한국통계학회논문집 제4권 3호
1997년 12월 pp. 611 - 616

불균형일원변량모형에서 분산성분비율의 추정[1)]

이 장 택[2)]

요 약

불균형일원변량모형에서 분산성분비율의 점추정에 관한 문제가 고려되어진다. 분산성분비율에 대한 새로운 추정량이 제안되며, 분산성분비율에 대한 여러 가지 점추정량과 제안된 추정량을 평균자승오차(MSE)의 관점에서 추정량들의 효율성을 모의실험을 통하여 살펴본다. 결론적으로 제안된 추정량은 수준의 수가 크고 불균형정도가 매우 심한 경우를 제외하고 다른 추정량들보다 훨씬 MSE 효율성이 높아짐을 알 수 있다.

1. 서론

반복수가 같지 않은 i번째 처리효과에 있어서 j번째 관측치에 대한 불균형일원변량모형은 다음과 같이 나타낼 수 있다.

$$y_{ij} = \mu + a_i + e_{ij}, \qquad i = 1, 2, \cdots, k, \; j = 1, \cdots, n_i \tag{1.1}$$

위의 모형에서 μ는 미지의 모수, a_i와 e_{ij}는 서로 독립이며, 평균이 0이고 분산이 각각 σ_a^2와 σ_e^2인 정규확률변수, n_i는 i번째 처리효과에 있어서의 관측치의 개수이다. 위와 같은 불균형일원변량모형에 대하여 주로 논의의 대상이 되는 분산성분의 함수는 분산성분비율 $\theta = \sigma_a^2/\sigma_e^2$와 급내상관계수 $\rho = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$이다. 분산성분비율 θ는 분산성분에 관한 신뢰한계를 구하려고 할 때, 신뢰한계에서 가장 빈번하게 나타나며, 동물학을 연구하는 학자들은 이 비율을 유전력이라고도 한다. 또한 생태학 연구에서는 생물학적 또는 환경학적 특성에 관하여 가족의 닮은 정도를 측정하는데 일반적으로 사용되어지며 심리학에서는 신뢰성이론 분야에서 매우 중요한 역할을 한다.

균형일원변량모형에 대한 분산성분비율의 추정문제는 많은 학자들에 의해 연구되어졌다. 여기에 대한 자세한 결과는 Das(1992) 또는 이장택(1996)의 논문에 상세히 소개되어 있다. 하지만 불균형일원변량모형인 경우에서의 분산성분비율의 추정문제에 대한 연구결과는 알려져 있는 것이 거의 없다. 왜냐하면 분산성분비율의 추정에 사용되는 분산분석표의 분산비통계량에 대한 정확한 분포를 알 수 없기 때문이다. 따라서 분산분석표의 부산물로써 구할 수 있는 분산분석추정량이나 분산성분비율의 추정량을 정확히 구할 수 있는 최우추정량이나 제한최우추정량이 널리 사용되고 있는 실정이다. 이와 같은 이유로 불균형일원변량모형에서의 분산성분비율의 추정문제에 대한 새로운

1) 이 연구는 1997년 단국대학교 대학연구비의 지원으로 연구되었음
2) (140-714) 서울시 용산구 한남동 산 8번지 단국대학교 전산통계학과 교수

접근방법이 필요하며, 본 논문에서는 분산성분비율에 대한 새로운 추정량을 제안한다. 제안된 추정량은 이장택(1996)이 고려한 균형일원변량모형에서의 분산성분비율의 추정량을 구하는 방법을 이용하여 자료가 불균형자료인 경우에 균형자료인 경우처럼 분산분석표에서의 분산성분비 통계량이 F-분포와 관계를 지을 수 있도록 Thomas와 Hultquist(1978)의 근사식을 이용하여 구하였다.

본 논문의 구성은 2절에서는 분산성분비율의 추정량으로서 사용되는 여러 가지 추정량들을 간략히 살펴보며, 3절에서는 분산성분비율의 새 추정량을 제시한다. 4절에서는 모의실험을 통하여 기존의 여러 가지 추정량과 새로 제안되는 추정량을 MSE 판정아래에서 그 효율성을 알아본다. 아울러 제안된 추정량이 최우추정량에 비해 효율성이 높아지는 조건들을 살펴본다. 그리고 끝으로 5절에서는 결론을 제시한다.

2. 여러 가지 추정량

이 절에서는 지금까지 θ의 점추정량으로써 많이 사용되는 추정량들을 알아보기로 한다.

A. 분산분석추정량 (ANOVA)

모형(1.1)에 대한 분산분석표를 작성하면 다음 [표2.1]과 같다.

[표2.1] 불균형일원변량모형의 분산분석표

요인	자유도	제곱합	제곱평균	평균제곱기대값
처리	$k-1$	$SSA=\sum n_i(\bar{y}_{i.}-\bar{y}_{..})^2$	$MSA=SSA/(k-1)$	$\sigma_e^2+n_0\sigma_a^2$
오차	$N-k$	$SSE=\sum(y_{ij}-\bar{y}_{i.})^2$	$MSE=SSE/(N-k)$	σ_e^2
총합	$N-1$	$SST=\sum(y_{ij}-\bar{y}_{..})^2$		

[표2.1]에서 사용된 표기는 $N=\sum_{i=1}^{k}n_i$는 총 관측치의 개수, $\bar{y}_{i.}=\sum_{j=1}^{n_i}y_{ij}/n_i$는 i번째 그룹의 평균이며, $\bar{y}_{..}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}/N$는 전체관측치의 총평균, 그리고 n_0는 $n_0=(N-\sum_{i=1}^{k}n_i^2/N)/(k-1)$로 정의된 가중평균관측수이다. 분산분석추정량은 분산분석표를 이용하여 제곱합을 구하고 이 값과 평균제곱기대값(EMS)을 같다고 놓고 분산성분을 추정한다. 따라서 분산분석표를 이용하여 구한 σ_e^2의 분산분석추정량은 $\hat{\sigma}_e^{\,2}=MSE$이고, σ_a^2의 추정량은 $\hat{\sigma}_a^{\,2}=(MSA-MSE)/n_0$로 표시된다. 그러므로 분산성분비율 θ의 분산분석추정량 $\hat{\theta}_A$는 각 분산성분에 분산분석추정량을 대입하여 구할 수 있고 $\hat{\theta}_A=(R-1)/n_0$와 같이 표시된다. 여기서 R는 분산분석에 사용되는 분산비통계량 $R=MSA/MSE$이다. 하지만 θ의 분산분석추정치는 음수가 될 수 있으며 이 경우 보통 0으로 판정한다.

B. 최우추정량 (ML)

θ의 최우추정량은 랜덤효과와 오차항이 정규분포를 따른다는 가정아래에서 최우추정법을 사용하여 구한 추정량이다. 그리고 이 추정치는 항상 0 이상의 값을 가지는 장점이 있는 반면에 반복수렴해이기 때문에 발산이 되는 경우에는 해를 구할 수가 없으며 또한 분산분석추정치를 구하는 경우보다 시간이 많이 걸린다. 일원변량모형에 θ의 최우추정량을 구하는 경우에는 σ_a^2와 σ_e^2의 최우추정량을 각각 구하여 θ의 정의에 대입함으로써 구할 수도 있고 또는 θ의 우도함수를 직접 최대화하여 구할 수도 있다.

C. 제한최우추정량 (REML)

θ의 제한최우추정량은 랜덤효과가 정규분포를 따른다는 가정아래에서 구하여진 추정량으로 고정효과부분과 랜덤효과부분을 분리하여 추정하는 것이 최우추정량과 다른 점이다. 이 추정량은 근본적으로 최소분산이차불편추정량과 비슷하며 추정값은 항상 0이상으로 나온다. SAS의 MIXED 절차의 디폴트 추정량으로 사용되며, 일반적으로 가장 많이 사용되는 추정량중의 하나이다.

3. 제안된 추정량

Thomas와 Hultquist(1978)는 불균형일원변량모형에서 분산성분에 대한 신뢰구간을 구하는 문제에 대하여 다음과 같은 통계량이 근사적으로 자유도가 $k-1$인 χ^2분포를 따르는 사실을 밝혔다.

$$W=(k-1)S_y^2/E(S_y^2), \quad (k-1)S_y^2=\sum_i(\bar{y}_{i.}-\bar{y}_{..})^2 \tag{3.1}$$

위 식에서 S_y^2은 처리평균을 이용하여 구한 표본분산이며 또한 기대값은 $E(S_y^2)=\sigma_a^2+\sigma_e^2/\lambda$와 같은데, 이 경우 λ는 $\lambda=k/(\sum_{i=1}^{k}1/n_i)$와 같이 정의되며, 이것은 n_i들의 조화평균이다. 그리고 그들은 이와 같은 근사가 분산성분비율의 값이 $\theta\geq 0.25$인 경우에 매우 정확하다고 주장하였다. 한편 SSE/σ_e^2는 자유도가 $N-k$ 인 카이제곱분포를 따르므로, 확률변수 $G=S_y^2/MSE$는 $(\theta+1/\lambda)F$와 같은 근사분포를 가지는데, 확률변수 F는 자유도가 (v_1, v_2)인 F분포를 따른다. 여기서 v_1과 v_2는 각각 $v_1=k-1$, $v_2=N-k$와 같다.

이 절에서 제안되는 추정량은 δG와 같은 형태에 국한한다. 왜냐하면 이장택(1996)의 논문에서처럼 여러 가지 타입의 추정량을 고려할 수 있으나, 모의실험을 통하여 알아본 바에 의하면 δG의 형태보다 MSE 효율성이 높지 않을뿐더러 수치적분등의 복잡한 계산이 필요하기 때문이다. 그리고 δG 형태의 추정량은 항상 음이 아닌 추정량을 제공한다. 따라서 $E(\delta G-\theta)^2$를 최소로 하는 δ값 δ^*를 δ에 대한 이차방정식을 이용하여 구하면 $\delta^*=K\theta/(\theta+1/\lambda)$, $K=v_1(v_2-4)(v_2(v_1+2))^{-1}<1$, 임을 알 수 있다. 하지만 δ^*는 실제로 θ의 함수이기 때문에 우리는 활용할 수가 없으며, 따라서 $E(\delta^*-\delta_0)^2$를 최소화하는 상수 δ_0를 선택하여 보자. 이 경우 δ_0의 값은 $\delta_0=E(\delta^*)$가 된다는 것을 우리는 알고 있으므로, δ^*의 기대값을 구하기 위하여 급내상관계수 ρ가 $(0,1)$에서 균일분포를 따르는 확률변수라고 가정하면, 따라서

$$E(\delta^*) = E(\frac{K\lambda\rho}{1+\rho(\lambda-1)}) = K\lambda\int_0^1 \frac{\rho}{1+\rho(\lambda-1)}d\rho$$
$$= \frac{K\lambda}{(\lambda-1)^2}\int_0^{\lambda-1}\frac{t}{1+t}dt = \frac{K\lambda}{(\lambda-1)^2}((\lambda-1)-\log(\lambda)).$$

그러므로 제안되는 θ의 추정량 $\hat{\theta}_{emp}$은 다음 식(3.2)과 같이 정의된다.

$$\hat{\theta}_{emp} = \frac{K\lambda}{(\lambda-1)^2}((\lambda-1)-\log(\lambda))G, \quad K = v_1(v_2-4)(v_2(v_1+2))^{-1} < 1 \qquad (3.2)$$

4. 모의실험

이 절에서는 모의실험을 통하여 3절에서 제안된 $\hat{\theta}_{emp}$의 MSE 효율성을 알아보기로 한다. 모의실험에 사용된 불균형실험계획의 모양은 표[4.1]에 표시 되어있다.

[표4.1] 모의실험에 사용된 불균형 실험계획의 모양

모양	k	n_i
P1	3	3, 5, 7
P2	3	1, 5, 9
P3	3	1, 7, 17
P4	6	3, 3, 5, 5, 7, 7,
P5	6	1, 1, 5, 5, 9, 9,
P6	6	1, 1, 7, 7, 17, 17
P7	9	3, 3, 3, 5, 5, 5, 7, 7, 7
P8	9	1, 1, 1, 5, 5, 5, 9, 9, 9
P9	9	1, 1, 1, 7, 7, 7, 17, 17, 17

[표4.1]에서 실험계획 P1부터 P3은 수준의 수가 3이며, P4부터 P6은 수준의 수가 6, P7부터 P9까지는 수준의 수가 9이다. 또한 불균형정도가 약한 실험계획은 P1, P4, P7이며, 중간인 실험계획은 P2, P5, P8, 매우 강한 실험계획은 P3, P6, P9이다. 모의실험은 통계 팩키지 SAS를 이용하였는데 난수발생을 위하여 난수생성함수 RANGAM와 RANNOR를 사용하였다. 또한 θ의 값은 급내상관계수 $\rho = \theta/(1+\theta)$를 이용하여 ρ의 값이 0.1부터 0.9까지 0.1간격으로 선택되어지는 것을 택하였다. 그리고 일반성을 잃지 않고 $\sigma_a^2+\sigma_e^2 = 1$과 $\mu = 0$을 가정하여서 y_{ij}값을 각각 생성하지 않고, 대신 $\bar{y}_{i.}$의 값을 RANNOR을 이용하여 평균이 0이고 분산이 $\rho+(1-\rho)/n_i$인 정규분포로부터 생성하였다. 또한 분산분석표의 *MSE* 값은 $(1-\rho)W/(N-k)$와 같이 표현된다. 여기서 W는 자유도가 $N-k$인 카이제곱확률변수이다. 각 추정량에 대하여 모든 MSE값은 10000번의 반복을 이용하여 구하였다. 다음 [표4.2]는 각 추정량들의 추정된 MSE값이다. [표4.2]에 표시된 *는 해

당되는 경우에 있어서 4개의 추정량 중에서 가장 MSE 효율성이 높은 추정량이며, MY로 표시된 추정량은 식(3.2)로 정의된 제안된 추정량이다.

[표4.2] 분산성분비율 추정량들의 추정된 MSE 값

실험계획	추정량	ρ								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
P1	ANOVA	0.243	0.458	0.828	1.704	3.460	6.871	15.238	37.868	180.75
	REML	0.255	0.484	0.838	1.732	3.462	6.653	14.906	37.012	173.36
	ML	0.093	0.190	0.349	0.745	1.532	2.955	6.700	16.858	79.253
	MY	0.019*	0.044*	0.109*	0.266*	0.614*	1.350*	3.335*	9.730*	49.460*
P2	ANOVA	0.318	0.577	1.030	2.109	4.241	8.750	19.244	47.145	229.43
	REML	0.626	0.970	1.337	2.567	4.702	8.253	17.273	40.704	155.01
	ML	0.195	0.330	0.491	1.022	1.987	3.583	7.707	18.508	82.954
	MY	0.072*	0.092*	0.142*	0.304*	0.654*	1.370*	3.342*	9.749*	49.523*
P3	ANOVA	0.119	0.290	0.647	1.369	2.898	6.031	13.990	35.545	179.59
	REML	0.299	0.503	0.867	1.514	2.812	5.170	11.525	28.958	124.67
	ML	0.082	0.162	0.326	0.633	1.266	2.407	5.495	14.090	62.154
	MY	0.081*	0.100*	0.162*	0.299*	0.612*	1.279*	3.107*	9.142*	45.713*
P4	ANOVA	0.055	0.129	0.257	0.493	1.000	2.031	4.421	12.540	61.329
	REML	0.058	0.135	0.260	0.493	0.983	1.985	4.278	11.720	57.291
	ML	0.034	0.087	0.175	0.339	0.683	1.384	2.986	8.196	39.961
	MY	0.017*	0.030*	0.066*	0.153*	0.358*	0.830*	2.032*	6.115*	31.595*
P5	ANOVA	0.062	0.146	0.302	0.586	1.196	2.442	5.416	15.757	77.830
	REML	0.086	0.202	0.353	0.650	1.227	2.369	4.891	12.798	58.611
	ML	0.043*	0.118	0.221	0.426	0.831	1.631	3.397	8.958	41.359
	MY	0.061	0.071*	0.095*	0.174*	0.370*	0.839*	2.065*	6.282*	32.863*
P6	ANOVA	0.032	0.091	0.215	0.461	0.980	2.045	4.768	14.316	71.581
	REML	0.044	0.118	0.245	0.464	0.914	1.716	3.637	10.182	46.542
	ML	0.022*	0.070*	0.157	0.315	0.637	1.233	2.635	7.348	33.662
	MY	0.065	0.073	0.100*	0.170*	0.350*	0.763*	1.864*	5.742*	29.654*
P7	ANOVA	0.032	0.072	0.153	0.300	0.573	1.141	2.570	7.329	33.875
	REML	0.034	0.073	0.154	0.293	0.561	1.118	2.455	6.852	31.780
	ML	0.024	0.055	0.120	0.231	0.445	0.885	1.949	5.407	25.252
	MY	0.018*	0.023*	0.049*	0.110*	0.258*	0.598*	1.507*	4.544*	23.783*
P8	ANOVA	0.035	0.081	0.175	0.356	0.679	1.370	3.143	9.097	42.098
	REML	0.042	0.090	0.194	0.363	0.682	1.319	2.789	7.450	33.180
	ML	0.026*	0.064	0.144	0.278	0.533	1.035	2.210	5.877	26.373
	MY	0.065	0.058*	0.076*	0.125*	0.265*	0.608*	1.556*	4.801*	25.800*
P9	ANOVA	0.018	0.054	0.134	0.280	0.594	1.210	2.929	8.393	40.960
	REML	0.022	0.058	0.137	0.264	0.520	1.042	2.284	5.979	27.601
	ML	0.015*	0.042*	0.103	0.206	0.416	0.835	1.848	4.832	22.382*
	MY	0.063	0.058	0.075*	0.118*	0.244*	0.557*	1.419*	4.340*	23.208

[표4.2]를 통해서 알 수 있는 사실은 거의 모든 경우에 제안된 추정량이 우수하다는 점인데, 식 (3.2)로 정의된 추정량은 수준의 개수가 크고 불균형정도가 매우 심한 경우에만 ML 추정량보다 MSE 효율성이 떨어진다. 이 사실은 직관과 일치하는 데, 제안된 추정량은 근사 F 분포를 이용하여 유도되었기 때문에 불균형정도가 심할수록 그 정밀도가 떨어진다고 할 수 있다. 그리고 MSE의 효율성에서 제안된 추정량과 비교가 될 수 있는 추정량은 ML 추정량뿐이므로 어떤 경우에 제안된 추정량이 ML 추정량에 비하여 효율적인지를 알아보기 위하여 두 추정량의 비율 값을 구한 것이 [표4.3]이다. [표4.3]의 비율 값은 모의실험을 통하여 나온 두 추정량의 MSE값을 이용하여 구한 평균값이다.

[표4.3] ML 추정량과 제안된 추정량의 MSE값의 비율

인자	ML / MV 의 비율								
불균형정도	하 (P1,P4,P7)			중 (P2,P5,P8)			상 (P3,P6,P9)		
	2.13679			1.97424			1.40914		
수준의 수	3			6			9		
	2.41664			1.66994			1.43360		
급내상관계수	0.3	0.4	0.2	0.5	0.6	0.7	0.1	0.8	0.9
	2.3321	2.3202	2.1394	2.1124	1.8436	1.6255	1.5107	1.4162	1.2604

[표4.2]를 통해서 알 수 있는 사실은 불균형정도가 약할수록, 수준의 수가 작을수록 제안된 추정량이 ML 추정량보다 효율적이며, 급내상관계수 ρ의 값에 대해서는 $0.2 \le \rho \le 0.5$인 경우에 2배이상 제안된 추정량이 효율적임을 알 수 있다.

5. 결론

이 논문에서는 불균형일원변량모형에 있어서 분산성분비율 θ에 대한 새로운 추정량을 고려하였다. 제안된 추정량은 모의실험을 통하여 수준의 개수가 크고 불균형정도가 매우 심한 경우를 제외하고 거의 모든 경우에 기존의 추정량보다 MSE 효율성이 높다는 사실이 판명되었다.

참 고 문 헌

[1] Das,K. (1992). Improved Estimation of the Ratio of Variance Components for a Balanced One-Way Random Effects Model, *Statistics & Probability Letters*, Vol. 13, 99-108.

[2] Loh,W.Y. (1986). Improved Estimators for Ratios of Variance Components, *Journal of the American Statistical Association*, Vol. 81, 699-702.

[3] 이장택 (1996). 균형일원변량모형에서 분산성분비율의 새로운 추정량, 한국통계학회논문집, 제3권, 2호, 43-51.

[4] Thomas,J.D. and Hultquist,R.A. (1978). Interval Estimation for the Unbalanced Case of the One Way Random Effects Model, The Annals of Statistics, Vol. 6, 582-587.