IRT에서 피험자능력 및 문항모수 추정 알고리즘 개발

박영선 [1] 진정언 [2] 차경준 [3]
이종성 [4] 박 정 [5] 김성훈 [6] 이원식 [7] 이재화 [8]

요 약

문항반응이론(item response theory; IRT)에서는 문항이 가지고 있는 특성을 기초로 피험자의 능력을 추정하고 동시에 각 문항별 문항특성곡선(item characteristics curve; ICC)을 이용하여 문항모수를 추정하게 된다. 이러한 추정알고리즘은 이미 외국에서는 상용화되어 활용되고 있는바, 국내에서 개발한 Any Assess의 활용 가능성을 모의실험을 통하여 그 결과의 신뢰성을 검증해 보았다.

주요용어: 문항반응이론, 문항특성곡선, Any Assess

1. 서론

일반적으로 교육·심리측정이론의 제2세대이론으로 불리는 문항반응이론(item response theory; IRT)에서는 문항이 가지고 있는 특성을 기초로 피험자의 능력을 추정하기 때문에 각 문항별 문항특성곡선(item characteristics curve; ICC)이 중요한 검사정보가 된다. ICC의 고유한 모수(item parameter)를 추정하기 위한 최대우도추정(maximum likelihood estimation; MLE)은 Lord(1953)가 처음으로 정규오자이브 모형(normal ogive model)을 대상으로 사용하였다. 또한, 로지스틱모형은 Birnbaum(1968)에 의해 연구 개발되었는데, 이후 MLE 절차를 위한 보다 단순화된 방법인 결합/조건/주변최대우도추정(joint/conditional/ marginal maximum likelihood estimation)과 베이지안 추정기법이 개발되어 피험자 수와 문항수가 적거나 잡음이 섞인 자료 등에서 신뢰성이 떨어지며, 모수추정치의 여러 왜곡현상(歪曲現象)을 줄이는데 획기적인 기여를 하였다(Craven & Wahba, 1977; Foutz, 1977; 이종성, 1990 참조).

이러한 추정기법 중 특히, 베이지안 추정알고리즘은 통계적으로도 매우 난해하며 상용화프로그램구현은 더더욱 많은 노력이 요구된다. 외국의 경우에 대표적인 것은 BILOG (Mislevy & Bock, 1990)와 LOGIST(Wingersky, Barton & Lord, 1982) 등이 있는데, 국내에서 저자 등이 개발한 'Any Assess'에 대해서 모의실험을 통하여 그 결과의 신뢰성을 검증해 보았다.

1) 한양대학교 자연과학연구소 연구교수
2) 한양대학교 BK21 기계사업단 연구원
3) 한양대학교 수학과 교수
4) 서경대학교 교양학부 석좌교수
5) 한국교육과정평가원 부연구위원
6) 동국대학교 교육학과 교수
7) (주)케이세스 기술연구소 선임연구원
8) (주)케이세스 대표이사

2. 능력모수추정 알고리즘

피험자의 문항에 대한 반응은 각 문항을 맞추거나 틀리는 이분형(binary)이고 문항점수(item score) u_{ij}는 문항을 맞은 경우에는 1, 틀린 경우에는 0으로 주어진다. 문항반응모형(ICC)은 피험자의 능력의 함수로써 문항을 맞출 확률로 표현된다. 즉, 피험자 j의 능력값 θ_j에 대하여 문항 i를 맞을 확률과 틀릴 확률은 각각 아래와 같이 표현된다.

$$P(u_{ij}=1|\theta_j)=P_i(\theta_j), \quad P(u_{ij}=0|\theta_j)=1-P_i(\theta_j)$$

그리고 피험자 능력추정기법으로 먼저, 최대 우도(Maximum Likelihood : ML) 방법은 아래의 로그 우도 방정식을 최대로 하는 모수를 추정하는 것이다.

$$\log L(\theta_j|\underline{U}_j)=\sum_{i=1}^{n}\{u_{ij}\log P_i(\theta_j)+(1-u_{ij})\log(1-P_i(\theta_j))\}$$

그러면 우도 방정식은 아래와 같고

$$\frac{\partial}{\partial\theta_j}\log L(\theta_j|\underline{U}_j)=\sum_{i=1}^{n}\frac{u_{ij}-P_i(\theta_j)}{P_i(\theta_j)(1-P_i(\theta_j))}\cdot\frac{\partial P_i(\theta_j)}{\partial\theta_j}=0$$

Fisher's scoring method에 의해 아래의 식이 수렴할 때까지 계산한다(Baker, 1992 참조).

$$\widehat{\theta}_j^{(t+1)}=\widehat{\theta}_j^{(t)}+[I(\widehat{\theta}_j^{(t)})]^{-1}\left(\frac{\partial}{\partial\theta_j}\log L(\widehat{\theta}_j^{(t)}|\underline{U}_j)\right)$$

여기서 $I(\theta_j)=\sum_{i=1}^{n}a_i^2P_i(\theta_j)(1-P_i(\theta_j))$이고 이 식을 Fisher의 정보함수라고 한다. 또한, ML 추정치의 표준오차(standard error)는 $S.E.(\widehat{\theta}_j)=\sqrt{1/I(\widehat{\theta}_j)}$이다.

그리고 사후분포의 최대값(Maximum A Posteriori : MAP) 기법은 아래와 같이 베이즈 정리(Bayes' theorem)의 형태에 기초하여 추정한다.

$$P(\theta_j|\underline{U}_j,\ \underline{\xi}_i)\ \propto\ L(\underline{U}_j|\theta_j,\ \underline{\xi}_i)g(\theta_j)$$

위의 식에 log를 취한 후 최대가 되는 능력모수를 MAP 추정치라 하고 그것은 아래의 식을 구한 해이다.

$$\frac{\partial}{\partial\theta_j}\log L(\underline{U}_j|\theta_j,\ \underline{\xi}_i)+\frac{\partial}{\partial\theta_j}\log g(\theta_j)=0$$

또한, MAP의 추정치 정확도 측도로서의 사후 표준 오차(posterior standard error)는

$$P.S.E.(\widehat{\theta}_j)=\sqrt{1/J(\widehat{\theta}_j)}$$

이고, 여기서 $J(\theta_j)=I(\theta_j)-\frac{\partial^2}{\partial\theta_j^2}\log g(\theta_j)$를 사후 정보(posterior information)라고 한다.

마지막으로 사후분포의 평균(Expectation A Posteriori : EAP)기법은 주변최대우도(Marginal Maximum Likelihood, MML) 방법에서와 같이 Gaussian 구적에 의해 근사적으로 계산한다(Baker, 1992 참조).

3. Any Assess 문항모수 추정 알고리즘

문항반응모형은 추정 대상인 문항 모수(item parameter)와 능력 모수(ability parameter)를 가진다. 문항 모수는 변별도(a_i), 난이도(b_i), 추측도(c_i)를 나타낸다. 이들 모수들의 추정은 다음과 같은 단계로 이루어진다.

Step1. 문항 반응 모수 a_i, b_i, c_i를 MML/ EM으로 추정

Step2. 능력 모수 θ_j를 ML, MAP 또는 EAP로 추정

위의 모수추정의 두 단계를 통합하여 JML(Joint Maximum Likelihood)라고 부른다.

실제로 능력 θ_j를 가진 피험자의 반응점수의 패턴은 $\underline{U}_j = (u_{1j}, u_{2j}, \cdots, u_{nj})$으로 나타나고 이것의 확률은 아래와 같이 표현된다.

$$P(\underline{U}_j|\theta_j) = \Pi_{i=1}^{n} P_i(\theta_j)^{u_{ij}}(1 - P_i(\theta_j))^{1-u_{ij}}$$

또한, 반응점수의 패턴의 주변우도방정식은 $P(\underline{U}_j) = \int P(\underline{U}_j|\theta_j)g(\theta_j)d\theta_j$이다. 여기서, 능력 θ_j는 연속함수 $g(\theta_j)$의 분포를 따른다. 이 주변우도방정식은

$$\overline{P}(\underline{U}_j) \approx \sum_{k=1}^{q} P(\underline{U}_j|X_k)A(X_k)$$

으로 근사적으로 계산되며, 여기서 X_k는 구적 점(quadrature point)이고 $A(X_k)$는 θ_j의 밀도함수 $g(X_k)$에 해당하는 가중치(weight)이다. 로그주변우도방정식은

$$\log L = \sum_{l=1}^{S} r_l \log \overline{P}(\underline{U}_l)$$

이다. 여기서 r_l은 N명의 피험자 중 같은 반응 점수패턴을 가진 빈도이고 S는 구분되는 패턴들의 그룹의 수이다. 위 식을 최대화하는 모수의 추정치를 주변우도추정치(Marginal Maximum Likelihood Estimate: MMLE)이라하고 이는 아래의 로그우도방정식의 해이다.

$$\sum_{k=1}^{q} \left(\frac{\overline{r_{ik}} - \overline{N_k} P_i(X_k)}{P_i(X_k)(1 - P_i(X_k))} \right) \frac{\partial P_i(X_k)}{\partial \begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix}} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

여기서 $\overline{r_{ik}} = \sum_{l=1}^{S} \frac{r_l u_{lj} P(\underline{U}_l|X_k)A(X_k)}{\overline{P}(\underline{U}_l)}$이고 $\overline{N_k} = \sum_{l=1}^{S} \frac{r_l P(\underline{U}_l|X_k)A(X_k)}{\overline{P}(\underline{U}_l)}$이다.

위의 식에서 Fisher's scoring method 와 EM algorithm(Bock & Aitkin, 1981)을 이용하여 구하는데 Expectation 과정에서 $\overline{r_{ik}}$, $\overline{N_k}$을 계산한 후 Maximization 과정을 실행한 후 문항 모수 a_i, b_i, c_i를 추정한다. Fisher's scoring method의 반복 계산과정 중 문항모수의 분산이 0이 되는 경우에 추정치가 무한대로 증가하는 현상이 나타난다. 이러한 현상을 방지하기 위해 문항모수를 MMAP (Marginal Maximum A Posteriori) 추정이라 부르는 베이즈(Bayes) 절차를 사용하여 추정한다. 베이즈 절차에 앞서 문항모수들에 대한 정보를 제공하는 사전(informative prior) 분포로 문항의 기울기 a_i는 lognormal 분포, 하한 점근선 모수 c_i는 베타분포를 갖는다(Swaminathan & Gifford, 1986 참조).

4. 모의실험

Any Assess의 알고리즘을 검증하기 위해 RESGEN(Muraki, 2000)을 이용하여 모의실험을 시도하였던 바, 문항수는 18문제, 피험자수는 1,000명으로 하였다.

4.1. 문항모수 추정

1, 2-모수모형에서는 초기값을 다음과 같이 사용하여 문항모수를 추정하였다.

$$b = 0.5 - p(\text{해당 문항의 정답률}), \quad a = \frac{corr(i,j)}{1 - corr(i,j)^2}$$

여기서 b는 고전검사이론의 p(정답률)를 기준으로 설정한 난이도의 초기값이며, 변별력모수 a의 초기값은 고전검사이론에서 문항의 변별도로서 피험자(i), 점수(j)에 대한 상관관계식 $corr(i,\ j)$를 이용하였다. 또한, 3-모수 모형에서의 초기값은 상기한 a, b와 더불어 c의 초기값으로 고전검사이론에서의 추측도로 설정하였다(성태제, 1994; 이종성, 1990 참조).

1,2-모수모형에서 실 모수와 BILOG 그리고 Any Assess의 난이도와 변별력을 비교분석

[표 1] 1, 2-모수에서 Real Parameter와 BILOG, Any Assess의 추정결과 비교

Item	1-모수			2-모수					
	Real	BILOG	Any Assess	Real		BILOG		Any Assess	
	Di.P.	Di.P.	Di.P.	Di.P.	Ds.P.	Di.P.	Ds.P.	Di.P.	Ds.P.
1	-2.291	-2.227	-2.131	-2.291	0.479	-2.276	0.484	-2.319	0.483
2	0.027	0.051	0.065	0.027	0.421	-0.043	0.386	-0.094	0.384
3	1.125	1.209	1.193	1.125	0.529	1.102	0.611	1.059	0.599
:	:	:	:	:	:	:	:	:	:
16	-1.500	-1.551	-1.481	-1.500	1.500	-1.611	1.210	-1.602	1.264
17	0.098	0.131	0.142	0.098	1.588	0.068	1.660	-0.008	1.706
18	1.673	1.673	1.652	1.673	1.542	1.736	1.462	1.697	1.412

[표 2] 3-모수에서 Real Parameter와 BILOG, Any Assess의 추정결과 비교

Item	Real Parameter			BILOG			Any Assess		
	Ds.P.	Di.P.	Gu.P.	Ds.P.	Di.P.	Gu.p.	Ds.P.	Di.P.	Gu.P.
1	0.479	-2.291	0.221	0.453	-2.408	0.295	0.469	-2.871	0.168
2	0.421	0.027	0.197	0.414	0.122	0.323	0.487	0.149	0.073
3	0.529	1.125	0.273	0.918	1.604	0.414	0.273	1.153	0.091
:	:	:	:	:	:	:	:	:	:
16	1.500	-1.500	0.050	1.600	-1.482	0.239	1.512	-1.568	0.069
17	1.588	0.098	0.187	1.276	0.219	0.220	1.544	-0.145	0.111
18	1.542	1.673	0.278	1.380	1.921	0.289	1.349	1.916	0.129

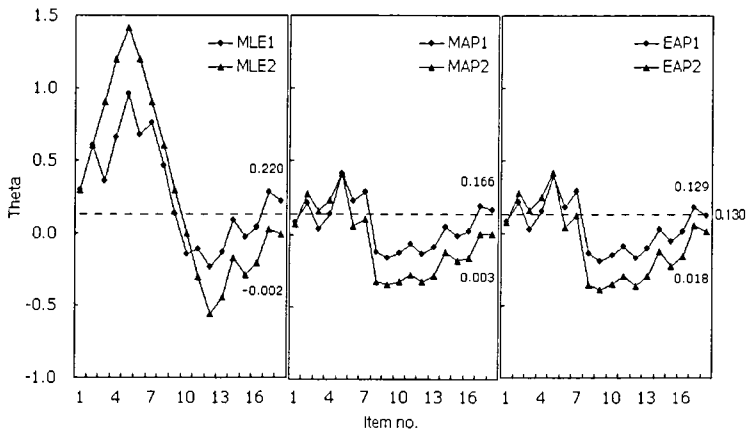
Ds.P.=Discriminating Power; Di.P.=Difficulty Parameter; Gu.P.=Guessing Parameter.

한 결과는 [표 1]과 같다. 그 결과를 살펴보면 대부분의 추정모수는 실 모수와 비슷한 결과를 보였으며 BILOG의 추정결과와도 유사한 추정치를 보였다. 또한, Root MSE(Root Mean Square Error)를 구한결과, 1-모수모형에서 BILOG는 0.063, Any Assess는 0.057 이었고, 2-모수모형에서는 먼저, 난이도에서 BILOG는 0.128, Any Assess는 0.136 이었으며 또한, 변별력에서는 각각 0.129, 0.125 로서 비교적 낮게 나타났다.

[표 2]의 3-모수모형에서는 변별력의 경우에 전체적으로 추정모수는 실 모수와 유사한 패턴을 보였으며, Root MSE 값이 BILOG는 0.236, Any Assess는 0.173 으로서 1,2-모수모형에서보다 다소 높게 나타났다. 또한, 난이도에서도 유사한 결과를 보였는데 Root MSE값은 BILOG는 0.298, Any Assess는 0.407로서 다소 높게 나타났다. 그리고 추측도를 비교한 결과에서는 BILOG는 0.089, Any Assess의 경우에서는 0.118로서 다소 큰 오차를 보였다.

4.2. 피험자의 능력 추정

[그림1]은 피험자 능력추정 알고리즘을 검정하기 위해 정/오답정보가 (1101101001101 10110)인 경우에 실모수를 이용한 MLE, MAP, EAP 방법별 추정값(MLE1, MAP1, EAP1)과 Any Assess에서 추정한 문항모수를 이용한 추정치(MLE2, MAP2, EAP2)를 비교한 그림이다. 그 결과, 실문항모수를 이용한 추정치가 Any Assess를 이용한 것에 비해 실능력값 0.130에 좀더 근사하였으며 대체로 MLE의 경우보다 MAP,EAP 알고리즘이 좀더 안정적인 경향이 있었다.



[그림 1] 실모수(1)와 Any Assess(2)의 추정값을 이용한 MLE, MAP, EAP 능력추정비교; Real ability=0.130.

5. 결론

이상에서 Any Assess의 문항반응이론과 관련한 능력모수와 문항모수 추정알고리즘의 신뢰성을 분석한 결과 다음과 같은 결론을 얻었다. 먼저, 피험자 능력추정방법에서는 MLE 방법보다는 베이지안 추정알고리즘이 좀더 적합한 추정치를 제공한다고 할 수 있으며 또

한, 문항 모수추정 알고리즘에서는 1, 2-모수모형에서와는 달리 3-모수모형에서 만족스러운 추정결과를 보이지는 않았으나 전체적으로 실 모수와 유사한 값을 얻을 수 있었다.

향후과제로서는 첫째, error의 감소문제로서 문항별 잡음(noise) 제거, data의 보정 등의 문제를 추정알고리즘에 포함시켜야 할 것이다. 둘째, 새로운 추정알고리즘 개발문제로서 좀더 손쉽고 단순하며 접근이 용이한 그리고 기법자체의 혁신성을 갖는 추정기법을 개발하여야 할 것이다.

참고문헌

[1] 이종성 (1990). 문항반응이론과 응용. 대광문화사.

[2] 성태제 (1994). 대학별고사를 위한 문항분석, 표준점수, 검사동등화. 한국통계학회논문집, 1, 206-214.

[3] Baker F.B. (1992) Item Response Theory : Parameter Estimation Technique.

[4] Birnbaum, A. (1968). Test scores, sufficient statistics, and the information structures of tests. In Lord F. M. & Novick M. R., Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

[5] Bock, R.D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. Psychometrika, 46, 443-459.

[6] Craven, P. & Wahba, G. (1977). Smoothing noisy data with spline functions : Estimating the correct degree of smoothing by the method of generalized cross-validation. Technical Report No. 445. Madison, Wis.:Department of Statistics, University of Wisconsin.

[7] Foutz, R.V. (1977). On the unique consistent solution to the likelihood equations. Journal of the American Statistical Association, 72, 147-148.

[8] Lord, F.M. (1953). Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 39, 247-264.

[9] Mislevy, R.J., & Bock, R.D. (1990). BILOG 3: Item analysis and test scoring: With binary logistic model. Mooresville, IN: Scientific Software, Inc.

[10] Muraki, E. (2000). RESGEN: A computer program to generate item response vectors. Princeton: ETS

[11] Swaminathan, H. and Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. Psychometrika, 51, 589-601.

[12] Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). LOGIST user's guide. Princeton, NJ:Educational Testing Service.