순서형 자료로 측정된 구조방정식모형 분석

윤상운*, 박정선**, 이태섭***
*연세대학교 응용통계학과, **연세대학교 입학관리처, ***안양대학교 정보통계학과

The Structural Equation Model with Ordinal Data

Sang Un Yun*, Chung Seon Park**, Tai Sup Lee***
*Dept. of Applied Statistics, Yonsei University, **Office of Admissions, Yonsei University
***Dept. of Statistics, Anyang University

Key Words : Structural Equation Model, Ordinal Data, Quantification

Abstract

This paper is concerned with the analysis of structural equation model(SEM) with the ordinal data such as Likert scale. The SEM is misused when the arbitrary scores allocated to the Likert scale are treated as quantitative data. The underlying distribution approaches have been studied to solve this problem, and the partial least squares(PLS) is also tried.

In this paper the quantification methods for the Likert scale are proposed to analyze the SEM. We assume that the Likert scale is an observation of the interval of the continuous underlying distribution, and the respondents have their own patterns in the response of some questions. Normal and beta distributions as the response patterns are considered to quantify the Likert scale. To compare the efficiency of the proposed method the bootstrap simulations are tried.

1. 서 론

대부분의 통계적인 분석 방법은 직접적인 측정을 통해 얻은 자료를 분석하는데 적절하다. 예를 들어 누군가 청소년의 체중 변화에 대해 연구하고자 한다면 그는 표본들을 대상으로 체중을 직접 측정하고 결과를 분석할 것이다. 그러나 사회과학이나 행동과학, 그리고 실용적인 마케팅 분석에서는 분석의 대상이 간접적으로 측정되고 이들 간의 구조적 관계를 모형화하는 문제가 중요시되고 있다. 이처럼 변수가 간접적으로 측정되고 이들 간의 구조적 설명 관계를 모형화하기 위해서 최근에 구조방정식모형(structural equation model)이 널리 이용되고 있다.

구조방정식모형은 직접적인 측정을 통해 얻은 자료들을 이용하여 간접적으로 측정되는 변수와의 관계를 설정할 수 있을 뿐만 아

니라 간접적으로 측정된 변수들 간의 구조적 관계도 모형화할 수 있다. 이처럼 직접 측정된 변수들과 간접적으로 생성된 변수들 간의 관계를 모형화할 수 있다는 장점으로 인해 구조방정식모형은 최근에 교육학, 마케팅, 심리학, 사회학, 경영학, 인구통계학, 생물학, 보건학, 그리고 유전학에 이르기까지 거의 모든 연구 분야에서 실용적으로 이용되고 있다(Hair et al, 1998).

구조방정식모형은 원칙적으로 양적 자료(quantitative data)를 분석 대상으로 하기 때문에 연속형으로 관찰된 경우 모형을 분석하는데는 아무런 어려움이 없다. 그러나 많은 실용적인 분석에서 자료는 리커트 척도(Likert scale)를 사용하여 측정한다. 특히, 교육학, 심리학, 또는 사회학과 같은 사회과학과 마케팅 분석에서는 대부분 측정의 편리성으로 인해 리커트 척도를 이용하고 있다. 리커트 척도로 조사된 자료는 실용적인 측면에서 양적인 자료로 취급되고 있으나 이는 엄격한 의미에서 양적인 자료가 아니라 순서형 자료로 간주되어야 한다. 물론 많은 통계적 방법들이 극단적으로 가정을 이탈하지만 않는다면 이런 분석 방법에 대해 강건한(robust) 특성을 보인다고 알려져 있지만(Olsson, 1979), 리커트 척도를 양적인 형태로 취급하여 구조방정식모형을 분석하는 것은 잘못된 분석 결과를 유발하게 된다.(Jöreskog and Sörbom, 1996a; Bollen, 1989). 이는 곧 변수들의 관계를 설명하는데 부적절한 결과를 얻게 된다는 것을 의미한다.

리커트 척도를 양적인 자료로 취급할 경우 발생되는 또 다른 문제는 응답자들의 개인적인 반응의 차이를 반영하지 못한다는 것이다. 응답자들은 각 개인들이 갖고 있는 기준에 따라 반응하는데 이 과정에서 똑같은 정도의 생각을 하면서도 어떤 사람은 보통으로, 어떤 사람은 극단적으로 표현하는 등의 차이를 보일 수 있다는 것이다. 또한, 리커트 척도는 모든 척도의 차이를 동일한 것으로 간주하므로 척도간의 차이는 그 절대적인 크기에 관계없이 모두 같게 된다. 예를 들어 만족도 조사에서 리커트 척도를 사용한 경우 '매우 만족'과 '만족'의 차이나 '매우 불만'과 '불만'의 차이는 모두 동일하게 간주된다.

이러한 문제는 구조방정식모형의 분석에서도 오랫동안 연구되어져 왔다. 먼저 2장에서는 순서형 자료를 갖는 구조방정식모형을 분석하는데 적용된 방법 중 대표적인 방법론이라 할 수 있는 다범주 상관계수[1]를 이용한 가중최소제곱법과 편최소제곱법(partial least squares)을 이용한 분석 방법에 대하여 간단히 살펴본다. 3장에서는 순서형 자료, 특히 리커트 척도를 수량화하는 알고리즘을 제시한다. 4장에서는 이를 구조방정식 모형에 적용할 경우의 효율성에 대하여 모의실험를 통하여 살펴보기로 한다.

2. 기존의 분석 방법

구조방정식모형은 어떤 현상이 나타내는 이론적 관계를 통계적으로 규명하는 다변량 분석기법이다. 일반적으로 구조방정식모형의 분석과정은 두 단계로 나눌 수 있는데 첫째는 일련의 구조 방정식을 통해 변수들 간의 인과관계를 표현하는 것이며, 두번째 과정은 이러한 구조적 관계를 관찰된 자료를 통해

1) polychoric correlation으로 통일된 용어가 없어 본논문에서는 다범주 상관계수로 사용함.

보다 분명한 이론적 관계로 개념화하는 것이다. 즉 구조방정식모형은 이론적 가설을 구조적 관계로 설정하고 관찰된 자료를 통하여 이를 통계적으로 확인하는 과정으로 나눌 수 있다. 구조방정식모형에 대한 자세한 이론적 내용은 Bollen(1989), Jöreskog and Sörbom (1996a), 조현철(1999) 등을 참고할 수 있다.

일반적인 구조방정식모형에서 구조모형과 측정모형은 각각 다음과 같이 정의된다.

구조모형(Structural Model)

$$\eta = B \cdot \eta + \Gamma \cdot \xi + \zeta \tag{1}$$

측정모형(Measurement Model)

$$y = \Lambda_y \cdot \eta + \varepsilon, \quad x = \Lambda_x \cdot \xi + \delta \tag{2}$$

여기서 ξ와 η는 각각 잠재 내생변수와 외생변수이며, x와 y는 관찰된 외생변수와 내생변수를 나타내는 벡터들이다. 또한 ε, δ, ζ는 모형의 오차이며, $B, \Gamma, \Lambda_y, \Lambda_x$는 각각 계수 행렬을 나타낸다. 구조방정식모형에서는 기본적으로 공분산 행렬을 이용해서 변수들간의 구조적 관계를 분석하는데 이는 관찰변수들이 연속형이어야 한다는 것을 의미한다. 따라서, 만약 자료를 리커트 척도로 수집하였다면 이는 엄격한 의미에서 질적 자료이므로 양적 자료로 취급하여 분석하는 경우에는 문제를 야기한다. 그러나 현실적으로는 리커트 척도와 같은 순서형 범주 자료를 마치 연속형인 것처럼 간주하고 각 범주에 임의의 수 값을 할당하여 양적 자료로 이용한다. 물론 많은 통계적 방법들이 극단적으로 가정을 이탈하지만 않는다면 이런 분석 방법을 통해서 비교적 안정적인 결과를 얻을 수 있다고 알려져 있지만 리커트 척도를 양적인 자료로 취급하는 것은 일반적으로 잘못된 분석 결과를 유발하게 된다(Olsson, 1979).

구조방정식모형을 분석하는 경우가 아니더라도 리커트 척도로 측정된 값은 척도에 점수를 할당하는 방법에 따라 서로 다른 결과를 보이기 때문에 연속형 자료에서처럼 평균, 분산, 공분산 등과 같은 통계량을 이용하는 것은 큰 의미를 갖지 못한다. 이러한 문제점에도 불구하고 실용적인 측면에서 리커트 척도는 연속형 자료로 간주한다. 즉 리커트 척도로 측정된 자료의 공분산 또는 상관계수를 계산하고 이를 바탕으로 최우추정법이나 일반화최소제곱법과 같은 추정방법을 이용하여 구조방정식모형을 추정한다. 이 경우에 문제는 관찰된 순서형 변수의 분포가 연속형이 아니므로, 모수에 대한 추정값뿐만 아니라 적합도 지수나 표준오차 등을 왜곡하는 결과를 초래하게 된다는 것이다(Jöreskog and Sörbom, 1996a). 따라서 리커트 척도를 이용하여 측정된 자료를 분석하는 데는 통상의 구조방정식모형 추정법과는 다른 절차를 필요로 한다.

여기서는 먼저 현재 시도되고 있는 방법 가운데 적용가능한 방법인 다범주 상관계수를 이용한 방법(Jöreskog, 1994)과 편최소제곱법을 이용한 방법을 소개하기로 한다. 특히, 편최소제곱법은 미국의 ACSI(Ameriacn Customer Satisfaction Index) 모형과 한국생산성본부에서 주관하는 국가고객만족도(NCSI) 모형을 분석하는데 이용되고 있다.

2.1 다범주 상관계수를 이용한 가중최소제곱법

Jöreskog(1994)은 순서형 변수로 조사된 자료를 이용하여 구조방정식모형을 분석하기 위해 단계적 접근방법을 제안하였다. 먼저 다범주 상관계수를 추정한 후, 다음으로 이를 이용하여 모형을 추정하였다. 다범주 상관계수를 추정하기 위해서는 Olsson(1979)의 2단계 최우추정법을 적용하였으며, 모형 추정을 위해서는 가중최소제곱법을 이용할 것을 제안했다. 또한 다범주 상관계수에 대한 점근적 공분산 행렬을 제시하였다. 가중최소제곱법을 적용하기 위해서는 가중치 행렬이 필요한데 일반적으로 다범주 상관계수 행렬의 점근적 공분산 행렬을 사용한다. 이러한 방법은 기본적으로 범주형 자료에 대한 Pearson의 관점에 의존하고 있다. Pearson은 범주형 변수들의 연관성을 얻기 위해 범주형 자료는 그 바탕에 연속형 분포가 존재하고 있으며, 범주형 자료는 이들 연속형 분포의 일정 구간이 관찰된 것으로 간주했다(Agresti 1990). 여기서 범주형 변수들의 상관계수를 다범주 상관계수라 한다.

순서형 변수 y를 k개의 범주에서 관찰한 변수라 정의하고 변수 y^*는 순서형 변수 y의 기저변수(underlying variable)라 정의하면, 변수 y와 y^* 사이에는 다음 관계가 성립한다.

$$y = \begin{cases} 1, & a_0 < y^* \le a_1 \\ 2, & a_1 < y^* \le a_2 \\ \vdots & \\ k, & a_{k-1} < y^* \le a_k \end{cases} \tag{3}$$

여기서 $a_i,\ i=0,1,\cdots,k,$는 기저분포에서 범주가 관찰되는 구간의 경계값으로 양 끝값인 a_0와 a_k는 각각 $-\infty$와 ∞로 정의한다. 이렇게 정의된 y는 우리가 실제로 관찰할 수 있는 순서형 변수의 형태를 취하는데 이의 대표적인 형태가 리커트 척도라 할 수 있다. 리커트 척도를 (3)과 같이 정의하면 식 (2)의 측정모형은 다음과 같이 수정된다.

$$y^* = \Lambda_y \cdot \eta + \varepsilon, \quad x^* = \Lambda_x \cdot \xi + \delta \tag{4}$$

식 (4)에서 잠재변수 η와 ξ는 관찰된 변수인 x, y와의 관계가 아니라 기저변수인 x^*, y^*와의 관계로 모형화 됨을 알 수 있다. 기저변수 x^*, y^*의 분포를 연속형 확률분포로 가정한다면 식 (4)의 측정모형은 구조방정식모형에서 요구하는 가정사항을 만족한다. 따라서 분석의 대상은 관찰변수인 x, y가 아니라 기저변수인 x^*, y^*이며, 이 기저변수 x^*, y^*의 공분산 또는 상관계수를 모형 추정에 이용할 수 있다. 특히 이렇게 구해진 상관계수를 다범주 상관계수라 한다.

구조방정식모형을 분석하는데 가중최소제곱법을 적용하기 위해서는 표본의 크기가 커야 한다. 표본의 크기가 크지 않을 경우 점근적 공분산 행렬을 정확히 추정하기 힘들어지며, 만약 표본의 크기가 적당히 크지 못할 경우에 가중최소제곱법을 적용하여 모수를 추정하는 것은 오히려 최우추정법이나 일반화최소제곱을 사용하는 것보다 나쁠 수 있다(Jöreskog and Sörbom, 1996a). 현재 구조방정식모형을 분석하는 데 가장 널리 사용되는 소프트웨어인 Lisrel에서는 이 방법을 이용한 모형 추정을 지원하고 있다. 물론 범주형 자료에 대한 상관계수를 계산하는데는 위에서 언급한 다범주 상관계수외에도 여러 가지를 고려할 수 있다. Lisrel에서도 스피어만의 순위상관계수, 캔달의 tau-b 상관계수 등 여러 가지를 제공하고 있으나, 효율성 측면에

서 다범주 상관계수가 다른 것 보다 비교적 우수하다.(Jöreskog and Sörbom, 1996b). Muthén(1984)과 Lee and Poon(1987), Lee, Poon and Bentler(1990a) 등은 기저분포가 다변량 정규분포인 경우에 대하여 연구했다.

2.2 편최소제곱법을 이용한 분석

편최소제곱법은 원래 Wold(1975)가 제안한 것으로 계량분석화학(chemometrics) 분야에서 다공선성 문제를 해결하기 위해 주성분회귀와 함께 적용되어 왔으며, 최근에는 그 유용성이 알려지면서 계량경제와 사회과학 등 여러 분야로 영역을 확대하고 있다(최엄문, 1998). Fornell과 Bookstein(1982)은 마케팅 자료를 분석하는데 있어 구조방정식모형의 추정 방법으로 PLS를 적용하였다. 특히 Fornell(1992, 1996)은 스웨덴과 미국의 고객만족모형을 분석하면서 구조방정식모형에 PLS를 적용하였다.

편최소제곱법은 관찰변수와 잠재변수의 관계를 선형식으로 정의하여 관찰하지 못한 잠재변수에 값을 할당하며, 이 할당된 값을 이용하여 모형을 추정한다. 이는 사실상 인자점수와 유사한 것으로 이해할 수 있다. 이처럼 잠재변수에 값을 할당한 후에는 회귀분석을 이용하여 정의된 구조방정식모형을 추정한다. PLS 방법을 이용하여 구조방정식모형을 추정하는 자세한 알고리즘은 Fornell and Bookstein(1982), Wold(1982) 등을 참조하라.

PLS 방법을 이용하여 구조방정식모형을 추정할 경우 마케팅 자료가 갖는 몇 가지 문제점을 해결할 수 있다(Fornell and Bookstein, 1982). 첫째는 많은 마케팅 자료가 정규성을 만족하지 못하는데 PLS는 정규성 가정을 요구하지 않는다. 구조방정식모형에서 최우추정법은 자료가 다변량 정규분포를 따른다는 가정을 필요로 하므로 이에 대한 확인이 필요하나, 최소제곱법과 마찬가지로 PLS는 이러한 분포의 가정을 필요로 하지 않는다. 두 번째로는 인자분석처럼 구조방정식모형을 추정할 경우에도 해가 존재하지 않거나 또는 적절치 못한 해(improper solution)가 존재하는 문제가 자주 발생하나, PLS를 적용할 경우에는 이런 문제를 피할 수 있다.

그러나 PLS는 잠재변수에 값을 할당하는 과정에서 반복적인 계산 절차를 필요로 하기 때문에 추정량에 대한 통계적 특성을 파악하기 어려우며, 이에 따라 추정량에 대해 대수적으로 표준오차를 계산하지 못하는 문제가 발생된다(Wold 1982). 반면에 잠재변수에 관찰된 값을 할당하고 이들 관계를 회귀모형화 함으로써 잠재변수에 대한 예측이 가능해진다.

3. 수량화 방법

현재 많은 자료가 실용적인 측면에서 리커트 척도로 측정되고 이를 마치 양적인 자료인 것으로 간주하여 분석한다. 그러나 엄밀한 의미에서 리커트 척도는 질적자료이며, 설사 양적자료로 간주할 수 있다 하더라도 몇 가지 문제에 대한 고려가 필요하다. 첫째는 응답자들은 각 개인들이 갖고 있는 기준에 따라 반응하는데 이 과정에서 똑같은 정도의 생각을 하면서도 다른 범주로 반응을 보일 수 있다는 것이다. 또 다른 문제로 리커트 척도는 통상적으로 등간 척도로 취급된다는 것이다. 리커트 척도로 응답한 경우 척도간의 차이는 그 인식의 강도에 관계없이

모두 동일한 것으로 간주된다. 예를 들어 만족도를 조사하는 경우에 만족의 강도에 관계없이 '매우 만족'과 '만족' 간의 차이는 '매우 불만'과 '불만'간의 차이와 항상 같은 것으로 평가한다. 따라서 본 논문에서는 리커트 척도를 양적인 크기로 전환하기 위해 개인 간에 존재하는 반응특성의 차이를 고려한다. 이렇게 수량화된 값을 이용하여 관찰변수들 간의 공분산 또는 상관계수를 계산한 후, 구조방정식모형을 추정하는 절차를 이용한다.

임의의 한 응답자는 n개의 질문에 대하여 각각 k개의 순서형 범주로 반응한다고 가정하자. 이 때, k개의 순서형 범주는 (3)에서 정의한 것처럼 그 기저분포의 한 구간이 관찰된 것으로 간주할 수 있다. 임의의 관찰변수 $X_i, i=1,2,\cdots,n$의 기저변수를 W_i라 하고, 기저변수 W_i는 동일한 연속형 분포함수 $F(w:\theta)$를 따른다고 가정하자. 이때 관찰변수인 X_i와 관찰 불가능 변수인 기저변수 W_i 사이에는 다음과 같은 관계가 성립한다. 여기서 편의상 첨자 i는 생략하여 표현하기로 한다.

$$X=\begin{cases} 1, & a_0 < w \le a_1 \\ 2, & a_1 < w \le a_2 \\ \vdots & \\ k, & a_{k-1} < w \le a_k \end{cases} \qquad (5)$$

여기서 a_i는 경계점(thresholds)이며, 경계점의 양쪽 끝값인 a_0와 a_k는 각각 $-\infty$와 ∞로 정의한다. 이처럼 리커트 척도는 연속형 기저분포의 어떤 한 구간이 관찰된 형태로 볼 수 있으며 리커트 척도에 대한 수량화 값은 자연스럽게 그 구간의 기대값으로 선택할 수 있을 것이다. 구간자료의 수량화 값으로 그 구간의 기대값을 선택하는 것은 구간으로 정의된 자료를 수량화 할 때 일반적으로 택하는 방법이기도 하다. 이렇게 선택된 수량화 값을 기저점수(underlying score)라 부르기로 하며 다음과 같이 정의한다.

$$w_x = E(W\,|\,x),\ \ x = 1,2,\cdots,k \qquad (6)$$

기저점수 w_x를 구하기 위해서는 먼저 기저변수 W의 조건부 확률밀도함수를 구해야 한다. 기저변수 W의 조건부 확률밀도함수는 다음과 같다.

$$f(w\,|x) = \frac{\Pr(W=w,\ X=x)}{\Pr(X=x)} = \frac{f_w(x)}{\pi_x} \qquad (7)$$

여기서 $f(w)$는 기저변수 W의 확률밀도함수이며, π_x는 범주 x가 관찰될 확률로서 다음과 같다.

$$\pi_x = F(a_x\,;\theta) - F(a_{x-1}\,;\theta) \qquad (8)$$

식 (7)에 주어진 조건부 확률밀도함수는 임의의 구간 x가 주어졌을 경우의 절삭분포(truncated distribution) 형태와 같음을 알 수 있다. 수량화 값인 기저점수 w_x는 기저변수 W의 조건부 기대값으로 다음과 같다.

$$w_x = \frac{1}{\pi_x}\int_{a_{x-1}}^{a_x} t \cdot f(t)\,dt \qquad (9)$$

리커트 척도의 수량화 값으로 식 (9)의 조건부 기대값을 이용하기 위해서는 기저분포함수의 모수 θ와 경계점 a_x를 추정해야 하며, 이는 최우추정법을 이용한다.

확률변수 r_x를 범주 x에서 관찰된 반응의 수라고 정의하면, 반응의 수 r_x는 다항분포를 따르므로 k개의 범주에서 각각 r_x의 반응이 관찰될 결합밀도함수에 대한 대수우도함수는 다음과 같다.

$$\log L(a,\theta) \propto \sum_{x=1}^{k} r_x \cdot \log\{F(a_x)-F(a_{x-1})\} \quad (10)$$

따라서 경계점 a_x와 모수 θ의 최우추정량은 다음의 방정식을 통해 구할 수 있다.

$$\frac{\partial \log L(a,\theta)}{\partial \theta}=0, \quad \frac{\partial \log L(a,\theta)}{\partial a}=0 \quad (11)$$

식 (11)을 각각의 경계점 a_x에 대해서 정리하면 다음과 같다.

$$\frac{\partial \log L(a,\theta)}{\partial a_x}=f(a_x)\left\{\frac{r_x}{F(a_x)-F(a_{x-1})} - \frac{r_x}{F(a_{x+1})-F(a_x)}\right\} \quad (12)$$

이와 유사한 문제에서 Olsson(1979)은 다범주 상관계수를 구하기 위해 2단계 추정법을 적용하였다. 본 연구에서도 경계점과 모수를 추정하기 위하여 Olsson의 방법을 적용한다. 첫 단계에서는 각 범주에서 추정된 확률값을 통해 누적확률을 구하고 그 누적확률 값에서 기저분포함수의 역함수로 경계점 a_x를 추정한 후, 추정된 경계점에서 기저분포의 모수 θ를 추정한다.

알고리즘 : 경계점 및 모수 추정

1단계 : 추정 누적확률 값에서 기저 분포함수의 역함수로 경계점 a_x를 다음과 같이 추정한다. 여기서 P_x는 범주 x의 누적 확률을 의미한다.

$$\widehat{a_x} = F^{-1}(P_x), \quad x=1,2,\cdots,k-1 \quad (13)$$

2단계 : 추정된 경계점 a_x에서 모수 θ를 다음과 같이 추정한다.

$$\frac{\partial \log L(a,\theta)}{\partial \theta}\Big|_{a=\hat{a}} = 0 \quad (14)$$

3.1 기저분포가 정규분포일 경우

앞서 제시한 수량화 알고리즘에서 기저분포가 정규분포일 경우에 수량화 값의 추정 문제를 살펴보자. 기저변수 W는 평균이 μ이고 분산이 σ^2인 정규분포를 따른다고 가정하면, 표준화 변수 Z는 표준정규분포를 따르게 된다. 이때, 범주 x에서 Z에 대한 조건부 확률밀도함수는 다음과 같으며, 이는 절삭된 정규분포(truncated normal distribution)와 같다.

$$g(z|x)=\frac{\phi(z)}{\sigma\{\Phi(z_x)-\Phi(z_{x-1})\}} \quad (15)$$

여기서 $z_x, x=1,2,\cdots,k$,는 경계점 a_x에서 표준화한 값이며, $\Phi(\cdot)$는 표준정규분포함수를, $\phi(\cdot)$는 표준정규확률밀도함수를 의미한다. 기저분포가 정규분포일 때 기저점수인 조건부 기대값은 식 (16)과 같다(Barr and Sherill, 1999 ; Schneider, 1986).

$$E(W|x)=\mu+\sigma(M_l-M_u), \quad x=1,2,\cdots,k \quad (16)$$

여기서, M_l과 M_u는 다음과 같다.

$$M_l = \frac{\phi(z_{x-1})}{\Phi(z_x) - \Phi(z_{x-1})}, \quad M_u = \frac{\phi(z_x)}{\Phi(z_x) - \Phi(z_{x-1})} \tag{17}$$

z_x는 경계점 a_x의 표준화 값으로 $(a_x - \mu)\ /\sigma$로 정의된다. 경계점의 양 끝값인 a_0와 a_k는 각각 $-\infty$와 ∞이므로 이 점에서의 확률밀도값은 0이며, 누적확률값은 각각 0과 1이 된다. 식 (16)으로부터 기저점수 w_x의 추정값을 얻기 위해서는 경계점 a_x와 기저분포의 모수인 μ, σ를 추정하여야 한다. 이 경우에 대수우도함수는 다음과 같다.

$$\log L(a, \mu, \sigma) \propto \sum_{x=1}^{k} r_x \cdot \log\left\{ \Phi\left(\frac{a_x - \mu}{\sigma}\right) - \Phi\left(\frac{a_{x-1} - \mu}{\sigma}\right) \right\} \tag{18}$$

기저분포가 정규분포일 경우 기저점수 w_x는 앞서 제시한 알고리즘을 적용하여 추정할 수 있는데 이에 대한 자세한 과정은 박정선(2001)을 참고할 수 있다.

3.2 기저분포가 베타분포인 경우

일반적으로 응답자의 반응은 대칭형만을 가정할 수 없다. 예를 들어 어떤 제품의 만족도 조사에서 응답자들은 자신의 경험에 의존하여 반응하는데 경험에 따라 긍정적인 면이 강할 수도 있고, 부정적인 면이 강하기도 할 것이다. 따라서 이러한 상황을 반영하기 위해서는 기저분포에 대한 가정을 바꿀 필요가 있다. 여기서는 기저분포로 모수 값에 따라 다양한 형태를 갖는 베타분포를 고려하기로 한다. 두 개의 모수 α, β를 갖는 베타분포의 확률밀도함수는 (19)와 같다.

$$f(w) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < w < 1 \tag{19}$$

(19)에서 $B(\alpha, \beta)$는 베타함수를 나타낸다. 베타분포는 모수 α, β 값에 따라 다양한 형태를 갖는다. $\alpha < \beta$이면 오른쪽으로 기울어진(right skewed) 형태를 취하며, $\alpha > \beta$이면 왼쪽으로 기울어진(left skewed) 형태를 갖고 α와 β가 같으면 분포는 좌우대칭 형태를 띤다(Johnson and Kotz, 1970).

범주 x에서 기저변수 W에 대한 조건부 확률밀도함수는 (20)과 같으며, 이는 절삭된 베타분포(truncated beta distribution)와 같다.

$$g(w \mid x) = \frac{1}{\pi_x} f(w), \quad a_{x-1} < w \le a_x \tag{20}$$

(20)에서 확률 π_x는 범주 x가 관찰될 확률로 (21)과 같다.

$$\pi_x = F(a_x) - F(a_{x-1}), \quad F(a_x) = \int_0^{a_x} f(t)dt$$
$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}[B(\alpha_x;\alpha,\beta) - B(\alpha_{x-1};\alpha,\beta)]$$
$$\text{단, } B(\alpha_x;\alpha,\beta) = \int_0^{a_x} t^{\alpha-1}(1-t)^{\beta-1}dt \tag{21}$$

(21)을 이용하면 기저분포가 베타분포일 경우의 수량화 값인 조건부 기대값은 (22)와 같이 구할 수 있다(Philips and Cho 2000).

$$E(W|x) = \frac{1}{\pi_x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot [B(\alpha_x;\alpha+1,\beta) - B(\alpha_{x-1};\alpha+1,\beta)] \tag{22}$$

기저점수 w_x를 추정하기 위해서는 경계점 a_x와 기저분포의 모수인 α, β를 추정하여야 하며, 이를 위해 알고리즘 3.1을 적용한다. 이 경우 대수우도함수는 (23)과 같다.

$$\log L(a, \mu, \sigma) = C + \sum_{x=1}^{k} r_x \cdot \log \pi_x \propto \sum_{x=1}^{k} r_x \cdot \log\{F(a_x;\alpha,\beta) - F(a_{x-1};\alpha,\beta)\} \tag{23}$$

여기서 $F(\alpha_x;\alpha,\beta)$는 불완전베타함수(incomplete beta function)로서 (24)와 같이 정의된다.

$$F(a_x;\alpha,\beta) = \int_0^{a_x} \frac{1}{B(\alpha,\beta)} t^{\alpha-1}(1-t)^{\beta-1}\,dt \equiv I_x(\alpha,\beta) \tag{24}$$

기저분포가 베타분포일 경우 기저점수 w_x는 앞서 제시한 알고리즘을 적용하여 추정할 수 있는데 이에 대한 자세한 과정은 박정선(2001)을 참고할 수 있다.

3.3 수량화 값 추정 상의 문제점

제안된 알고리즘을 적용해서 실제로 리커트 척도를 수량화하기 위해서는 몇 가지 문제점들이 해결되어야 한다. 첫 번째 문제는 리커트 척도를 사용한 조사에서는 흔히 응답자들이 하나의 범주에 집중하여 반응함에 따라 어떤 범주에서는 응답이 하나도 관찰되지 않는 경우에 발생된다. 이는 범주형 변수를 관찰하는 경우에 흔히 발생되는데 이 경우에는 각 범주에는 0.5를 할당하는 것이 일반적이다. 또 다른 문제는 응답자가 응답을 모두 하나의 범주로 집중하는 경우인데 만약 기저분포를 정규분포로 적용하면 표준편차가 0이 되어 기저분포의 모수를 추정하지 못하는 문제가 발생된다. 이 경우에는 두 가지 해결책을 생각할 수 있는데 한 가지는 표본 표준편차 값으로 강제로 0보다 큰 값을 임의로 할당하는 것이며, 다른 하나는 수량화 과정을 거치지 않고 수량화 값으로 리커트 척도값을 그대로 사용하는 것이다.

또한 수량화를 위해서 기저분포의 모수값을 반복법을 적용한 최우추정법으로 추정하는데 이 경우에 초기값을 어떻게 선택하느냐에 따라 최우추정량이 변하게 되는 문제가 발생된다. 그러나 구조방정식모형을 분석하기 위해서는 수량화 값을 이용하여 계산한 상관계수를 사용하므로 절대적인 값의 변화는 큰 영향을 미치지 않으리라고 판단된다. 반복 추정법을 적용할 때 초기값으로는 리커트 척도를 양적 자료로 변환시킨 값을 사용하는 것이 적절할 것이다.

4. 붓스트랩 모의실험을 통한 효율성 비교

리커트 척도를 수량화하여 구조방정식모

형을 분석할 경우 나타나는 추정량의 효율성을 알아보기 위하여 모의실험을 통해 추정량의 성질을 살펴보기로 한다. 본 논문에서는 리커트 척도로 자료를 조사할 때 개인의 응답에 차이가 있다고 가정하였다. 이러한 개인의 반응 차이를 확률 모형화하는 것이 어렵기 때문에 본 연구에서는 실제 자료를 이용한 붓스트랩 모의실험을 이용하였다. 붓스트랩 표본은 1000번 실험하며, 표본 추출률을 10%, 20%, 30%로 변화시켜 추정량의 움직임을 관찰한다. 추정량의 평균제곱오차(mean square error)를 비교하여 효율성을 평가하며, 구조방정식의 추정을 위해서는 최우추정법을 이용한다.

효율성은 수량화 방법을 적용하여 구조방정식모형을 추정한 경우와 리커트 척도를 양적자료로 취급하여 분석한 경우 - 이를 리커트 방법이라 부르기로 한다 -를 비교하기로 하며, 수량화 방법의 경우에는 기저분포로 정규분포를 사용한 경우와 베타분포를 사용한 경우로 분리하기로 한다. 기저분포를 정규분포를 사용한 경우를 정규분포방법, 베타분포를 사용한 경우를 베타분포방법이라고 각각 부르기로 한다.

4.1 실험모형

모형에 사용된 변수는 표 1과 같으며, 각 항목들은 5점 척도로 측정되었다. 실험에 사용된 측정모형과 구조모형은 각각 다음과 같다.

측정모형 :

$$\underset{\sim}{y} = \Lambda_y \cdot \eta + \varepsilon \qquad \underset{\sim}{x} = \Lambda_x \cdot \xi + \delta$$

구조모형 :

$$\eta = \mathrm{B} \cdot \eta + \Gamma \cdot \xi + \zeta$$

여기서,

$$\Lambda_x(LX) = \begin{bmatrix} \lambda_{11}^x & \lambda_{21}^x & \lambda_{31}^x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{42}^x & \lambda_{52}^x & \lambda_{62}^x & \lambda_{72}^x & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{83}^x & \lambda_{93}^x & \lambda_{103}^x & 0 \\ 0 & \lambda_{24}^x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^T$$

$$\Lambda_y(LY) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathrm{B}(BE) = \begin{pmatrix} 0 & \beta_{12} \\ 0 & 0 \end{pmatrix},$$

$$\Gamma(GA) = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} \\ 0 & 0 & \gamma_{23} & \gamma_{24} \end{pmatrix}$$

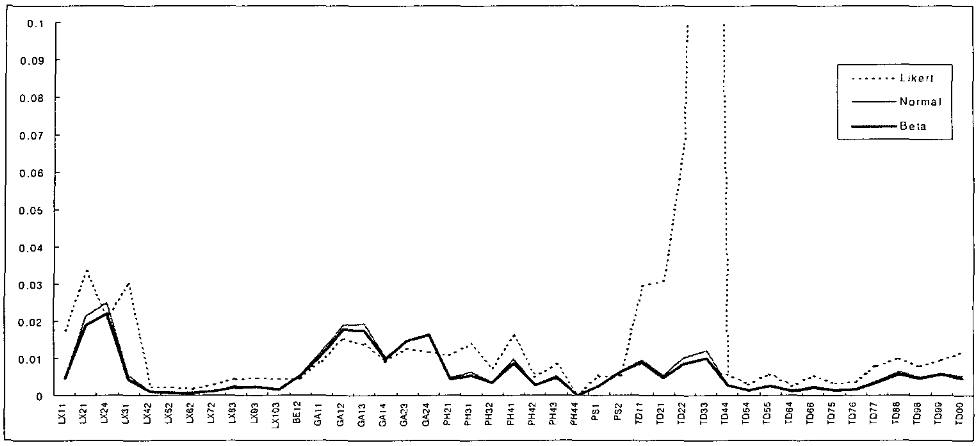
$$\Phi(PH) = \begin{bmatrix} 1 & & & \\ \phi_{21} & 1 & & \\ \phi_{31} & \phi_{32} & 1 & \\ \phi_{41} & \phi_{42} & \phi_{43} & 1 \end{bmatrix}$$

$$\Psi(PS) = \begin{bmatrix} \varphi_1 & \\ 0 & \varphi_2 \end{bmatrix}$$

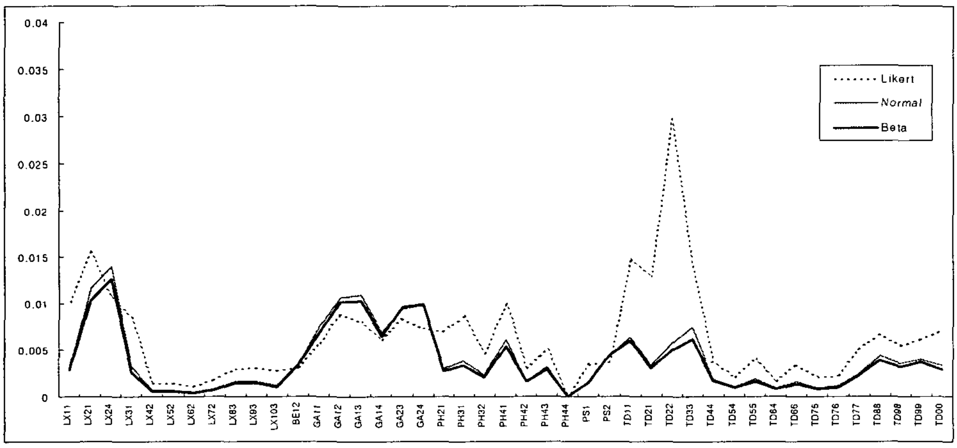
$$\Theta_\delta(TD) = \begin{bmatrix} \theta_{11}^\delta & & & & & & & & & \\ 0 & \theta_{22}^\delta & & & & & & & & \\ 0 & 0 & \theta_{33}^\delta & & & & & & & \\ 0 & 0 & 0 & \theta_{44}^\delta & & & & & & \\ 0 & 0 & 0 & \theta_{54}^\delta & \theta_{55}^\delta & & & & & \\ 0 & 0 & 0 & \theta_{64}^\delta & 0 & \theta_{66}^\delta & & & & \\ 0 & 0 & 0 & 0 & \theta_{75}^\delta & \theta_{76}^\delta & \theta_{77}^\delta & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{88}^\delta & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{98}^\delta & \theta_{99}^\delta & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{00}^\delta \end{bmatrix}$$

4.2 추정결과 비교

실험모형에 대하여 표본 추출율을 10% (n



<그림 1> 최우추정량의 평균제곱오차, B=1000, n=140



<그림 2> 최우추정량의 평균제곱오차, B=1000, n=210

=70), 20%(n=140), 30%(n=210)로 변화시켜 최우추정법을 적용할 경우에 나타나는 추정값의 변화를 살펴보았다. 이 중 표본 추출율을 10%로 한 경우에는 모형이 식별되지 않는 경우가 많아 제외하고 표본 추출율이 20%, 30%인 경우에 대해서만 분석하였다. 표본 추출율이 10%인 경우에는 모형의 식별가능성을 비교하였다. 추정방법에 따른 추정량의 정확성을 평가하기 위해서는 평균제곱오차(MSE)를 비교한다. 그림 1과 그림 2는 세 가지 방법에 따른 추정량의 평균제곱오차를 표본 추출율에 따라 보여주고 있다. 이에 대한 자세한 결과는 부록의 표 1과 표 2에 제시되어 있다.

계수행렬의 경우 측정모형의 계수인 LX와 관련된 추정량의 평균제곱오차는 수량화 방법이 리커트 방법보다 작아 더 정확한 것으로 나타났으나, 구조모형의 계수인 BE와

<표 1> 평균제곱오차 : B=1000, n=140(추출율=20%)

모수	리커트	정규분포	베타분포	모수	리커트	정규분포	베타분포
LX11	0.01750	0.00485	0.00445	PH42	0.00491	0.00278	0.00265
LX21	0.03338	0.02126	0.01880	PH43	0.00840	0.00514	0.00474
LX24	0.02062	0.02495	0.02210	PH44	0.00000	0.00000	0.00000
LX31	0.03012	0.00538	0.00428	PS1	0.00528	0.00259	0.00241
LX42	0.00221	0.00097	0.00089	PS2	0.00555	0.00645	0.00644
LX52	0.00224	0.00091	0.00078	TD11	0.02959	0.00960	0.00896
LX62	0.00172	0.00072	0.00060	TD21	0.03085	0.00538	0.00486
LX72	0.00289	0.00121	0.00108	TD22	0.06898	0.01031	0.00868
LX83	0.00465	0.00246	0.00212	TD33	0.41574	0.01212	0.01007
LX93	0.00485	0.00238	0.00218	TD44	0.00588	0.00291	0.00272
LX103	0.00451	0.00183	0.00155	TD54	0.00302	0.00164	0.00142
BE12	0.00457	0.00541	0.00518	TD55	0.00599	0.00284	0.00250
GA11	0.00899	0.01187	0.01066	TD64	0.00260	0.00136	0.00119
GA12	0.01537	0.01878	0.01764	TD66	0.00557	0.00245	0.00206
GA13	0.01392	0.01926	0.01759	TD75	0.00349	0.00155	0.00129
GA14	0.00915	0.01030	0.00968	TD76	0.00355	0.00175	0.00154
GA23	0.01266	0.01463	0.01488	TD77	0.00786	0.00377	0.00341
GA24	0.01190	0.01605	0.01626	TD88	0.01004	0.00639	0.00561
PH21	0.01100	0.00478	0.00429	TD98	0.00801	0.00505	0.00444
PH31	0.01396	0.00629	0.00551	TD99	0.00962	0.00594	0.00557
PH32	0.00723	0.00373	0.00333	TD00	0.01167	0.00529	0.00458
PH41	0.01608	0.00984	0.00866				

GA 추정량의 경우에는 리커트 방법의 평균제곱오차가 작은 것으로 나타났다. 공분산행렬과 관련된 추정량의 평균제곱오차는 전체적으로 수량화 방법을 적용한 경우에 작아서 더 정확한 것으로 볼 수 있다. 특히 리커트 방법을 적용할 경우 TD11, TD21, TD22, TD33의 값이 수량화 방법의 경우보다 상당히 큰 값으로 관찰되었다.

수량화 방법의 경우 정규분포 방법이나 베타분포 방법에서 값의 크기는 거의 차이가 없는 것으로 나타났으나 베타분포 방법이 약간 작은 값을 갖는다. 전체적으로 수량화 방법을 적용할 경우 평균제곱오차가 리커트 방법보다 작아 더 정확한 추정 결과를 제공하는 것으로 볼 수 있다.

앞서 지적한 바와 같이 표본 추출율이

<표 2> 평균제곱오차 : B=1000, n=210(추출율=30%)

모수	리커트	정규분포	베타분포	모수	리커트	정규분포	베타분포
LX11	0.01021	0.00317	0.00295	PH42	0.00296	0.00167	0.00160
LX21	0.01558	0.01158	0.01040	PH43	0.00498	0.00316	0.00289
LX24	0.01067	0.01395	0.01252	PH44	0.00000	0.00000	0.00000
LX31	0.00845	0.00327	0.00259	PS1	0.00347	0.00159	0.00151
LX42	0.00141	0.00062	0.00057	PS2	0.00376	0.00443	0.00444
LX52	0.00148	0.00060	0.00051	TD11	0.01474	0.00633	0.00599
LX62	0.00106	0.00047	0.00038	TD21	0.01287	0.00340	0.00309
LX72	0.00185	0.00080	0.00070	TD22	0.02974	0.00571	0.00496
LX83	0.00303	0.00165	0.00145	TD33	0.01403	0.00749	0.00618
LX93	0.00306	0.00160	0.00147	TD44	0.00380	0.00188	0.00176
LX103	0.00281	0.00117	0.00099	TD54	0.00205	0.00112	0.00097
BE12	0.00306	0.00358	0.00342	TD55	0.00401	0.00189	0.00164
GA11	0.00569	0.00752	0.00692	TD64	0.00163	0.00089	0.00079
GA12	0.00895	0.01060	0.01015	TD66	0.00344	0.00160	0.00133
GA13	0.00800	0.01094	0.01024	TD75	0.00211	0.00095	0.00077
GA14	0.00613	0.00687	0.00643	TD76	0.00227	0.00116	0.00102
GA23	0.00835	0.00946	0.00965	TD77	0.00509	0.00249	0.00223
GA24	0.00740	0.00979	0.00995	TD88	0.00675	0.00434	0.00389
PH21	0.00702	0.00304	0.00274	TD98	0.00543	0.00353	0.00317
PH31	0.00866	0.00380	0.00335	TD99	0.00621	0.00399	0.00375
PH32	0.00460	0.00239	0.00213	TD00	0.00709	0.00338	0.00293
PH41	0.00996	0.00612	0.00540				

10%(n=70)일 때는 많은 경우에 모형이 식별되지 않았으므로 이 경우에는 방법에 따른 모형의 식별율을 비교한다. 여기서 모형 식별율이란 모형에 있는 모든 모수를 적절하게 식별한 모형의 비율을 의미한다. 최우추정법을 적용할 경우에 모형식별율은 기저분포로 베타분포를 이용한 경우가 75.3%로 가장 높으며, 이어서 정규분포를 사용한 경우가 73.9%를 식별하였다. 반면에 리커트 방법을 이용할 경우 모형 식별율은 63.6%로베타분포 방법이나 정규분포 방법보다 상당히 낮게 나타났다.

5. 결 론

현재 실용적인 측면에서 많은 자료가 리커

트 척도를 이용하여 측정되고 이를 양적인 자료로 간주하여 분석하고 있다. 그러나 엄밀한 의미에서 리커트 척도는 범주형 자료이므로 이러한 분석 방법은 수정되어야 할 것이다. 이는 구조방정식모형을 분석할 때도 마찬가지로 본 논문에서는 기존에 제시된 두 가지 방안을 비교 설명하였으며, 새로운 방안으로 리커트 척도를 수량화하는 방법을 제안하였다.

모의실험 결과 일부의 경우를 제외하고는 수량화 방법을 적용한 방법이 리커트 척도를 양적 자료로 취급하는 경우보다 효율성과 정확성에서 앞서는 것을 확인하였다. 구조방정식모형의 모수 행렬 중 특히 측정모형의 계수 행렬 및 공분산 행렬과 관련된 추정량의 평균제곱오차 값이 리커트 척도를 양적 자료로 간주하여 분석하는 경우보다 작게 추정되어 보다 효율적이고 정확한 결과를 얻을 수 있었다. 그러나 구조모형의 계수 행렬과 관련된 모의실험 결과에서는 리커트 척도를 양적인 형태로 사용할 경우가 수량화 방법을 사용하는 경우보다 평균제곱오차를 작게 하는 것으로 나타났다. 또한, 표본 추출율이 작을 경우에는 모형을 식별할 수 없는 경우가 발생하였는데 이들의 모형 식별율을 비교한 결과 수량화 방법을 적용하여 추정한 경우가 리커트 척도를 그대로 이용한 경우보다 높은 모형 식별율을 보이는 것을 확인하였다.

그러나 이러한 결과는 모의실험에 의한 결과이므로 일반화하여 적용하는데는 어려움이 있으리라고 판단된다. 또한 계산상의 어려움으로 인해 모의실험 과정에서 모형 추정방법 중 다범주 상관계수를 이용한 가중최소제곱법과 편최소제곱법을 이용한 추정방법을 비교 대상에서 제외하였다. 이로 인해 본 연구의 결과를 기존의 방법과 비교할 수가 없었으며, 이는 향후 추가적인 연구과제로 삼아야 할 것이라고 판단된다.

참고문헌

[1] 박정선(2001), “리커트 척도의 수량화를 이용한 구조방정식모형 분석”, 연세대학교 대학원 박사학위논문.

[2] 조현철(1999), Lisrel에 의한 구조방정식 모델. 서울 : 석정.

[3] 최엄문(1998), “편최소제곱법과 그의 응용에 관한 연구”, 서울대학교 대학원 박사학위논문.

[4] Agresti, Alan(1990), *Categorical Data Analysis*. New York : John Wiley and Sons.

[5] Barr, Donald R., and E. Todd Sherill.(1999), "Mean and Variance of Truncated Normal Distributions". *The American Statistician*, Vol. 53, No. 4, pp. 357-361.

[6] Bollen, Kenneth A.(1989), *Structural Equations with Latent Variables*. New York : John Wiley and Sons.

[7] Fornell, Claes (1992), "A National Customer Satisfaction Barometer : The Swedish Experience". *J. of Marketing*, Vol. 56(January), pp. 6-21.

[8] ———(1996), "The American Customer Satisfaction Index : Nature, Purpose, and Findings". *J. of Marketing*, Vol. 60(October) pp. 7-18.

[9] ———, and Fred L. Bookstein(1982), "Two Structural Equation Models : LISREL and PLS Applied to Consumer Exit-Voice Theory". *J. of Marketing*

Research vol XIX(November) pp. 440-452.

[10] Hair Jr., Joseph F and R. E. Anderson, R. L. Tatham, and W. C. Black(1998). *Multivariate Data Analysis* 5th Ed. New Jersey : Prentice-Hall.

[11] Johnson, Norman, L., and Samuel Kotz(1970). *Continuous Univariate Distributions-2.* Boston : Houghton Mifflin Company.

[12] Jöreskog, Karl(1994), "On the Estimation of Polychoric Correlations and their Asymptotic Covariance Matrix". *Psychometrika*, Vol. 59 pp. 381-389.

[13] ———, and Dag Sörbom(1996a), *LISREL 8 : User's Reference Guide.* Chicago : Scientific Software International.

[14] ———, and ———(1996b), *PRELIS 2 : User's Reference Guide.* Chicago : Scientific Software International.

[15] Lee, Sik-Yum and Wai-Yin Poon(1987). "Two-Step Estimation of Multivariate Polychoric Correlation". Communications of Statistics-Theory and methods, Vol. 16, No. 2 : 307-320.

[16] ———, and P. M. Bentler(1990). "A Three-Stage Estimation Procedure for Structural Equation Models with Polytomous Variables". *Psychometrika,* Vol.55 : 45-51.

[17] Muthén, Bengt(1984). "A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators". *Psychometrika*, Vol. 49 : 115-132.

[18] Olsson, Ulf(1979), "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient". *Psychometrika,* Vol. 44 pp. 443-460.

[19] Philips, Michael D. and Byung-Rae Cho(2000). "A Nonlinear Model for Determining the Most Economic Process Mean under Beta Distribution". *International J. of Reliability, Quality and Safety Engineering*, Vol. 7, No. 1 : 61-74.

[20] Schneider, Helmut(1986). *Truncated and Censored Samples from Normal Populations.* New York : Marcel Dekker.

[21] Wold, Herman(1975), "Soft Modeling by Latent Variables ; The Non-Linear Iterative Partial Least Squares Approach", in *Perspectives in Probability and Statistics* edited by J. Gani, Lodon : Academic Press. Quoted in 최엽문(1998).

[22] ———(1982), "System Under Indirect Observation Using PLS" in *A second generation of multivariate analysis* edited by Claes Fornell pp. 325-347. New York : Praeger Publishers.