The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.6.733
31(6), 733–749

A new sample selection model for overdispersed count data

Sung Eun Jo[a] · Jun Zhao[a] · Hyoung-Moon Kim[a,1]

[a]Department of Applied Statistics, Konkuk University

(Received August 31, 2018; Revised October 8, 2018; Accepted October 31, 2018)

Abstract

Sample selection arises as a result of the partial observability of the outcome of interest in a study. Heckman introduced a sample selection model to analyze such data and proposed a full maximum likelihood estimation method under the assumption of normality. Recently sample selection models for binomial and Poisson response variables have been proposed. Based on the theory of symmetry-modulated distribution, we extend these to a model for overdispersed count data. This type of data with no sample selection is often modeled using negative binomial distribution. Hence we propose a sample selection model for overdispersed count data using the negative binomial distribution. A real data application is employed. Simulation studies reveal that our estimation method based on profile log-likelihood is stable.

Keywords: sample selection bias, Heckman's sample selection model, overdispersed data, negative binomial regression, Poisson regression

1. 서론

표본선택은 어떤 연구에서 관심이 있는 종속변수를 일부분만 관측할 수 있어 일어난다. 즉 관심이 있는 종속변수(Y)는 다른 관측되지 않은 확률변수(U)가 어떤 실수의 부분집합에 속할때만($0 < P(U \in C) < 1, C \in \mathbb{R}$) 관측이 된다. Y와 U가 독립이라면 관측된 Y만을 사용하여 모형의 모수를 추정해도 아무런 문제가 발생되지 않는다. 그러나 두 확률변수의 상관관계가 존재하면 관측된 Y만을 사용한 통계적 추론은 편의된 결과를 야기한다. 이 문제는 표본선택(sample selection) 또는 비임의결측자료(data missing not at random; MNAR) (Rubin, 1976)라고 알려져 있다.

이러한 표본선택모형은 관심이 있는 종속변수가 연속형인 경우 주로 연구가 많이 진행되었다 (Heckman, 1979; Vella, 1998; Wooldridge, 2010; Greene, 2012). 이산형 자료로의 확장은 현재까지 포아송 자료와 이항자료에 대한 모형들이 존재한다 (Boyes 등, 1989; Terza, 1998; Greene, 2012). 하지만 이때 평균의 함수를 고려시 하나의 랜덤요인이 추가로 나타나 로그우도함수의 계산시 적분을 매 단계마다 해야 하는 단점이 존재한다. 또한 이 모형에 대한 계산 방법도 아주 단순한 형태로 제시된 알고리즘만 존재하고 있다. 따라서 이런 문제점을 해결할 수 있고 여러 다양한 이산형자료에 대해 적용할 수 있는 새로운 모형에 대한 필요성이 대두된다.

This paper was supported by Konkuk University in 2016.
[1]Corresponding author: Department of Applied Statistics, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea. E-mail: hmkim@konkuk.ac.kr

최근 Azzalini 등 (2018)에서 이항자료와 포아송 자료의 경우 분포조정(distribution modulation)을 이용하여 새로운 표본선택모형을 제시하였다. 하지만 이는 과대산포를 가지는 가산자료 분석에는 적절하지 않다. 따라서 본 연구에서는 기존에 제시되었던 포아송 분포를 기초로 한 표본선택 모형 방법에 산포모수를 추가하여 음이항 표본선택 모형으로 확장 발전시켰다. 표본선택이 없는 경우에도 과대산포 자료의 경우 음이항 회귀분석이 포아송 회귀분석에 비해 더 유연하다는 사실이 알려져있다 (Agresti, 2013). 따라서 과대산포와 표본선택이 동시에 일어나는 경우 음이항 표본선택모형이 포아송 표본선택모형에 비해 더 유연해질것으로 기대된다.

본 연구의 2절에서는 연속형 자료에 관한 전통적인 헤크만 표본선택모형을, 3절에서는 기존의 이산자료에 관한 표본선택모형들에 대해 소개하였다. 4절에서는 포아송 표본선택모형에 산포모수를 추가하여 과대산포를 가지는 자료 분석에 적절한 음이항 표본선택모형을 제시하였고, 이에 대한 최대우도추정방법을 유도하였다. 5절에서는 실제 자료를 이용하여 포아송 표본선택모형과 음이항 표본선택모형으로 각각 분석한 후에 두 결과를 비교하였다. 6절에서는 모의실험을 통해 음이항 표본선택모형에서의 제외제약에 대한 중요성을 알아보았다. 제외제약이란 표본선택여부를 결정하는 회귀모형에서 사용되는 공변량 중 적어도 하나 이상의 공변량이 주요 관심사에 대한 회귀모형에서 사용되는 공변량에서 제외되는 것을 의미한다. 마지막으로 7절에서 결론을 유도하였다.

2. 연속형 자료에 관한 표본선택모형

Heckman (1976, 1979)의 여성 근로자들의 임금 결정요인에 대한 연구에서 근로자의 임금과 나이, 교육정도 등의 결정요인간의 선형회귀분석 모형이 소개되었다. 이 연구에서 표본선택모형이 사용되었다. 왜냐하면 임금이 어떤 최소한의 수준보다 낮은 경우에는 여성 근로자들이 노동시장에 참여하지 않기 때문이다. 여기서 최소한의 임금수준을 희망임금(reservation wage)이라 하고, 이는 고정되어 있는 값이 아니라 각각의 근로자의 특성에 따라 다르게 나타난다. 이로 인해 임금수준이 희망임금보다 낮다면 노동시장에 참여하지 않기 때문에 임금변수의 값은 없고, 결정요인들의 값만 관측되는 경우가 발생하게 된다. 이러한 경우 관측된 임금 자료만을 사용하여 최소제곱법을 이용한 회귀모형의 계수에서는 선택편향(selection bias)이 나타난다. 왜냐하면 관측되지 않은 임금변수의 값이 직관적으로 임금범위의 가장 낮은 부분에 놓이기 때문이다. 이런 임금변수의 값이 관측되지 않는 경우의 편향은 종속변수인 임금변수와 선택을 조정한 변수들간의 의존에 의해서 나타난다. 이러한 표본선택문제는 경제학, 생물통계학, 사회학 등 여러 분야의 응용문제에서 등장한다.

표본선택모형은 표본선택여부를 결정하는 회귀모형

$$U_i = w_i^T \gamma + u_i$$

과 주요 관심사에 대한 회귀모형

$$Y_i = x_i^T \beta + \epsilon_i$$

으로 구성된다. x_i는 p차원 벡터, w_i는 q차원 벡터이다. (Y_i, U_i)의 i번째 자료($i = 1, 2, \ldots, n$)는 다음과 같은 분포를 따른다.

$$\begin{pmatrix} Y_i \\ U_i \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_i \\ \tau_i \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right), \tag{2.1}$$

여기에서 $\mu_i = x_i^T \beta$이고 $\tau_i = w_i^T \gamma$이다. 문제는 Y_i의 관측이 U_i가 0보다 크거나 같은 경우에만 가능하다는 것이다. 주어진 정보로는 U_i의 크기는 구할수 없고, 부호만 관측할 수 있다. 따라서 분산은 식별

불가능(unidentifiable)하여 1로 일반성을 잃지 않고 가정하였다. 이진변수 D_i는 아래 같이 정의된다.

$$D_i = \begin{cases} 1, & U_i \geq 0\text{인 경우,} \\ 0, & \text{그 외의 경우.} \end{cases} \tag{2.2}$$

즉, Y_i의 관측은 D_i가 1인 경우에만 가능하다. Y_i가 관측되는 경우 Y_i의 조건부 기대값은 다음과 같이 구할 수 있다 (Greene, 2012).

$$\begin{aligned} E[Y_i|D_i = 1] &= E\left[Y_i|U_i \geq 0\right] \\ &= E\left[Y_i|u_i \geq -w_i^T\gamma\right] \\ &= E\left[x_i^T\beta + \epsilon_i|u_i > -w_i^T\gamma\right] \\ &= x_i^T\beta + \rho\sigma\lambda\left(w_i^T\gamma\right). \end{aligned}$$

여기에서 $\lambda(w_i^T\gamma) = \phi(w_i^T\gamma)/\Phi(w_i^T\gamma)$이며 이는 역밀스비(inverse Mills ratio)를 나타내며 ϕ는 표준정규분포의 확률밀도함수, Φ는 표준정규분포의 누적분포함수를 나타낸다. 최소제곱법을 이용하여 관측된 자료를 분석하면 $\rho \neq 0$인 경우에 변수 λ에 대한 계수를 제외하므로 추정된 $\hat{\beta}$는 불일치추정량이다. 변수 λ를 포함한 표본선택모형의 모수는 최대우도추정 (Wooldridge, 2010)과 헤크만의 2단계추정 (Greene, 2012) 등을 이용한 모수적 방법이 있다. 이러한 모수적인 방법 외에도 준모수적인 방법과 비모수적인 방법이 제시되었다 (Vella, 1998).

식 (2.1)에서 가정된 이변량정규분포하의 로그 우도함수는 아래와 같다.

$$\log L = \sum_{d_i=1} \log[\Phi(\tau_i) f(Y_i|D_i = 1)] + \sum_{d_i=0} \log[1 - \Phi(\tau_i)].$$

여기에서

$$\begin{aligned} P(D_i = 1) &= \Phi\left(w_i^T\gamma\right) = \Phi(\tau_i), \\ f(y_i|D_i = 1) = f(y_i|U_i \geq 0) &= \frac{1}{\Phi(\tau_i)\sigma}\phi(z)\Phi\left(\frac{\tau_i + \rho z}{\sqrt{1-\rho^2}}\right), \quad z \in \mathbb{R}, \end{aligned} \tag{2.3}$$

그리고 $z = (y_i - \mu_i)/\sigma$이다. $\rho = 0$인 경우 식 (2.3)은 $N(\mu_i, \sigma^2)$의 확률밀도함수가 된다.

3. 이산형 자료의 표본선택모형

이 장에서는 기존에 제시된 몇 가지 이산형 자료에 관한 표본선택모형들을 설명한다.

3.1. 이변량 프로빗 표본선택모형

주요 관심사에 대한 회귀모형식에서 종속변수가 이산형으로 주어질 수 있다. Boyes 등 (1989)에서 분석한 모형은 다음과 같다.

$$Y_i = \begin{cases} 1, & i\text{의 채무불이행이 발생한 경우,} \\ 0, & \text{그 외의 경우,} \end{cases}$$

$$U_i = \begin{cases} 1, & i\text{의 대출이 승인된 경우,} \\ 0, & \text{그 외의 경우.} \end{cases}$$

Greene (1992)에서 Y_i는 신용카드 대출 채무 불이행 여부이고, U_i는 신용카드 신청 승인 여부이다. 여기서 Y_i는 $U_i = 1$인 경우에만 관측된다. 오차의 분포로 다음의 이변량 정규분포를 가정한다.

$$\begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{3.1}$$

왜냐하면 Y_i에 관한 값의 크기를 구할 수 없어 식 (2.1)에서 $\sigma = 1$이 되었기 때문이다. 다음의 세 가지 경우가 표본에서 관측되며 이에 관한 비조건부 확률은 다음과 같다.

$$\begin{aligned} U_i = 0 &: P(U_i = 0|x_i, w_i) = \Phi\left(-w_i^T\gamma\right), \\ Y_i = 0,\, U_i = 1 &: P(Y_i = 0, U_i = 1|x_i, w_i) = \Phi_2\left(-x_i^T\beta, w_i^T\gamma, -\rho\right), \\ Y_i = 1,\, U_i = 1 &: P(Y_i = 1, U_i = 1|x_i, w_i) = \Phi_2\left(x_i^T\beta, w_i^T\gamma, \rho\right), \end{aligned}$$

여기에서 $\Phi_2(a, b, \rho)$는 가정된 오차의 분포 (3.1)의 누적분포함수를 나타낸다. 표본에서 가능한 위 세 가지 확률들을 이용하여 우도함수를 구하고 이를 최대화하는 모수들을 추정할 수 있다 (Greene, 2012). 이때의 우도함수는 아래와 같다.

$$\log L = \sum_{Y_i=1, U_i=1} \ln \Phi_2\left(x_i^T\beta, w_i^T\gamma, \rho\right) + \sum_{Y_i=0, U_i=1} \ln \Phi_2\left(-x_i^T\beta, w_i^T\gamma, -\rho\right) + \sum_{U_i=0} \ln \Phi\left(-w_i^T\gamma\right).$$

3.2. 비선형 표본선택모형

헤크만 모형의 논리를 기반으로 하여 이산형 자료에 대한 비선형 표본선택모형이 제시되었다 (Greene, 2012; Terza, 1998). 먼저 프로빗 선택 방정식은 아래와 같다.

$$\begin{aligned} U_i &= w_i^T\gamma + u_i, \\ D_i &= \begin{cases} 1, & U_i \geq 0\text{인 경우}, \\ 0, & \text{그 외의 경우}. \end{cases} \end{aligned}$$

비선형 표본선택 모형은 아래와 같이 주어진다.

$$\begin{aligned} &\mu_i|\epsilon_i = x_i^T\beta + \sigma\epsilon_i, \quad \epsilon_i \sim N(0, 1), \\ &y_i|x_i, \quad \epsilon_i \sim \text{ 확률질량함수 } g(y_i|x_i, \epsilon_i) = f\left(y_i|x_i^T\beta + \sigma\epsilon_i\right), \\ &y_i, x_i\text{는 } D_i = 1\text{일 때만 관측됨}. \end{aligned}$$

오차항의 분포는 식 (3.1)에 주어진 분포와 동일한 이변량 정규분포를 가정한다. 포아송 회귀모형을 예로 들면 조건부 평균 함수는 $E(y_i|x_i) = \lambda_i = \exp(x_i^T\beta + \sigma\epsilon_i) = \exp(\mu_i)$이다. 로그우도함수를 구하면

$$\log L = \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} \left[(1 - D_i) + D_i f\left(y_i|x_i^T\beta + \sigma\epsilon_i\right)\right] \Phi\left[(2D_i - 1)\left(w_i^T\alpha + \tau\epsilon_i\right)\right] \phi(\epsilon_i) d\epsilon_i$$

이며, 여기에서 $\alpha = \gamma/\sqrt{1-\rho^2}$, $\tau = \rho/\sqrt{1-\rho^2}$이다. 로그우도함수의 최대화는 구적법(quadrature)이나 시뮬레이션을 이용하였다. 하지만 이때 평균의 함수를 고려시 하나의 랜덤요인(ϵ_i)이 추가로 나타나 로그우도함수의 계산시 적분을 매 단계마다 해야 하는 단점이 존재한다. 다른 모형과의 차이는 어떻게 $f(y_i|x_i^T\beta + \sigma\epsilon_i)$을 설정하냐의 문제이다 (Terza, 1998).

3.3. 분포조정을 이용한 새로운 표본선택모형

이번 절에서는 Azzalini 등 (2018)에서 소개된 분포조정을 이용한 표본선택 방법을 소개한다. 위에서도 이항자료나 포아송 자료에 대한 표본선택모형들이 소개되었으나, 예를 들면, 3.2절에서 제시한 방법은 Heckman 표본선택모형의 아이디어를 비선형 모형으로 확장하였고 여러 이산형 모형에 적용할 수 있다는 장점이 존재하지만 로그우도함수의 계산시 적분을 매 단계마다 해야 하는 단점 또한 존재한다. 따라서 유연한 모수적 방법을 사용하여 새로운 표본선택모형을 제시할 필요성이 대두되었다.

식 (2.3)에서 주어진 분포는 확장된 왜곡정규분포이다 (Azzalini와 Capitanio, 2014). 분포조정에 의한 일반적인 분포를 유도하기에 앞서 확장된 왜곡정규분포를 도출하는 무상관 확률변수들을 먼저 소개한다. 이는 추후 이론 전개과정을 쉽게 하는 기초가 된다. 새로운 확률변수들 Z와 T를 아래와 같이 정의하자.

$$Z = \frac{Y-\mu}{\sigma},$$
$$T = \alpha Z - \left(1+\alpha^2\right)^{\frac{1}{2}}(U-\tau), \quad \alpha = \rho\left(1-\rho^2\right)^{\frac{1}{2}}.$$

편리를 위해서 i번째 자료를 의미하는 아래 첨자는 생략하고 나타내겠다. (Y,U)의 분포가정에서 상관관계가 존재하지만 (Y,T)는 무상관이다. 또한 $\mathrm{corr}(T,Y) = \mathrm{corr}(T,Z) = 0$인 특징을 갖고 있다. 식 (2.2)의 $D=1$인 사건은

$$T \le \alpha Z + \tau\left(1+\alpha^2\right)^{\frac{1}{2}}$$

와 동등하고, $D=1$인 사건을 조건부로 관측되는 y에 대한 밀도함수는

$$f(y|D=1) = \frac{1}{\Phi(\tau)}\left[\frac{1}{\sigma}\phi(z)\Phi\left(\tau\left(1+\alpha^2\right)^{\frac{1}{2}} + \alpha z\right)\right], \quad z = \frac{y-\mu}{\sigma},\ z \in \mathbb{R} \tag{3.2}$$

이다. 식 (3.2)는 식 (2.3)과 정확히 일치한다. 차이점은 식 (2.3)은 상관된 이변량 정규분포의 가정에서 유도되었고 식 (3.2)는 독립인 이변량 정규분포의 가정에서 유도되었다. $\alpha = 0$인 경우 식 (3.2)는 $N(\mu,\sigma^2)$의 확률밀도함수가 된다.

만약 기저 밀도함수 f가 평균이 0이고 분산이 σ^2인 정규분포를 따르고, $G_0 = \Phi$이고 $h(y) = \tau(1+\alpha^2)^{1/2} + \alpha(y-\mu)/\sigma$이면 식 (3.2)은 아래와 같이 일반화된 형태로 다시 쓰여질 수 있다. 즉,

$$f(y|D=1) = \frac{1}{\pi} f(y) G_0\{h(y)\} \tag{3.3}$$

이며 이 때 정규화 상수 π는

$$\pi = \int_{\mathbb{R}} f(y) G_0\{h(y)\} dy \tag{3.4}$$

로 나타낼 수 있다. 식 (3.3)의 기저 밀도함수 f는 변화인자인 $G_0\{h(y)\}$에 의해서 조절된다. G_0는 단변량 분포함수이고, $h(y)$는 실수함수이다. 이산형의 경우에 f는 확률질량함수를 나타내고 식 (3.4)의 적분은 합으로 바꾸어야한다. 기저 밀도함수 f가 대칭인 분포일때 식 (3.3)에서 나타나는 분포를 대칭-조정된 분포(symmetry-moduated distribution)라 칭한다.

Y는 밀도함수 f를 갖는 확률변수이고, T는 분포함수 G_0를 갖는 Y와 독립인 변수를 나타낸다고 하자. 또한 f로부터 표본 추출된 y값이 $T \le h(y)$인 사건을 조건부로 관측된다고 가정하자. 이때 f로부터 생

성된 y는

$$\begin{aligned} P(D=1|Y=y) &= P(T \le h(y)|Y=y) \\ &= G_0\{h(y)\} \\ &= G(y) \end{aligned}$$

의 조건부확률로 관측된다. $G(y)$의 표현은 $Y=y$의 조건부분포함수라는 의미를 나타낸 표현이다. 무조건부로 f로부터 값이 관측될 확률은

$$\begin{aligned} \pi = P(D=1) &= E_Y(P(T \le h(y)|Y=y)) \\ &= E_Y(G_0\{h(y)\}) \end{aligned}$$

이다. 조건부확률인 $G(y)$는 대부분의 상황에서 y의 연속함수로의 가정이 합리적이다. 연속성은 G_0와 h에서도 비슷하게 가정되며 다양한 G_0와 h를 사용할 수 있다. 로그 우도함수는 다음과 같이 구해진다.

$$\begin{aligned} \log L &= \sum_{d_i=1} \log[P(D_i=1) \times f(y_i|D_i=1)] + \sum_{d_i=0} \log P(D_i=0) \\ &= \sum_{d_i=1} \log[f(y_i) \times P(D_i=1|y_i)] + \sum_{d_i=0} \log P(D_i=0) \\ &= \sum_{d_i=1} \log[f(y_i)G(y_i)] + \sum_{d_i=0} \log(1-\pi_i) \end{aligned} \tag{3.5}$$

$\log L$은 f와 G의 모수에 의존한다. 식 (3.5)의 최적화는 최대우도추정을 통해서 얻을 수 있다.

3.3.1. 이항 반응변수 종속변수 Y_i가 1과 0의 값을 갖는 이항확률변수라 하자. 성공의 확률은 공변량 x_i의 함수로 다음과 같이 나타낸다.

$$\mu_i = E(Y_i) = P(Y_i=1) = P_0\left(x_i^T\beta\right),$$

여기서 P_0는 실수에서의 분포함수이다. 보통 로지스틱분포와 정규분포의 분포함수를 사용한다. 즉,

$$\begin{aligned} P_0(u) &= \frac{\exp(u)}{1+\exp(u)}, \\ P_0(u) &= \Phi(u) \end{aligned}$$

이다. 이때 μ_i에 대한 로짓 모형과 프로빗 모형이 유도된다. 로지스틱분포, 정규분포 함수 외에 다른 선택도 가능하다. 이항확률변수 Y_i의 확률질량함수는

$$f(y) = (1-\mu_i)^{1-y}\mu_i^y, \quad y=0,1$$

이다. 식 (3.4)의 정규화상수 π_i의 계산은 적분을 합으로 바꾸고, 다음과 같이 구할 수 있다.

$$\pi_i = (1-\mu_i)G(0) + \mu_i G(1) \tag{3.6}$$

또한 $T \sim N(0,1)$이고 $h(y) = \tau_i + \alpha\mu_i^{-1}y$인 경우는 다음과 같이 선형으로 나타낼 수 있다. 기타 다양한 $h(y)$의 선택은 Azzalini 등 (2018)을 참조하면 된다.

$$G(y) = G_0\{h(y)\} = \Phi\left(\tau_i + \alpha\mu_i^{-1}y\right), \tag{3.7}$$

여기서 α는 y에 의존하여 조절되는 모수이다. 이 식은 $\mathbb{E}\{\mu_i^{-1}y\} = 1$인 의미로 표준화한 형식이다. 여기서 $\eta_i = \alpha/\mu_i$로 나타내겠다. 또 다른 방법으로는 $T \sim \text{Exp}(1)$인 경우이다. 이 경우에는 변수가 양수여야 되는 조건이 있으므로 위의 $h(y)$에 지수를 취한 $\exp(\tau_i + \eta_i y)$를 이용하여 나타내면

$$G(y) = 1 - \exp\{-\exp(\tau_i + \alpha y)\} \quad \text{혹은} \quad G(y) = 1 - \exp\{-\exp(\tau_i + \eta_i y)\} \tag{3.8}$$

이 된다. 여러 다양한 $G(y)$의 선택이 가능하며 Azzalini 등 (2018)을 참조하기 바란다. 따라서 이항 변수에서 표본선택이 일어날때 사용가능한 새로이 유도된 분포는 아래와 같다.

$$f(y_i|D_i = 1) = \frac{1}{\pi_i}(1-\mu_i)^{1-y_i}\mu_i^{y_i}\, G(y_i), \quad y_i = 0, 1. \tag{3.9}$$

여기에서 π_i는 식 (3.6) 그리고 $G(y_i)$는 식들 (3.7)과 (3.8)에 나타나 있다. $\alpha = 0$인 경우 식 (3.9)에 나타난 확률질량함수는 $G(y_i)$가 y_i에 의존하지 아니하여 베르누이 분포의 확률질량함수가 된다.

3.3.2. 포아송 반응변수 종속변수 Y_i가 가산자료일 경우 포아송 분포를 가정할 수 있다. Y_i의 평균을 μ_i라고 하면

$$Y_i \sim \text{Poi}(\mu_i)$$

이다. 포아송 분포의 평균을 공변량으로 나타내면 그 식은

$$\mu_i = \log\left(x_i^T\beta\right) \tag{3.10}$$

이며 다른 형태의 식도 가능하다. Y_i에 대한 확률질량함수는

$$f(y) = \frac{e^{-\mu_i}\mu_i^y}{y!}, \quad y = 0, 1, 2, \ldots$$

이다.

선택방법은 이항확률변수에서 소개된 방법을 사용한다. 정규화 상수 π는 무한합으로 나타난다. 이는 적당한 큰 값 K까지 절단한 합으로 적절히 근사될 수 있다.

$$\pi_i \approx \sum_{k=0}^{K} \frac{e^{-\mu_i}\mu_i^k}{k!} G(k). \tag{3.11}$$

K의 선택은 Y_i의 최대값보다 큰 값으로 정하는 것이 적절하다. 따라서 포아송 반응변수에서 표본선택이 일어날때 사용가능한 새로이 유도된 분포는 아래와 같다.

$$f(y_i|D_i = 1) = \frac{1}{\pi_i}\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\, G(y_i), \quad y_i = 0, 1, 2, \ldots, \tag{3.12}$$

여기서 π_i는 식 (3.11)에서 $K = \infty$ 그리고 $G(y_i)$는 식 (3.7)과 (3.8)에 나타나 있다. $\alpha = 0$인 경우 식 (3.12)에 나타난 확률질량함수는 $G(y_i)$가 y_i에 의존하지 아니하여 포아송 분포의 확률질량함수가 된다.

4. 과대산포 이산자료에 대한 표본선택모형

4.1. 음이항 표본선택모형

포아송 분포에 기초한 표본선택모형은 기저 확률질량함수 f를 포아송 분포로 가정한다. 따라서 포아송

분포의 단점(즉, 평균과 분산이 동일함)이 제안된 표본선택모형의 단점으로 제기될 수 있다. 이러한 문제를 해결하기 위해 포아송 분포를 포함하는 더 큰 확률질량함수 f(예를 들면 음이항분포)를 기저함수로 하여 표본선택모형을 구한다면 이는 자료에 흔히 나타나는 과대산포 문제 또한 해결할 수 있을 것이다.

포아송 분포 하에서 평균이 λ_i인 것은 상수이거나 계층 내에서 동질적인 것을 가정한다. 그러나 λ_i에 대해 특별한 분포를 가정하면 계층 내에서 이질성을 허락하게 된다. 예로 λ_i가 평균이 $E(\lambda_i) = \mu_i$이고, 분산이 $\mathrm{Var}(\lambda_i) = \mu_i^2\psi^{-1}$인 감마분포임을 가정하자. 또한 $Y_i|\lambda_i$는 조건부평균 $E(Y_i|\lambda_i) = \lambda_i$를 갖는 포아송 분포임을 가정한다. 이 때 Y_i의 주변분포는

$$\begin{aligned} P(Y_i = y_i) &= \int P(Y_i = y_i|\lambda_i) f(\lambda_i) d\lambda_i \\ &= \frac{\Gamma(y_i+\psi)}{\Gamma(y_i+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\mu_i}\right)^{\psi} \left(\frac{\mu_i}{\psi+\mu_i}\right)^{y_i}, \quad y_i = 0, 1, 2, \ldots \end{aligned}$$

의 확률을 갖는 음이항 분포임을 보일 수 있다 (Agresti, 2013). 이 분포의 평균은 $E(Y_i) = \mu_i$이고, 분산은 $\mathrm{Var}(Y_i) = \mu_i + \mu_i^2\psi^{-1}$이다. 여기서 ψ^{-1}은 산포모수이다. 만약 ψ^{-1}가 0이라면 $E(Y_i) = \mathrm{Var}(Y_i)$이고, 결과적으로 Y_i는 포아송 분포를 따른다. 만약 $\psi^{-1} > 0$이라면 $E(Y_i) < \mathrm{Var}(Y_i)$이므로 과대산포를 나타내는 분포가 된다.

평균을 공변량으로 나타낸 식은 (3.10)과 같다. 표본선택방법은 포아송 분포의 경우와 동일한 방법을 사용한다. 정규화 상수 π는 무한합으로 나타난다. 포아송 분포에서와 같은 방법으로 K를 선택하여 K까지 절단한 합으로 다음과 같이 정규화 상수를 근사시킬 수 있다.

$$\pi_i = \sum_{k=0}^{K} \frac{\Gamma(k+\psi)}{\Gamma(k+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\mu_i}\right)^{\psi} \left(\frac{\mu_i}{\psi+\mu_i}\right)^{k} G(k). \tag{4.1}$$

따라서 음이항 분포를 기저확률질량함수로 구한 새로운 분포는 아래와 같다.

$$f(y_i|D_i = 1) = \frac{1}{\pi_i} \frac{\Gamma(y_i+\psi)}{\Gamma(y_i+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\mu_i}\right)^{\psi} \left(\frac{\mu_i}{\psi+\mu_i}\right)^{y_i} G(y_i), \quad y_i = 0, 1, 2, \ldots, \tag{4.2}$$

여기에서 π_i는 식 (4.1)에서 K가 ∞인 경우이며 $G(y_i)$는 식들 (3.7)과 (3.8)에 나타나 있다. $\alpha = 0$인 경우 식 (4.2)에 나타난 확률질량함수는 $G(y_i)$가 y_i에 의존하지 아니하여 음이항 분포의 확률질량함수가 된다.

4.2. 음이항 표본선택모형에 대한 최대우도측정

로그우도 함수를 수치적으로 최대화시키는 값을 구하기 위하여 α에 대한 프로파일 로그우도 함수를 사용하였다.

$$\log L_p(\alpha) = \log L\left(\alpha, \hat{\theta}(\alpha)\right)$$

여기서 $\hat{\theta}(\alpha) = (\beta^T, \gamma^T, b)^T$, $b = \log(\psi)$이다. $\hat{\theta}(\alpha)$는 주어진 α에 대해 $\log L$값이 최대값을 갖는 θ를 선택한다. 최대우도추정량 $\hat{\alpha}$는 $\log L_p(\alpha)$이 최대값을 갖는 지점의 α값이다. 이때 $\hat{\theta} = \hat{\theta}(\hat{\alpha})$이라고 나타내자. ψ를 변환한 b를 사용하는 이유는 추후에 Newton-Raphson 방법을 이용할 때 비제약 최대화를 하기 위함이다. 변환을 이용하여도 b가 단조이고, 가역변환이기 때문에 로그우도함수가 최대값을 갖는 b값에 해당하는 ψ 또한 로그우도함수가 최대값을 갖는다.

θ의 초기치를 구하는 방법은 α의 초기치를 0으로 고정하고 Y와 D 각각의 분리된 회귀분석을 이용한다. β와 b는 종속변수 Y의 음이항 회귀분석을 통해 초기값을 구하고, γ는 식 (2.2)의 모형구조에 의하여 종속변수 D의 이항회귀모형을 통해 초기값을 구한다. 여기서 구해진 예측값인 β, γ와 b를 모든 최적화의 시작값으로 사용한다.

주어진 α에 대하여 수치적인 최적화를 통해 $\theta(\alpha)$를 구하게 된다. 이 때 너무 세밀한 간격의 α를 주면 계산 시간이 증가할 가능성이 있다. 아래 경사도의 명확한 수식이 있으면 최적화 알고리즘에서 효율성이 증가한다.

$$\frac{d}{d\theta}\log L(\alpha,\theta).$$

로그우도함수의 일차와 이차 도함수 계산의 일반적인 대수적 표현을 부록에 첨부하였다. 여기서 적절한 f, G_0와 h를 사용하면 된다. 부록에서 구한 수식들은 다른 f, G_0와 h에 대해서도 성립하는 일반적인 수식이다. 이들을 이용하여 Newton-Raphson 방법을 사용하여 모수들을 추정하였다.

일반적인 점근적 분포에 관한 이론(윌크스의 정리(Wilks' theorem))을 사용하면, α의 신뢰구간을 구할 수 있고, 이는

$$2[\log L_p(\hat{\alpha}) - \log L_p(\alpha)] \le \chi^2_{q,1} \tag{4.3}$$

이다. 여기서 $\chi^2_{q,1}$은 카이제곱 분포의 분위수를 뜻한다. 만약 식 (4.3)의 신뢰구간이 $\alpha = 0$을 포함한다면, 우도비검정에서 $H_0 : \alpha = 0$을 기각하지 못한다. $\alpha = 0$인 경우는 표본선택이 전혀 일어나지 않았다는 것을 나타낸다. 이때는 관측된 자료만을 가지고 분석하면 된다. 예를 들면, 음이항 표본선택모형의 경우 $\alpha = 0$이라면 관측된 자료만을 가지고 음이항 회귀분석을 적용하면 된다. 왜냐하면 $\alpha = 0$인 경우 식 (4.2)에 나타난 확률질량함수는 $G(y_i)$가 y_i에 의존하지 아니하여 음이항 분포의 확률질량함수가 되기 때문이다. Heckman 모형의 경우는 관측된 자료만을 사용하여 최소제곱추정법을 적용하면 되고 이때 추정된 회귀계수들은 일치추정량이고 불편추정량이 된다. 왜냐하면 $\rho = 0$인 경우 식 (2.3)은 $N(\mu_i, \sigma^2)$의 확률밀도함수가 되기 때문이다.

5. 자료분석

Riphahn 등 (2003)의 독일 건강 보험 제도의 사용자 선호도 및 사용에 관한 연구에서 사용된 자료를 Greene (2012)은 Example 19.13에서 다시 분석하였다. Greene (2012)에서는 종속변수 Y를 방문여부에 따라 이항반응변수로 변환하여 분석하였지만 여기에서는 종속변수 Y를 방문횟수 그대로 이산형 반응변수로 보고 분석하겠다. 본 분석에서는 Greene (2012)에서 사용된 동일한 공변량을 사용하고 표본 중 레알슐레(Realshule)를 졸업한 인원들을 대상으로 하여 포아송 표본선택모형과 음이항 표본선택모형을 이용하여 분석하였다. 표본선택여부를 결정하는 회귀모형에서 사용된 종속변수는 공공건강보험에 가입하였는지의 여부(insured in public health insurance = 1; otherwise = 0)이며 독립변수들은 나이(age), 교육수준(education = years of schooling) 그리고 성별(female = 1; male = 0)이다. 주관심사에 대한 회귀모형에서 사용된 종속변수는 각 개인이 주어진 년도에 병원에 방문한 횟수이며 독립변수는 나이(age), 소득(income), 미성년자존재여부(kids = children under age 16 in the household = 1; otherwise = 0), 교육수준(education = years of schooling), 그리고 결혼여부(married = 1; otherwise = 0)이다. 표본선택방법은 두가지 방법을 사용한다. 이를 위해 G 함수를 두 가지 유형으로 하여 (A)는 식 (3.7), (B)는 식 (3.8)을 나타낸다. 이에 부합하는 전체 표본의 수는 1,546개이며 이중 표본선택된 자료의 수는 1,347로서 약 87%의 자료가 선택되었다. 이 표본에서 특이한 점은 종속변수로 사용된 각 개

Table 5.1. Poisson model (A), $T \sim N(0,1)$, $h(y) = \tau + \eta y$

반응변수의 로그선형모형						
	one	age	income	kids	education	married
$\hat{\beta}$	1.5188	0.0162	−1.4118	−0.1653	−0.0375	−0.1069
std.err	0.2682	0.0013	0.1212	0.0378	0.0229	0.0382
선택모형						
	one	age	education	female		
$\hat{\gamma}$	1.9180	−0.0055	−0.0671	0.4466		
std.err	0.7105	0.0036	0.0603	0.0830		
최대화된 $\log L$와 $\hat{\alpha}$						
$\hat{\alpha}$	−0.032					
$\log L$	−5383.313					

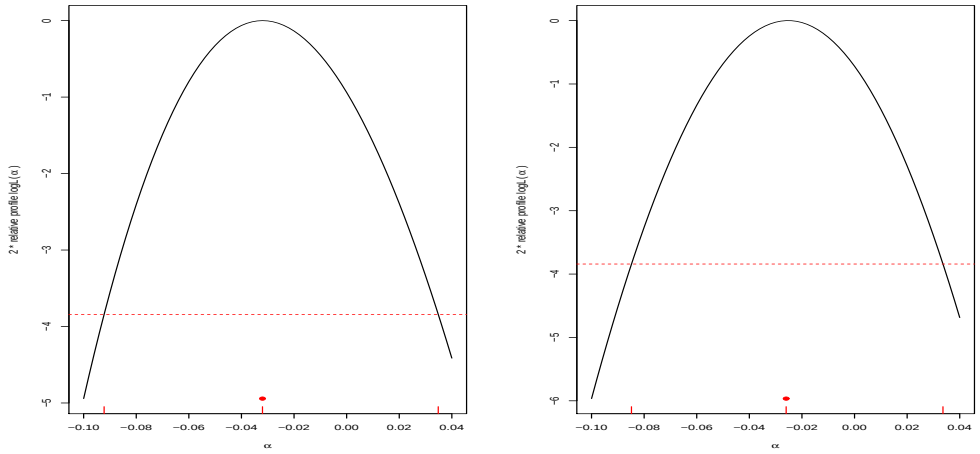
Table 5.2. Poisson model (B), $T \sim \text{Exp}(1)$, $h(y) = \exp(\tau + \eta y)$

반응변수의 로그선형모형						
	one	age	income	kids	education	married
$\hat{\beta}$	1.5197	0.0162	−1.4117	−0.1653	−0.0376	−0.1069
std.err	0.2321	0.0012	0.1056	0.0332	0.0199	0.0344
선택모형						
	one	age	education	female		
$\hat{\gamma}$	1.3463	−0.0050	−0.0522	0.3577		
std.err	0.5557	0.0030	0.0471	0.0664		
최대화된 $\log L$와 $\hat{\alpha}$						
$\hat{\alpha}$	−0.026					
$\log L$	−5383.149					

인이 주어진 년도에 병원에 방문한 횟수에 대한 표본평균과 표본분산은 각각 3.0067과 24.8997이다. 이는 모든 자료를 대상으로 한것이 아닌 표본선택된 자료만을 대상으로 한것이므로 정확히 과대산포를 가진다고 할수는 없으나 어느 정도 과대산포가 존재할것으로 추측가능하다.

5.1. 포아송 표본선택모형

이 자료를 먼저 Azzalini 등 (2018)에서 제시된 포아송 반응변수에 관한 표본선택모형을 이용하여 분석하였다. 결과는 Table 5.1과 Table 5.2에 요약하였고, 두 가지 유형에 대한 프로파일 로그우도함수의 그래프는 Figure 5.1에 나타내었다. 두 테이블에 나타난 모수와 그에 대한 표준오차 값을 살펴보면 두 가지 유형에서 비슷한 결과를 나타내었다. $\hat{\gamma}$에 대한 추정결과는 차이가 조금 있지만 관심 있는 변수에 대한 추정 값인 $\hat{\beta}$는 매우 유사한 결과가 나왔다. α의 경우에는 (A)는 $\hat{\alpha} = -0.032$, (B)는 $\hat{\alpha} = -0.026$ 값을 얻었다.이는 T의 분포와 $h(y)$를 다르게 하더라도 모수의 추정값과 그에 대한 표준오차값들은 많이 달라지지 않는다는 의미이다. 이외의 로버스트성에 관한 모의실험과 결과는 Azzalini 등 (2018)를 참조하면 된다. 식 (4.3)을 통하여 포아송 표본선택모형을 이용한 α의 신뢰구간을 구해보면 (A)의 경우 $(-0.0922, 0.0348)$, (B)는 $(-0.0848, 0.0336)$으로 구해진다. 두 신뢰구간에서 공통으로 0을 포함하므로 귀무가설 $H_0 : \alpha = 0$를 기각하지 못하는 것으로 판명되었다. 즉 표본선택의 효과가 없다는 것인데 이는 이 자료의 종속변수의 분산이 평균보다 매우 큰 이유 때문인 것으로 추측된다. 이 자료의 종속변수인 각 개인이 주어진 년도에 병원에 방문한 횟수의 표본평균과 표본분산은 각각 3.0067과 24.8997이다.



(A) $T \sim N(0,1),\ h(y) = \tau + \eta y$ (B) $T \sim \text{Exp}(1),\ h(y) = \exp(\tau + \eta y)$

Figure 5.1. Profile likelihoods for Poisson models.

Table 5.3. Negative binomial model (A), $T \sim N(0,1)$, $h(y) = \tau + \eta y$

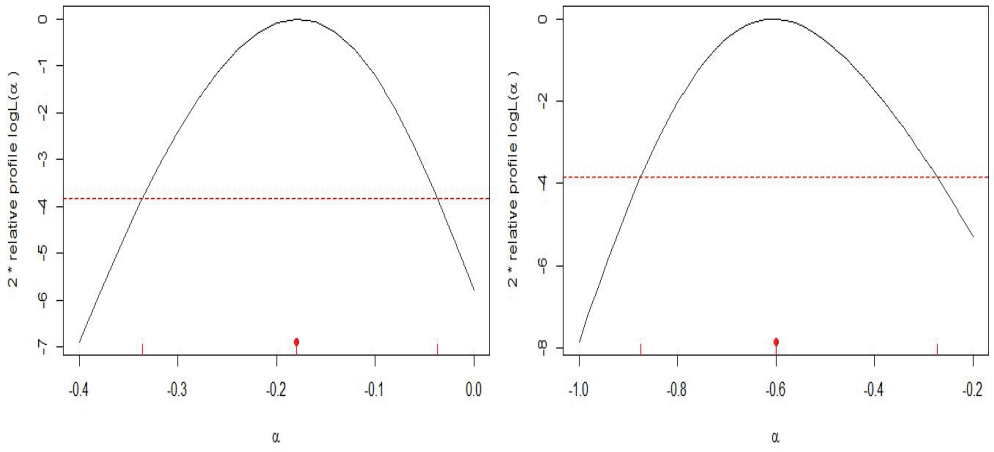
반응변수의 로그선형모형						
	one	age	income	kids	education	married
$\hat{\beta}$	2.0541	0.0154	−1.5260	−0.1410	−0.0675	−0.1082
std.err	0.7257	0.0036	0.2762	0.0895	0.0627	0.0944
선택모형						
	one	age	education	female		
$\hat{\gamma}$	2.0392	−0.0057	−0.0643	0.5315		
std.err	0.7516	0.0037	0.0638	0.0869		
척도모수						
$\hat{b} = \log(\hat{\psi})$			−0.6943			
std.err			0.0539			
최대화된 $\log L$와 $\hat{\alpha}$						
$\hat{\alpha}$			−0.18			
$\log L$			−3478.494			

5.2. 음이항 표본선택모형

동일한 자료를 이용하고 음이항 표본선택모형을 적용하여 분석한 결과를 Table 5.3과 Table 5.4에 요약하였다. 두 가지 유형에 대한 프로파일 로그우도 함수의 그래프는 Figure 5.2에 나타내었다. 포아송 표본선택모형의 경우에서와 마찬가지로 두 가지 유형에서 비슷한 결과를 얻었다. Table 5.3과 Table 5.4에 나타난 θ와 그에 대한 표준오차 값도 비슷한 값을 가진다. 식 (4.3)를 통하여 음이항 분포 모형의 α에 관한 신뢰구간들은 구해보면 (A)의 경우 $(-0.3361, -0.0368)$, (B)의 경우 $(-0.8744, -0.2727)$으로 구해진다. 두 경우 모두 0인 값을 포함하지 않으므로 귀무가설 $H_0 : \alpha = 0$을 기각할 수 있다. 즉 표본선택모형이 효과가 있다는 의미이다. 표본선택에 사용된 종속변수(공공건강보험에 가입하였는지의 여부)가 주요 관심사에 대한 회귀모형에 사용된 종속변수(각 개인이 주어진 년도에 병원에 방문한 횟수)

Table 5.4. Negative binomial model (B), $T \sim \text{Exp}(1)$, $h(y) = \exp(\tau + \eta y)$

반응변수의 로그선형모형						
	one	age	income	kids	education	married
$\hat{\beta}$	2.1490	0.0155	−1.4823	−0.1470	−0.0503	−0.1146
std.err	0.7169	0.0035	0.2575	0.0847	0.0620	0.0898
선택모형						
	one	age	education	female		
$\hat{\gamma}$	2.1415	−0.0042	−0.0741	0.8307		
std.err	1.0940	0.0050	0.0936	0.1257		
척도모수						
$\hat{b} = \log(\hat{\psi})$			−0.7982			
std.err			0.0523			
최대화된 $\log L$와 $\hat{\alpha}$						
$\hat{\alpha}$			−0.6			
$\log L$			−3475.101			



(a) $T \sim N(0, 1)$, $h(y) = \tau + \eta y$ (b) $T \sim \text{Exp}(1)$, $h(y) = \exp(\tau + \eta y)$

Figure 5.2. Profile likelihoods for negative binomial models.

에 영향을 미친다는 의미이다. 연속형 자료에 관한 표본선택모형과 비슷하게 두 변수들은 서로 상관관계가 있다는 의미이다. 식으로 살펴보면 $\alpha = 0$인 경우 식 (4.2)에 나타난 음이항 표본선택모형에 사용되는 확률질량함수는 식들 (3.7)과 (3.8)에 나타난 $G(y_i)$가 y_i에 의존하지 아니하여 음이항 분포의 확률질량함수로 바뀐다.

만약 음이항 표본선택모형의 두 유형 중 한 가지를 선택해야 한다면 최대화된 로그우도를 이용해서 Akaike information criteria (AIC) 값을 구하여 비교한다. 로그우도가 (A)는 −3478.49, (B)는 −3475.10로 최대화되므로 AIC 값은 각각 6980.99, 6974.20이다. 따라서 둘 중 AIC가 작은 (B)를 선택하면 된다. 또한 동일한 자료의 포아송 표본선택모형의 분석 결과와 비교하자면 최대화된 로그우도는 유형 (A)의 경우에 −5383.31이며 유형(B)의 경우에 −5383.15이므로 이를 이용하여 AIC를 구하면 각각 10788.63와 10788.30이다. 따라서 AIC 기준으로 음이항 표본선택모형을 이용하는 것이 적절하며 두 유형중 하나를 선택해야 한다면 유형 (B), 즉 표본의 선택에서 지수분포를 사용한 모형이 최적이다.

Table 6.1. Simulation study: negative binomial response with exclusion restriction ($\alpha = -0.5$)

Type	n	$\beta_0 = 0.5$	$\beta_0 = 1.5$	$\gamma_0 = 1$	$\gamma_1 = 1$	$\gamma_2 = 1.5$	$b = 0.5$	$\alpha = -0.5$
A	500	0.4932	1.5164	1.0855	1.0979	1.5760	0.5081	−0.5347
		0.0691	0.0655	0.1001	0.1146	0.1429	0.1412	0.9510
	1000	0.4956	1.5147	1.0433	1.0646	1.5346	0.5029	−0.5095
		0.0487	0.0460	0.0686	0.0788	0.0978	0.0996	0.9550
B	500	0.4621	1.5086	0.9439	0.9766	1.5002	0.5426	−0.4473
		0.0664	0.0627	0.1018	0.1135	0.1521	0.1333	0.9320
	1000	0.4636	1.5075	0.9075	0.9557	1.4640	0.5331	−0.4305
		0.0469	0.0442	0.0698	0.0783	0.1047	0.0939	0.8790

6. 모의실험

헤크만 모형과 관련된 모형의 중요한 쟁점은 x_i와 w_i의 공변량 집합에서의 제외제약 조건이 있냐의 문제이다. 제외제약이란 w_i에서의 공변량 중 적어도 하나의 공변량이 x_i에서 제외되는 것을 의미한다. 역으로 설명하면 w_i는 x_i에 포함되지 않는 하나 이상의 공변량을 가지고 있다는 것을 의미한다. 이 제외제약이 없는 경우에 헤크만 모형의 식별성은 가정된 분포의 함수적 형태에 의존한다. 특별히 이변량 정규분포를 가정한 헤크만 모형의 경우 회귀계수의 식별성(identifiability)은 역밀스비의 비선형성으로부터 얻어질수있다. 문제는 역밀스비는 지지(support)의 많은 부분에서 선형성이 일어난다. 따라서 이 문제를 경감하기 위해 계량경제학 문헌에서는 주관심사의 회귀식에 포함되지 않은 적어도 하나의 추가적인 변수가 표본선택의 회귀식에 포함된다는 제외제약 조건을 부과한다. 따라서 새로이 제시된 음이항 표본선택 모형에서도 제외제약 조건의 유무에 따라 최대우도추정의 결과가 어떻게 변하는지 알아보기위해 모의실험을 하였다.

모의실험에는 음이항 반응변수를 사용하고 두가지 유형의 선택방법

$$\text{(A) } T \sim N(0,1),\ h(y) = \tau + \eta y; \qquad \text{(B) } T \sim \text{Exp}(1),\ h(y) = \exp(\tau + \eta y)$$

을 이용하여 실험하였다. 제외제약을 가지는 경우

$$\mu_i = x_i^T \beta = 0.5 + 1.5 x_i, \qquad \tau_i = w_i^T \gamma = 1 + x_i + 1.5 w_i$$

으로 가정하였다. 여기서 x_i와 w_i는 독립적으로 $N(0,1)$로부터 추출된다. 제외제약이 없는 경우에는 w_i를 제거한

$$\mu_i = x_i^T \beta = 0.5 + 1.5 x_i, \qquad \tau_i = w_i^T \gamma = 1 + x_i$$

으로 가정한다. 우선 제외제약조건의 유무에 따라 두가지 유형의 표본선택방법을 사용하였고 각각의 경우에 α가 $-0.5, -0.1$; 표본의 크기가 $n = 500, 1000$인 경우로 총 16가지의 실험환경을 만들었다. 각각의 실험환경마다 $N = 1000$회씩 반복하여 실험하였다. 실험결과는 Table 6.1부터 Table 6.4에 나타내었다. 각각의 상황에서 해당하는 Table들의 첫 줄은 N개의 모수들의 추정치들의 평균을 나타내었고, 두 번째 줄은 α를 제외하고 N개의 표준오차들의 평균을 나타내었다. 여기에서 α의 경우는 우도함수를 기초로 구한 신뢰구간이 실제 모수의 값을 포함하는 비율을 나타낸다.

표본의 개수 n이 500인 경우 보다 1,000인 경우에 표준오차 값이 더 작게 나타났다. α의 값에 관계없이 전체적으로 제외제약이 없는 경우보다 제약이 있는 경우에 더 좋은 추정치를 얻었다. 이를 통해 제외제약의 중요성을 알 수 있다.

Table 6.2. Simulation study: negative binomial response without exclusion restriction ($\alpha = -0.5$)

Type	n	$\beta_0 = 0.5$	$\beta_0 = 1.5$	$\gamma_0 = 1$	$\gamma_1 = 1$	$b = 0.5$	$\alpha = -0.5$
A	500	0.5028	1.5105	1.0990	1.0957	0.4934	−0.5586
		0.0672	0.0647	0.0833	0.0977	0.1358	0.9700
	1000	0.4985	1.5145	1.0471	1.0592	0.4952	−0.5196
		0.0474	0.0456	0.0568	0.0672	0.0960	0.9690
B	500	0.4349	1.5407	0.9267	0.9793	0.5476	−0.4269
		0.0647	0.0631	0.0806	0.1006	0.1286	0.9190
	1000	0.4214	1.5486	0.8576	0.9331	0.5427	−0.3750
		0.0459	0.0448	0.0539	0.0681	0.0907	0.8360

Table 6.3. Simulation study: negative binomial response with exclusion restriction ($\alpha = -0.1$)

Type	n	$\beta_0 = 0.5$	$\beta_0 = 1.5$	$\gamma_0 = 1$	$\gamma_1 = 1$	$\gamma_2 = 1.5$	$b = 0.5$	$\alpha = -0.1$
A	500	0.5047	1.5079	1.0849	1.0884	1.5598	0.4998	−0.1269
		0.0710	0.0677	0.1060	0.1191	0.1436	0.1306	0.9570
	1000	0.5044	1.5092	1.0498	1.0625	1.5345	0.4956	−0.1075
		0.0503	0.0479	0.0732	0.0824	0.0992	0.0920	0.9500
B	500	0.4932	1.4991	1.0379	1.0320	1.5573	0.5201	−0.1006
		0.0670	0.0644	0.1136	0.1259	0.1641	0.1255	0.9610
	1000	0.4961	1.4968	1.0189	1.0245	1.5331	0.5104	−0.0982
		0.0474	0.0454	0.0783	0.0878	0.1134	0.0884	0.9580

Table 6.4. Simulation study: negative binomial response without exclusion restriction ($\alpha = -0.1$)

Type	n	$\beta_0 = 0.5$	$\beta_0 = 1.5$	$\gamma_0 = 1$	$\gamma_1 = 1$	$b = 0.5$	$\alpha = -0.1$
A	500	0.5279	1.4842	1.1040	1.0736	0.4883	−0.1488
		0.0660	0.0645	0.0866	0.1026	0.1236	0.9360
	1000	0.5175	1.4951	1.0526	1.0439	0.4920	−0.1250
		0.0471	0.0459	0.0594	0.0707	0.0870	0.9430
B	500	0.5122	1.4905	1.0917	1.0657	0.5110	−0.1433
		0.0623	0.0621	0.0931	0.1201	0.1197	0.9440
	1000	0.5059	1.4953	1.0409	1.0333	0.5036	−0.1209
		0.0443	0.0440	0.0626	0.0820	0.0843	0.9590

7. 결론

Azzalini 등 (2018)에서 소개된 포아송 표본선택모형에 산포모수를 추가하여 음이항 표본선택모형을 새로이 제시하였다. 가산반응변수를 갖는 자료를 포아송 표본선택모형과 음이항 표본선택모형으로 각각 분석하였고, 두 분석의 AIC 값을 기준으로 과대산포를 갖는 자료의 경우에는 음이항 표본선택을 이용하는 것이 더 적절하다는 결론을 얻었다. 표본선택이 일어났는지에 대한 가설검정은 우도비 검정을 이용하여 귀무가설 $H_0 : \alpha = 0$을 이용하여 검정할수 있으며 또한 α에 관한 신뢰구간을 구할 수 있다. Heckman 모형의 경우 제외제약 조건의 유무에 따라 모수의 추정이 영향을 받으므로 새로이 제시된 음이항 표본선택모형에 대해서도 제외제약 조건의 유무와 다른 몇가지 상황을 추가하여 모의실험을 시행하였다. 시행 결과 제외제약조건의 유무에 관계없이 어느 정도 적절한 추정치를 얻었으며 제외제약이 있는 경우가 없는 경우보다 더 좋은 추정치를 주었다. 이 결과는 Heckman 모형의 모의실험결과와 유사한 것이다.

부록: 스코어 함수와 헤시안 행렬

$$f(y_i;\mu_i,\psi)=\frac{\Gamma(y_i+\psi)}{\Gamma(\psi)\Gamma(y_i+1)}\left(\frac{\psi}{\mu_i+\psi}\right)^{\psi}\left(\frac{\mu_i}{\mu_i+\psi}\right)^{y_i},\quad y_i=0,1,2,\ldots$$

$$b=\log(\psi),\ \psi=e^b$$

$$f(y_i;\mu_i,b)=\frac{\Gamma(y_i+e^b)}{\Gamma(e^b)\Gamma(y_i+1)}\left(\frac{e^b}{\mu_i+e^b}\right)^{e^b}\left(\frac{\mu_i}{\mu_i+e^b}\right)^{y_i},\quad y_i=0,1,2,\ldots$$

$$\pi_i=\sum_{y_i=0}^{\infty}f(y_i)G(y_i)$$

$$\begin{aligned}\log L&=\sum_i l_i=\sum_{d_i=1}\log\left[f(y_i)G_0\{h(y_i)\}\right]+\sum_{d_i=0}\log(1-\pi_i)\\&=\sum_{d_i=1}\left[\log f(y_i)+\log G_0\{h(y_i)\}\right]+\sum_{d_i=0}\log(1-\pi_i)\end{aligned}$$

$$s(\beta_j)=\frac{\partial\log L}{\partial\beta_j}=\sum_{i=1}^{N}\left[d_i\frac{\partial\log f(y_i)}{\partial\mu_i}+d_i\frac{g_0\{h(y_i)\}}{G_0\{h(y_i)\}}\frac{\partial h(y_i)}{\partial\mu_i}-(1-d_i)\frac{\partial\pi_i/\partial\mu_i}{1-\pi_i}\right]x_{ij}\mu_i$$

$$s(\gamma_h)=\frac{\partial\log L}{\partial\gamma_h}=\sum_{i=1}^{N}\left[d_i\frac{g_0\{h(y_i)\}}{G_0\{h(y_i)\}}\frac{\partial h(y_i)}{\partial\tau_i}-(1-d_i)\frac{\partial\pi_i/\partial\tau_i}{1-\pi_i}\right]w_{ih}$$

$$s(b)=\frac{\partial\log L}{\partial b}=\sum_{i=1}^{N}\left[d_i\frac{\partial\log f(y_i)}{\partial b}-(1-d_i)\frac{\partial\pi_i/\partial b}{1-\pi_i}\right]$$

$$\begin{aligned}H(\beta_j,\beta_h)=\frac{\partial^2\log L}{\partial\beta_j\partial\beta_h}&=\sum_{i=1}^{N}\frac{\partial}{\partial\beta_h}\left(\frac{\partial l_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_j}\right)\\&=\sum_{i=1}^{N}\left[\left\{d_i\left\{\frac{\partial^2\log f(y_i)}{\partial\mu_i^2}+\frac{g'_0\{h(y_i)\}G_0\{h(y_i)\}-g_0\{h(y_i)\}^2}{G_0\{h(y_i)\}^2}\left(\frac{\partial h(y_i)}{\partial\mu_i}\right)^2\right.\right.\right.\\&\quad\left.+\frac{g_0\{h(y_i)\}}{G_0\{h(y_i)\}}\frac{\partial^2h(y_i)}{\partial\mu_i^2}\right\}-(1-d_i)\frac{(\partial^2\pi_i/\partial\mu_i^2)(1-\pi_i)+(\partial\pi_i/\partial\mu_i)^2}{(1-\pi_i)^2}\left.\right\}\mu_i\\&\quad+\left\{d_i\frac{\partial\log f(y_i)}{\partial\mu_i}+d_i\frac{g_0\{h(y_i)\}}{G_0\{h(y_i)\}}\frac{\partial h(y_i)}{\partial\mu_i}-(1-d_i)\frac{\partial\pi_i/\partial\mu_i}{1-\pi_i}\right\}\left.\right]x_{ij}x_{ih}\mu_i\end{aligned}$$

$$\begin{aligned}H(\beta_j,\gamma_h)=\frac{\partial^2\log L}{\partial\gamma_h\partial\beta_j}&=\sum_{i=1}^{N}\frac{\partial}{\partial\gamma_h}\left(\frac{\partial l_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_j}\right)\\&=\sum_{i=1}^{N}\left[d_i\left\{\frac{g'_0\{h(y_i)\}}{G_0\{h(y_i)\}}-\left(\frac{g_0\{h(y_i)\}}{G_0\{h(y_i)\}}\right)^2\right\}\frac{\partial h(y_i)}{\partial\tau_i}\frac{\partial h(y_i)}{\partial\mu_i}+d_i\frac{g_0\{h(y_i)\}}{G_0\{h(y_i)\}}\frac{\partial^2h(y_i)}{\partial\tau_i\partial\mu_i}\right.\\&\quad\left.-(1-d_i)\frac{(\partial^2\pi_i/\partial\tau_i\partial\mu_i)(1-\pi_i)+(\partial\pi_i/\partial\mu_i)(\partial\pi_i/\partial\tau_i)}{(1-\pi_i)^2}\right]w_{ih}x_{ij}\mu_i\end{aligned}$$

$$\begin{aligned}H(\beta_j,b)=\frac{\partial^2\log L}{\partial b\partial\beta_j}&=\sum_{i=1}^{N}\frac{\partial}{\partial b}\left(\frac{\partial l_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_j}\right)\\&=\sum_{i=1}^{N}\left[d_i\frac{\partial^2\log f(y_i)}{\partial b\partial\mu_i}-(1-d_i)\frac{(\partial^2\pi_i/\partial b\partial\mu_i)(1-\pi_i)+(\partial\pi_i/\partial b)(\partial\pi_i/\partial\mu_i)}{(1-\pi_i)^2}\right]x_{ij}\mu_i\end{aligned}$$

$$
\begin{aligned}
H(\gamma_j,\gamma_h) = \frac{\partial^2 \log L}{\partial\gamma_h\partial\gamma_j} &= \sum_{i=1}^{N} \frac{\partial}{\partial\gamma_h}\left(\frac{\partial l_i}{\partial\tau_i}\frac{\partial\tau_i}{\partial\gamma_j}\right) \\
&= \sum_{i=1}^{N}\left[d_i\left\{\frac{g_0'\{h(y_i)\}}{G_0\{h(y_i)\}} - \left(\frac{g_0\{h(y_i)\}}{G_0\{h(y_i)\}}\right)^2\right\}\left(\frac{\partial h(y_i)}{\partial\tau_i}\right)^2 + d_i\frac{g_0\{h(y_i)\}}{G_0\{h(y_i)\}}\frac{\partial^2 h(y_i)}{\partial\tau_i^2}\right. \\
&\quad \left. - (1-d_i)\frac{(\partial^2\pi_i/\partial\tau_i^2)(1-\pi_i) + (\partial\pi_i/\partial\tau_i)^2}{(1-\pi_i)^2}\right] w_{ij}w_{ih}, \\
H(\gamma_j, b) = \frac{\partial^2 \log L}{\partial b\partial\gamma_j} &= \sum_{i=1}^{N} \frac{\partial}{\partial b}\left(\frac{\partial l_i}{\partial\tau_i}\frac{\partial\tau_i}{\partial\gamma_j}\right) \\
&= \sum_{i=1}^{N}\left[-(1-d_i)\frac{(\partial^2\pi_i/\partial b\partial\tau_i)(1-\pi_i) + (\partial\pi_i/\partial b)(\partial\pi_i/\partial\tau_i)}{(1-\pi_i)^2}\right] w_{ij}, \\
H(b,b) = \frac{\partial^2 \log L}{\partial b^2} &= \sum_{i=1}^{N} \frac{\partial}{\partial b}\left(\frac{\partial l_i}{\partial b}\right) \\
&= \sum_{i=1}^{N}\left[d_i\frac{\partial^2 \log f(y_i)}{\partial b^2} - (1-d_i)\frac{(\partial^2\pi_i/\partial b^2)(1-\pi_i) + (\partial\pi_i/\partial b)^2}{(1-\pi_i)^2}\right].
\end{aligned}
$$

References

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed), Wiley.

Azzalini, A. and Capitanio, A. (2014). *The Skew-Normal and Related Families*, IMS Monographs series.

Azzalini, A., Kim, H. M., and Kim, H. J. (2018). Sample selection models for discrete and other non-Gaussian response variables, *Statistical Methods & Applications*, accepted

Boyes, W., Hoffman, D., and Low, S. (1989). An econometric analysis of the bank credit scoring problem, *Journal of Econometrics*, **40**, 3–14.

Greene, W. H. (1992). A Statistical Model for Credit Scoring, *NYU Working Paper*, **EC-92-29**, Available at SSRN: https://ssrn.com/abstract=1867088.

Greene, W. H. (2012). *Econometric Analysis* (7th ed), Pearson Education Ltd.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models, *Annals of Economic and Social Measurement*, **5**, 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error, *Econometrica*, **47**, 153–161.

Riphahn, R. T., Wambach, A., and Million, A. (2003). Incentive effects in the demand for health care: a bivariate panel count data estimation, *Journal of Applied Econometrics*, **18**, 387–405.

Rubin, D. B. (1976). Inference and missing data, *Biometrika*, **63**, 581–592.

Terza, J. (1998). Estimating count data models with endogenous switching: sample selection and endogenous treatment effects, *Journal of Econometrics*, **84**, 129–154.

Vella, F. (1998). Estimating models with sample selection bias: a survey, *The Journal of Human Resources*, **33**, 127–169.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed), MIT Press, Cambridge.

과대산포 가산자료의 새로운 표본선택모형

조성은[a] · 조준[a] · 김형문[a,1]

[a]건국대학교 응용통계학과

(2018년 8월 31일 접수, 2018년 10월 8일 수정, 2018년 10월 31일 채택)

요 약

어떠한 연구에서 관심의 대상이 되는 관찰치가 부분적으로 관측 가능할 때 표본선택의 문제가 일어난다. 이러한 자료를 분석하기 위해 헤크만은 표본선택 모형을 개발하였고 이변량 정규분표의 가정 하에 최대우도방법을 사용하여 모수를 추정하였다. 최근 이항자료와 포아송 자료에 대한 표본선택모형이 제안되었다. 이를 분포조정에 기초하여 과대산포 자료에 대한 모형으로 확장하고자 한다. 표본선택이 없는 과대산포 자료는 흔히 음이항 분포로 분석되어진다. 따라서 음이항 분포를 이용하고 분포조정을 도입한 과대산포 자료에 대한 새로운 모형을 제시하고자 한다. 실제 자료를 이용하여 분석을 하였다. 모의실험 결과 프로파일 우도함수를 이용하여 모수에 대해 추정한 결과는 안정적이다.

주요용어: 표본선택 편향, 헤크만 표본선택모형, 과대산포 자료, 음이항 회귀, 포아송 회귀

이 논문은 2016년도 건국대학교 KU학술연구비 지원에 의한 논문임.

[1]교신저자: (05029) 서울특별시 광진구 능동로 120, 건국대학교 응용통계학과. E-mail: hmkim@konkuk.ac.kr