boldsymbol{\gamma}\right)\right]^{I(y_i=j)} \cdot \pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})\pi(\boldsymbol{\delta})$$

이다.

결합사후분포에서 $I(y_i = 0)$일 때를 살펴보면 우도함수의 부분이 $s = 0\,(s^* \le 0)$일 때와 $s = 1\,(s^* > 0)$이면서 $\tilde{y} = 0\,(y^* \le 0)$인 부분의 합으로 표현됨을 확인할 수 있다.

모수에 대한 결합사후분포가 복잡한 형태를 띠므로 Markov chain Monte Carlo(MCMC) 기법인 깁스 표본기법을 이용하여 모수를 추정하고자 하는데 이에 필요한 조건부 사후분포는 다음과 같다.

S^*의 조건부 사후분포를 구해보면 $y_i = 0$일 때

$$\pi(s_i^*|y_i = 0, \text{else}) \propto I(s_i^* \le 0)\phi(s_i^* - \mathbf{x_i}'\boldsymbol{\beta}) + I(s_i^* > 0)\phi(s_i^* - \mathbf{x_i}'\boldsymbol{\beta})\phi(y_i^* - \mathbf{z_i}'\boldsymbol{\gamma})I(y_i^* \le 0)$$
$$\propto w(\boldsymbol{\beta}, \boldsymbol{\gamma}) \cdot \phi(s_i^* - \mathbf{x_i}'\boldsymbol{\beta})I(s_i^* \le 0) + [1 - w(\boldsymbol{\beta}, \boldsymbol{\gamma})] \cdot \phi(s_i^* - \mathbf{x_i}'\boldsymbol{\beta})I(s_i^* > 0)$$

이며, 이 때

$$w(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1 - \Phi(\mathbf{x_i}'\boldsymbol{\beta})}{1 - \Phi(\mathbf{x_i}'\boldsymbol{\beta}) + \Phi(\mathbf{x_i}'\boldsymbol{\beta})\Phi(-\mathbf{z_i}'\boldsymbol{\gamma})}$$

이다. 만약 $y_i \neq 0$일 때는 $\pi(s_i^*|y_i \neq 0, \text{else}) \sim N(\mathbf{x_i}'\boldsymbol{\beta}, 1)I(s_i^* > 0)$이다.

$\boldsymbol{\beta}, \boldsymbol{\gamma}, Y^*$의 조건부 사후분포는 $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_n})$, $\mathbf{s}^* = (s_1^*, \ldots, s_n^*)'$라 할 때 다음과 같이 주어진다.

$$\pi(\boldsymbol{\beta}|\text{else}) \propto \prod_{i=1}^{n} \phi(s_i^* - \mathbf{x_i}'\boldsymbol{\beta}) \cdot \pi(\boldsymbol{\beta})$$
$$\sim N(\mu_\beta^\pi, \Sigma_\beta^\pi), \quad \Sigma_\beta^\pi = \left(\mathbf{X}\mathbf{X}' + \Sigma_\beta^{-1}\right)^{-1},\ \mu_\beta^\pi = \Sigma_\beta^\pi\left(\mathbf{X}\mathbf{s}^* + \Sigma_\beta^{-1}\boldsymbol{\beta}_0\right)$$
$$\pi(y_i^*|y_i = j, s_i^* > 0, \text{else}) \sim N(\mathbf{z_i}'\boldsymbol{\gamma}, 1)I(\delta_{j-1} \le y_i^* < \delta_j), \quad i = 1, \ldots, n,$$
$$\pi(\boldsymbol{\gamma}|s_i^* > 0, \text{else}) \propto \prod_{i=1, s_i^*>0}^{n} \phi(y_i^* - \mathbf{z_i}'\boldsymbol{\gamma}) \cdot \pi(\boldsymbol{\gamma})$$
$$\sim N(\mu_\gamma^\pi,\ \ \Sigma_\gamma^\pi), \Sigma_\gamma^\pi = \left(\mathbf{z}_{(1)}'\mathbf{z}_{(1)} + \Sigma_\gamma^{-1}\right)^{-1}, \mu_\gamma^\pi = \Sigma_\gamma^\pi\left(\mathbf{z}_{(1)}'\mathbf{y}_{(1)}^* + \Sigma_\gamma^{-1}\right).$$

표 **4.1.** 음주의 소비빈도 요약

반응변수(Y)		N	%
없다	0	1079	54.0%
월2∼3회	1	386	19.3%
주1∼2회	2	320	16.0%
주3∼4회	3	146	7.3%
거의매일	4	69	3.4%

$\mu_\gamma^\pi, \Sigma_\gamma^\pi$ 식에서 $\mathbf{y}_{(1)}^*, \mathbf{z}_{(1)}$은 $s_i^* > 0$를 만족하는 i에 대응하는 $y_i^*, \mathbf{z}_i$로 구성된다.

끝으로, $\boldsymbol{\delta}$의 조건부 사후분포는

$$\begin{aligned}\pi(\delta_j|s_i^*>0,\text{else}) &\propto \pi(\delta_j)\cdot\prod_{i=1,s_i^*>0}^{n}[I(\delta_{j-1}\le y_i^*<\delta_j)]^{I(y_i=j)}\cdot[I(\delta_j\le y_i^*<\delta_{j+1})]^{I(y_i=j+1)}\\ &\propto \pi(\delta_j)\cdot I\left(\max_{i:y_j=j,s_i=1}y_i^*<\delta_j<\min_{i:y_i=j+1,s_i=1}y_i^*\right)\\ &\sim N(\delta_{0j},\sigma_j^2)I\left(\max\left(\delta_{j-1},\max_{i:y_i=j,s_i=1}y_i^*\right)\le\delta_j<\min\left(\delta_{j+1},\min_{i:y_i=j+1,s_i=1}y_i^*\right)\right)\end{aligned}$$

이다.

4. 한국인의 음주실태자료 분석

통계청에서 주관한 '2008년도 보건, 가족 부문 조사'는 2008년 6월 24일부터 7월 2일까지 실시되었으며 전국 약 20,000 표본 가구의 만 15세 이상 가구원 42473명을 대상으로 하였다. 본 연구에서는 42473명 중에서 랜덤으로 2000명을 뽑아 분석에 사용하였다.

설문지에서 "지난 1년(2007.6.24 ∼ 2008.6.23)동안 술을 한 잔 이상 마신 적이 있습니까? 있다면, 얼마나 자주 드셨습니까?" 라는 질문에 대한 응답을 반응변수 Y로 두었다. 2000명의 자료에 대한 반응변수를 요약하면 표 4.1과 그림 4.1과 같이 정리할 수 있다. 표 4.1에서 보이는 것처럼 응답자의 절반이상인 54%가 비음주자이며 자료의 절반이상이 0값을 차지하고 있음을 알 수 있다.

본 분석은 이와 같은 비음주자를 절대적으로 음주를 하지 않는 절대적 비음주자(genuine non-drinker)와 음주 소비에 대한 잠재성을 가졌으나 조사대상 기간 동안에는 음주를 하지 않은 잠재적 음주자(zero consumption potential drinker), 두 그룹으로 나누어 본다. 잠재적 음주자는 미래에 어떠한 다른 변화가 발생했을 때 언제든 음주를 할 가능성이 있다고 보는 그룹이다.

설명변수로는 총 11개의 변수를 사용하였으며 표 4.2와 같이 정리하였다. 조사에서 직업을 고용원을 둔 사업주, 임금 근로자, 자영업자로 분류하였고, 이 세 가지 모두에 해당되지 않는 경우는 직업이 없는 무직으로 간주하였다. 또한 본 분석에서는 Becker와 Murphy (1988)의 Rational addiction theory에 따라 음주와 같이 중독성이 있는 재화의 경우, 시간이 경과함에 따라 얻게 되는 장기적인 효과가 기대되므로 분석에 사용되는 연속형 변수인 나이와 교육정도의 두 가지 변수에 대해서 선형적인 효과 외에도 비선형적인 효과를 고려해 2차 항 까지도 사용하였다.

본 분석에서는 1단계 프로빗 모형에서의 설명변수 $\mathbf{x}$와 2단계 OP 모형에서의 설명변수 $\mathbf{z}$를 동일하게 상수항과 표 4.2의 11개 변수를 사용하였다.

모수에 대한 사전분포의 설정은 다음과 같다. 먼저 관측된 반응 변수 값이 0보다 큰 경우에 $S = 1$로 놓

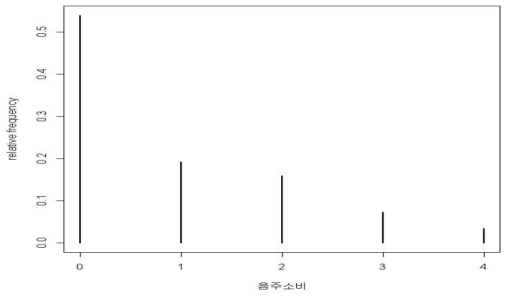

그림 **4.1.** Y의 상대도수

표 **4.2.** 분석에 사용된 설명변수의 요약

변수	값	변수	값
나이	연속형 변수	교육년수	연속형 변수
지역	광역시 = 1, 그 외 = 0	사업주	사업주 = 1, 그외 = 0
소득	로그소득, 연속형 변수	근로자	근로자 = 1, 그외 = 0
		자영업자	자영업자 = 1, 그외 = 0
성별	남자 = 1, 여자 = 0	나이^2	나이∗나이
혼인상태	배우자있음 = 1, 그외 = 0	교육^2	교육년수∗교육년수

고, 반응변수 값이 0인 경우는 그 중 임의로 절반을 $S = 1$로 놓았다. 참고로, S는 실제적, 잠재적 음주자를 포함한 음주자에 대해서는 1값을, 절대적 비음주자에 대해서는 0값을 갖는 이항변수이므로 비음주자의 절반을 잠재적 음주자로 초기에 설정한 것이다. 다음, S를 반응변수로 프로빗 모형을 적합 시켜 얻은 $\boldsymbol{\beta}$의 추정치를 $\boldsymbol{\beta}$의 사전평균으로, 추정된 $\boldsymbol{\beta}$의 분산에 10을 곱하여 $\boldsymbol{\beta}$의 사전분산으로 선택하였다. 또한 $S = 1$인 경우의 자료만을 이용하여 OP 모형을 적합 시킨 후 얻은 $\boldsymbol{\gamma}$의 추정치를 $\boldsymbol{\gamma}$의 사전평균으로, 추정된 $\boldsymbol{\gamma}$의 분산에 10을 곱하여 $\boldsymbol{\gamma}$의 사전분산으로 선택하였다. 끝으로, 위의 OP 모형에서 얻은 경계값 $\boldsymbol{\delta}$의 추정치를 $\boldsymbol{\delta}$의 사전평균으로, $\boldsymbol{\delta}$의 사전분산은 10을 사용하였다. 단, $\delta_0 < \delta_1 < \cdots < \delta_{J-1} < \infty$ 제한조건을 만족하도록 하였다.

베이지안 추론을 위해 통계 소프트웨어 R을 이용하여 ZIOP 모형의 깁스표본알고리즘을 구현하였다. 깁스표본 알고리즘에서 $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$의 초기치로는 각자의 사전평균을 사용하였다. 깁스표본기법에서 5,000번의 burn-in time 이후 발생된 70,000 개의 표본을 취하여 추론에 사용하였으며 모수 $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ 표본의 경로그림을 그려 수렴성을 확인하였다.

표 4.3은 ZIOP 모형의 모수 $\boldsymbol{\beta}$와 $\boldsymbol{\gamma}$의 추정값과 표준오차를 보여주고 있다. 5% 유의수준으로 유의하게 0과 다른 추정치는 진하게 표시하였다. 절대적 비음주자와 음주자를 구분하는 프로빗 모형의 계수 추정치 $\boldsymbol{\beta}$를 보면, 나이, 성별, 근로자, 자영업자, 나이^2, 교육^{2}이 절대적 비음주자와 음주자를 구분하는데 유의한 영향을 미치는 변수임을 알 수 있다. 나이가 증가할수록, 그리고 남자가 여자보다 음주자가 될 가능성이 크다는 것을 의미한다. 또한 세 종류의 직업 중 근로자와 자영업자가 무직자 보다 음주자가 될 가능성이 크다.

음주횟수에 대한 OP 모형의 계수 $\boldsymbol{\gamma}$의 추정 값을 보면, 성별과 사업자의 설명변수에서 유의한 결과를 보이고 있다. 성별에 대한 계수추정치는 1.2545로 이는 잠재적 음주자와 실제적 음주자를 포함한 음주자의 소비 형태를 보면, 남자가 여자보다 음주횟수가 상당히 많다는 것을 의미한다. 또한 사업자가 무직자보다 음주소비를 많이 하는 경향이 있다는 것을 의미한다.

표 **4.3.** ZIOP 모형의 $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$의 사후평균과 표준오차

Variable	$\hat{\beta}$(SE)	$\hat{\gamma}$(SE)
상수항	−2.8559 (0.5585)	0.8838 (0.5407)
나이	**0.1033** (0.0213)	−0.0248 (0.0186)
지역	0.0113 (0.1129)	0.0397 (0.0765)
소득	0.0236 (0.0269)	−0.0206 (0.0167)
성별(남 = 1)	**0.3224** (0.1284)	**1.2545** (0.0913)
혼인상태(배우자있음 = 1)	−0.2018 (0.1626)	0.1754 (0.1025)
교육년수	−0.0143 (0.0626)	−0.0161 (0.0570)
사업자	0.4509 (0.3815)	**0.4387** (0.1924)
근로자	**0.5117** (0.1400)	0.1810 (0.0961)
자영업자	**0.6382** (0.2121)	0.2009 (0.1296)
나이^2	**−0.0010** (0.0002)	0.0001 (0.0002)
교육^2	**0.0064** (0.0029)	−0.0019 (0.0022)

표 **4.4.** $\boldsymbol{\delta}$와 Zero Inflation 값의 사후추정치

$\boldsymbol{\delta}$ (delta)	
delta 1	0.8833 (0.0483)
delta 2	1.7556 (0.0620)
delta 3	2.4667 (0.0788)
Zero Inflation	**0.304** (0.0228)

표 **4.5.** 음주소비의 상대도수

범주	관측값	ZIOP 예측값	OP 예측값
$Y = 0$	0.5395	0.5347	0.5514
$Y = 1$	0.1930	0.1951	0.1926
$Y = 2$	0.1600	0.1619	0.1565
$Y = 3$	0.0730	0.0724	0.0666
$Y = 4$	0.0345	0.0356	0.0326

표 4.4는 경계값 $\boldsymbol{\delta}$의 추정치와 Zero-inflation 값의 추정치를 보여준다. $\boldsymbol{\delta}$의 추정 치를 보면 $\delta_0 < \delta_1 < \delta_2 < \delta_3$의 순서가 매우 유의함을 볼 수 있다. 깁스표본으로부터 Zero-inflation 값을 추정한 결과, 표 4.4에서 보인 바와 같이 추정값은 30.4%로, 전체 비음주자의 약 30%가 절대적인 비음주자임을 의미하며 따라서 본 자료는 영 과잉 특성이 매우 뚜렷함을 보여주고 있다.

영 과잉 특성을 반영한 ZIOP 모형과 그렇지 않은 OP 모형에서 음주 소비의 각 범주에 대한 예측값을 비교하면 표 4.5와 같다. ZIOP 모형은 모든 범주에서 예측 상대도수가 관측된 상대도수와 거의 일치하지만 OP 모형의 경우 0 범주에서 상당한 차이를 보여주고 있다.

ZIOP 모형과 OP 모형을 비교하기 위하여 Spiegelhalter 등 (2002)이 제안한 DIC를 계산해 보았다. $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta})$일 때 식 (2.1)에 주어진 우도함수로부터 $D(\boldsymbol{\theta}) = -2\log(p(y|\boldsymbol{\theta}))$를 정의하고 이의 사후기대치 $\bar{D} = E_{\boldsymbol{\theta}|y}[D(\boldsymbol{\theta})]$와 $\boldsymbol{\theta}$의 사후기대치를 대입한 $D(\hat{\boldsymbol{\theta}})$를 계산하면 $\bar{D}$는 모형의 적합정도, $pD = \bar{D} - D(\hat{\boldsymbol{\theta}})$는 모형의 복잡성에 대한 측도로 볼 수 있다. DIC는 이 둘을 결합한 DIC $= \bar{D} + pD$으로 모형 선택의 기준으로 삼을 수 있다. 표 4.6에 주어진 두 모형에 대한 DIC 값을 보면 ZIOP 모형이 자료에 더 적절한 모형임을 알 수 있다.

다음으로, ZIOP 모형에서 관심 있는 확률에 대한 각 변수의 한계효과를 정리하면 표 4.7과 같다. 설명

표 **4.6.** ZIOP 모형과 OP 모형의 적합도통계량

Model	Dbar	Dhat	pD	DIC
OP	4411.9	4396.9	15.0	4426.9
ZIOP	4330.0	4308.1	21.9	4373.8

표 **4.7.** ZIOP 모형의 한계효과

	절대적 비음주자	잠재적 음주자	
	$P(S=0)$	$P(S=1,=0)$	$P(Y=0)$
나이	**−0.0330 (0.0066)**	**0.0173 (0.0063)**	**−0.0157 (0.0047)**
지역	−0.0040 (0.0361)	−0.0095 (0.0282)	−0.0135 (0.0230)
수입	−0.0075 (0.0086)	0.0079 (0.0061)	0.0004 (0.0058)
성별	**−0.1034 (0.0428)**	**−0.2804 (0.0398)**	**−0.3839 (0.0227)**
혼인상태	0.0619 (0.0491)	−0.0690 (0.0402)	−0.0072 (0.0298)
교육년수	0.0045 (0.0199)	0.0027 (0.0199)	0.0073 (0.0128)
사업자	−0.1075 (0.0789)	−0.0782 (0.0541)	**−0.1857 (0.0599)**
근로자	**−0.1554 (0.0407)**	0.0007 (0.0344)	**−0.1547 (0.0265)**
자영업자	**−0.1632 (0.0437)**	−0.0058 (0.0439)	**−0.1690 (0.0352)**

	$P(Y=1)$	$P(Y=2)$	$P(Y=3)$	$P(Y=4)$
나이	**0.0112 (0.0023)**	**0.0078 (0.0017)**	**0.0025 (0.0006)**	**0.0007 (0.0002)**
지역	**0.2536 (0.0127)**	**0.1794 (0.0105)**	**0.0572 (0.0061)**	**0.0174 (0.0030)**
수입	0.0026 (0.0029)	0.0018 (0.0020)	0.0006 (0.0007)	0.0002 (0.0002)
성별	**0.2213 (0.0131)**	**0.2611 (0.0128)**	**0.1344 (0.0111)**	**0.0680 (0.0085)**
혼인상태	**0.2448 (0.0132)**	**0.1795 (0.0106)**	**0.0592 (0.0063)**	**0.0187 (0.0032)**
교육년수	−0.0016 (0.0068)	−0.0011 (0.0047)	−0.0003 (0.0015)	−0.0001 (0.0004)
사업자	**0.2609 (0.0382)**	**0.2566 (0.0263)**	**0.1137 (0.0239)**	**0.0499 (0.0180)**
근로자	**0.2815 (0.0153)**	**0.2155 (0.0132)**	**0.0741 (0.0086)**	**0.0244 (0.0046)**
자영업자	**0.2948 (0.0210)**	**0.2385 (0.0191)**	**0.0869 (0.0145)**	**0.0306 (0.0083)**

변수 x_k의 한계효과를 계산할 때 나머지 설명변수들의 값은 평균값으로 고정하였다.

우리의 관심은 절대적 비음주자와 잠재적 음주자의 확률에 대한 각 변수의 한계효과이다. 앞서 언급한 바와 같이 특정 설명변수의 $P(Y=0)$에 대한 한계효과는 그 변수의 절대적 비음주자와 잠재적 음주자에 대한 한계효과를 더한 것이다.

표 4.7의 한계효과를 살펴보면 다음과 같다. $P(Y=0)$에 대한 나이의 한계효과가 유의하게 0보다 작으므로 나이가 많아질수록 비음주자가 될 가능성이 낮다는 의미이다. 그런데 절대적 비음주자 $P(S=0)$와 잠재적 음주자 $P(S=1,\tilde{Y}=0)$에 대한 한계효과를 보면 두 한계효과의 부호가 반대인 것으로 나타나며 나이가 많아질수록 절대적 비음주자가 될 가능성을 낮아지고 잠재적 음주자가 될 가능성은 커진다는 것을 의미한다. 따라서 나이가 많아질수록 비음주자가 될 가능성이 낮아지는 것은 절대적 비음주자가 될 가능성이 낮아지기 때문이며, 잠재성은 가지고 있지만 조사대상 기간에만 음주를 하진 않는 잠재적 음주자가 될 가능성은 오히려 나이가 많아질수록 높아진다. $P(Y=j), j=1,2,3,4$에 대한 나이의 한계효과를 보면 모두 유의한 양의 값을 가지므로 나이가 많아질수록 음주 횟수가 많아짐을 의미한다.

지역 설명변수를 보면, 비음주자 확률에 대한 한계효과는 유의하지 않지만 $P(Y=j), j=1,2,3,4$에 대한 한계효과는 모두 유의하게 0보다 크므로 광역시 주민들이 상대적으로 음주 횟수가 많음을 의미한다.

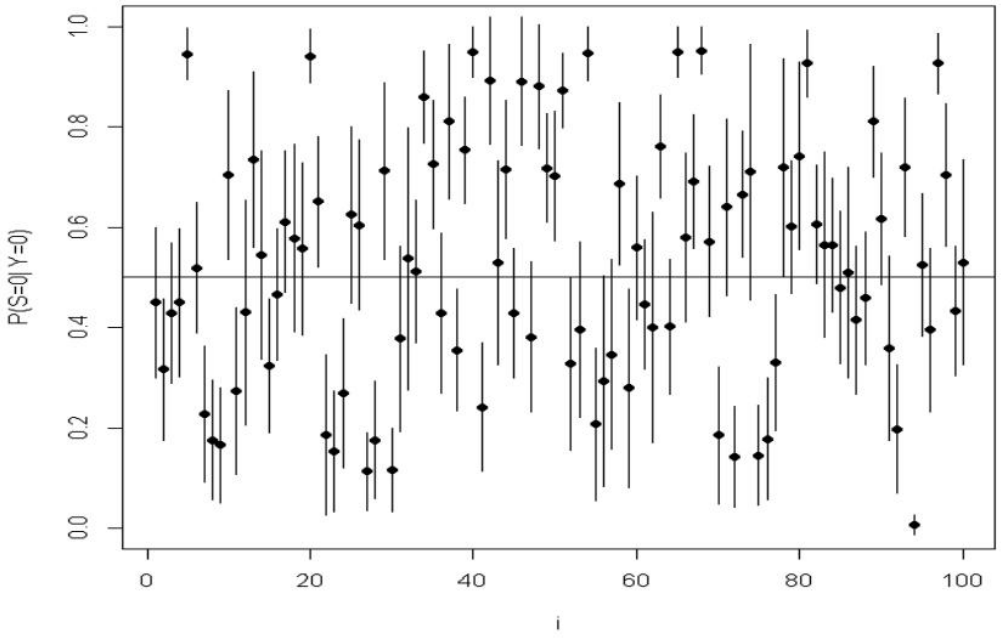

그림 4.2. 비음주자가 절대적 비음주자일 확률

성별의 경우 남자가 여자보다 절대적 비음주자 그리고 잠재적 음주자가 될 가능성이 모두 유의하게 낮고 따라서 비음주자가 될 가능성이 유의하게 낮다. 또한 $P(Y = j)$, $j = 1, 2, 3, 4$에 대한 성별의 한계효과가 모두 유의하게 0보다 크므로 남자가 여자보다 유의하게 음주횟수가 많음을 의미한다.

사업자의 경우 절대적 비음주자나 잠재적 음주자가 될 한계효과는 유의하지 않으나 이 둘을 합한 비음주자 한계효과는 유의하게 0 보다 작으므로 사업자는 무직에 비하여 비음주자가 될 가능성이 낮고 또한 $P(Y = j)$, $j = 1, 2, 3, 4$에 대한 한계효과를 볼 때 사업자가 무직에 비하여 음주횟수가 많음을 의미한다.

근로자, 자영업자의 경우 매우 흥미로운 현상을 발견하게 된다. 근로자나 자영업자의 비음주자 확률 $P(Y = 0)$에 대한 한계효과를 보면 유의하게 0보다 작으므로 근로자나 자영업자는 무직에 비하여 비음주자가 될 가능성이 낮다. 그러나 비음주자를 절대적 비음주자와 잠재적 음주자로 분류하여 보면, 비음주자 한계효과가 유의하게 0보다 작게 나오는 것은 절대적 비음주자 한계효과가 유의하게 0보다 작기 때문이고 잠재적 음주자 한계효과는 0과 유의하게 다르지 않다. 따라서 근로자와 자영업자가 잠재적 음주자가 될 가능성은 무직자와 유의하게 다르지 않으며 단지 절대적 비음주자가 될 가능성이 유의하게 낮기 때문에 비음주자가 될 가능성이 낮게 나오는 것이다.

각 직업의 $P(Y = j)$, $j = 1, 2, 3, 4$에 대한 한계효과를 보면, 예상대로 모든 직업에서 무직자보다 음주소비 횟수가 많음을 보여준다.

수입과 교육년수는 모든 한계효과에서 유의하지 않다. 수입과 교육년수가 한 사람의 사회경제적 지위를 대체적으로 잘 반영하는 변수들임을 고려하면, 우리나라 국민들의 음주 여부와 음주소비 형태가 사회경제적 지위에 따라 다르지 않음을 의미한다.

그림 4.2는 비음주자 $(Y_i = 0)$ 중 처음 100명에 대하여 절대적 비음주자일 확률 $P(S_i = 0|Y_i = 0)$과

이 확률에 대한 95% 사후구간을 그림으로 표시한 것이다. 사후구간이 0.5 보다 큰 비음주자들은 절대적 비음주자로, 사후구간이 0.5 보다 작은 비음주자들은 잠재적 음주자로 판정할 수 있을 것이다. 이렇듯 각 비음주자에 대하여 절대적 비음주자인지 잠재적 음주자인지 판정할 수 있다면, 모든 비음주자에 대하여 하나의 정책이나 태도 등을 유지하는 것이 아니라 각 비음주자에 맞는 맞춤형 정책이나 태도 등을 적용하는 것이 도움이 될 것이다.

5. 결론 및 요약

본 논문에서는 영 과잉 자료의 특성을 나타내는 음주자료에 대하여 ZIOP 모형을 설정하고 MCMC 기법을 사용한 베이지안 분석을 적용하였다. 본 연구에서 적용한 ZIOP 모형은 두 단계의 일반선형 모형을 설정하는데, 먼저 프로빗 모형을 사용하여 비음주자를 절대적 비음주자와 잠재적 비음주자로 분류하고 두 번째 단계에서는 잠재적 음주자와 실제적 음주자를 합한 전체 음주자에 대하여 OP 모형을 사용하여 분석하였다.

베이지안 모수추정결과를 통해 $\boldsymbol{\beta}$는 나이, 성별, 근로자, 자영업자, 나이^2, 교육^2 설명변수에서 유의하다는 결론을 얻었고 $\boldsymbol{\gamma}$는 성별과 사업자 설명변수에서 유의하다는 결론을 얻었다. 유의하게 추정된 $\boldsymbol{\beta}$값은 나이가 증가할수록, 남자가 여자 보다, 근로자와 자영업자가 무직자 보다 음주자가 될 가능성이 크다는 것을 보여주었고 추정된 $\boldsymbol{\gamma}$값은 남자가 여자 보다, 사업자가 무직자 보다 음주시 음주 횟수가 더 많다는 것을 보여주었다.

각 설명변수의 관심 있는 확률에 대한 한계효과를 살펴본 결과, 나이 변수는 비음주자 중에서 절대적 비음주자와 잠재적 음주자에 서로 상반되는 영향을 미침을 알 수 있었다. 또한 근로자와 자영업자의 경우 비음주자에 대한 음의 한계효과의 원인이 절대적 비음주자에 대한 음의 한계효과 때문이며 잠재적 음주자에 대한 한계효과는 유의하게 0과 다르지 않음을 알 수 있었다. 이상의 예를 통하여 비음주자를 절대적 비음주자와 잠재적 음주자로 구분하여 분석하는 것이 매우 유용함을 보여주었다.

음주 자료에서는 나타나지 않았지만, 절대적 비음주자와 잠재적 음주자에 대한 한계효과가 반대 부호를 가지며 크기가 비슷할 때 서로 상쇄효과로 인하여 비음주자의 한계효과가 유의하지 않게 나타날 가능성이 있는 점을 고려할 때 비음주자의 구분이 매우 중요함을 알 수 있다.

또한 마코브 체인 몬테칼로 로부터 얻은 사후 표본을 가지고 각 비음주자에 대하여 절대적 비음주자일 확률과 그 사후구간을 구할 수 있으므로, 비음주자들을 절대적 비음주자와 잠재적 음주자로 구분할 수 있으며 이 구분에 따라 적합한 음주 정책을 수립하고 시행하는데 매우 유용할 것이다.

Harris와 Zhao (2007)에서는 담배소비 자료에서 절대적 비흡연자와 (잠재적, 실제적) 흡연자를 구분하는 정규 잠재변수 S^*와 흡연자의 담배소비정도를 구분하는 정규 잠재변수 Y^*가 서로 독립인 모형(ZIOP)과 서로 관련성을 가정한 ZIOPC 모형을 둘 다 고려한 결과 두 모형 사이에 거의 차이가 없다는 결론을 얻었다. 즉, 각 개인의 흡연에 대한 참여여부와 참여시 소비정도는 서로 관련이 없다는 것이다. 우리는 음주 소비에서도 비슷한 양상을 띨 것으로 판단하여 본 논문에서는 ZIOPC 모형을 고려하지 않고 ZIOP 모형만을 고려하였다. ZIOPC 모형의 베이지안 분석은 본 논문에서 제시된 ZIOP 모형에 대한 깁스알고리즘을 변형하여 수행할 수 있으므로 이에 대한 연구는 후속 연구에서 살펴보기로 하겠다.

참고문헌

김승수, 정슬기 (2009). 대학생 음주실태와 문제음주 변화 추이, 2006년과 2008년 비교, <한국 알코올과학회지>, **10**, 75–88.

박재홍, 이민경, 장용언 (2010). 근로자의 직무스트레스가 우울에 미치는 영향: 음주량과 음주빈도의 조절효과, <한국 알콜과학회지>, **11**, 1–14.

서경현, 양승애 (2010). 중학생의 음주 지식 및 태도와 음주행위 요인 간의 관계, <한국 알코올과학회지>, **11**, 15–26.

윤명숙, 김성혜, 채완순 (2010). 노인의 음주 및 정신건강 특성이 자살생각에 미치는 영향, <한국 알코올과학회지>, **11**, 27–44.

천성수, 손애리 (2008). 한국인의 사인별 알코올기여도 산출, <한국 알코올과학회지>, **9**, 1–12.

Becker, G. S. and Murphy, K. M. (1988). A theory of rational addiction, *Journal of Political Economy*, **96**, 675–700.

Gurmu, S. and Dagne, G. A. (2009). *Bayesian Approach to Zero-Inflated Ordered Probit Models, with an Application*, Georgia State University, University of South Florida.

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.

Rodrigues, R. (2003). Bayesian analysis of zero-infalted distributions, *Communications in Statistics*, **32**, 281–289.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B*, **64**, 583–649.

Bayesian Analysis of Korean Alcohol Consumption Data Using a Zero-Inflated Ordered Probit Model

Man-Suk Oh[1] · Hyun Tak Oh[2] · Semi Park[3]

[1]Department of Statistics, Ewha Womans University
[2]Department of Business Administration, Chonbook National University
[3]Department of Statistics, Ewha Womans University

(Received January 13, 2012; Revised February 21, 2012; Accepted April 12, 2012)

Abstract

Excessive zeroes are often observed in ordinal categorical response variables. An ordinary ordered Probit model is not appropriate for zero-inflated data especially when there are many different sources of generating 0 observations. In this paper, we apply a two-stage zero-inflated ordered Probit (ZIOP) model which incorporate the zero-flated nature of data, propose a Bayesian analysis of a ZIOP model, and apply the method to alcohol consumption data collected by the National Bureau of Statistics, Korea. In the first stage of a ZIOP model, a Probit model is introduced to divide the non-drinkers into genuine non-drinkers who do not participate in drinking due to personal beliefs or permanent health problems and potential drinkers who did not drink at the time of the survey but have the potential to become drinkers. In the second stage, an ordered probit model is applied to drinkers that consists of zero-consumption potential drinkers and positive consumption drinkers. The analysis results show that about 30% of non-drinkers are genuine non-drinkers and hence the Korean alcohol consumption data has the feature of zero-inflated data. A study on the marginal effect of each explanatory variable shows that certain explanatory variables have effects on the genuine non-drinkers and potential drinkers in opposite directions, which may not be detected by an ordered Probit model.

Keywords: Zero-inflation, Markov chain Monte Carlo, Posterior distribution, Ordinal categorical data.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science, and Technology (No. 2011-0004589).
[1]Corresponding author: Professor, Department of Statistics, Ewha Womans University, Seoul 120-750, Korea. E-mail: msoh@ewha.ac.kr