2003 Proceedings of the Spring Conference, Korean Statistical Society

IRT 모수 추정에서 초기값에 관한 연구

박영선[1], 차경준[2], 장창원[3]

요 약

문항반응이론(IRT)에서 문항특성곡선(ICC)의 모수를 추정하는 경우에 발생되는 초기값(initial value) 문제를 비선형 로지스틱모형을 선형 회귀모형으로 근사화하여 해결하고자 하였다. 특히, 신규 또는 잡음이 섞인(local fluctuation) 문항의 직접적인 평가와 소규모집단별 검사가 이루어질 수 있는 현실적 문제에서 모수추정의 대안으로서 그 의의가 있을 수 있다.

주요용어 : 문항반응이론, 문항특성곡선, 초기값 문제

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(1952)가 처음으로 정규오자이브 모형(normal ogive model)을 대상으로 사용하였다. 로지스틱함수는 Birnbaum(1968)에 의해 연구 개발되었는데 그는 MLE 절차를 위한 보다 단순화된 방법을 제시하였다. 그러나, 여러 가지 이유에서 반응이 없는 경우에 결합 최대우도추정(Joint Maximum Likelihood Estimation; JMLE)절차 역시 'missing' 자료를 용납하지 않는다.

MLE는 비선형 우도방정식을 풀어야만 하는데 이는 반복적인 Newton-Raphson기법을 적용하여 해결할 수 있다. 그러나, 이 알고리즘은 초기값(initial value)에 민감하여 우도방정식의 해 근방의 초기값이 아니면 반복적인 과정에서 해에 수렴하지 못하는 단점을 가지고 있다. 더욱이, 각 피험자가 비협조적인 경우와 피험자와 문항수가 적은 경우에는 더욱 모수추정에 신뢰할 만한 결과를 얻을 수 없다(Mislevy 와 Wu, 1988; Lord, 1980; Samejima, 1973).

본 연구에서는 이러한 경우에 근사 추정기법을 도입하여 초기값 문제를 보완하였으며 ICC 근사모수를 추정하였다.

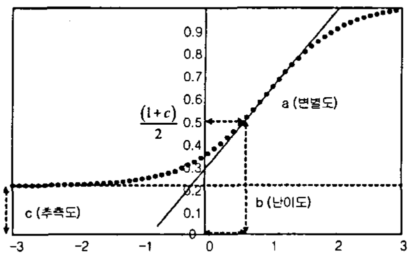
2. 문항반응이론

2-1 문항특성곡선

고전 검사이론에서 해결하지 못하는 여러 문제들을 문항반응이론(IRT)에서 해결할 수 있다는 이론과 실증적 연구가 많이 보고되고 있다(Lord, 1980; Hambleton, 1983; De Ayala 등, 2001).

현재 이용되고있는 IRT의 개념은 잠재적 공간의 단일차원성(unidimensionality of latent space)

1) (133-792) 서울특별시 성동구 행당동 17, 한양대학교 자연과학대학 연구교수
2) (133-792) 서울특별시 성동구 행당동 17, 한양대학교 수학과 교수
3) (135-961) 서울특별시 강남구 포이동 814-8, (주) 케이세스 기술연구소 부장



<그림 1> 문항반응곡선의 3모수 정의.
변별력(discrimination power) a=1.07;
곤란도(difficulty parameter) b=0.60;
추측모수(guessing parameter) c=0.23.

과 지역독립성(local independence)을 가정하고, 피험자의 능력 θ와 문항모수의 관계를 확률함수로 규정하는 문항특성곡선을 사용한다. 모수에는 곤란도 모수(difficulty parameter), 변별력(discrimination power) 그리고 추측모수(guessing parameter)로 구성된다(이종성, 1990 참조).

ICC 연구에서 여러 모형들이 연구된바 있으나, 대표적인 것으로 정규오자이브 모형(normal ogive)과 로지스틱 모형(logistic model)이 있는데 로지스틱 함수가 정규 오자이브 함수보다 오류(error)에 덜 민감하기 때문에 더 많은 분야에서 활용되고 있다(Baker, 1992 참조).

특성치가 θ인 피험자가 문항 j에 정답을 할 확률을 $P(\theta)$라 하면 3-모수 로지스틱함수는

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-1.7 a_j(\theta - b_j)}}$$

이고, 여기서 각각 a_j, b_j, c_j는 문항 j의 변별력, 곤란도, 추측모수이다. 특히, <그림 1>에서와 같이 변별도 a는 변곡점 $\left(b, \frac{1+c}{2}\right)$에서의 기울기 $0.425a(1-c)$와 관계된다(Lord, 1980).

2-2 모형 검증

피험자의 능력을 정확하게 측정할 수 없기 때문에 일반적인 검증절차는 모형으로부터 예측을 하고 그러한 예측치가 거의 정확한지 알아보기 위해서 관찰한 자료와 비교하는 것이다. 또한, 문항반응이론(IRT)에 근거한 타당도를 검증하는 방법이 있다.

먼저, 일차원성의 가정에 대한 만족도는 문항에 대한 인자분석(factor analysis)과 신뢰도 계수(Cronbach alpha)로서 검증할 수 있으며(Reckase, 1979; Hambleton 과 Swaminathan , 1985), 극소독립성은 일차원성으로부터 귀결된다(Lord, 1980). 또한, 모형의 적합도 검정방법으로는 Wright 등(1976)이 제안한 분산분석 방법과 Waller(1981)의 우도비 검정이 사용된다.

3. 문항모수의 추정

3-1 추정기법과 문제점

문항반응이론에 의거 분석을 하려면 적합한 문항반응모형을 선택하고 그것에 따라 각 문항의 모수를 추정해야 한다. 추정방법으로는 결합/조건/주변 최대우도추정(joint/conditional/marginal maximum likelihood estimation)과 베이지안 추정(Bayesian estimation) 방법들이 사용되고 있다.

이러한 추정기법과 상용화된 프로그램들은 다량의 문항과 많은 피험자를 대상으로 한 신뢰성 있는 모수추정에 역점을 두고 있다. 그것은 측정 및 추정오차를 줄이고 심리측정의 왜곡현상(歪曲現象)을 줄인다는 측면에서 당연시된다.

실제로, MLE에서 우도방정식은 비선형 방정식으로서 반복적인 계산절차에 의해 구해지는데, Samejima(1973)는 만일 문항수가 적다면 하나이상의 근을 가질 수 있으며, 20문항이상이면 중다근은 발생하지 않는다고 하였다. 또한 Foutz(1977)의 공리에 의하면 문항수가 충분히 많으면

근 θ의 유일성은 보장된다. 그러나 Looney 와 Spray(1992)의 연구결과에 의하면, 표본이 크기나 검사길이가 늘어나면 측정의 표준오차는 감소하지만 연습이나 피로 그리고 검사를 반복함으로서 지역의 독립성가정이 지켜지지 않는 모순된 결과가 도출되기도 한다. 또한, 추정 알고리즘 측면에서 볼 때, ICC는 단조증가함수이어야 한다는 가정 역시 오차와 무관하지 않으며 매끄러움(smoothness)에도 관계된다. 그리고 국소 변동(local fluctuation)이 많은 경우에는 더욱 그러하다. 이러한 추정문제의 특성은 관찰자료에서의 작은 변화가 근에서는 큰 변화로 나타난다는 것이다(Craven 과 Wahba, 1977; Stocking 등, 1973 참조).

본 연구에서는 이러한 문항특성곡선의 내재된 문제점으로 인한 모수 추정문제에서 특히, MLE 추정절차에서의 'Newton-Raphson Algorithm' 의 초기값(intial value)문제를 근사 해(approximate root)을 구하여 대체함으로써 여러 추정 난제들을 완화하고자 하였다.

3-2. ICC 근사 모수 추정

본 연구에서는 문항특성곡선(ICC)을 피험자의 능력 θ는 -3~3 에서 정의하고 20개로 그룹화하여 각 구간별 정답률 $P(\theta_i)$를 계산하였으며 모형은 로지스틱 함수를 사용하였다. 그리고 초기값으로 대체될 ICC 근사 모수 추정과정은 단순선형회귀분석을 이용한 ICC의 최대경사(slope) 추정단계와 모수 a, b, c의 근사 모수 추정단계, 끝으로 MLE 추정단계로 구성된다.

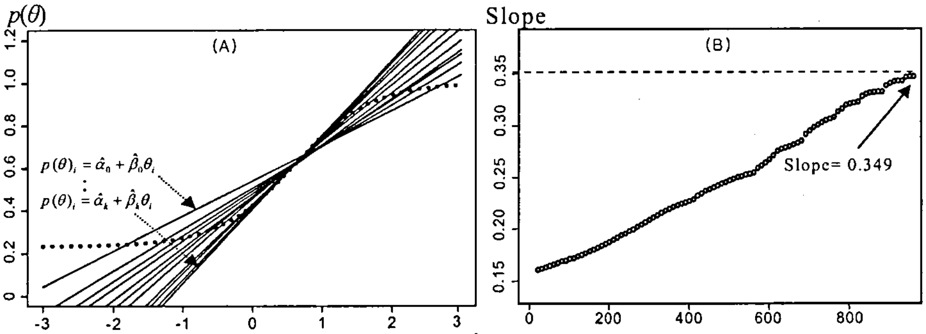
Step1. (Slope 추정단계)

i) a_0 : 능력값을 $\theta_0 \le \theta_1 \le \cdots \le \theta_{n-1} \le \theta_n$이라 할 때, 단순선형회귀직선 $P(\theta_i) = \hat{\alpha}_0 + \hat{\beta}_0\theta_i$를 추정하여 $\hat{\beta}_0 \equiv a_0$로 둔다.

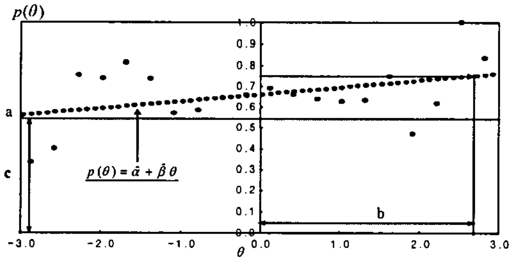
ii) a_1 : $\Delta_1 = P(\theta_1) - P(\theta_0)$, $\Delta_{n-1} = P(\theta_n) - P(\theta_{n-1})$이라 할 때, $\Delta_1 \le \Delta_{n-1}$이면 $P(\theta_0)$을 제거하고 $n-1$개의 $P(\theta_i)$를 가지고 선형회귀직선 $P(\theta_i) = \hat{\alpha}_1 + \hat{\beta}_1\theta_i$ 를 추정하여 $\hat{\beta}_1 \equiv a_1$이라고 하자(만일, $\Delta_1 \ge \Delta_{n-1}$이면 $P(\theta_n)$을 제거한다).

이 과정을 k번 반복하여 $a_k = \hat{\beta}_k$라 하면, 수열 $\{a_n\}$은 $\lim_{n\to\infty} a_n = \beta$ (최대경사)이 된다.

실제로 <그림 2>는 <그림 1>에서 예시된 함수를 실험한 것으로서 로지스틱함수를 이용하여 1,000개의 점에 대해 k=950 회 반복 추정한 것이다. (A)는 추정절차를 설명한 것이고, (B)는 Slope=0.350 에 수렴하고 있음을 보여주고 있다.



<그림 2> 선형 회귀를 이용한 점근적 최대 경사의 추정.
(A) 반복적인 점근적 최대경사 추정과정; (B) 점근적 최대경사의 추정.



<그림 3> 선형회귀모형을 이용한 문항특성곡선의 근사 모수 추정. (사례: item=6; N=1,000중에서 편의된 n=600개의 표본).

Step2. (초기값 단계)

로지스틱 모형에서 정의된 모수 a, b, c는 하한 점근선(lower asymptote)은 추측모수 c로서 $c=\frac{P(\theta_0)+P(\theta_1)+P(\theta_2)}{3}$를 선택한다. 또한, 로지스틱함수의 변곡점은 $(b, \frac{1+c}{2})$이므로 Step1에서 구한 최대경사 $\beta=0.425\,a(1-c)$ 이므로 변별력 모수 $a=\frac{\beta}{0.425(1-c)}$ 로서 구한다. 이때, 문항곤란도 b는 $\frac{1+c}{2}$ 이다.

Step3. (MLE 추정 단계 : New-Raphson algorithm)

Step2 에서 구한 초기값 $(a_0, b_0, c_0)=(\frac{\beta}{0.425(1-c)}, \frac{1+c}{2}, \frac{P(\theta_0)+P(\theta_1)+P(\theta_2)}{3})$ 을 이용하여 Newton-Raphson algorithm을 오차 < 0.1 범위에서 반복 시행한다.

또한, 더욱이 ICC 근사 모수 추정과정에서 초기값을 이용하였을 때에도 수렴하지 않는 경우(예를 들면, bad item)에는 직접 선형회귀모형을 적용하여 ICC 각 모수를 정의할 수 있다.

<그림 3>의 경우는 피험자 1,000명을 대상으로 실증사례 item=6에서 편의된 600 명만을 추출하여 ICC 모수추정을 위 단계에서 초기값을 구하여 실시하였던 바, 수렴하지 않았다. 따라서 선형회귀모형이 $P(\theta)=\hat{\alpha}+\hat{\beta}\theta$로 추정되었다면 <그림 3>과 같이 모수 a, b, c를 다음과 같이 근사적으로 정의할 수 있다.

$$\text{변별력모수 } a=\frac{\hat{\beta}}{0.425(1-c)},\ \text{난이도모수 } b=\frac{1+c}{2},\ \text{추측모수 } c=\hat{\alpha}+\hat{\beta}(-3).$$

4. 실증 사례

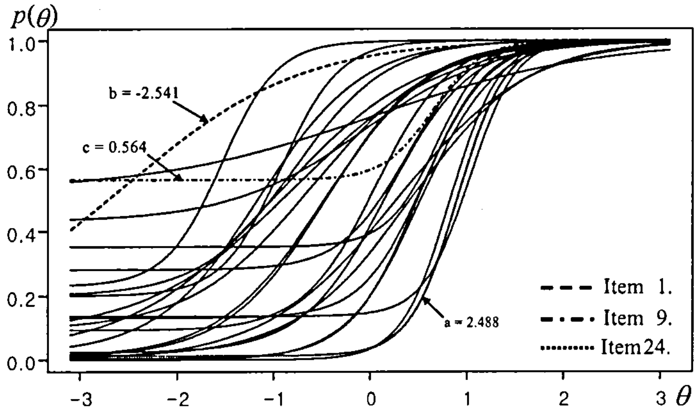
4-1. 분석자료 및 분석도구

본 연구에서 분석한 검사자료는 2000년도 국제수학검정평가원(한국지식기반평가연구회)에서 실시한 중학교 2학년생을 대상으로 수학문제 중에서 5지 선다형 25개 문항을 분석하였다. 피험자수는 1,000명이었다.

분석도구로는 문항반응이론에 의한 문항분석은 (주)케이세스가 독자적으로 개발한 'Any Assess' 와 'New Item Analysis System'을 사용하였으며 그밖에 요인분석과 일부 문항의 추정 모수를 비교하기 위하여 SAS (Statistical Analysis System, V.8.12, USA)를 이용하였다.

4-2. 결과

<그림 4>는 모든 문항을 추정하여 그린 문항특성곡선들이다. 변별력이 가장 높았던 문항



<그림 4> 로지스틱모형을 이용하여 추정된 ICC. Converge=3, 4, 5, 11, 12, 15, 16, 17, 21, 22; Conditional converge=1, 2, 6, 8, 9, 10, 13, 14, 18, 19, 20.

<표 1> 추정된 문항모수의 통계량

	Item parameter			Inter-correlation			
Statistics	a	b	c		a	b	c
Mean	1.469	-0.131	0.142	a	·	·	·
Std dev.	0.619	0.871	0.170	b	0.541	·	·
Minimum	0.512	-2.541	0.001	c	0.107	0.675	·
Maximum	2.488	1.028	0.564				

고유값; 요인1=23.08 (92.34%), 요인2=0.944 (0.038%); Cronbach alpha=0.995.

은 24번으로 $a=2.488$이었으며, 난이도가 가장 낮았던 문항은 1번 문항으로 $b=-2.541$ 그리고 추측도가 가장 높았던 문항은 9번 문항으로서 $c=0.564$로 나타났다. 특히, 초기값을 추정하여 모수를 추정한 경우(;conditional converge)가 단순히 추정 알고리즘에 의존한 경우보다 훨씬 많았다는 사실로서 우리의 논제는 가치가 있어 보인다.

전체 문항에 대한 기초통계량은 <표 1>과 같다. 추가분석은 일차원성 가정의 만족여부를 판단하기 위해 실시하였는데 제1인자 vs. 제2인자 고유값의 비가 23.08 vs. 0.94로서 Reckase (1979)가 제시한 기준치를 초과하여 만족하였으며 신뢰도 역시 (Cronbach alpha=0.995) 일차원성 기본 가정을 만족된다고 하겠다.

5. 결론

수학적으로 비선형 방정식의 근을 구하는 것은 존재성(existence)과는 별개의 문제이다. 특히, 외국에서 뿐 만 아니라 국내에서도 문항반응이론을 접목하여 교육 및 심리측정 그리고 기타 여러 과학분야에서 활용과 그 가치를 인정받고 있는 시점에서 본 연구의 의의를 둘 수 있다고 하겠다.

본 연구의 결론으로서 첫째, 신규문항에 대한 모수추정을 성공적으로 수행할 수 있었다. 이는 문항특성곡선의 근사 모수추정치를 초기값으로 하여 다소 잡음이 포함된 경우에 효과적이라고 할 수 있다. 둘째, 문항반응특성곡선은 선형회귀모형으로 근사화 시킴으로써 소규모 집단검사가 가능할 수 있다. 물론 신뢰성 있는 모수추정은 될 수 없으나 현실적으로 대규모 집단검사가 불가능한 경우에 대안으로써 활용할 수 있겠다. 셋째, 단일문항평가가 이루어짐으로 해서 구조적으로 원활하고 단순한 'item-DB'구축이 가능할 수 있겠다. 또한, 신뢰성이 낮은 문항에 대해서는 'feedback'과정을 거쳐 신뢰도를 높일 수 있는, 재 추정방법을 사용한다면 검사의 직접적인 청정(purification) 효과도 가능하리라 기대된다.

참고문헌

이종성 (1990역). 문항반응 이론과 적용. 서울. 대광문화사.

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.

Baker F. B. (1992). Item Response Theory : Parameter Estimation Technique. Number 129 in Statistics, textbooks and monographs. Marcel Dekker, Inc., New York.

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.

De Ayala R. J., Plake B. S., Impara J. C. (2001). The Impact of Omitted Responses on the Accuracy of Ability Estimation in Item Response Theory. Journal of Educational measurement Fall, 38, 213-234.

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

Hambleton, R. K. (1983). Applications of item response theory. Vancouver: Educational Research Institute of British Columbia, 175-195.

Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston: Kluwer Academic Publishers.

Lord, F. M. (1952). The relationship of the reliability of multiple choice items to the distribution of item difficulties. Psychometrika, 18, 181-194.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Looney, M, A., & Spray, J. A. (1992). Effects of violating local independence on IRT parameter estimation for the Binomial Trials model. Research Quarterly for Exercise and Sport, 63, 356-359.

Mislevy, R. J., & Wu, P. (1988). Inferring examinee ability when some item responses are missing (RR 88-48-ONR). Princeton, NJ: Educational Testing Service.

Reckase, M. D. (1979). Unifactor latent trait models appled to multifactor tests: Results and implications. Journal of Educational statistics, 4, 207-230.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychomtrika, 38, 221-233.

Stocking, M., Wingersky, M. S., Lees, D. M., Lennon, V., & Lord, F. M. (1973). A program for estimating the relative efficiency of tests at various ability levels, for equating true scores, and for predicting bivariate distributions of observed socres. Research Memorandum 73-24. Princeton, N.J. : Educational Testing Service.

Waller, M. I. (1981). A procedure for comparing logistic latent trait models. Journal of Educational Measurement, 18, 119-125.

Wright, B. D., Mead, R., & Draba, R. (1976). Detecting and Correcting item bias with a logistic response model. Research Memorandom No.22, Chicago: University of Chicago, Statistical Laboratory. Department of Education.