제로팽창 이변량 음이항 회귀모형에서 산포모수에 대한 가설검정

신지은[1], 장동민[1], 정병철[2]

January 29, 2024

[1]서울시립대학교 통계데이터사이언스학과 석박사통합과정
[2]서울시립대학교 통계학과 교수

목차

1. 서론

1. 서론

계수자료 (count data)

- 정해진 시간 안에 어떤 사건이 일어날 횟수
- 0건, 0개와 같이 0의 값이 과도하게 발생하는 경우가 많음 (영과잉; zero-inflation)
- 영과잉 현상은 평균보다 분산을 더 커지게 함 (과산포)
- 포아송 분포(평균과 분산이 동일) 대신 음이항 분포(과산포성 반영) 고려함

1. 서론

제로팽창 이변량 계수자료

- Y_1: 의사에게 상담받은 사람의 수, $\bar{Y}_1 = 0.302, \sigma_1 = 0.978$
- Y_2: 의사가 아닌 전문가에게 상담받은 사람의 수, $\bar{Y}_2 = 0.215, \sigma_2 = 0.965$
- $(0, 0)$-cell의 확률: 0.737, 피어슨 상관계수 $= 0.148$, 스피어만 상관계수: 0.109

	Y_2												
Y_1	0	1	2	3	4	5	6	7	8	9	10	11	Total
0	3826	196	57	9	17	2	3	22	3	5	1	0	4141
1	670	66	18	4	6	2	4	8	2	1	0	1	782
2	148	11	4	1	1	1	1	3	0	2	0	2	174
3	25	2	2	0	0	0	1	0	0	0	0	0	30
4	19	1	1	0	0	0	0	2	1	0	0	0	24
5	7	1	0	0	1	0	0	0	0	0	0	0	9
6	10	0	1	0	1	0	0	0	0	0	0	0	12
7	6	1	1	0	0	0	1	2	0	0	1	0	12
8	4	0	0	0	0	1	0	0	0	0	0	0	5
9	1	0	0	0	0	0	0	0	0	0	0	0	1
Total	4716	278	84	14	26	6	10	37	6	8	2	3	5190

Table 1: 이변량 계수자료의 예: 호주 건강서베이 데이터에서 Y_1와 Y_2의 교차표

1. 서론

기존의 이변량 음이항 분포 (Wang, 2003):

$$f(Y_1 = y_1, Y_2 = y_2) = \frac{\Gamma(y_1 + y_2 + \tau)}{\Gamma(\tau) y_1! y_2!} \frac{\mu_1^{y_1} \mu_2^{y_2} \tau^{\tau}}{(\mu_1 + \mu_2 + \tau)^{y_1 + y_2 + \tau}},$$

여기서 $\tau > 0$는 산포모수이다.

- 모형의 한계점: 두 반응변수의 양(+)의 상관만을 반영하며 음(-)의 상관을 허용하지 않음, 서로 다른 산포를 반영하지 못함

1. 서론

Sarmanov 분포족 기반의 이변량 음이항 분포 (Famoye, 2010):

$$
\begin{aligned}
f(Y_1 = y_1, Y_2 = y_2) \\
&= \left[\prod_{k=1}^{2} \frac{\Gamma(y_k + \tau_1^{-1})}{\Gamma(y_k + 1)\Gamma(\tau_k^{-1})} \left(\frac{\tau_k \mu_k}{1 + \tau_k \mu_k} \right)^{y_k} \left(\frac{1}{1 + \tau_k \mu_k} \right)^{\tau_k^{-1}} \right] \\
&\quad \times \left[1 + \omega \prod_{k=1}^{2} (e^{-y_k} - c_k) \right], \quad y_k = 0, 1, 2, \ldots, \quad k = 1, 2.
\end{aligned}
$$

- $\omega \in \mathbb{R}$ is multiplicative factor parameter.
- $c_k = \mathbb{E}(e^{-Y_k}) = (1 + (1 - e^{-1})\tau_k \mu_k)^{-\tau_k^{-1}}$ are mixing functions.

Why the Sarmanov family is used?

- 두 반응변수 간 양과 음의 상관 모두 반영하는 모형 (Marshall & Olkin, 1990)
- Copula 모형보다 수식적으로 더 간결하고 모수 추정에 소요되는 시간이 더 적게걸림 (Hofer and Leitner, 2012)

본 연구에서는

- (0,0)-cell에서 영과잉이 발생하는 서로 다른 산포를 가진 이변량 계수자료를 (Faroughi and Ismail, 2017)가 제안한 제로팽창 이변량 음이항 회귀모형으로 모형화
- 이 모형에서 산포모수에 대한 검정을 시행
- 특히 스코어 검정에 대한 유도를 중점으로 다루며 LR 검정과 비교함

2. 제로팽창 이변량 음이항 회귀모형

Bivariate negative binomial regression model based on the Sarmanov family (Famoye, 2010):

$$f(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}) = \left[\prod_{k=1}^{2} \frac{\Gamma(y_{ik} + \tau_1^{-1})}{\Gamma(y_{ik}+1)\Gamma(\tau_k^{-1})} \left(\frac{\tau_k \mu_{ik}}{1 + \tau_k \mu_{ik}} \right)^{y_{ik}} \left(\frac{1}{1 + \tau_k \mu_{ik}} \right)^{\tau_k^{-1}} \right] \times \left[1 + \omega \prod_{k=1}^{2} (e^{-y_{ik}} - c_{ik}) \right], \quad y_{ik} = 0, 1, 2, \ldots, \quad k = 1, 2.$$

- $\omega \in \mathbb{R}$ is multiplicative factor parameter.
- $c_{ik} = (e^{-Y_{ik}}) = (1 + (1 - e^{-1})\tau_{ik}\mu_{ik})^{-\tau_k^{-1}}$ are mixing functions.

Regression setup for mean parameters:

$$\log(\mu_{ik}) = \boldsymbol{x}_{ik}\boldsymbol{\beta}_k, \quad k = 1, 2.$$

- $\boldsymbol{x}_{ik} = (x_{ik}^{(1)}, \ldots, x_{ik}^{(p_k)})$ are the p_k-dimensional vectors of predictors.
- $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{p_k k})^T$ are the p_k-dimensional vectors of coefficient parameters.

Bivariate zero-inflated negative binomial regression model (BZINB) regression model is defined by adding the probability of an "extra zero" ϕ_i (Faroughi and Ismail, 2017):

$$f_{\text{BZINB}}(Y_1 = y_{i1}, Y_2 = y_{i2})$$
$$= \begin{cases} \phi_i + (1-\phi_i)\prod_{k=1}^{2}(1+\tau_k\mu_{ik})^{-\tau_k^{-1}}\left[1+\omega\prod_{k=1}^{2}(1-c_{ik})\right], & \text{if } (0,0) \\ (1-\phi_i)f(y_{i1},y_{i2}), & \text{if o.w.} \end{cases}$$

Regression setup for zero probability:

$$\log\frac{\phi_i}{1-\phi_i} = \boldsymbol{z}_i^T\boldsymbol{\gamma}$$

- $\boldsymbol{z}_i = (z_i^{(1)}, \ldots, z_i^{(q)})$ are the q-dimensional vectors of predictors.
- $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T$ are the q-dimensional vectors of coefficient parameters.

Property of Sarmanov BZINB (Lee, 1996):

$$
\begin{aligned}
\mathbb{E}(Y_{ik}) &= (1-\phi_i)\mu_{ik} \\
\mathrm{Var}(Y_{ik}) &= (1-\phi_i)(1+\mu_{ik}\tau_k+\phi_i\mu_{ik})\mu_{ik} \\
\mathrm{Cov}(Y_{i1},Y_{i2}) &= (1-\phi_i)(\omega c_{i1}c_{i2}R_{i1}R_{i2}+\phi_i\mu_{i1}\mu_{i2}) \\
\mathrm{Corr}(Y_{i1},Y_{i2}) &= \frac{\omega c_{i1}c_{i2}R_{i1}R_{i2}+\phi_i\mu_{i1}\mu_{i2}}{\sqrt{\mu_{i1}\mu_{i2}(1+\tau_1\mu_{i1}+\phi_i\mu_{i1})(1+\tau_2\mu_{i2}+\phi_i\mu_{i2})}}
\end{aligned}
$$

with $R_{ik} = -\mu_{ki}(1-e^{-1})(1+\tau_k\mu_{ki})(1+(1-e^{-1})\tau_k\mu_{ki})^{-1}, k = 1, 2.$

3. 가설검정

Hypothesis:

$$H_0 : \tau_1 = \tau_2 = 0 \quad \text{vs.} \quad H_1 : \text{not } H_0. \tag{1}$$

- Since the dispersion parameters τ_1, τ_2 are nonnegative, the H_1 is to be one-sided test ($\tau_1 > 0$ or $\tau_2 > 0$).
- The BZINB reduces to the BZIP when the parameter $\tau_k \to 0, k = 1, 2$ and still $\phi_i > 0$.

BZIP regression model:

$$f_{\text{BZIP}}(Y_1 = y_{i1}, Y_2 = y_{i2})$$
$$= \begin{cases} \phi_i + (1-\phi_i)\prod_{k=1}^{2} e^{-\mu_{i1}-\mu_{21}} \left[1 + \omega \prod_{k=1}^{2}(1-c_{ik})\right], & \text{if } (0,0) \\ (1-\phi_i)\frac{e^{-\mu_{i1}-\mu_{21}}\mu_{i1}^{y_{i1}}\mu_{21}^{y_{i2}}}{y_{i1}!y_{i2}!}, & \text{if o.w.,} \end{cases}$$

with $\mathbb{E}(Y_{ik}) = (1-\phi_i)\mu_{ik}$ and $\text{Var}(Y_{ik}) = (1-\phi_i)(1+\phi_i\mu_{ik})\mu_{ik}$.

Score statistic:

$$T = S(\hat{\boldsymbol{\theta}})^T I(\hat{\boldsymbol{\theta}})^{-1} S(\hat{\boldsymbol{\theta}}) \tag{2}$$

- S is score function
- I is information matrix
- $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T, \hat{\boldsymbol{\gamma}}^T, \hat{\omega}, \tau_1 = 0, \tau_2 = 0)^T$ is the maximum likelihood estimators under H_0
- In two-sided test, $T \sim \chi^2(df = 2)$

The log-likelihood function of BZINB:

$$
\begin{aligned}
\log L &= \sum_{i=1}^{n} \ell_i \\
&= \sum_{i=1}^{n} \mathbb{1}(0,0) \log \left\{ \phi_i + (1-\phi_i) \prod_{k=1}^{2} (1+\tau_k \mu_{ik})^{-\tau_k^{-1}} \left[1 + \omega \prod_{k=1}^{2} (1 - c_{ik}) \right] \right\} \\
&+ \sum_{i=1}^{n} [1 - \mathbb{1}(0,0)] \left\{ \log(1-\phi_i) + \sum_{k=1}^{2} \log \left[\frac{\Gamma(y_{ik} + \tau_k^{-1})}{\Gamma(y_{ik}+1)\Gamma(\tau_k^{-1})} \right. \right. \\
&\times \left. \left(\frac{\tau_k \mu_{ik}}{1+\tau_k \mu_{ik}} \right)^{y_{ik}} \left(\frac{1}{1+\tau_k \mu_{ik}} \right)^{\tau_k^{-1}} + \log \left[1 + \omega (e^{-y_{1i}} - c_{1i})(e^{-y_{2i}} - c_{2i}) \right] \right\} \quad \backslash fr
\end{aligned}
$$

Score function:

$$
S(\hat{\boldsymbol{\theta}}) = \left. \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}
$$

Expected information matrix:

$$I(\hat{\boldsymbol{\theta}}) = \mathbb{E}\left[-\frac{\partial \log L^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right] \tag{3}$$

Since it is difficult to calculate an exact form of the expected information matrix, the observed information matrix is alternatively used.

- The (j, k)-element of the observed information matrix under H_0 is calculated by:

$$I(\hat{\boldsymbol{\theta}})_{jk} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta_j}\frac{\partial \ell_i}{\partial \theta_k}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad j, k = 1, \ldots, p_1 + p_2 + q + 3,$$

Note. One-sided score test

Little studies for one-sided score tests with multivariate parameters have been done.

- Test for overdispersion ($H_0 : \tau = 0$ vs $H_1 : \tau > 0$) in ZINB (Ridout, 2001)
- Test for zero-inflation ($H_0 : \phi = 0$ vs $H_1 : \phi > 0$) in BZIP (Lee et. al., 2009)
- (Famoye, 2010) induced the one-sided test for two parameters in the Sarmanov BNB model but it does not mention the approximate distribution.

We alternatively adopt the method proposed by (King and Wu, 1997)

- It consists of the sum of score functions related only to parameters of interest.
- The approximate distribution of score statistic is known as $N(0, 1)$.

3. 가설검정

Suppose $\boldsymbol{\theta} = (\boldsymbol{\tau}^T, \boldsymbol{\eta}^T)^T$ where $\boldsymbol{\tau} = (\tau_1, \tau_2)^T$ and $\boldsymbol{\eta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, \omega)^T$. Testing $H_0 : \boldsymbol{\tau} = 0$ against $H_1 : \boldsymbol{\tau} > \mathbf{0}$. The score function proposed by (King and Wu, 1997) is:

$$T^{KW} = \frac{S^{KW}(\hat{\boldsymbol{\theta}})}{\sqrt{d^T I_{\boldsymbol{\tau}\boldsymbol{\tau}}^{-1} d}} \sim N(0,1), \tag{4}$$

with

$$S^{KW}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{2} \frac{\partial \log L(\boldsymbol{\theta})}{\partial \tau_i}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad I(\hat{\boldsymbol{\theta}}) = I(\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}} = \begin{bmatrix} I_{\boldsymbol{\eta}\boldsymbol{\eta}} & I_{\boldsymbol{\eta}\boldsymbol{\tau}} \\ I_{\boldsymbol{\eta}\boldsymbol{\tau}}^T & I_{\boldsymbol{\tau}\boldsymbol{\tau}} \end{bmatrix}, \tag{5}$$

where $d = (1,1)^T$.

Likelihood ratio (LR) test:

$$LR = -2(\log L(\text{res}) - \log L(\text{unres})) \sim \frac{1}{4}\chi^2(0) + \frac{1}{2}\chi^2(1) + \frac{1}{4}\chi^2(2)$$

- $\log L(\text{res})$ is log-likelihood in restricted model (BZINB)
- $\log L(\text{unres})$ is log-likelihood in unrestricted model (BZIP)

Since the parameters are on the boundary, LR approximately follows a mixture chi-square distribution (Chernoff, 1954).

4. Simulation study

Simulation setting

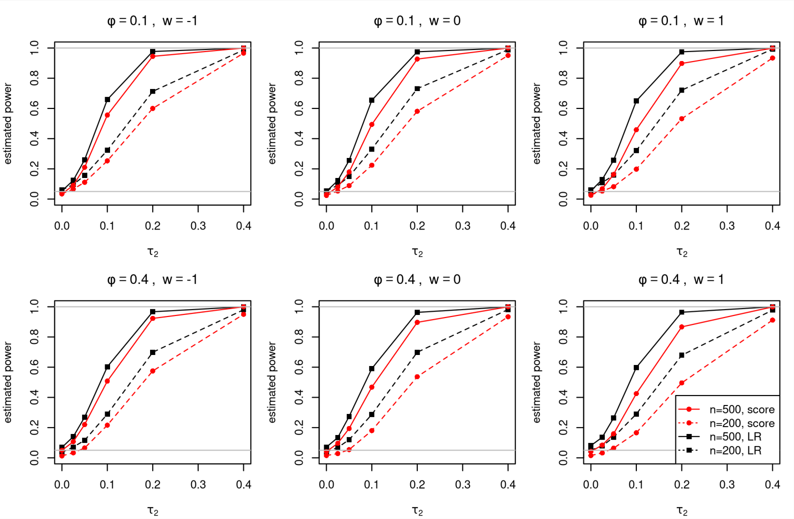
- $\mu_{1i} = \exp(\beta_{10} + X_{1i}\beta_{11}), \mu_{2i} = \beta_{20} + X_{2i}\beta_{21}$, where X_{1i} and $X_{2i} \sim U(0,1)$ and $(\beta_{10}, \beta_{11}) = (0.2, 0.4), (\beta_{20}, \beta_{21}) = (0.4, 0.8)$
- the marginal means of Y_{1i} and Y_{2i}, ranged in value (1.221, 1.822) and (1.491, 3.320), respectively.
- $\log \frac{\phi_i}{1-\phi_i} = \gamma_0 + Z_i\gamma_1$, where $Z_i \sim U(0,1)$ and $\gamma_0 = -1.2$, $\gamma_1 \in \{-0.997, -0.186, 0.352, 0.794, 1.2\}$
- ϕ_i lie at $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ on mean.
- $\omega \in \{-1, 0, 1\}$.
- Repeat 4,000 times.
- 검정 지표로 명목 유의수준 (0.05, 0.1)과 검정력을 추정함

Simulation 1. Comparison for estimated nominal confidence level.

		$n = 100$				$n = 200$			
		score		LR		score		LR	
ω	ϕ	0.05	0.1	0.05	0.1	0.05	0.1	0.05	0.1
-1	0.1	0.023	0.061	0.056	0.096	0.030	0.072	0.063	0.103
	0.2	0.024	0.070	0.048	0.086	0.029	0.070	0.056	0.100
	0.3	0.024	0.064	0.052	0.087	0.031	0.071	0.054	0.098
	0.4	0.026	0.062	0.048	0.083	0.029	0.071	0.052	0.093
	0.5	0.030	0.074	0.050	0.083	0.027	0.067	0.055	0.094
0	0.1	0.021	0.060	0.058	0.097	0.028	0.065	0.060	0.103
	0.2	0.022	0.065	0.054	0.092	0.022	0.066	0.061	0.105
	0.3	0.020	0.058	0.054	0.095	0.028	0.064	0.058	0.100
	0.4	0.027	0.061	0.048	0.090	0.027	0.063	0.055	0.095
	0.5	0.031	0.068	0.058	0.093	0.026	0.062	0.053	0.094
1	0.1	0.020	0.051	0.061	0.106	0.026	0.060	0.063	0.110
	0.2	0.026	0.062	0.059	0.097	0.021	0.066	0.064	0.104
	0.3	0.021	0.056	0.068	0.102	0.026	0.064	0.064	0.106
	0.4	0.027	0.068	0.057	0.097	0.025	0.061	0.056	0.100
	0.5	0.037	0.080	0.069	0.109	0.024	0.059	0.055	0.093

4. Simulation study

Simulation 2. Comparison for estimated power.



5. Application

Australian health survey data

- Y_1: 의사에게 상담받은 사람의 수, $\bar{X}_1 = 0.302, \sigma_1 = 0.978$
- Y_2: 의사가 아닌 전문가에게 상담받은 사람의 수, $\bar{X}_2 = 0.215, \sigma_2 = 0.965$
- $(0, 0)$-cell의 확률: 0.737, 피어슨 상관계수 $= 0.148$
- 검정 결과: 귀무가설 기각 (BZINB가 BZIP에 비해 적합도가 높음)

$$T^{KW} = 18.115(p < 0.001),$$
$$LR = 1539.844(p < 0.001)$$

6. 결론

6. 결론

- 영과잉이 있는 서로 다른 산포를 가진 이변량 계수자료를 Faroughi and Ismail (2017)의 방법으로 모형화
- King and Wu (1997)이 제안한 스코어 검정과 LR 검정을 비교함
- 스코어 검정은 명목 유의수준을 과소추정, LR 검정은 비교적 적절히 유지함
- 검정력도 스코어 검정이 LR 검정에 비해 낮게 나타남
- 해당 데이터에서는 LR검정이 더 나은 성능을 보임
- 계산적 측면에서는 스코어 검정이 더 효율적