응용통계연구 (2009)
22(2), 271–280

Sarmanov형 이변량 일반화이항모형의 적합

이주용[1] · 김기영[2]

[1]산은자산운용(주) 리스크관리팀, [2]고려대학교 통계학과

(2008년 12월 접수, 2009년 2월 채택)

요 약

급내/급간상관이 동시에 존재하는 이변량 이항자료에 대한 모형으로 Danaher과 Hardie (2005)는 베타이항분포를 제안한바 있다. 그러나 이 모형은 베타분포에 따르는 성공확률을 통해 급내상관을 묘사하므로 그 적용범위가 양의 급내상관을 가지는 자료에 제한된다. 이 연구에서는 보다 더 넓은 범위의 급내상관에 대해 유용성을 가지는 일반화가법/승법이항모형과 확장베타이항모형 등에 Sarmanov형식의 이변량 확장을 고려하고 이들을 기존 모형과 적합도의 측면에서 비교한다. 실제자료인 주식자료와 소비자패널자료에 이변량 일반화이항모형들을 적용한 결과, B-mB와 B-ebB의 성능이 우수한 것으로 나타나며, 그 중 상대적으로 넓은 허용범위의 급내상관을 가지는 B-mB가 선호된다고 할 수 있다.

주요용어: 급내/급간상관, 일반화이항분포, Sarmanov형 이변량 분포.

1. 서론

급내(intra-class)상관관계를 가지는 베르누이 변수들의 합을 통상적 이항분포로 처리하면 그의 변이성이 이 모형에 의해 적절히 포착되지 않아(과산포) 분산이 과소평가되는 문제가 일어난다. 하나의 대안으로 고려될 수 있는 베타이항모형 (Skellam, 1948; Williams, 1975; Griffiths, 1973; 김병수 등, 1995)은 베타분포에 따르는 성공확률을 가정하고 이를 통해 급내상관을 묘사함으로 모형의 적용범위가 양의 급내상관을 가지는 자료에 제한된다. Danaher과 Hardie (2005)는 급내상관과 급간(inter-class)상관이 동시에 존재하는 이변량 이항자료에 대한 모형으로 Sarmanov형 이변량 베타이항분포를 고려한 바 있다. 이 연구에서는 베타이항모형보다 급내상관의 더 넓은 범위를 포괄하는 일반화가법/승법이항모형 (Altham, 1978)과 확장베타이항모형 (Prentice, 1986) 등의 이용을 제안하고 모형적합도를 비교 검토한다. 실제자료에 대해 이변량 일반화이항모형들을 적용한 분석결과를 제공한다.

2. Sarmanov형 이변량 모형

주변확률함수 $f_i(y_i)$, $i = 1, 2$를 가지는 두 확률변수 (Y_1, Y_2)에 대한 Sarmanov (1966)형 결합확률함수는 $\int \phi_i(t) g_i(t) dt = 0$, $1 + \omega\phi_1(y_1)\phi_2(y_2) \geq 0$을 만족하는 $\phi_i(y_i)$와 ω에 대해 $f(y_1, y_2) = f_1(y_1) f_2(y_2)\{1 + \omega\phi_1(y_1)\phi_2(y_2)\}$로 정의된다. 이는 각 $f_i(y_i)$에 합성함수 $\phi_i(y_i)$와 두 변수 간 급간상관에 관여하는 모수 ω를 도입하여 다양한 형식의 이변량분포를 생성하는데 유용하게 이용할 수 있으며, FGM분포군 (Farlie, 1960)도 Sarmanov형 분포군에 포함된다. Sarmanov형 분포군에서 급간(inter-class)상관계수 $\mathrm{Corr}(Y_1, Y_2) = (\omega\nu_1\nu_2)/(\sigma_1\sigma_2)$, 단, $\sigma_i^2 = \mathrm{Var}(Y_i)$, $\nu_i = E[Y_i\phi_i(Y_i)]$, $i = 1, 2$이다. 이

이 연구는 2007년도 고려대학교 특별연구비 지원에 의해 수행되었음.
[2]교신저자: (136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 교수. E-mail: kykim@korea.ac.kr

연구에서 고려하는 합성함수는 Lee (1996)에 의해 제안된 $\phi_i(y_i) = y_i - \mu_{y_i}$, $\mu_{y_i} = E(Y_i)$, $i = 1, 2$로서, 이 경우 $\mathrm{Corr}(Y_1, Y_2) = \omega\sigma_1\sigma_2$가 되어, ω의 부호는 두 변수 간 급간상관의 방향을 그리고 $\omega = 0$은 곧 독립(무상관)인 경우를 의미하게 된다. 이와 같은 Sarmanov 형식으로 베타이항모형, 가법/승법이항모형, 확장베타이항모형 등의 일변량 일반화이항모형을 이변량 모형으로 확장시킬 수 있다.

2.1. Sarmanov형 이변량 일반화이항모형

베르누이 변수 $Z_{ij}(j = 1, 2, \ldots, n_i; i = 1, 2)$에서 성공확률을 π_i로, 그리고 이들 간 (공통)급내상관계수 $\mathrm{Corr}(Z_{ij}, Z_{ij'}) = \rho_i, j \neq j'$로 표기한다. 이 때, 각 $X_i = \sum_{j=1}^{n_i} Z_{ij}$에 대한 (일변량)일반화이항모형을 Sarmanov형 이변량의 경우로 확장한 모형은 다음과 같다.

2.1.1. 이변량 베타이항모형(Bivariate Beta Binomial: B-bB) 이는 π_i에 모수$(\alpha_i, \beta_i > 0)$를 가지는 베타분포를 가정하고, $X_i|\pi_i$의 조건부확률을 이항분포를 통해 모형화한 것으로, 이때 X_i는 다음과 같은 베타이항분포에 따른다.

$$f_i(x_i) = \binom{n_i}{x_i} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \frac{\Gamma(\alpha_i + x_i)\Gamma(n_i + \beta_i - x_i)}{\Gamma(\alpha_i + \beta_i + n_i)}, \quad x_i = 0, 1, \ldots, n_i \tag{2.1}$$

이때, $E(X_i) = n_i\pi_i^b = \mu_i^b$, $\mathrm{Var}(X_i) = n_i\pi_i^b(1 - \pi_i^b)\{1 + (n_i - 1)/(1 + \alpha_i + \beta_i)\}$, $\pi_i^b = \alpha_i(\alpha_i + \beta_i)^{-1}$. 이와 같은 베타이항모형에서 각 변수별 (공통)급내상관계수 $\rho_i^b = (1 + \alpha_i + \beta_i)^{-1}$로서 그 값은 항상 양(+)의 값으로 가정되어있어 음의 급내상관의 경우 그 유용성이 제한된다고 할 수 있다 (Kupper와 Haseman, 1978).

Danaher과 Hardie (2005)는 식 (2.1)의 베타이항분포를 다음과 같이 급간상관이 있는 두 변수의 경우로 확장하였다. 두 성공확률 (π_1, π_2)가 모수 $(\alpha_1, \alpha_2, \beta_1, \beta_2, \omega_b)$를 가지는 Sarmanov형 이변량 베타분포에 따른다고 가정하고, $X_1, X_2|\pi_1, \pi_2$의 조건부 분포를 이변량 이항분포를 통해 모형화할 때 이로부터 유도된 (X_1, X_2)의 결합확률함수(이변량 베타이항모형: B-bB)는 다음과 같다.

$$f(x_1, x_2) = f_1(x_1) f_2(x_2) \left\{ 1 + \omega_b \frac{(x_1 - \mu_1^b)(x_2 - \mu_2^b)}{(\alpha_1 + \beta_1 + n_1)(\alpha_2 + \beta_2 + n_2)} \right\}. \tag{2.2}$$

2.1.2. 이변량 가법이항모형(Bivariate additive Binomial: B-aB) 일반화가법이항모형 (Altham, 1978)은 Z_{ij}들 간 교호효과를 가법형식으로 처리하여 급내상관을 다룬다. 여기서 2차까지의 교호효과만을 고려하고, Z_{ij}들 간 3차 이상의 결합확률이 대칭임을 가정하여 모형을 간소화할 때, X_i의 주변분포는 다음과 같다.

$$f_i(x_i) = \binom{n_i}{x_i} (\pi_i^a)^{x_i} (1 - \pi_i^a)^{n_i - x_i} \left[\frac{\gamma_i^a}{2} \left\{ \frac{x_i(x_i - 1)}{\pi_i^a} + \frac{(n_i - x_i)(n_i - x_i - 1)}{1 - \pi_i^a} \right\} - \frac{\gamma_i^a n_i(n_i - 1)}{2} + 1 \right] \tag{2.3}$$

단, $1 - \pi_i^a = p_{i00} + p_{i01}$, $p_{i11} = \gamma_i^a\pi_i^a + (1 - \gamma_i^a)(\pi_i^a)^2$, $p_{i00} = \gamma_i^a(1 - \pi_i^a) + (1 - \gamma_i^a)(1 - \pi_i^a)^2$,
$p_{iab} = \Pr(Z_{ij} = a, Z_{ij'} = b)$, $j \neq j'$; $a, b = 0, 1$, $x_i = 0, 1, \ldots, n_i$.

여기서 모수 $\gamma_i^a = \rho_i^a$, 즉, 가법이항모형에서의 급내상관계수를 나타내며, 이 값은 식 (2.3)이 확률함수가 되기 위해 다음 조건을 충족해야 한다.

$$-\min\left(\frac{1 - \pi_i^a}{\pi_i^a}, \frac{\pi_i^a}{1 - \pi_i^a}\right) \leq \rho_i^a \leq 1, \quad \frac{-2}{n_i(n_i - 1)} \min\left(\frac{1 - \pi_i^a}{\pi_i^a}, \frac{\pi_i^a}{1 - \pi_i^a}\right) \leq \rho_i^a \leq U(n_i, 1 - \pi_i^a) \tag{2.4}$$

$$\text{단, } n_i \to \infty \text{일 때 } U(n_i, 1-\pi_i^a) \simeq 2\left\{n_i + \frac{(1-2\pi_i^a)^2}{4\pi_i^a(1-\pi_i^a)}\right\}^{-1}.$$

이와 같은 모형에 대한 식 (2.3)은 Z_{ij}들 간의 모든 가능한 차수의 급내상관관계를 반영하는 확률함수 (Bahadur, 1961)에서 3차 이상의 급내상관을 무시하고 짝지은(2차) 급내상관이 모두 같을 경우에 얻어지는 근사확률분포(상관이 있는 이항모형, Kupper와 Haseman (1978); 김병수 등 (1995))와 동일하다. 여기서 $E(X_i) = n_i\pi_i^a = \mu_i^a$, $\mathrm{Var}(X_i) = n_i\pi_i^a(1-\pi_i^a)\{1+(n_i-1)\gamma_i^a\}$이며, 이를 확장한 Sarmanov형 이변량 일반화가법이항모형(B-aB)에 대한 결합확률함수는 다음과 같다.

$$f(x_1, x_2) = f_1(x_1)f_2(x_2)\left\{1 + \omega_a(x_1-\mu_1^a)(x_2-\mu_2^a)\right\}. \tag{2.5}$$

2.1.3. 이변량 승법이항모형(Bivariate multiplicative Binomial: B-mB) 베르누이 시행들 간의 교호효과를 승법형식으로 정의하여 급내상관을 처리한 일반화승법이항모형 (Altham, 1978)은 가법모형에서와 같이 Z_{ij}에 대한 3차 이상의 결합확률들의 대칭성과 2차 교호효과만을 가정할 경우 Z_{ij}들의 결합확률은 지수족의 형태를 취하며, 이에 대한 일반화승법이항모형에 대한 확률함수는 다음과 같이 표현된다.

$$f_i(x_i) = \binom{n_i}{x_i}\frac{(\pi_i^m)^{x_i}(1-\pi_i^m)^{n_i-x_i}(\gamma_i^m)^{x_i(n_i-x_i)}}{d_i(\pi_i^m, \gamma_i^m, n_i)}, \tag{2.6}$$

$$\text{단, } x_i = 0, 1, \ldots, n_i; i = 1, 2,\ d_i(\pi_i^m, \gamma_i^m, n_i) = \sum_{x_i=0}^{n_i}\binom{n_i}{x_i}(\pi_i^m)^{x_i}(1-\pi_i^m)^{n_i-x_i}(\gamma_i^m)^{x_i(n_i-x_i)}.$$

이때, $\mu_i^m = E(X_i) = n_i\pi_i^m\{\pi_i^m+(1-\pi_i^m)\pi_i^m\}^{n_i-1}d_i[\pi_i^m/\{\pi_i^m+(1-\pi_i^m)\gamma_i^m\}, \gamma_i^m, n_i-1]/d_i(\pi_i^m, \gamma_i^m, n_i)$, $E(X_i^2) = \mu_i^m + n_i(n_i-1)(\pi_i^m)^2\{\pi_i^m + (1-\pi_i^m)(\gamma_i^m)^2\}^{n_i-2}d_i[\pi_i^m/\{\pi_i^m+(1-\pi_i^m)(\gamma_i^m)^2\}, \gamma_i^m, n_i - 2]/d_i(\pi_i^m, \gamma_i^m, n_i)$.

여기서 모수 γ_i^m는 독립적 베르누이의 경우에는 1, 양의 급내상관의 경우에는 1보다 작은 값을 그리고 음의 급내상관의 경우에는 1보다 큰 값을 취한다. 특히 이 모형에서 3차 이상의 급내상관관계를 무시하고 공통급내상관계수를 ρ_i^m이라 하면 모수

$$\gamma_i^m = \frac{\pi_i^m(1-\pi_i^m)(1-\rho_i^m)}{\sqrt{\{(1-\pi_i^m)^2+\rho_i^m\pi_i^m(1-\pi_i^m)\}\{(\pi_i^m)^2+\rho_i^m\pi_i^m(1-\pi_i^m)\}}}$$

의 관계를 가지는데, 여기서 분모의 제곱근 안의 식은 양이 되어야 하므로 다음 조건이 요구된다.

$$\left\{(1-\pi_i^m)^2+\rho_i^m\pi_i^m(1-\pi_i^m)\right\}\left\{(\pi_i^m)^2+\rho_i^m\pi_i^m(1-\pi_i^m)\right\} > 0. \tag{2.7}$$

다음은 이를 확장한 Sarmanov형 이변량 일반화승법이항모형(B-mB)의 결합확률함수이다.

$$f(x_1, x_2) = f_1(x_1)f_2(x_2)\left\{1 + w_m(x_1-\mu_1^m)(x_2-\mu_2^m)\right\}. \tag{2.8}$$

2.1.4. 이변량 일반화확장베타이항모형(Bivariate extended beta Binomial: B-ebB) 앞서 고려한 베타이항모형이 음의 급내상관에 대해 가지는 제약성을 완화하기 위한 하나의 방안으로 Prentice (1986)는 식 (2.1)을 재모수화(re-parameterization)하여 얻은 다음 형식의 확장베타이항모형을 고려하였다.

$$f_i(x_i) = \binom{n_i}{x_i}\prod_{k=0}^{x_i-1}(\pi_i^e+k\gamma_i^e)\prod_{k=0}^{n_i-x_i-1}(1-\pi_i^e+k\gamma_i^e)\Big/\prod_{k=0}^{n_i-1}(1+k\gamma_i^e), \tag{2.9}$$

표 2.1. 베르누이 변인수와 성공확률에 따른 모형별 급내상관계수의 범위

베르누이 변인수	성공확률	모형별 급내상관계수			
		B-bB	B-aB	B-mB	B-ebB
2	0.1	>0	$-0.1\sim0.5$	$-0.1\sim0.9$	$-0.1\sim0.9$
	0.2	>0	$-0.2\sim0.7$	$-0.2\sim0.9$	$-0.2\sim0.9$
	0.3	>0	$-0.4\sim0.9$	$-0.4\sim0.9$	$-0.4\sim0.9$
	0.4	>0	$-0.6\sim0.9$	$-0.6\sim0.9$	$-0.6\sim0.9$
	0.5	>0	$-0.9\sim0.9$	$-0.9\sim0.9$	$-0.9\sim0.9$
	0.6	>0	$-0.6\sim0.9$	$-0.6\sim0.9$	$-0.6\sim0.9$
	0.7	>0	$-0.4\sim0.9$	$-0.4\sim0.9$	$-0.4\sim0.9$
	0.8	>0	$-0.2\sim0.7$	$-0.2\sim0.9$	$-0.2\sim0.9$
	0.9	>0	$-0.1\sim0.5$	$-0.1\sim0.9$	$-0.1\sim0.9$
3	0.1	>0	$0.0\sim0.4$	$-0.1\sim0.9$	$0.0\sim0.9$
	0.2	>0	$0.0\sim0.5$	$-0.2\sim0.9$	$-0.1\sim0.9$
	0.3	>0	$-0.1\sim0.6$	$-0.4\sim0.9$	$-0.1\sim0.9$
	0.4	>0	$-0.2\sim0.6$	$-0.4\sim0.9$	$-0.2\sim0.9$
	0.5	>0	$-0.3\sim0.6$	$-0.4\sim0.9$	$-0.3\sim0.9$
	0.6	>0	$-0.2\sim0.6$	$-0.4\sim0.9$	$-0.2\sim0.9$
	0.7	>0	$-0.1\sim0.6$	$-0.4\sim0.9$	$-0.1\sim0.9$
	0.8	>0	$0.0\sim0.5$	$-0.2\sim0.9$	$-0.1\sim0.9$
	0.9	>0	$0.0\sim0.4$	$-0.1\sim0.9$	$0.0\sim0.9$
4	0.1	>0	$0.0\sim0.3$	$-0.1\sim0.9$	$0.0\sim0.9$
	0.2	>0	$0.0\sim0.4$	$-0.2\sim0.9$	$0.0\sim0.9$
	0.3	>0	$0.0\sim0.4$	$-0.3\sim0.9$	$-0.1\sim0.9$
	0.4	>0	$-0.1\sim0.4$	$-0.3\sim0.9$	$-0.1\sim0.9$
	0.5	>0	$-0.1\sim0.5$	$-0.3\sim0.9$	$-0.2\sim0.9$
	0.6	>0	$-0.1\sim0.4$	$-0.3\sim0.9$	$-0.1\sim0.9$
	0.7	>0	$0.0\sim0.4$	$-0.3\sim0.9$	$-0.1\sim0.9$
	0.8	>0	$0.0\sim0.4$	$-0.2\sim0.9$	$0.0\sim0.9$
	0.9	>0	$0.0\sim0.3$	$-0.1\sim0.9$	$0.0\sim0.9$

여기서 (π_i^e, γ_i^e)가 다음 관계를 가지는 실수이면 식 (2.9)는 확률함수의 조건을 충족하게 된다.

$$\gamma_i^e \geq \max\left[-\frac{\pi_i^e}{n_i-1}, -\frac{1-\pi_i^e}{n_i-1}\right] \tag{2.10}$$

이때 급내상관계수 $\rho_i^e = \gamma_i^e/(1+\gamma_i^e)$가 되므로 이 모형은 자연히 급내상관계수가 음인 일부영역을 포함하게 된다. 이를 확장한 Sarmanov형 이변량 일반화확장베타이항모형(B-ebB)의 결합확률함수는 $E(X_i) = n_i\pi_i^e = \mu_i^e$, $i = 1, 2$에 대하여 다음 형식을 가진다.

$$f(x_1, x_2) = f_1(x_1)f_2(x_2)\{1 + w_e(x_1 - \mu_1^e)(x_2 - \mu_2^e)\}. \tag{2.11}$$

이 이외에도 일반화가법이항모형과 성공확률에 가정한 베타분포를 혼합하여 3개의 모수를 가지는 '상관이 있는 베타이항모형'(Beta-correlated Binomial model, Paul (1987)) 등이 제안된 바 있으나, 이 연구에서는 고려하지 않는다.

2.2. 모형별 적용가능한 급내상관의 범위

앞서 고려한 일반화이항모형에서는 각 X_i를 구성하는 n_i개의 Z_{ij}들 간의 급내상관계수가 동일하다고

표 3.1. 외국인투자자의 순매수횟수와 주가상승횟수

주가의 상승횟수(X_1)	외국인투자자의 순매수횟수(X_2)				
	0	1	2	3	계
0	9	18	11	5	43
1	35	53	71	44	203
2	14	40	44	64	162
3	2	3	20	19	44
계	60	114	146	132	452

가정하므로 Z_{ij}들에 관한 크기 $n_i \times n_i$ 등상관행렬에 대해 기본적으로 요구되는 양정치성에 따라 공통 급내상관계수는 일단 $[-1/(n_i - 1), 1]$의 범위에 제약된다. 이와 더불어 모형별 적용가능한 급내상관계수의 범위는 각 확률모형에 대한 조건(식 (2.4), (2.7), (2.10))에 제약되어있어(B-bB 제외) 베르누이 변인들의 수(n_i)와 각 모형에서의 베르누이 성공확률에 연계된 모수 $[\pi_i^b, \pi_i^a, \pi_i^m, \pi_i^e]$에 따라 다르게 된다. 그러므로 실제자료를 적합하기 위해 선정된 특정모형이 가지는 타당성은 급내상관계수의 크기가 중요한 기준중의 하나가 된다고 할 수 있다.

표 2.1은 베르누이 변인수와 성공확률의 조합별로 계산한 급내상관계수의 적용가능한 범위를 나타내고 있다. 이 표에서는 제한된 크기의 베르누이 변인수를 고려하고 있지만 성공확률의 모든 가능한 경우에 걸쳐 급내상관계수의 범위는 그 크기가 일관되게 B-mB > B-ebB > B-aB의 순서를 유지하고 있음을 관찰할 수 있다. 구체적으로 모형 B-mB는 B-ebB에 비해 상한은 비슷한 수준이나 하한의 영역이 좀 더 확대되어 있고, 모형 B-aB에 비해서는 상, 하한 모두 크게 확대되어있다. 전반적으로 보아 모형 B-mB가 가장 넓은 범위의 급내상관을 수용하고 있다고 하겠다. 또 베르누이 변인수가 커짐에 따라, 그리고 같은 베르누이 변인의수 내에서는 성공확률이 상하한값 어느 방향으로 크게 치우친 값을 가짐에 따라 허용 가능한 급내상관계수의 범위는 세 모형 모두 작아지는 경향을 볼 수 있다.

3. 자료분석

다음은 두개의 실제자료를 위에서 고려한 4개 모형: B-bB, B-aB, B-mB, B-ebB에 적합한 결과이다. 규정된 모수의 ML추정치는 각 모형에 대한 로그가능도함수 $\sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} n_{x_1 x_2} \ln f(x_1, x_2)$ (단, $n_{x_1 x_2}$는 셀(x_1, x_2)의 관찰도수)에 Newton방법(EXCEL: Solver)을 사용하여 계산하였다. 이 방법에 따라 특정모형에 의해 추정된 확률분포는 $n_1 = n_2 = 1$로 놓았을 때 얻어지는 다음 관계를 통해 예측에 이용할 수 있다.

$$P(X_1 = 1) = P(X_1 = 1 | X_2 = 0) + \{P(X_1 = 1 | X_2 = 1) - P(X_1 = 1 | X_2 = 0)\} P(X_2 = 1). \quad (3.1)$$

3.1. 주식자료

국내 주식시장에서 외국인 주식거래가 있는 종목들에 대해 당일의 순매수 여부가 다음 시점에서의 주가 등락과 가지는 관계를 통하여 종합주가지수의 행보를 예측하는 것은 의미 있는 일이다. 표 3.1의 '주식자료: 증권선물거래소, 2006'는 증권거래소에 상장되어 있는 종목 중 연속한 3일(2006년 9월 6/7/8일) 동안 관리종목이나 외국인투자자의 거래가 발생하지 않았던 것을 제외한 총 452개를 대상으로 동기간 중 외국인투자자의 순매수횟수($X_2 = Z_{21} + Z_{22} + Z_{23}$)와 동년월 7/8/11일의 3일간(9월 9일은 토요일로 증시폐장) 주가상승횟수($X_1 = Z_{11} + Z_{12} + Z_{13}$)에 관한 자료이다. 여기서 각 일자별 외국인투자자의 순매수여부를 나타내는 Z_{2j}는 순매수량이 0 이상인 경우 1, 순매도인 경우에는 0을; 그리고 주가상

표 **3.2.** 주식자료에 대한 이변량 일반화이항모형의 적합

모형		B-bB	B-aB	B-mB	B-ebB
로그 가능도비 통계량(LL)		−1127.592	−1123.028	−1122.813	−1122.824
	LL_0(ω[3]= 0; 독립)	−1144.044	−1139.950	−1139.734	−1139.747
	$2(LL - LL_0)$	33.169	33.845	33.842	33.845
적합도 검정	카이제곱 통계량	22.380	15.291	14.607	14.690
	유의확률	0.013	0.122	0.147	0.144
	자유도	10	10	10	10
AIC		2265.183	2256.056	2255.626	2255.649
SBC		2285.752	2276.624	2276.194	2276.217
추정치(ML)		$\hat{\alpha}_1 = 26949.1$	$\hat{\pi}_1^a = 0.487$	$\hat{\pi}_1^m = 0.485$	$\hat{\pi}_1^e = 0.487$
		$\hat{\alpha}_2 = 2.069$	$\hat{\pi}_2^a = 0.590$	$\hat{\pi}_2^m = 0.564$	$\hat{\pi}_2^e = 0.591$
		$\hat{\beta}_1 = 28674.2$	$\hat{\gamma}_1^a = -0.077$	$\hat{\gamma}_1^m = 1.185$	$\hat{\gamma}_1^e = -0.072$
		$\hat{\beta}_2 = 1.436$	$\hat{\gamma}_2^a = 0.203$	$\hat{\gamma}_2^m = 0.694$	$\hat{\gamma}_2^e = 0.260$
		$\hat{\omega}_b = 98682$	$\hat{\omega}_a = 0.306$	$\hat{\omega}_m = 0.305$	$\hat{\omega}_e = 0.305$

표 **3.3.** 적합된 B-mB모형에 의해 추정된 확률($n_1 = n_2 = 1$의 경우)

주가의 상승(X_1)	외국인투자자의 순매수(X_2)		
	0	1	계
0	0.243	0.272	0.515
1	0.193	0.292	0.485
계	0.436	0.564	1

승여부에 관한 Z_{1j}는 전날대비 주가가 보합 내지 상승하는 경우에는 1, 하락한 경우 0의 값을 가진다. 이때 (X_1, X_2) 간에는 급간상관을 그리고 (Z_{i1}, Z_{i2}, Z_{i3}) 간의 급내상관을 상정할 수 있다.

이 자료에 대한 각 이변량 일반화이항모형별 적합결과가 표 3.2에 주어진다. 우선 적합도 검정통계량의 유의확률(p-값)을 보면 B-aB, B-mB, B-ebB 모형의 경우 각각 0.122, 0.147, 0.144로 거의 비슷한 수준으로 유효히 적합되고 있으며, AIC, SBC 통계량 역시 같은 상황을 보이고 있다. 또한 이들에 대한 ML 추정치 $\hat{\gamma}_1^a < 0$, $\hat{\gamma}_1^m > 1$, $\hat{\gamma}_1^e < 0$임을 보아 주가상승여부를 나타내는 $Z_{1j}, j = 1, 2, 3$ 간에는 음의 급내상관계가 있음을 알 수 있다. 이에 반하여 B-bB 모형은 유의확률이나 AIC, SBC의 측면에서 모형적합도가 현저히 떨어지고 있으며, 특히 $\hat{\alpha}_1$, $\hat{\beta}_1$이 매우 큰 이유도 Z_{1j}들 간에 실제 음의 급내상관을 B-bB 모형이 허용하는 급내상관계수의 하한(0)에 가깝도록 추정하였기 때문인 것으로 판단된다. 이와는 달리 외국인투자자의 순매수여부(Z_{2j})에서는 $\hat{\gamma}_2^a > 0$, $\hat{\gamma}_2^m < 1$, $\hat{\gamma}_2^e > 0$의 값으로부터 양의 급내상관이 있음을 확인할 수 있으며, 특히 이 경우 B-bB 모형에서 급내상관계수는 $\hat{\alpha}_2$, $\hat{\beta}_2$을 통해 적정 추정됨을 볼 수 있다. 한편, 4개 모형 모두에 걸쳐 (X_1, X_2) 간의 급간상관은 이에 연계된 모수 ω_b, ω_a, ω_m, ω_e 각각을 0으로 제약했을 경우에 대한 로그 가능도비 $2(LL - LL_0)$가 모두 $3.84\,(= \chi_1^2(0.05))$보다 크므로 유의한 급간상관관계가 존재한다고 할 수 있다.

표 3.2에서와 같이 적합된 모형은, 예를 들어, 452개 종목에 대한 9월 11일 외국인투자자의 순매수종목 비율이 주어질 때 그 다음 시점인 9월 12일 주가등락의 비율을 예측하는 데에 이용할 수 있다. 여기서는 모형적합도가 가장 양호한 일반화승법이항모형(B-mB)으로부터 예측한 결과만을 제시한다. 식 (2.8)에서 n_1과 n_2를 1로 고정하고 표 3.2의 추정치를 통해 확률분포를 계산한 결과는 표 3.3과 같다. 실제 증

[3] ω는 X_1, X_2간의 독립성을 나타내는 모수로서 각 모형별로 w_b, w_a, w_m, w_e에 해당한다.

표 3.4. 연속 4회 식품점방문에서 베이컨과 계란구매횟수[4]

베이컨구매횟수(X_1)	계란구매횟수(X_2)					
	0	1	2	3	4	계
0	254	115	42	13	6	430
1	34	29	16	6	1	86
2	8	8	3	3	1	23
3	0	0	4	1	1	6
4	1	1	1	0	0	3
계	297	153	66	23	9	548

권선물거래소의 9월 11일 자료에 따르면 총 452개 종목에 대한 외국인투자자의 실제 순매수종목수는 236개(순매수종목비율은 52.2%)로 나타났다. 이때 9월 12일의 주가상승비율 $P(X_1 = 1)$은 식 (3.1)과 표 3.3로부터,

$$\frac{0.193}{0.463} + \left(\frac{0.292}{0.564} - \frac{0.193}{0.436}\right) \times 0.522 = 0.482$$

로 추정된다. 즉, 이는 약 218개(452의 48.2%)의 주가상승종목이 있을 것으로 예상되며, 이에 따라 50%를 기준으로 주가하락종목이 더 많아 결과적으로 종합주가지수는 하락할 가능성이 높다고 단순 추정할 수 있다. 실제로 9월 12일 452개 중 주가상승을 기록한 종목수는 221개이고, 이날 종합주가지수는 20.81p 하락하였다.

3.2. 베이컨과 계란구매에 관한 소비자패널자료

Danaher과 Hardie (2005)는 소비자패널자료에서 총 548명으로부터 연속 4회의 식품점방문동안 베이컨구매횟수($X_1 = \sum_{j=1}^{4} Z_{1j}$)와 계란구매횟수($X_2 = \sum_{j=1}^{4} Z_{2j}$)에 관한 자료 표 3.4를 B-bB 모형을 통해 적합한 바 있다.

표 3.5의 적합결과는 B-bB, B-mB, B-ebB의 적합[5]이 양호함(유의확률과 AIC, SBC 참조)을 나타내고 있다. 이에 반해 모형 B-aB가 부적합함은 가법이항모형이 지나치게 강한 급내상관이 있거나, 0의 빈도비율이 높은 자료에 대해 부적절할 수 있다는 실증분석보고 (Altham, 1978)에 비춰 이 자료의 경우 총 빈도수의 46%가 셀 $(0,0)$에 집중되고 있다는 사실이 가능한 주원인으로 지목할 수 있다[6] 한편, 급간상관의 측면에서, 모수 ω_b, ω_m, ω_e에 대한 $2(LL - LL_0)$는 모두 26.1로 $3.84(= \chi_1^2(0.05))$보다 크므로 베이컨과 계란의 구매횟수사이에 유의한 상관관계가 인정되며, 추정 상관계수는 약 0.22로 나타났다.

표 3.6은 $n_1 = n_2 = 1$일 때 표 3.5의 추정치를 통해 확률분포를 계산한 결과이다. 계란할인판매를 시작하면 경험상 그 기간 중 계란구매확률이 평소의 두 배 $0.638(= 2 \times 0.319)$라면 베이컨구매확률은 식 (3.1)과 표 3.6에 따라,

$$\frac{0.120}{0.681} + \left(\frac{0.080}{0.319} - \frac{0.120}{0.681}\right) \times 0.683 = 0.224$$

로 예상된다. 이는 베이컨구매확률이 표 3.6의 0.201에서 11.4%p$(= (0.224 - 0.201)/0.201)$ 증가함을 뜻하므로, 식품점관리자가 이런 정보에 근거하여 계란할인판매기간 동안 베이컨재고량을 평소보다 11.4%p 증가시킨다면 그 기간 동안 베이컨품절로 인한 잠재적인 손실을 피할 수 있다는 것이다.

[4] 식품점방문은 구입총액이 5달러 이상인 구매활동을, 계란의 구매는 보통 계란 한 꾸러미의 구매를 의미한다.
[5] 급내상관계수가 양일 때 B-bB와 B-ebB의 적합결과는 동일하다.
[6] 이 자료에서 급내상관계수는 두 변수 모두 $0.21 \sim 0.34$ 정도로 추정된다.

표 3.5. 베이컨과 계란구매자료에 대한 이변량 일반화이항모형의 적합[7]

모형		B-bB	B-aB	B-mB	B-ebB
로그 가능도비 통계량(LL)		−949.966	−954.561	−949.292	−949.966
적합도 검정	$LL_0(\omega = 0;$ 독립$)$	−963.019	−967.606	−962.346	−963.019
	$2(LL - LL_0)$	26.107	26.091	26.108	26.107
	카이제곱 통계량	3.350	13.085	1.989	3.350
	유의확률	0.764	0.042	0.921	0.764
	자유도	6	6	6	6
AIC		1909.931	1919.122	1908.585	1909.931
SBC		1931.462	1940.654	1930.116	1931.462
추정치(ML)		$\hat{\alpha}_1 = 0.271$	$\hat{\pi}_1^a = 0.136$	$\hat{\pi}_1^m = 0.214$	$\hat{\pi}_1^e = 0.137$
		$\hat{\alpha}_2 = 0.706$	$\hat{\pi}_2^a = 0.235$	$\hat{\pi}_2^m = 0.319$	$\hat{\pi}_2^e = 0.232$
		$\hat{\beta}_1 = 1.707$	$\hat{\gamma}_1^a = 0.304$	$\hat{\gamma}_1^m = 0.367$	$\hat{\gamma}_1^e = 0.506$
		$\hat{\beta}_2 = 2.333$	$\hat{\gamma}_2^a = 0.207$	$\hat{\gamma}_2^m = 0.596$	$\hat{\gamma}_2^e = 0.329$
		$\hat{\omega}_b = 10.588$	$\hat{\omega}_a = 0.439$	$\hat{\omega}_m = 0.441$	$\hat{\omega}_e = 0.441$

표 3.6. 적합된 B-mB모형에 의해 추정된 확률($n_1 = n_2 = 1$의 경우)

베이컨의 구매(X_1)	계란의 구매(X_2)		
	0	1	계
0	0.561	0.239	0.799
1	0.120	0.080	0.201
계	0.681	0.319	1

4. 결론

급내/급간상관이 동시에 존재하는 이변량 이항자료에 대한 모형으로 4개의 Sarmanov형 B-bB, B-aB, B-mB, B-ebB 등을 고려하였다. 이 경우 적합하고자 하는 모형의 유용성은 자료 내 급내상관계수의 크기와 부호에 따라 제한적임을 발견할 수 있었다. 특히 B-bB는 급내상관이 음인 경우 정당성을 가지지 못한 모형이라 할 수 있으며, 비교적 간단한 개념의 B-aB는 2차원 빈도자료에서 영셀의 비중에 대해 민감함이 발견되었다. 그러나 급내상관이 극단적으로 크거나 빈도가 어느 한쪽으로 크게 치우쳐 있는 자료가 아닌 한 4개 모형 모두 동등한 수준의 좋은 적합을 가짐이 확인되었다. 각 모형이 수용하는 급내상관의 크기는 베르누이 변인의 개수와 성공확률에 연계된 모수의 크기에 의존하고 있다. 이런 점에서 상대적으로 넓은 허용범위의 급내상관을 가지는 B-mB가 선호된다고 할 수 있다. 실제자료의 적합에서 음의 급내상관이 있는 주식자료의 경우 B-aB, B-mB, B-ebB가 B-bB에 비해 우수한 성능을 가지는 것으로 나타났다. 한편, 양의 급내상관이 있는 베이컨과 계란구매자료의 경우 B-bB, B-mB, B-ebB는 모두 비슷한 수준의 우수한 적합을 보였으나, 영셀의 비중이 큰 B-aB 모형은 현격히 열등한 결과를 나타냈다. 그러나 각 변수에 포함된 베르누이 변인의 개수가 점차적으로 증가하면 이에 따른 공통급내상관계수가 자연히 [0,1]의 범위에 있게 되므로, 이 경우 4개의 모형 중 어떤 것을 사용하더라도 큰 차이는 없을 것으로 기대된다. 일반적으로 B-bB나 B-ebB는 베르누이 성공확률에 대해 가공적 베타분포를 가정하고 있다. 반면, B-aB와 B-mB는 급내상관의 영향을 모형화하는데 있어 각 시점간의 차이를 고려하지 않고, 모든 가능한 2차 교호작용을 하나의 모수로 설명하는 구조를 갖고 있다. 실제로 급내상관의 정도가 각 시점의 차이에 따라 다르게 나타나는 경우 그 성능이 떨어지는 것으로 나타났다. 시점간의 차이

[7]셀기댓값이 5이하인 셀의 경우가 많아 빈도표의 4, 5번째 행과 5번째 열을 인접셀에 병합한 후 적합하였다.

에 의존하여 서로 다른 급내상관이 존재하는 경우 모수를 절약하면서 이를 적절히 제어할 수 있는 모형의 개발이 향후 연구과제이다.

참고문헌

김병수, 오경주, 박철용 (1995). 가산자료(count data)의 과산포 검색: 일반화 과정, <응용통계연구>, **8**, 147–161.

Altham, P. M. E. (1978). Two generalizations of the binomial distribution, *Applied Statistics*, **27**, 162–167.

Bahadur, R. R. (1961). *A Representation of the Joint Distribution of Responses to n Dichotomous Items*, In Studies in Item Analysis & Prediction, Stanford University Press, Stanford.

Danaher, P. J. and Hardie, G. S. (2005). Bacon with your eggs? Applications of a new bivariate beta-binomial distribution, *The American Statistician*, **59**, 282–286.

Farlie, D. J. G. (1960). The performance of some correlation coefficients for a general bivariate distribution, *Biometrika*, **47**, 307–323.

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease, *Biometrics*, **29**, 637–648.

Kupper, L. L. and Haseman, J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments, *Biometrics*, **34**, 69–76.

Lee, M. L. T. (1996). Properties and applications of the Sarmanov family of bivariate distributions, *Communications in Statistics-Theory and Methods*, **25**, 1207–1222.

Paul, S. R. (1987). On the beta-correlated binomial distribution - A three parameter generalization of the binomial distribution, *Communication in Statistics-Theory and Methods*, **16**, 1473–1478.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors, *Journal of the American Statistical Association*, **81**, 321–327.

Sarmanov, O. V. (1966). Generalized normal correlation and two-dimensional Frèchet classes, *Doklady (Soviet Mathematics)*, **168**, 596–599.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of a success as variable between the sets of trials, *Journal of the Royal Statistical Society*, **B10**, 257–261.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity, *Biometrics*, **31**, 949–952.

Fitting Bivariate Generalized Binomial Models of the Sarmanov Type

JooYong Lee[1] · KeeYoung Kim[2]

[1]Risk Management Team, KDB Asset Management Co., Ltd.;
[2]Department of Statistics, Korea University

(Received December 2008; accepted February 2009)

Abstract

For bivariate binomial data with both intra and inter-class correlation, Danaher and Hardie (2005) proposed a bivariate beta-binomial model. However, the model is limited to the situation where the intra-class correlation is strictly positive. Thus it might be seriously inadequate for data with a negative intra-class correlation. Several authors have considered generalized binomial distributions covering a wider range of intra-class correlation which could relax the possible model restrictions imposed. Among others there are the additive/multiplicative and the beta/extended beta binomial model. In this study, bivariate models of the Sarmanov (1966) type are formed by combining each of those univariate models to take care of the inter-class correlation, and are evaluated in terms of the goodness-of-fit. As a result, B-mB and B-ebB are fitted, successfully, to real data and that B-mB, which has a wider permissible range than B-ebB for the intra-class correlation is relatively preferred.

Keywords: Intra/Inter-class correlation, generalized binomial distributions, Sarmanov type of bivariate models.

This work was supported by 2007 Korea University Research Grant.

[2]Corresponding author: Professor, Department of Statistics, Korea University, 5-1 Anamdong, Seoul 136-701, Korea. E-mail: kykim@korea.ac.kr