예방의학회지 : 제31권 제4호
Korean J. of Preventive Medicine
Vol. 31, No. 4, November. 1998

보건조사연구에서 다변량결측치가 내포된 자료를 효율적으로 분석하기 위한 통계학적 방법

김동기, 박은철*, 손명세*, 김한중*, 박형욱*, 안재형, 임종건, 송기준

연세대학교 의과대학 의학통계학과, 예방의학교실*

= Abstract =

Statistical Methods for Multivariate Missing Data in Health Survey Research

Dong Kee Kim, Eun-Cheol Park*, Myongsei Sohn*, Han Joong Kim*, Hyung Uk Park*, Chae Hyung Ahn, Jong Gun Lim, Ki Jun Song

Department of Biostatistics and Department of Preventive Medicine and Public Health, Yonsei University College of Medicine*

Missing observations are common in medical research and health survey research. Several statistical methods to handle the missing data problem have been proposed. The EM algorithm (Expectation-Maximization algorithm) is one of the ways of efficiently handling the missing data problem based on sufficient statistics. In this paper, we developed statistical models and methods for survey data with multivariate missing observations. Especially, we adopted the EM algorithm to handle the multivariate missing observations. We assume that the multivariate observations follow a multivariate normal distribution, where the mean vector and the covariance matrix are primarily of interest. We applied the proposed statistical method to analyze data from a health survey. The data set we used came from a physician survey on Resource-Based Relative Value Scale(RBRVS). In addition to the EM algorithm, we applied the complete case analysis, which uses only completely observed cases, and the available case analysis, which utilizes all available information. The residual and normal probability plots were evaluated to access the assumption of normality. We found that the residual sum of squares from the EM algorithm was smaller than those of the complete-case and the available-case analyses.

*본 연구는 1997년도 한국과학재단 핵심전문연구비(No : KOSEF 971-0105-026-1)로 이루어졌음.

Key words : Missing data; Multivariate normal data; EM algorithm; Biostatistics; Resource-Based Relative Value Scale

Ⅰ. 서 론

보건조사연구의 목적 중 하나는 연구대상의 특성이나 의견을 파악하는 일이다. 이를 위하여 흔히 사용하는 방법은 우편조사나 면접조사에 의해 설문자료를 얻는 것이다. 이러한 조사연구결과가 신뢰할 만하다는 것을 입증하려면 충분한 양의 자료가 필수적이다. 또한 자료가 충분할 뿐만 아니라 살펴보고자 하는 제반변수가 모두 완전히 관찰(complete observation)되어야 한다. 그러나 보건조사연구에서는 충분한 양의 대상에 대해 제반 변수를 완전하게 관찰하기란 쉬운 일이 아니다. 따라서 자료가 불완전하게 관찰(incomplete observation)되어 부분적으로 결측치(missing value)가 얻어지는 경우가 많다.

보건조사연구에서 결측치의 문제점은 흔히 발생하는 중요한 문제임에도 불구하고 이를 해결하기 위한 효과적인 통계학적인 방법은 그리 활발하게 연구되지 못하였다. 그 이유는 결측치를 갖게되면 이에 대한 적절한 통계학적 분석방법을 적용하기가 용이하지 못하기 때문이다.

지금까지의 보건조사연구에서는 결측치가 내포되어 있으면 결측치가 내포된 조사대상을 제외하고 제반변수에 대해 완전하게 관찰된 대상만을 사용하여 분석하였다. 이 방법은 사용상 편리함과 분석방법의 제한으로 인하여 통상적으로 적용되어 왔지만 결측치를 포함한 대상을 제외시킴으로 나타나는 표본규모의 감소 등에 따른 통계학적 검정력의 약화(loss of power), 보고자 하는 주요 변수와 관련된 모수 추정치(parameter estimate)의 편의(bias), 비효율성(inefficiency) 등의 문제는 이미 널리 알려져 왔다(Beale and Little, 1975; Little and Rubin, 1986).

이러한 문제의 대처방안으로 결측치에 대한 연구는 그간 대치법(imputation method)을 중심으로한 단순한 형태에 대해 이루어져 왔다. 이러한 기존의 연구는 Dempster 등(1976)의 EM algorithm에 의해 단순한 형태의 결측치에 대한 통계학적 해결은 물론 분산성분모형(variance component models)과 요인분석 등 제반 연구에 널리 적용되어 있다(Dempster 등, 1976, Little and Rubin, 1987). Little(1992)은 회귀분석에서 독립변수에 결측치가 있을 경우에 EM algorithm의 적용에 대해 연구하였으며, Kim과 Taylor(1995)는 이 EM algorithm을 사용하였을 때 우도비검정과 우도신뢰구간을 구하는 방안에 대하여 연구하였다.

이 연구에서는 보건조사연구에 결측치를 해결하기 위한 통계학적 방법을 제시하고자 한다. 특히 EM algorithm의 분석틀 안에서 다변량결측치를 내포한 보건의학조사자료의 통계학적 분석에 초점을 맞춘다. 또한 실제 보건조사 연구자료에 적용하여 그 결과를 분석하고자 한다.

Ⅱ. 자료 및 연구방법

1. 다변량 결측자료의 예

김한중 등(1997)은 현행의료보험 진료수가체계 전반에 관한 개선방안을 검토하고자 한국표준의료행위분류와 한국표준치과의료분류별 상대가치의 개발에 관한 연구를 하였다. 이를 위하여 각 전문과목별로 의료행위의 진료에 관한 진료업무량, 시간, 기술적 노력, 정신적 노력, 그리고 스트레스의 항목에 대하여 설문조사를 실시하였는데 중요한 문제 중의 하나는 응답율이 낮고 자료의 기입율은 낮다는 점이다. 〈표 1〉은

표 1. 보건조사연구에서 특정진료과의 상대가치에 관한 각 문항별 설문조사 결과

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	2.176	2.789	2.653	2.69	2.653	2.342	2.447
2	*	*	2.398	2.477	2.477	2.301	*
3	2.342	2.845	2.415	2.602	2.477	2.477	2.699
4	2	*	2.255	2.447	*	2.267	2.274
5	*	*	2.322	2.362	*	2.23	2.255
6	2.255	*	2.279	2.362	2.362	2.255	2.255
7	*	2.519	2.322	2.398	2.398	1.845	1.903
8	2	*	2.531	2.602	2.602	2.301	2.398
9	1.699	2.681	2.519	2.699	2.74	2.602	2.653
10	*	*	2.415	*	1.903	2.079	2.301
11	*	*	2.544	*	2.602	*	2.477
12	3	*	2.544	2.602	2.477	2.301	2.477
13	1.699	2.477	2.477	2.602	2.602	2.477	2.477
14	*	2.398	2.362	2.477	2.447	2.398	2.398
15	2.477	2.724	2.699	2.771	2.778	2.602	2.903
16	*	*	*	2.447	2.447	2.301	*
17	2.477	2.653	2.602	2.699	2.699	2.477	2.477
18	*	*	2.322	2.362	*	2.176	*
19	*	2.519	2.322	2.398	2.398	1.845	1.903
20	*	2.519	2.322	2.398	2.398	1.845	1.903

단, *는 결측치를 나타냄
X_1 = 외래재진, 심층진료
X_2 = 입원중기(급성최장염, 입원익일－퇴원전날) 기초진료
X_3 = 협의진료 : 재왕절개술환자의 소화불량으로 입원, 초진, 기초진료
X_4 = 협의진료 : 당뇨병 입원환자, 재진, 기초진료
X_5 = 응급환자 : 좌하복부통증, 기초진료
X_6 = 식도운동검사 : 풍선유발검사
X_7 = 식도운동검사 : 산 재거능 검사

특정진료과에서 7개 의료행위에 대한 설문에서 얻은 결과이다. 연구의 주 목적은 설문결과 각 의료행위에 대한 평균점수와 표준편차를 추정하는 일이다. 6명의 응답자는 7개 항목 모두 응답하였으나 나머지 응답자는 하나이상의 의료행위에 응답을 하지 않음으로 인하여 결측치가 발생하였다. 이 결측치를 내포한 자료를 제거하여 분석하면 사용할 수 있는 자료의 수가 너무 적어질 뿐만 아니라 제외된 자료가 분석에 사용된 자료와 특성에 있어서 크게 차이나면 분석결과의 신뢰성이 크게 떨어지게 된다.

2. 결측치의 보완

본 연구에서는 이 결측치를 제거하여 분석하는 방법보다는 이 결측치를 통계학적으로 보완하여 분석하는 기법을 제시하고자 한다. 결측치를 보완하여주는 통계학적인 방법은 그 동안 대치법(imputation method)을 사용하는 단순한 방법을 크게 벗어나지 못하였다. 그러나 Dempster 등(1977)의 EM algorithm의 개발로 결측치에 대한 통계학적 분석방법으로 획기적인 발전을 가져왔다. 본 연구에서는 결측치를 보완하여주는 방법으로 EM algorithm을 중점적으로 적용하고자한다. 결측자료에 대한 기본적인 가정은 임의결측(Missing at Random, MAR)을 가정한다.

최대우도추정법(maximum likelihood estimation method)은 관심을 가지고 있는 변수에 관련된 모수(parameter)를 추정하는 대표적인 통계학적 추정방법이다. 이 최대우도추정법을 보건조사연구의 결측치에 적용할 수 있는데 이 추정방법을 얼마나 쉽게 적용할 수 있는 가는 그 얻어진 결측치가 어떤 형태를 지니는 가에 좌우된다. 즉 결측치가 어떤 특정한 형태를 지녀서 우도(likelihood)를 분리(factorization)시킬 수 있으면 최대우도추정법을 쉽게 적용할 수 있고 그렇지 않으면 쉽게 적용할 수 없게 된다. 또한 어떤 특성한 통계학적 모형에 있어서는 우도의 분리가 가능하더라도 각각 분리된 우도부분에 속한 모수가 서로 분리되지 못하기 때문에 최우추정치를 얻지 못하는 경우가 많다. 그러므로 결측치가 내포된 경우에 우도와 추정이 필요한 모수를 분리시키는 작업이 그 동안 관심의 대상이 되어왔고 그 해결방법의 하나로 EM(Expectation-Maximization) algorithm이 제시되었다(Dempster 등, 1976). 이 EM algorithm은 특히 결측치가 내포되어 있는 경우에 최대우도추정법으로 그 효율성과 안정성 등이 증명되었다(Wu, 1983). 이 EM algorithm은 기존의 결측치를 해결하는 단순한 방법인 대치법(imputation method)을 확대하여 반복계산(iteration)을 가능하게 하였다. 즉 (1) 결측치를 추정한 값으로

대치하며 (2) 모수를 추정하고 (3) 추정된 모수가 사실이라는 가정하에 결측치를 재추정하며 (4) 모수를 재추정한다. 이러한 과정을 반복 추정된 모수가 일정한 값으로 접근할 때까지 계속하게 된다. 이런 반복계산은 크게 두 가지의 단계로 나누어지는데 그 중 하나는 E-step(Expectation Step)이고, 다른 하나는 M-step (Maximization Step)이다. E-step에서는 추정된 모수의 값과 관찰된 값이 주어졌다는 조건하에 결측치의 조건부 기대값(conditional expectation)을 구하며 M-step에서는 결측치가 모두 대치되었다는 가정하에 최우추정치를 구하게 된다. 이 EM algorithm은 단순하면서도 효율적인데, 그 이유는 통계학적 모형에서 유효하게 쓰이는 충분통계량(sufficient statistics)에 기반을 두었기 때문이다. 또한 M-step의 과정에서도 결측치가 모두 대치되었다는 가정하에 완전자료(complete data)에서의 최대우도추정방법을 그대로 적용할 수 있게 되기 때문이다.

3. 이변량 정규모형의 경우 EM algorithm의 적용

결측치를 내포한 자료 중에서 단순한 모형은 이변량정규분포모형이다. 첫 번째 변수 Y_1은 모두 관찰되었고, 다른 변수 Y_2는 결측치를 포함하고 있는 경우를 가정해 보겠다. 이변량정규분포모형에서 추정해야 될 모수는 $\theta=(\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{12})$이다. 여기서 μ_1과 μ_2는 Y_1과 Y_2의 평균이며, $\sigma_{11}, \sigma_{22}, \sigma_{12}$는 Y_1의 분산, Y_2의 분산, Y_1과 Y_2의 공분산이다. 이변량정규분포모형을 위한 EM algorithm은 다음의 충분통계량에 기초하고 있다.

$$S_1=\sum_{i=1}^{n} Y_{i1},\ S_2=\sum_{i=1}^{n} Y_{i2}$$

$$S_{11}=\sum_{i=1}^{n} Y_{i1}^2,\ S_{22}=\sum_{i=1}^{n} Y_{i2}^2,\ S_{12}=\sum_{i=1}^{n} Y_{i1}Y_{i2}$$

EM algorithm의 E-step과 M-step은 다음과 같다.

(E-step)

Y_{i1}은 모두 관찰되었고 Y_{i2}는 결측치를 내포한 경우를 가정하여 보자. E-step에서는Y_{i2}가 결측치인 경우에 Y_{i1}을 독립변수로 사용하여 Y_{i2}에 관한 회귀분석 결과를 사용하게 된다.

$$E(Y_{i2}|Y_{i1}, \theta)=\beta_{20.1}+\beta_{21.1}Y_{i1}$$
$$E(Y_{i2}^2|Y_{i1}, \theta)=(\beta_{20.1}+\beta_{21.1}Y_{i1})_2+\sigma_{22.1}$$
$$E(Y_{i2}Y_{i1}|Y_{i1}, \theta)=(\beta_{20.1}+\beta_{21.1}Y_{i1})Y_{i1}$$

이 때, $\beta_{20.1}$, $\beta_{21.1}$, $\sigma_{22.1}$은 Y_{i1}을 독립변수로 사용한 Y_{i2}에 관한 회귀분석결과의 절편, 기울기 그리고 분산이다.

(M-step)

M-step에서는 E-step에서 구한 조건부기대값을 사용하여 다음 단계의 θ를 구한다.

$$\hat{\mu_1}=\frac{S_1}{n},\ \hat{\mu_2}=\frac{S_2}{n},\ \hat{\sigma_{11}}=\frac{S_{11}}{n}-\hat{\mu_1^2},\ \hat{\sigma_{22}}=\frac{S_{22}}{n}-\hat{\mu_2^2},$$

$$\hat{\sigma_{12}}=\frac{S_{12}}{n}-\hat{\mu_1}\hat{\mu_2}$$

위 E-step과 M-step을 모수 θ가 한 점으로 수렴할 때까지 반복 계산한다.

4. 다변량 정규모형에 EM algorithm의 적용

이 절에서는 3절에서 설명한 이변량 정규분포모형을 다변량 정규분포모형으로 확대하고자 한다. 표 1과 같은 다변량자료에서 EM algorithm으로 추정하고자 하는 모형은 다음과 같다. 각과에서 행위 1, 행위 2, …, 행위 p는 p차원의 모수 μ와 $\sum$를 갖는 다변량 정규분포를 가정한다. 여기서 μ는 p차원의 평균벡터이고, $\sum$는 $p \times p$차원의 공분산행렬이다. Y_{i1}, Y_{i2}, …, $Y_{ip}(i=1, 2, \cdots, n)$을 p개 행위에 대한 i번째 의사의 응답이라고 할 때, EM algorithm은 다음의 충분통계량(sufficient statistics)에 기초하고 있다.

표 2. 상대가치 설문 문항별 평균 및 표준편차의 추정치

변수	EM algorithm		완전자료분석				가능한 자료분석			
	μ^{EM}	σ^{EM}	μ^{C}	$\frac{\mu^{C}-\mu^{EM}}{\sigma^{EM}}$	σ^{C}	$\frac{\sigma^{C}-\sigma^{EM}}{\sigma^{EM}}$	μ^{A}	$\frac{\mu^{A}-\mu^{EM}}{\sigma^{EM}}$	σ^{A}	$\frac{\sigma^{A}-\sigma^{EM}}{\sigma^{EM}}$
X1	2.356	0.704	2.145	−0.300	0.363	−0.485	2.212	−0.204	0.394	−0.440
X2	2.526	0.189	2.695	0.895	0.128	−0.325	2.612	0.459	0.147	−0.223
X3	2.432	0.128	2.561	1.008	0.109	−0.148	2.437	0.040	0.132	0.034
X4	2.516	0.139	2.679	1.170	0.066	−0.529	2.522	0.049	0.137	−0.016
X5	2.469	0.195	2.658	0.968	0.108	−0.444	2.498	0.145	0.200	0.028
X6	2.275	0.221	2.496	1.003	0.097	0.560	2.270	0.023	0.231	0.047
X7	2.353	0.250	2.609	1.024	0.178	−0.291	2.365	0.047	0.276	0.103
	절대값의 평균			0.910		0.397		0.138		0.127

$$\sum_{i=1}^{n} Y_{ij} \quad (j=1, \cdots, p)$$

$$\sum_{i=1}^{n} Y_{ij}Y_{ik} \quad (j, k=1, \cdots, p)$$

EM algorithm은 초기치를 준 후에 계속 E step과 M-step을 반복 계산하여 더 이상 추정된 모수의 값의 변화가 없을 때까지 진행하게 된다. EM algorithm의 E-step과 M-step은 다음과 같다.

$$E(\sum_{i=1}^{n} Y_{ij} | Y_{obs}, \theta^{(t)}) = \sum_{i=1}^{n} Y_{ij}^{(t)}, \; j=1, 2, \cdots, p$$

$$E(\sum_{i=1}^{n} Y_{ij}Y_{ik} | Y_{obs}, \theta^{(t)}) = \sum_{i=1}^{n} (Y_{ij}^{(t)}Y_{ij}^{(t)} + C_{jki}^{(t)}), \; j, k=1, 2, \cdots, p$$

여기에서 $Y_{ij}^{(t)}$는 관찰한 경우는 Y_{ij}를 그대로 사용하며 결측치가 있는 경우에 $E(Y_{ij}|Y_{obs,i}, \theta^{(t)})$를 계산한다. 그리고 $C_{jki}^{(t)}$는 j 혹은 k번수 히나리도 관찰한 경우에는 0이고 둘 다 모두 결측치인 경우에는 $COV(Y_{ij}, Y_{ik}|Y_{obs,i}, \theta^{(t)})$를 계산하여 구한다. 이들은 조건다변량정규분포(conditional mutivariate normal distribution)의 평균벡터와 공분산행렬에 해당하며 자세한 식은 Anderson(1984)의 조건다변량정규분포의 정의(36쪽)를 참조하기 바란다.

(M-Step)

M-step에서는 $(t+1)$번째의 값을 구하는데 완전자료의 경우 충분통계량을 이용한다.

$$\mu_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} Y_{ij}^{(t)}, \; j=1, 2, \cdots, p$$

$$\sigma_{jk}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} [(Y_{ij}^{(t)} - \mu_j^{(t+1)})(Y_{ik}^{(t)} - \mu_k^{(t+1)}) + C_{jki}^{(t)}],$$
$$j=1, 2, \cdots, p$$

위의 E-Step과 M-Step을 반복하여 μ와 $\sum$가 더 이상 변화가 없을 때까지 계속한다.

Ⅲ. 연구결과

표 2는 본 연구에서 제시한 EM algorithm을 표 1의 보건조사 연구에서 얻어진 자료에 적용한 결과이다. 표 1의 자료는 7개의 변수로 구성되므로 7차원의 다변량정규분포를 가정하였다. 그러므로 7개의 평균과 21개의 분산과 공분산을 추정해야 한다. 표 2의 두 번째 열과 세 번째 열이 EM algorithm의 결과 평균과 표준편차의 추정치이다. 네 번째 열에서 일곱 번째 열은 완전자료분석(complete data analysis)의 결과이며, 여덟 번째에서 끝까지는 가능한 자료분석(available data analysis)의 평균과 표준편차의 추정결과이다. 여기서 완전자료분석이란 주어진 7개 변수가 모두 관찰된 자료만 가지고서 모수를 추정한 분석방법이며, 가능한

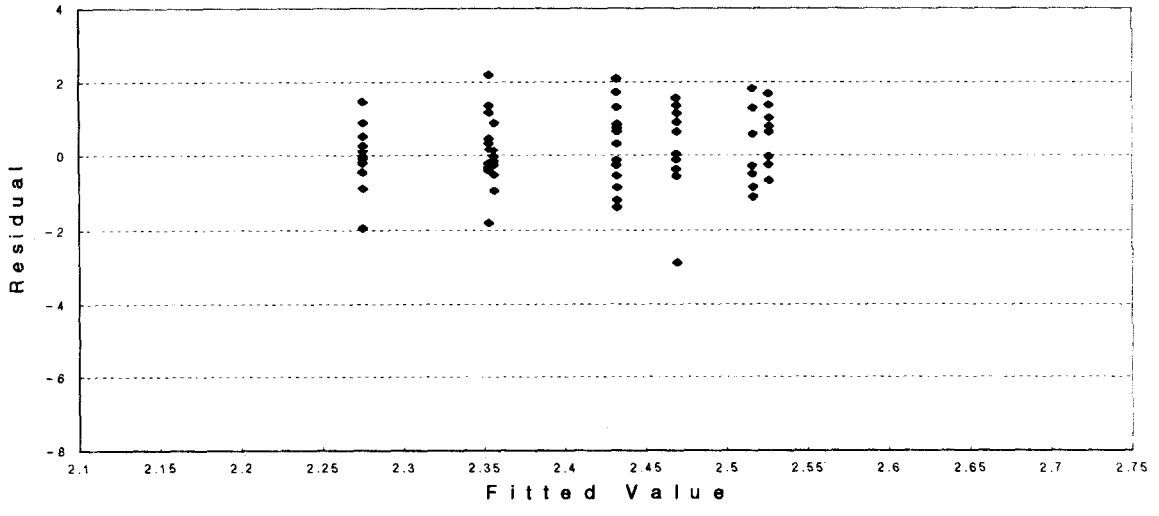

그림 1. EM algorithm을 적용한 경우의 잔차도표

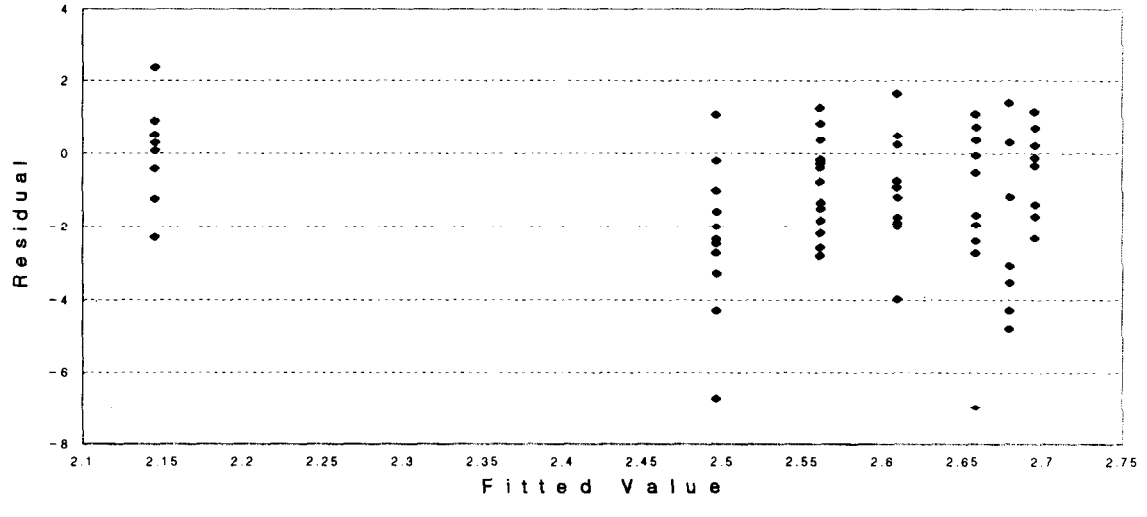

그림 2. 완전자료분석을 적용한 경우의 잔차도표

자료분석이란 결측치가 있는 경우 결측치가 포함된 변수만 해당 평균과 분산공분산을 계산하는데 제외하고, 나머지 사용가능한 자료는 모두 사용하여 분석한 방법이다.

그림 1에서 그림 3은 EM algorithm, 완전자료분석, 가능한 자료분석을 각각 표 1의 자료에 적용하였을 때 계산되어진 잔차도표 (residual plot)의 결과이다. 여기서 잔차는 추정된 표준편차를 이용한 표준화 잔차를 사용하였다. EM algorithm의 결과와 가능한 자료분석 결과가 완전자료분석 결과보다 좀 더 균등한 결과를 보이고 있다. 잔차가 얼마나 균등하가를 간단히 수량화하기 위해 잔차의 절대값과 평균의 추정치와의 Spearman 순위상관계수를 구하였다(Carroll과 Ruppert, 1988, p147). 완전자료분석결과 상관계수가

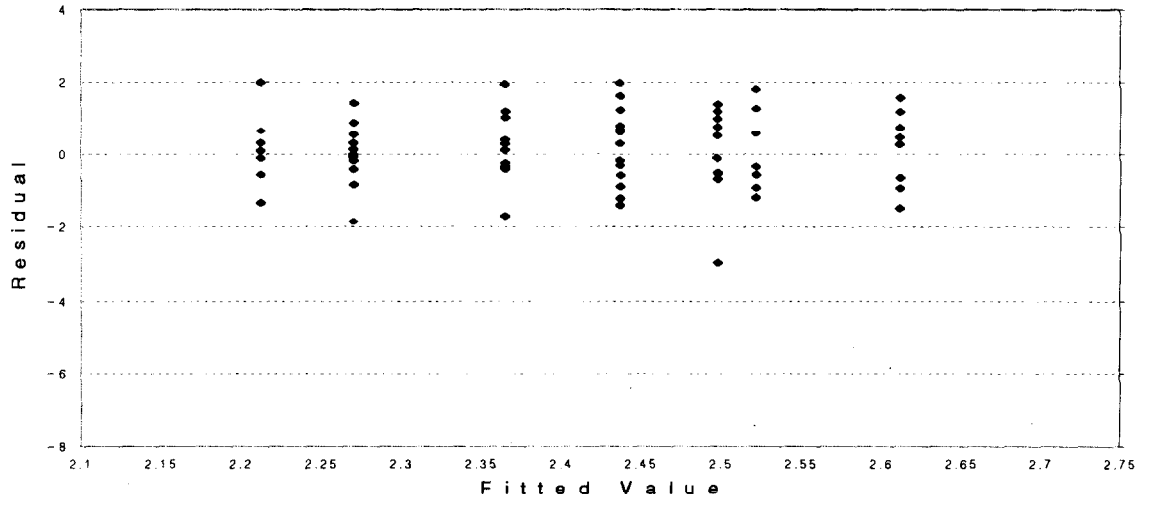

그림 3. 가능한 자료분석을 적용한 경우의 잔차도표

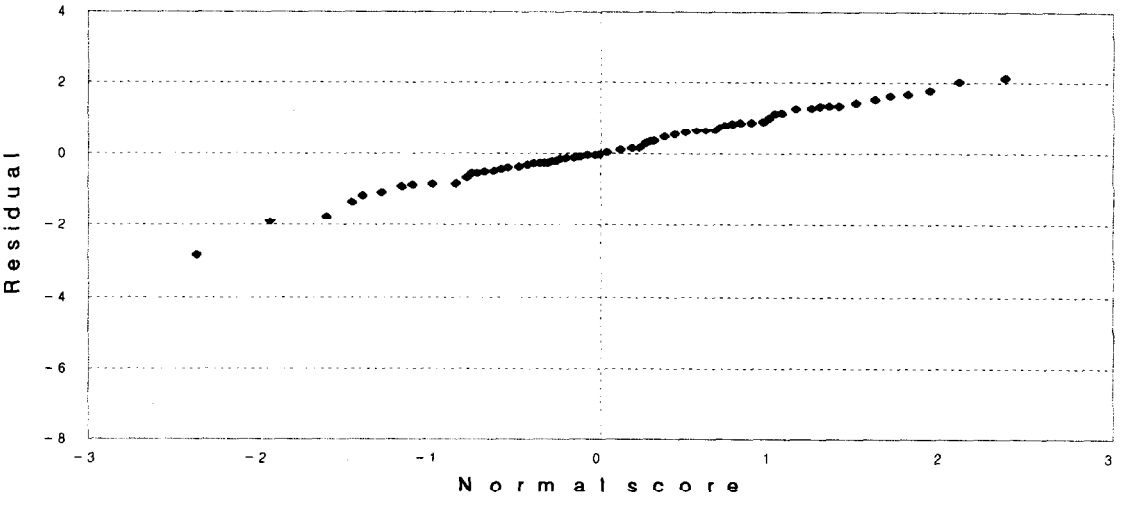

그림 4. EM algorithm을 적용한 경우의 정규확률도표

0.063 (p=0.510), 가능한 자료분석 결과 상관계수는 0.131 (p=0.173), EM algorithm의 결과 상관계수는 0.028(p=0.769)이었다. 따라서 EM algorithm의 결과에 있어서 잔차와 평균의 추정치와의 상관관계가 가장 낮았다. 또한 잔차가 추정치로부터 얼마나 멀리 떨어져 있는가를 간단히 수량화하기 위하여 표준화잔차의 제곱합(residual sum of squares)을 구해보면, 완전자료분석의 경우 잔차제곱합이 622.66, 가능한 자료분석의 경우 잔차제곱합이 103.0, EM algorithm의 경우 잔차제곱합이 97.90 이었다.

그림 4에서 그림 6은 위 세 가지 방법의 적용 결과 얻어진 잔차의 정규확률도표(normal probability plot)이다. 완전자료분석 방법의 경우 정규확률도표가 직선식에서 멀리 떨어졌으며 가능한 자료분석 방법과 EM

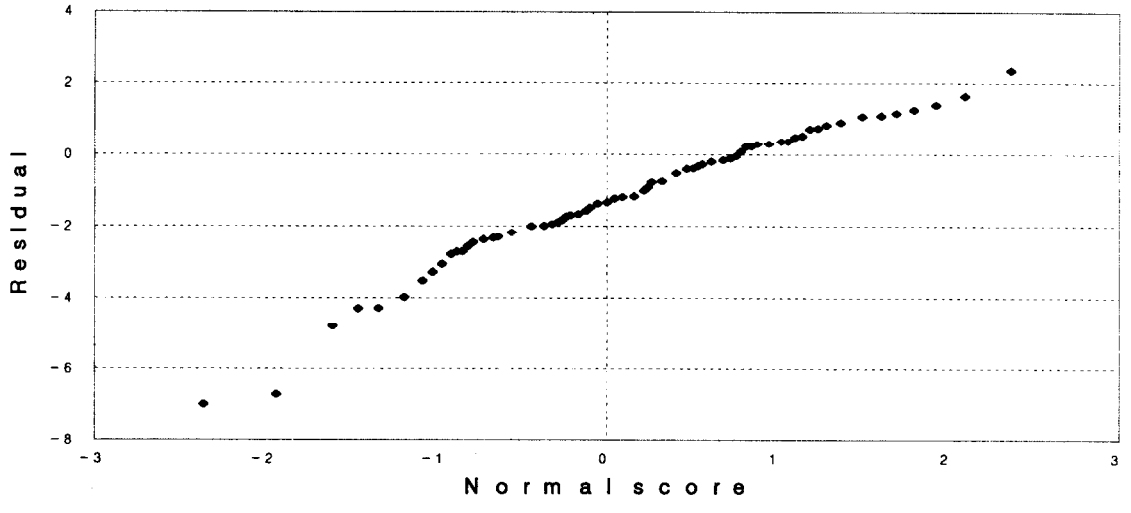

그림 5. 완전자료분석을 적용한 경우의 정규확률도표

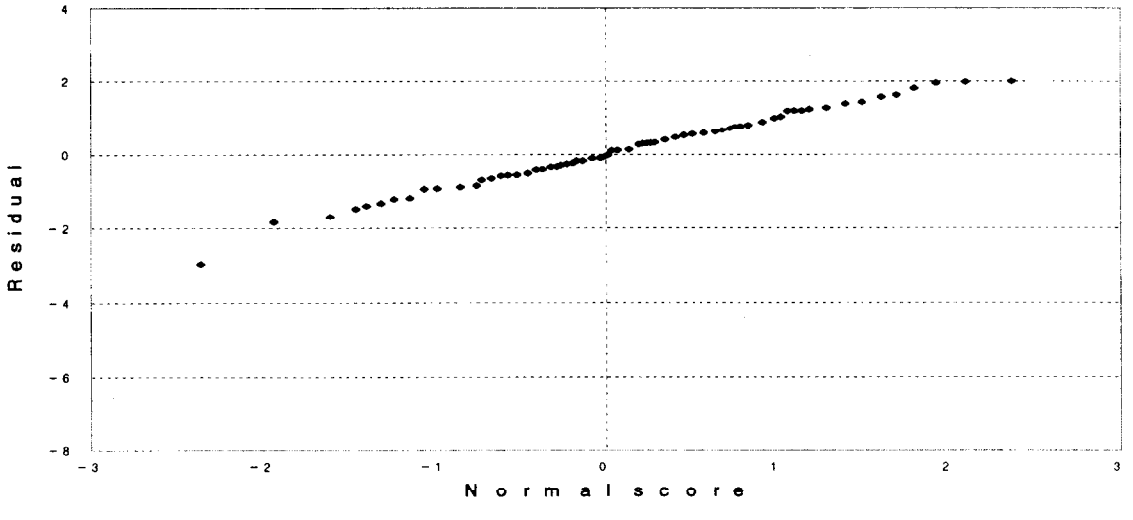

그림 6. 가능한 자료분석을 적용한 경우의 정규확률도표

algorithm의 분석결과가 좀 더 직선식에 가까와서 자료의 정규성 가정에 근접함을 알 수가 있다. 세 가지 방법으로 추정된 평균을 크기순서대로 나열해 보면, 완전자료분석의 경우 X1, X6, X3, X7, X5, X4, X2 이고 가능한 자료분석의 경우 X1, X6, X7, X3, X5, X4, X2 이고 EM algorithm의 경우는 X6, X7, X1, X3, X5, X4, X2 이었다. 완전자료분석과 가능한 자료분석은 X3 와 X7이 바뀐것외에는 큰 변화가 없었으나 EM algorithm의 경우는 X1이 X6과 X7보다 크게 추정되었다. 이는 EM algorithm의 경우에 결측자료를 고려하여 추정할 때 다른 변수의 관찰값에 영향을 받는 사실로 설명할 수 있다. X1에서 X5까지는 진료행위에 관한 항목(상관계수; $\gamma_{12}=0.57$, $\gamma_{23}=0.62$, $\gamma_{34}=0.94$, $\gamma_{45}=0.94$)이고 X6와 X7은 검사행위에 관한 항목(γ_{67}

=0.94)으로써 각 행위군내에서는 상관계수가 높았다. X1은 결측치의 비율이 높았는데 EM algorithm으로 추정하게 되면 타 진료행위의 값들이 전반적으로 높으므로 이에 영향을 받아 X1의 평균이 높게 추정이 되었다. 그러나 진료행위군내에서 그리고 검사행위군내에서는 각 행위별로 순위가 바뀌지 않았다.

표 2에는 완전자료분석 결과와 가능한 자료분석 결과가 EM algorithm에 비해서 얼마나 차이가 나는 가를 계산한 결과이다. 완전자료분석은 평균의 경우에는 EM algorithm의 결과와의 차이가 표준편차의 약 91%정도이었으며, 표준편차의 경우에는 약 39.7% 변동이 있었다. 가능한 자료의 경우에 있어서는 EM algorithm과의 차이가 평균의 경우 약 13.8%, 표준편차의 경우 약 12.7% 가량 변동이 있어서 완전자료경우보다 변동폭이 적었다. 여기서 가능한 자료분석 결과가 완전자료분석 결과보다 EM algorithm결과와 가까운 것을 알 수가 있다.

Ⅳ. 고 찰

그간 결측치가 내포되어 있는 문제에 대해서 사회과학 등에서는 대치법을 중심으로 연구가 진행되어 왔으나, 실제 결측치가 자주 발생되는 보건학 연구에서는 최근에서야 관심을 갖게 되었다. 본 연구에서는 결측치가 내포된 자료를 효과적으로 분석하기 위한 통계학적 방법을 제시하며 이를 실제 보건학 연구에 응용 시도하여 보고자 하였다.

본 연구에서는 보건학 연구에서 나타난 결측치를 분석하기 위한 통계학적 방법을 제시하였다. 특히 결측치에 대한 효율적인 분석방법인 EM algorithm을 도입함으로써 결측치에 대한 통합적인 분석들을 제시하였다. 또한 실제 보건조사자료에 직접 적용함으로써 그 결과를 종합 분석하여 비교하였다. 의료행위 상대가치 개발을 위한 조사 연구에서 나타난 결측치를 EM algorithm을 적용하여 추정하였다. 또한 완전자료분석방법(complete data analysis)과 가능한자료분석방법(available data analysis)을 비교하였다. 다변량정규분포를 가정하였을 때 EM algorithm을 사용한 분석결과가 다른 두 가지 방법보다 잔차와 정규확률도표에서 정규분포가정에 적절한 결과를 얻었다. 또한 잔차제곱합을 계산하였을 때 EM algorithm의 결과가 두 방법보다 적게 얻어졌다. 가능한 자료분석방법이 완전자료분석방법보다는 잔차도표와 정규확률도표에서 정규분포가정에 가까운 결과를 가져왔다.

이 연구결과는 보건조사연구에서 관찰된 자료의 통계학적 분석에 활용이 가능하다. 특히 결측치가 내포된 경우에도 통계학적 분석을 가능하게 하여 준다. 결측치가 흔히 나타나는 보건학 연구 중에서 특히 보건조사자료 분석에 활용될 수 있다. 이 연구결과는 보건조사연구에 기초하여 결측자료에 분석에 적용하였으나 그 결과는 보건학 연구에만 국한된 것은 아니다. 보건학 연구뿐만 아니라 사회학, 심리학 등 제반 사회과학에서도 결측치가 있는 경우의 통계학적 분석방법으로 활용될 수 있다.

본 연구에서는 완전자료분석방법, 가능한자료분석방법, 그리고 EM algorithm의 결과를 잔차분석을 통하여 비교하였다. 그러나 자료가 보건조사연구에서 얻은 특정한 자료라는 점과 비교도구로 쓰인 잔차분석이 제반 방법간의 비교를 위한 일반적인 도구는 아니라는 점에서 이 결과를 그대로 일반화시키기에는 연구의 한계가 있다. 제반 방법간의 좀더 정밀한 비교를 위하여서는 다양한 결측자료의 특성에 따라 각 방법의 추정치가 어떤지를 살펴보는 모의실험연구(simulation study)가 필수적이라 판단되며 본 저자들은 이를 준비중에 있다.

EM algorithm은 다변량결측치가 존재할 때에도 최대우도추정법에 근거한 효율적인 분석 방법으로 여러 연구에서 제시되고 있으나 이 방법은 많은 반복계산을 수행해야만 추정치를 얻을 수 있다는 제약이 있다. 또한 아직까지 이 방법을 사용하여 쉽게 다변량 정규분포에서의 평균과 분산을 구하는 프로그램이 SAS, SPSS등 널리 쓰이는 소프트웨어에서는 상용화되지 못하였기에 널리 사용되지 못하고 있다. 그러나 현재까지의 기술의 발전속도로 본다면 머지 않아 이 방법의 상용화는 가까운 시일내에 이루어질 것으로 판단된다.

참고문헌

김한중, 손명세, 박은철, 김영삼, 박형욱, 박웅섭, 최귀선, 최귀선, 안엉량, 김주희, 이대희, 강 임옥, 심지윤, 김동기, 임종건, 안재형, 염용권, 양동현, 안인환, 이윤태, 이용균, 명희봉, 김지홍, 이경태, 이상철, 최수미, 권호근, 이영희, 김권수, 조본경, 손정일, 김영남, 김명기, 홍미희, 이종기, 조인숙. 의료보험 수가구조개편을 위한 상대가치개발. 연세대학교 보건정책 및 관리연구소, 한국보건의료관리연구원, 1997. 10

Anderson TW(1984). An Introducton to Multivariate Statistical Analysis, 2nd ed. John Wiley & Sons. 1984

Beale EML, Little RJA. Missing values in multivariate analysis. Journal of Royal Statistical Society 1975: B37 : 129-146

Carroll RJ, Ruppert D. Transformation and Weighting in Regression. Chapman and Hall. 1988

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm(with discussion). Journal of Royal Statistical Society 1976; B39 : 1-38

Hsiao WC, Braun P, Becker E, Causino N, Couch NP, DeNicola M, Dunn D, Kelly NL, Ketcham T, Sobol A, Verrilli D, Yntema DB. A National Study of Resource-Based Relative Value Scales for Physician Services : Final Report. Harvard School of Public Health, Cambridge, Massachusetts, 1988

Hsiao WC, Braun P, Becker ER, Dunn DL, Kelly NL, Yntema DB. A National Study of Resource-Based Relative Value Scales for Physician Services : Phase II. Final Report. Harvard School of Public Health, Cambridge, Massachusetts, 1990

Hsiao WC, Yntema DB, Braun P, Spencer C. Measurement and analysis of intraservice work. JAMA 1988; 260(28) : 2361-2370

Kim DK, Taylor JMG. The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameter. Journal of the American Statistical Association 1995; 90 : 708-716

Little RJA. Regression with missing X's; A review. Journal of the American Statistical Association 1992; 87 : 1227-1237

Little RJA and Rubin DB. Statistical Analysis with Missing Data. John Wiley & Sons, New york, 1987

Louis TA. Finding the observed information matrix when using the EM algorithm. Journal of Royal Statistical Society 1982; B44 : 226-233

Meng X, Rubin DB. Using EM to obtain asymptotic variance-covariance matrices : SEM algorithm. Journal of the American Statistical Association, 1991; 89 : 899-909

Wu CFJ. On the convergence properties of the EM algorithm. Annals of Statistics 1983; 11 : 95-103