The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.5.859
25(5), 859–864

Analysis of Food Poisoning via Zero Inflation Models

Hwan Sik Jung[1] · Byung Jip Kim[2] · Sinsup Cho[3] · In-Kwon Yeo[4]

[1]Department of Statistics, Seoul National University
[2]Department of Statistics, Seoul National University
[3]Department of Statistics, Seoul National University
[4]Department of Statistics, Sookmyung Women's University

(Received May 30, 2012; Revised August 14, 2012; Accepted September 18, 2012)

Abstract

Poisson regression and negative binomial regression are usually used to analyze counting data; however, these models are unsuitable for fit zero-inflated data that contain unexpected zero-valued observations. In this paper, we review the zero-inflated regression in which Bernoulli process and the counting process are hierarchically mixed. It is known that zero-inflated regression can efficiently model the over-dispersion problem. Vuong statistic is employed to compare performances of the zero-inflated models with other standard models.

Keywords: Negative binomial regression, Poisson regression, Vuong statistic.

1. 서론

임의의 기간 동안 발생한 사건의 건수가 설명변수에 영향을 받는 경우 그 관계식은 흔히 포아송회귀모형(Poisson regression)에 의해 설명된다. 하지만 평균과 분산이 같아야 하는 포아송분포의 특징은 실증분석에서 발생하는 과산포(over-dispersion), 즉 표본분산이 평균에 비해 큰 현상을 설명하지 못하는 단점이 있다. 과산포가 심각한 경우 대체 모형으로 음의 이항회귀모형(negative binomial regression)이 사용되기도 한다. 사건 발생건수에 대한 실증자료 분석에서 과산포와 더불어 발생하는 문제는 모형에서 기대되는 0의 관측값 빈도보다 많은 0이 관측되는 것이다. 관측값에 0이 많으면 과산포 문제가 더욱 심화되고 모형의 적합도가 떨어진다. 이러한 문제를 해결하는 방안으로 Lambert (1992)는 영과잉모형(zero-inflated model)을 제안했다.

식품의약품안전청의 통계에 따르면 우리나라에서 발생하는 60% 이상의 식중독은 세균이나 바이러스 때문에 발생하고 있다고 한다. 주요 원인균으로는 노로바이러스, 병원성 대장균, 살모넬라, 장염비브리오균, 황색포도상구균이 있는데 이들 원인균은 대부분 기온이나 습도와 같은 기후에 영향을 받는 것으로 알려져 있다. 이 논문에서는 기후요인이 식중독 발생에 어떻게 영향을 주는지를 주요 원인균별로 분석하고자 한다. 분석에 사용된 자료는 2005년 1월부터 2010년 5월까지 원인균별로 신고된 전국의 주별 식중독 발생건수와 해당 주의 전국 60개 기상관측소에서 관측된 평균기온과 평균습도, 일사량이다. 원

[4]Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Chongpa-dong 2ga, Yongsan-gu, Seoul 140-742, Korea. E-mail: inkwon@sm.ac.kr

인균별로 나누어 보는 이유는 기온, 습도, 일사량에 대한 원인균 활성화 정도가 다르고 이에 따른 식중독 평균 발생건수에 차이가 있기 때문이다. 원인균별로 식중독 발생건수를 나누어 분석할 때의 문제는 발생건수가 0인 관측값이 많이 있다는 것이다. 이 논문에서는 관측값에 0이 많이 발생하는 경우 영과잉 모형을 이용하여 문제를 해결해 보고 Vuong 통계량을 이용하여 표준 포아송회귀모형이나 음의 이항회귀모형에 근거한 분석결과와 비교하는 방안에 대해 알아보고자 한다.

2. 영과잉 회귀모형

임의의 사건이 t시점에서 발생한 횟수를 Y_t라고 하고 Y_t의 기대값 $E(Y_t) = \mu_t$는 설명변수 $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})^T$에 영향을 받는다고 하자. 포아송회귀모형이나 음의 이항회귀모형에서는 이 관계식을 연결함수 $g(\cdot)$을 이용하여 다음과 같이 설명한다.

$$g(\mu_t) = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_p x_{pt} = \boldsymbol{\beta}^T \boldsymbol{x}_t,$$

여기서 $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$는 추정해야 할 모수이며 Y_t들은 독립이고 포아송분포 또는 음의 이항분포를 따른다는 가정 하에서 최대가능도 추정법을 이용하여 추정한다.

어떤 자료에서는 포아송분포나 음의 이항분포에서 기대할 수 있는 0의 빈도보다 훨씬 많은 수의 0이 관측되어 기존 포아송회귀분석이나 음의 이항회귀분석으로는 적합도가 떨어지는 경우가 있다. 이러한 자료를 분석하기 위해 Lambert (1992)는 영과잉모형을 제안하였다. Agarwal 등 (2002)에 의하면, 기본 자료생성분포가 포아송분포라고 할 때, 영과잉모형에서는 t 시점에서의 자료 관측이 다음과 같은 자료생성과정으로 이루어진다는 가정에서 모형설계가 이루어진다.

- 단계 1 (베르누이 난수생성): 0일 확률을 ϕ_t로 하고 0이면 자료로 사용하고 1이면 단계 2로 감.
- 단계 2 (포아송 난수생성): 평균이 μ_t인 포아송 난수를 발생하여 자료로 사용함.

이러한 자료생성과정을 거쳐 얻어진 자료의 확률질량함수는 다음과 같이 쓸 수 있다.

$$P(Y_t = y) = f(y; \phi_t, \mu_t) = \begin{cases} \phi_t + (1-\phi_t) f^*(y; \mu_t), & y = 0, \\ (1-\phi_t) f^*(y; \mu_t), & y = 1, 2, \ldots, \end{cases}$$

여기서 $f^*(y; \mu)$는 평균이 μ인 포아송확률질량함수를 의미한다.

만약 베르누이 확률 ϕ_t도 설명변수 $\boldsymbol{x}_t$에 영향을 받는다고 하면, 이진자료에 대한 일반화선형모형에서 가정했던 것처럼 임의의 연결함수 $h(\cdot)$에 대해 관계식을 다음과 같이 가정할 수 있다.

$$h(\phi_t) = \alpha_0 + \alpha_1 x_{1t} + \cdots + \alpha_p x_{pt} = \boldsymbol{\alpha}^T \boldsymbol{x}_t,$$

여기서 $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)^T$ 또한 추정해야 할 모수이다. 일반적으로 이진자료에 대한 연결함수 $h(\cdot)$는 분포함수의 역수인 위수(quantile) 함수가 많이 사용되고 있는데 로지스틱분포인 경우 로짓, 표준정규분포인 경우 프로빗, 극한값 분포인 경우 역로그로그 연결함수가 된다.

위와 같은 가정을 만족할 때, 모수 $\boldsymbol{\alpha}$와 $\boldsymbol{\beta}$의 로그가능도함수는 다음과 같다.

$$\begin{aligned} l(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{y}) = & \sum_{t=1}^{n} I(y_t = 0) \log \left(\phi_t(\boldsymbol{\alpha}, \boldsymbol{x}_t) + (1 - \phi_t(\boldsymbol{\alpha}, \boldsymbol{x}_t)) f^*(y_t; \mu_t(\boldsymbol{\beta}, \boldsymbol{x}_t)) \right) \\ & + \sum_{t=1}^{n} I(y_t > 0) \left\{ \log(1 - \phi_t(\boldsymbol{\alpha}, \boldsymbol{x}_t)) + \log(f^*(y_t; \mu_t(\boldsymbol{\beta}, \boldsymbol{x}_t))) \right\}, \end{aligned}$$

여기서 I는 지시함수를 나타내고 $\phi(\boldsymbol{\alpha}, \boldsymbol{x}) = h^{-1}(\boldsymbol{\alpha}^T\boldsymbol{x})$이고 $\mu(\boldsymbol{\beta}, \boldsymbol{x}) = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x})$를 의미한다. 로그가능도함수를 회귀모수 $\boldsymbol{\alpha}$와 $\boldsymbol{\beta}$로 미분하여 추정방정식을 구하고 Newton-Raphson 방법이나 Quasi-Newton 방법과 같은 수치해석학적인 방법을 통해 로그가능도함수를 최대로 만드는 추정값을 구한다. 자세한 내용은 Miller (2007) 또는 SAS (2008)를 참조하면 된다.

단계 2의 자료생성이 포아송분포 하에서 이루어지는 경우 영과잉모형의 평균과 분산은

$$\begin{aligned} E(Y_t) &= \mu_t(1-\phi_t), \\ \mathrm{Var}(Y_t) &= \mu_t(1-\phi_t)(1+\mu_t\phi_t) > E(Y_t) \end{aligned}$$

이며 이것은 일반적인 포아송분포에서의 평균과 분산의 관계 $E(Y_t) = \mu_t = \mathrm{Var}(Y_t)$ 때문에 다루기 어려운 과산포 문제를 영과잉모형으로 해결할 수 있다는 것을 의미한다. 단계 2에서 포아송분포 대신 음의 이항분포가 사용되었다면, 임의의 $\gamma > 0$에 대해, 일반화선형모형에서 재모수화된(re-parameterized) 확률질량함수는 다음과 같이 주어지고

$$f^*(y;\mu) = \frac{\Gamma(y+1/\gamma)}{y!\Gamma(1/\gamma)} \left(\frac{1}{1+\gamma y}\right)^{\frac{1}{\gamma}} \left(\frac{\gamma\mu}{1+\gamma\mu}\right)^y$$

평균과 분산은 다음과 같다.

$$\begin{aligned} E(Y_t) &= \mu_t(1-\phi_t), \\ \mathrm{Var}(Y_t) &= \mu_t(1-\phi_t)(1+\mu_t\gamma+\mu_t\phi_t) > E(Y_t). \end{aligned}$$

일반적인 음의 이항분포에서 평균이 $E(Y_t) = \mu_t$일 때 평균과 분산의 관계가 $\mathrm{Var}(Y_t) = \mu_t(1+\mu_t\gamma)$인데 영과잉모형은 보다 융통성 있게 과산포를 설명할 수 있다는 것을 의미한다.

3. 표준모형과 영과잉 모형의 비교

편의를 위해 포아송회귀모형과 음의 이항회귀모형을 표준모형이라고 하자. 영과잉모형과 표준모형은 비내포(non-nested) 모형이기 때문에 일반적인 로그가능도비 검정으로 비교할 수 없다. 이러한 비내포 모형들을 비교하는 연구들이 많이 진행되고 있는데 이 논문에서는 Kullback-Leibler 정보를 이용한 Vuong (1989)의 방법을 이용하여 비교하고자 한다.

확률질량함수 $f_T(y;\theta)$가 Y_t의 실제 분포의 것이고 $g(y;\theta_*)$를 비교하고자 하는 확률질량함수라고 하면 Kullback-Leibler 정보는 다음과 같은 기대값으로 정의된다.

$$\mathrm{KL}(g; f_T) = E_T\left[\log\left(\frac{f_T(Y;\theta)}{g(Y;\theta_*)}\right)\right] = \sum_y f_T(y;\theta)\log\left(\frac{f_T(y;\theta)}{g(y;\theta_*)}\right).$$

두 비내포 모형 M_0와 M_1이 있고 각 모형의 확률질량함수가 f_0이고 f_1이라고 하면, Vuong (1989)는 두 모형은 차이가 없다는 귀무가설을

$$H_0:\ E_0\left[\log\left(\frac{f_0(Y;\theta_0)}{f_1(Y;\theta_1)}\right)\right] = 0$$

로 정하고 f_0이 f_1보다는 우수하다는 대립가설을

$$H_1:\ E_0\left[\log\left(\frac{f_0(Y;\theta_0)}{f_1(Y;\theta_1)}\right)\right] > 0 \quad \text{또는} \quad E_0\left[\log\left(\frac{f_1(Y;\theta_1)}{f_0(y;\theta_0)}\right)\right] < 0$$

Table 4.1. Parameter estimation of Poisson regression and negative binomial regression

원인균	노로바이러스				살모렐라균			
회귀모형	포아송		음의 이항		포아송		음의 이항	
모수	추정값	유의확률	추정값	유의확률	추정값	유의확률	추정값	유의확률
절편	−0.196	0.447	0.404	0.012	0.674	0.374	0.694	0.407
평균기온	−0.026	0.001	−0.039	0.001	0.112	0.000	0.117	0.000
평균습도					−0.046	0.001	−0.047	0.002
일사량	0.154	0.021						
γ			1.520	0.000			0.282	0.181

로 설정한다. 여기서 E_0는 f_0를 기반으로 기대값을 계산한다는 것을 나타낸다. $Z_t = \log(f_0(Y_t;\hat{\theta}_0) - \log(f_1(Y_t;\hat{\theta}_1))$이라고 할 때, 이 가설을 검정하기 위한 Vuong (1989)는 검정통계량은 다음과 같이 정의될 수 있음을 보였다.

$$V = \frac{\sqrt{n}\bar{Z}}{S_Z},$$

여기서 $\bar{Z}$와 S_Z는 Z_t들의 표본평균과 표본표준편차를 의미한다. 이 통계량은 가설 $E[Z_t] = 0$을 검정하기 위한 검정통계량으로 점근적으로 표준정규분포를 따른다는 것을 Vuong (1989)는 보였다. 이 논문에서는 영과잉모형을 M_0라고 하고 표준모형을 M_1으로 설정하여 분석하였다.

4. 식중독 지표분석

Choi 등 (2008)는 2004년부터 2006년까지의 수도권 자료를 이용하여 일 최고기온, 습도, 월효과가 일별 식중독 발생에 어떻게 영향을 주는지를 로그선형모형으로 설명하였으며 이를 통해 식중독발생지수를 개발하였다. 이 논문에서는 2005년 1월부터 2010년 5월까지 보고된 주별 식중독 발생건수와 해당 기간 동안의 기상 자료와의 관계를 알아보고자 한다. 식중독은 여러 가지 원인균에 의해 발생하는데 이 원인균의 활성도는 기온, 습도, 일사량 등에 영향을 받는 것으로 알려져 있다. 대부분의 원인균은 기온이 상승할수록 활성도가 높아지지만 노로바이러스의 경우 오히려 떨어지는 경향이 있어 식중독 발생건수를 분석하는데 있어 원인균별로 분류할 필요가 있다. 이때 발생하는 문제가 관측 자료에 0이 많이 포함되는 것이다.

이 분석에서는 원인균이 노로바이러스와 살모렐라균인 경우에 대해 표준 포아송회귀모형과 음의 이항회귀모형으로 적합시켰을 때와 영과잉모형으로 적합시켰을 때를 비교해 본다. 음의 이항회귀모형이 포아송회귀모형 보다 좋은지를 비교할 때는 과산포 여부에 대한 가설검정을 통해 이루어진다. 이는 음의 이항회귀분석에서 $\gamma = 0$인지 $\gamma > 0$인지를 가설검정하는 것을 의미한다. 이 가설검정은 로그가능도비 검정을 통해서도 비교할 수 있는데, $\hat{L}_P$와 $\hat{L}_N$가 포아송회귀모형과 음의 이항회귀모형 가정 하에서 최대가능도 추정량의 로그가능도 함수라고 하자. Chernoff (1954)에 의하면 귀무가설 하에서 로그가능도비 검정통계량 $G = -2(\hat{L}_P - \hat{L}_N)$은 점근적으로 0에서 0.5의 확률을 가지고 0보다 큰 경우 $0.5\chi^2_1$를 따르는 분포를 따른다. 그러므로 $100\alpha\%$ 유의수준으로 검정하고자 한다면 기각역의 임계값은 χ^2_1에서 $1-2\alpha$에 해당되는 위수가 된다. 이와 관련된 연구는 Lawless (1987)와 Cameron과 Trivedi (1986) 등을 참조하기 바란다.

Table 4.1은 각각의 모형에서 10% 유의수준에서 유의한 회귀모수에 대해 추정한 결과이다. γ의 값이 0에 가까울수록 조건부 평균과 조건부 분산이 같다는 것을 의미하며 이를 통해서 모형 선택을 할 수 있

Table 4.2. Parameter estimation of zero-inflated Poisson regression and zero-inflated negative binomial regression

원인균	노로바이러스				살모렐라균			
회귀모형	포아송		음의 이항		포아송		음의 이항	
모수	추정값	유의확률	추정값	유의확률	추정값	유의확률	추정값	유의확률
절편	0.085	0.696	−0.409	0.170	0.793	0.307	0.778	0.338
평균기온					0.054	0.037	0.058	0.026
평균습도					−0.031	0.031	−0.032	0.035
일사량	0.177	0.007	0.184	0.045				
절편	−0.856	0.003	−10.348	0.071	3.145	0.021	3.234	0.025
평균기온	0.059	0.001	0.443	0.064	−0.618	0.100	−0.683	0.085
γ			1.163	0.000			0.145	0.429

Table 4.3. Comparison of standard model and zero-inflated model (Vuong Statistic)

변수	Vuong 통계량	유의확률	비고
살모넬라	−2.663	0.003	포아송분포
노로바이러스	−1.853	0.032	음의 이항분포

다. 이 표에서 볼 수 있듯이 노로바이러스에 대한 음의 이항회귀모형에서 $H_0 : \gamma = 0$은 1% 유의수준에서도 기각된다. 또한 로그가능도비 검정을 통해서도 비교할 수 있는데

$$G = -2\left(\hat{L}_P - \hat{L}_N\right) = -2(-433.09 + 375.54) = 115.1$$

이며 p-값이 거의 0에 가까워 과산포 문제가 확실히 있는 것을 확인할 수 있으며 음의 이항회귀모형이 포아송회귀모형보다 더 적합하다는 것을 알 수 있다. 노로바이러스의 경우 평균기온에 영향을 받고 있으며 온도가 상승하면 식중독 평균발생건수가 감소하는 것으로 나타났다.

살모렐라균의 경우 음의 이항회귀모형의 γ가 0이라는 귀무가설을 기각시키지 못하고 있으며 로그가능도비검정에서도 포아송회귀모형과 음의 이항회귀모형의 차이가 없는 것으로 나타났다. 살모렐라균에 의한 식중독 평균발생건수는 평균기온이 높아지면 증가하는 반면 습도가 높아지면 감소하는 경향이 있는 것으로 나타났다.

Table 4.2는 영과잉회귀모형을 적용했을 때 결과로 아래 부분의 추정값은 로지스틱 모형을 적용했을 때 식중독 발생건수가 0일 확률에 대한 설명변수의 영향력을 표시한 것으로 두 원인균 모두 평균기온에만 유의하게 영향받는 것으로 분석되었다. 표준모형에서와 같이 노로바이러스는 음의 이항회귀모형으로 살모렐라균은 포아송회귀모형으로 분석하는 것이 적절한 것으로 나타났다. 노로바이러스에 의한 식중독 발생건수의 경우 평균기온이 상승할수록 발생하지 않을 확률이 높아지며 발생하는 경우에는 일사량이 많을수록 평균발생건수가 증가하는 경향이 있다. 살모렐라균의 경우 평균기온이 낮아질수록 발생하지 않을 확률이 커지고 발생하는 경우 평균기온이 높아질수록, 평균습도가 낮아질수록 평균발생건수가 증가하는 것으로 분석되었다.

Table 4.3은 각각의 변수에서 Vuong 통계량을 통해서 표준모형과 영과잉모형을 비교한 결과이다. Vuong 통계량값이 음의 방향으로 유의하다는 것은 영과잉모형이 더 적합하다는 의미이고 양의 방향으로 유의하다면 표준모형이 더 적합하다는 의미한다. Table 4.3에서는 Table 4.1과 Table 4.2의 분석결과를 토대로 노로바이러스의 경우에는 음의 이항회귀분석을 바탕으로 표준모형과 영과잉모형에 대한 비교를 했고 살모넬라의 경우에는 포아송회귀분석을 바탕으로 비교하였다. 두 경우 모두 데이터에 0이 상대적으로 많아 5% 유의수준에서 영과잉 모형이 더 적합하다는 것을 알 수 있다.

5. 결론

원인균 별 식중독 자료와 같이 사건의 발생 건수에 0이 상대적으로 많이 관측되는 경우에는 표준 일반 화선형모형보다는 영과잉모형을 사용하여 분석하는 것이 더 적절하다. 또한 과산포의 문제가 함께 발생한 경우에는 포아송분포보다는 음의 이항분포를 이용하여 영과잉모형을 사용하여 분석하는 것이 더 적절하다. 하지만 영과잉모형을 이용하여 사건의 발생 빈도를 예측하고자 할 때, 발생건수가 0일 확률이 아닐 확률보다 높게 나타나는 경우가 많아 정확히 예측하기가 조금은 어려워 보인다. 이를 적절하게 조절하기위해서는 일정한 분계점(threshold)을 설정하여 사건발생빈도를 예측하는 방법을 적용할 수 있으나 이런 분계점 역시 주관적인 판단에 의해 설정되는 경우가 많아 보다 더 연구가 필요한 분야이다.

References

Agarwal, D. K., Gelfand, A. E. and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data, *Environmental and Ecological Statistics*, **9**, 341–355.

Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests, *Journal of Applied Econometrics*, **1**, 29–53.

Chernoff, H. (1954). On the distribution of the likelihood ratio, *Annals of Mathematical Statistics*, **25**, 573–578.

Choi, K., Kim, B., Bae, W., Jung, W. and Cho, Y. (2008). Developing the index of foodborne disease occurrence, *The Korean Journal of Applied Statistics*, **21**, 649–658.

Lambert, D. (1992). Zero-inflated Poisson regression models with an application to defects in manufacturing, *Technometrics*, **34**, 1–14.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression, *The Canadian Journal of Statistics*, **15**, 209–225.

Miller, J. M. (2007). *Comparing Poisson, Hurdle, and ZIP model fit under varying degrees of skew and zero-inflation*, University of Florida, DAI-A 68/06, Dec 2007.

SAS Institute Inc. (2008). *SAS/ETS User's Guide (Version 9.2, Chap.10, The COUNTREG Procedure)*, SAS Institute Inc., Cary, NC, USA.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypothesis, *Econometrica*, **57**, 307–334.