Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.3 pp. 579~590

Application of SOLAS to the Multiple Imputation for Missing Data

Sung-Ho Moon[1] · Hyun-Jeong Kim[2] · Jae-Kyoung Shin[3]

Abstract

When we analyze incomplete data, i.e., data with missing values, we need treatment for the missing values. A common way to deal with this problem is to delete the cases with missing values. Various other methods have been developed. Among them are EM algorithm and regression algorithm which can estimate missing values and impute the missing elements with the estimated values. In this paper, we introduce multiple imputation software SOLAS which generates multiple data sets and imputes with them.

Keywords : Hot-deck, EM-algorithm, NORM, Single/Multiple Imputation, SOLAS

1. 서 론

결측값이란 실험이나 조사에서 어떠한 이유에 의해 관측되어야 할 값을 얻지 못하는 경우를 말하며, 이런 현상은 조사한 자료의 다변량분석을 행할 때 자주 접하게 된다. 일반적으로 자료분석에서 다루어지는 데이터는 모든 원소들이 관측된 경우이다. 그러나 결측값을 갖는 자료의 경우 어떠한 처리를 하면 좋은가에 대해서 많은 방법들이 연구되어 왔다.

결측값이 존재하는 경우의 자료분석을 위한 몇 가지 편의적인 방법으로서, 관측된 값만을 이용하는 방법, 관측된 값에서 샘플링하여 결측값에 대치하는 방법, 평균값 또는 조건부 기대값으로 결측값을 대치하는 방법 등이 있다(김현정 등, 2001).

위의 방법들은 결측값을 포함한 자료를 분석할 때 자주 이용되는 방법들이지만 결측값 자체에 대한 정보를 얻을 수 없다는 단점도 있다. 이에 대한 개선된 방법으로서

1) (608-738) 부산시 남구 우암동 산55-1, 부산외국어대학교 정보통계학과 부교수
E-mail: shmoon@pufs.ac.kr
2) (617-736) 부산시 사상구 괘법동 산1-1, 신라대학교 교양학부 전임강사
3) (641-773) 경남 창원시 사림동 9번지, 창원대학교 통계학과 부교수

Little and Rubin(1987)과 Johnson and Wichern(1992)은 관측값에 근거해서 EM 알고리즘(Dempster et. al., 1977)을 이용한 최우추정법(Orchard and Woodbury, 1972)을 권장하고 있으며, 가장 일반적인 방법으로서는 회귀직선모형에 의한 대치법도 있다. 그러나 EM 알고리즘을 이용한 최우추정법과 회귀직선 모형에 의한 대치법들은 결측값에 대해 하나의 값만으로 대치(일대치, single imputation)하므로 편의적일 수도 있으며, 결측데이터의 결측률이 큰 경우 정확성의 정도도 문제시 될 수 있다.

본 연구에서는 위에서 지적한 방법의 문제점들을 개선하여 정확성의 정도를 높이는 방법으로서 결측값에 대해 여러 개의 값으로 대치하는 다대치(multiple imputation)용 S/W인 SOLAS의 사용방법에 대해 소개하고 이를 적용한 실례를 보인다.

물론 SOLAS 이외에 CAT(로그선형모형 가정하의 범주형 자료의 다대치용 S/W ; Schafer, 1997, Chapters 7-8), MIX(일반위치모형(general location model) 가정하의 연속형·범주형 자료의 다대치용 S/W ; Schafer, 1997, Chapter 9), PAN(다변량 선형 혼합효과모형 가정하의 패널자료나 집락자료에 대한 다대치용 S/W ; Schafer, 1999)과 NORM(EM 알고리즘을 이용하여 결측값에 대한 초기값을 구한 후 Data Augmentation 알고리즘에 의한 다대치용 S/W ; Schafer, 1997, Chapter 5)등이 있으나, 본 논문에서는 다대치 문제를 SOLAS를 이용해서 다루고자 한다.

2장에서는 SOLAS에 관한 소개와 사용방법 등에 관한 설명을 다루며, 3장의 수치예에서는 주어진 결측 데이터를 1, 2장에서 언급한 기존의 대치 방법에 의한 결과와 SOLAS를 이용해서 얻은 결과를 비교한다.

2. 결측값 대치를 위한 SOLAS의 사용방법

먼저, Sort Monotone과 Non-monotone 결측자료에 대해 SOLAS에 포함되어 있는 데이터를 사용한 예를 통해 SOLAS의 사용방법을 소개하고, 여기에서 사용한 데이터를 이용한 다대치의 결과를 3장에서 다른 소프트웨어를 사용한 결과와 비교하기로 한다.

2.1. 사용 데이터

수치 예는 SOLAS에 포함되어 있는 MI_TRIAL.MDD 데이터(subdirectory SAMPLES 안에 있다)를 사용한다. 데이터는 아래의 11개 변수에 대한 50명 환자의 임상실험 데이터이다.

- OBS - 관측자수
- SYMPDUR - 증상의 연속기간
- AGE - 환자의 연령
- MeasA_0, MeasA_1, MeasA_2, MeasA_3 - 환자의 반응변수 MeasA의 기본값(0)과 기본값에서부터 측정된 1개월, 2개월, 3개월 후의 반응값
- MeasB_0, MeasB_1, MeasB_2, MeasB_3 - 환자의 반응변수 MeasB의 기본값(0)과 기본값에서부터 측정된 1개월, 2개월, 3개월 후의 반응값

변수 OBS, SYMPDUR, AGE, MeasA_0과 MeasB_0은 전 개체가 관측되었고, 나머

지 6개의 변수는 결측값을 포함하고 있다. 먼저, 이 결측 패턴을 보기 위해 다음 조작을 행한다.

(i) 데이터가 있는 화면 View에서 Missing Pattern을 선택하고 Use All을 누른다.
(ii) 결측 데이터 패턴 화면의 View에서 View Monotone Pattern을 선택하고, 다시 View에서 Variable list를 선택하면 그림 1이 표시된다. 여기서 데이터가 단조로운 패턴으로 정렬(소트)된 후에도 시간이 경과되는 측정의 시간구조는 유지된다. 따라서 이 데이터의 결측 패턴은 시점추이에 대하여 단조롭다.
(iii) 결측 패턴 화면에서 File로부터 Close를 선택하고 종료한다.

2.2. Predictive Model Based Method - Example

아래의 순서대로 Predictive Model에 의한 방법을 사용하여 이 데이터의 결측값을 대입한다.

(i) Analyze 메뉴에서 Multiple Imputation과 Predictive Model Based Method를 선택한다.
(ii) Specify Predictive Model 화면이 표시된다. 이 화면은 시트가 2개이며, Base setup 과 Advanced Options이라고 하는 tab이 있다. 대입변수(결측변수)를 선택하면 Non-Monotone과 Monotone tab이 표시된다.

(1) Base Setup

Base Setup tab을 선택하면 대입변수를 지정할 수 있다. 또 Predictive Model (예측모형)에 있어 공변량으로 사용하고 싶은 변수를 지정할 수도 있다(그림 2).

(i) Variables to Imputed 필드에 변수 MeasA_1, MeasA_2, MeasA_3, MeasB_1, MeasB_2, MeasB_3을 끌어다 놓는다.
(ii) Fixed Covariates 필드에 SYMPDUR, AGE, MeasA_0, MeasB_0을 끌어다 놓는다.
(iii) 이 데이터에는 그룹변수가 없기 때문에 Grouping 필드는 공백으로 둔다.

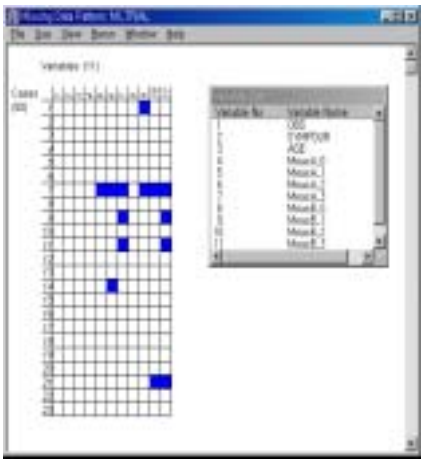

그림 1 결측패턴

그림 2 변수 지정

(2) Non-Monotone

Non-Monotone tab을 선택하면 데이터의 비단조 패턴의 결측값을 대입하기 위해 사용하는 Predictive Model의 공변량을 추가하거나 제거할 수 있다. 각각의 대입변수에 대해 공변량을 추가하거나 제거할 때는 +와 - 표시를 사용한다(그림 3).

각각의 대입변수에 대해 공변량의 리스트는 Base Setup tab에서 지정된 고정 공변량(Fixed Covariates)과 다른 모든 대입변수까지 작성 가능하다. 변수는 이 공변량 리스트로부터 Variable 필드에 끌어다 놓음으로써 증감이 가능하다. 단 대입변수에 따라 공변량 리스트 중 변수가 모두 최종적으로 확정되는 예측 모형에는 사용하지 못 할 수도 있다.

처음 시스템은 변수를 소트(정렬)하여 결측 패턴이 가능한 한 단조롭게 되도록 한다. 또 대입변수의 각 결측값에 대해서 공변량 리스트에서 어떤 변수를 예측변수로 사용할 수 있을지를 찾아낸다. 디폴트는 모든 공변량 예측모형을 대상으로 한다. 단 고정 공변량으로 변수를 체크하지 않으면, 그 변수는 예측모형의 독립변수의 Step-Wise 선택으로 제거할 수 있다. 마지막으로 결측값을 대입하기 위해 사용되는 예측모형의 자세한 부분은 View의 Multiply Imputed Data의 화면에서 선택되는 Regression Output에 표시된다. 이 화면은 Specify Predictive Model의 OK 버튼을 누르면 표시된다.

그림 3 비단조 패턴

그림 4 단조 패턴

(3) Monotone

Monotone tab을 누르면 단조패턴의 결측값을 대입하기 위해 사용된 Predictive Model 안의 공변량을 추가하거나 제거 할 수 있다(그림 4).

여기서 공변량의 리스트에 변수의 추가 또는 제거를 행할 때는 +와 - 표시를 사용한다. 각 대입변수에 대한 Base Setup tab의 고정 공변량으로 지정된 변수와 다른 모든 대입변수로부터 구성된다. 변수는 공변량 리스트와 변수 필드사이에 끌어다 놓음으로써 증감이 가능하다. 단 어떤 변수가 특정 대입변수에 대해 공변량 리스트 안에 표시되어 있어도 최종적인 예측 모형에는 사용되지 않을 수도 있다.

처음 시스템은 변수를 정렬하여 결측 패턴이 가능한 단조롭게 되도록 한다. 다음으로 대입변수의 좌측에 있는 변수만을 공변량으로 사용한다. 마지막으로 결측값을 대입하기 위하여 사용된 예측모형의 자세한 것은 Regression Output에 표시된다.

(4) Advanced Options

Advanced Options tab를 선택하면 회귀모형 또는 판별분석 모형(regression/discriminant model) 설정을 위한 화면이 표시된다(그림 5).

(a) Tolerance

Tolerance 필드의 수치는 수치계산의 정도를 제어하기 위한 것이다. 이 수치는 역 행렬을 계산할 때 어느 정도이면 특이행렬인가를 판단하기 위해 사용된다. 구체적으로 다른 독립변수에 대해서 기여율 R^2이 (1-Tolerance)의 값을 넘을 것 같은 변수는 독립변수로 사용할 수 없도록 체크한다.

(b) Stepping Criteria(변수선택을 위한 기준)

여기서는 예측모형의 독립변수의 Step-wise 선택을 위한 기준으로 F-to-Enter와 F-to-Remove의 수치를 지정하는데, 보다 많은 변수를 모형에 포함시키려면 F-to-Enter 값을 작게 잡아야한다. 또한 F-to-Remove의 값은 F-to-Enter 수치보다 작게 취해서는 안된다. 설정을 확인하고 OK를 누른다. 다중대입의 데이터

화면이 표시되고 대입된 수치는 파란색으로 표시된다.

여러 개의 대입값에 따라 생성된 복수 개의 데이터 집합을 분석한 결과를 종합적으로 정리하는 방법에 대해서는 SOLAS 메뉴얼(1999)의 “Analyzing Multiple Imputed Data sets”에 상세한 설명이 있으므로 참조하기 바람.

그림 5 Advanced Option

그림 6 변수 지정

2.3. Propensity Score법의 실행순서

여기서는 propensity score법을 다음과 같이 사용하여 데이터의 결측값을 대입하는 순서를 설명한다.

(i) Analyze 메뉴에서 Multiple Imputation(다중대입)과 Propensity Score Method를 선택한다.

(ii) Specify Propensity Method 화면이 표시된다. 여기서는 Base Setup과 Advanced Options라는 2개의 시트가 표시된다. 대입(결측)변수를 선택하면 Non-Monotone(비단조)과 Monotone(단조) tab 그리고, Donor Pool tab이 표시된다.

(1) Base Setup

Base Setup tab을 선택하여 대입변수를 지정한다. 다음으로 결측지표를 로지스틱 회귀식으로 모형화할 때 공변량으로 사용하고 싶은 변수를 지정한다(그림 6).

(i) Variable to Impute(대입변수) 필드에 변수 MeasA_1, MeasA_2, MeasA_3, MeasB_1, MeasB_2, MeasB_3를 끌어다 놓는다.

(ii) Fixed Covariates(고정 공변량) 필드에 SYMPDUR, AGE, MeasA_0, MeasB_0을 끌어다 놓는다.

(iii) 데이터는 그룹변수를 포함하고 있지 않기 때문에 Grouping 필드는 공란으로 둔다.

(2) Non-Monotone

Non-Monotone tab을 선택하면 데이터의 비단조 패턴의 결측값을 대입하기 위해 사용하는 로지스틱 회귀모형의 공변량을 추가하거나 제거할 수 있다(그림 7). 각각의 대입변수에 대해 공변량 리스트의 변수를 추가, 제거할 때는 +와 - 표시를 사용한다.

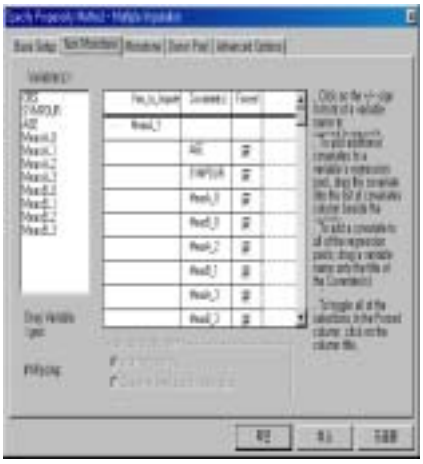

그림 7 Nonmonotone Pattern

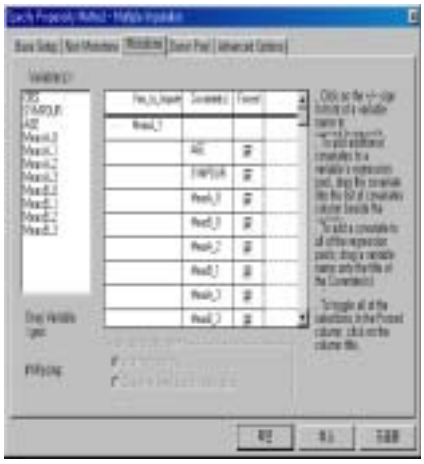

그림 8 Monotone Pattern

각각의 대입변수에 대해 공변량 리스트는 Base Setup tab에서 지정된 고정 공변량(Fixed Covariates)과 다른 모든 대입변수에서 작성 가능하다. 변수는 이 공변량 리스트에서 Variable 필드에 끌어다 놓음으로써 증감이 가능하다. 단 대입변수에 따라서는 공변량 리스트 안의 변수가 전부 최종적으로 사용되지 않을 수도 있다.

(3) Monotone

Monotone tab을 선택함으로써 데이터의 단조패턴의 결측값을 대입하기 위해 사용하는 로지스틱 회귀모형의 공변량을 추가하거나 제거할 수 있다(그림 8).

여기서 공변량 리스트에 관한 변수의 추가 또는 제거를 행할 때는 +와 - 표시를 사용한다. 각각의 대입변수에 대해 공변량 리스트는 Base Setup tab에서 지정된 고정 공변량(Fixed Covariates)과 다른 모든 대입변수에서 작성 가능하다. 변수는 이 공변량 리스트에서 Variable 필드에 끌어다 놓음으로써 증감이 가능하다. 단 대입변수에 따라서는 공변량 리스트 안의 변수가 전부 최종적으로 사용되지 않을 수도 있다.

(4) Donor Pool

Donor Pool tab을 선택하면 Donor Pool 페이지가 표시된다(그림 9). 사용자는 propensity score에 관해 데이터의 부분집합을 정의할 수 있으며 랜덤추출단계에서 제어를 할 수 있다.

그림 9 Dornor Pool

그림 10 Advanced Options

Propensity Score의 부분집합을 정의하기 위한 아래의 옵션이 있다.

- Divide propensity score into *c* subsets : propensity score의 값에서 c개의 부분집합을 작성한다. c의 디폴트값은 5이다.
- Use c closest cases : 부분집합에 포함되어야 하는 대입 개체의 앞 뒤 개체를 지정할 수 있다.
- Use d% of the dataset closest cases : 위의 개체 수를 백분율로 지정할 수 있다.

각각의 대입변수에 대해 refinement variable을 사용할 수 있다. 변수는 Variables의 리스트 상자에서 Refinement-Variable 열에 끌어다 놓는다. refinement variables를 사용하면 시스템은 수치에 근접하는 개체만을 포함하도록 부분집합의 개체를 줄여간다. donor pool에서 사용하는 refinement variable 수도 지정할 수 있다. 이 예에서는 디폴트값을 사용하여 분석한다.

(5) Advanced Options(추가옵션)

Advanced Options를 선택하면 Advanced Options의 화면이 표시되는데, 사용자는 대입분석과 로지스틱 회귀모형에 관해 상세히 제어할 수 있다(그림 10).

(i) Randomization(랜덤화)

Main Seed Value는 propensity의 부분집합 내에서의 랜덤추출에 사용된다. 디폴트값은 12345이다. 이 필드는 공백으로 두거나 0으로 설정하면 lock time이 사용된다.

(ii) Regression Options(회귀모형의 계산을 위한 옵션)

Model Tolerance 필드의 수치는 계산정도를 제어하기 위한 것이다. 이 값은 역행렬이 특이행렬이 되지 않도록 하기 위한 옵션으로 사용된다. 여기서는 다른

변수에 대한 기여율 R^2이 (1-Tolerance)를 초과할 것 같은 변수는 모형에 따른 독립변수에서 제거한다.

(iii) 그 외의 설정

위의 설정 외에 로지스틱 회귀모형의 추정은 최우추정법으로 행하기 위해 반복계산에 관한 여러 종류의 설정값이 요구된다.

2.4. 다중대입법의 출력

Propensity Score법과 Predictive Method Based Model을 선택한 경우의 다중대입법의 출력은 표준으로 5개의 시트가 있다. Data Page의 View 메뉴에서 Imputation Report, Regression Output, Missing Pattern 중의 어느 것을 선택할 수 있다. 대입법에 추가된 다른 분석이 Analyze 메뉴에서 행해진 경우에는 Combined 탭이 새롭게 추가되며, 그 결과를 알 수 있다.

(1) Data Pages

다중대입법의 출력으로서, 대입된 수치가 파랗게 반전된 데이터 화면이 5페이지 정도 표시된다. 이 예에서 분석한 경우의 처음 데이터 화면이 그림 11이다. 여기서는 2개의 수치가 대입값으로 강조되어 있다.

Imputation Report와 Regression Output은 로지스틱 회귀분석, 일반적인 회귀분석, 다중대입법에 사용한 설정 등에 관한 결과를 정리한 것이다.

(2) Imputation Report

대입 레포트에서는 다중대입법에 사용된 모수의 설정 등을 정리하여 보고하고 있다(그림 12). 예를 들면 랜덤추출에 사용된 seed value와 대입된 결측값의 수 등 모든 것이 기록되어 있다. 주된 내용은

- 다중대입법에 관계되는 모수값
- 대입값에 사용된 방정식
- 대입값의 논리적 성질과 정당성을 판단하기 위한 진단통계량

이다.

View 메뉴에서 Imputation Report(그림 12)와 Regression Output, Missing Pattern을 선택할 수 있다.

대입된 데이터에 대해 적용된 통계분석 결론은 5개의 중간해석 결과를 종합하여 구하여지게 된다. 각 화면에 중간결과가 표시되기 때문에 사용자는 최종적인 결과가 구하여진 것을 보게 된다.

그림 11 데이터 페이지

그림 12 Imputation Report

3. 수치 예와 토론

수치 예로서 여성 건강조사의 세 변수(X_1: sweat rate, X_2: sodium content, X_3: potassium content)에 대한 20명의 자료($p=3$, $n=20$)에 적용시켜 보았다. 이 자료는 표 3.1과 같으며 결측값에 관한 연구를 위해 세 변수의 상호관계를 고려하여 두 변수에 5개의 결측값을 만들었는데, *가 붙어있는 값이 결측값을 의미한다.

먼저 미지의 분포에 대한 공분산 행렬의 최대고유값을 추정한다고 가정하자(Efron, 1994).

$$\theta = \text{maximum eigenvalue of } \hat{\Sigma}$$

여기서 $\hat{\Sigma}$는 어떤 분포로부터 유도된 공분산 행렬이다.

이와 같은 결측데이터가 있는 경우 먼저 가장 일반적인 방법으로 결측 개체를 생략하여 관측된 개체만을 이용하여 얻은 추정값은 $\hat{\theta}=193.70$, 결측변수의 관측 자료만을 이용한 샘플링(hot-deck)으로 대치한 경우 모수 추정값은 $\hat{\theta}=173.84$이었다. 또한 결측변수의 관측된 데이터의 평균값을 대치(mean imputation)하여 $\hat{\theta}=170.40$을 얻었다. 관측된 개체만을 이용하여 얻은 추정값은 다른 두 방법보다 실제 추정값에 근사하였는데, 이는 결측 변수의 자료값들 중 낮거나 높은 값들이 결측되어 비교적 낮은 분산을 보이고 있기 때문이다. 실제의 경우는 이러한 결측 정보를 얻을 수 없으므로 결측데이터에 대한 대치적 방법으로는 부적합하다. 위의 세 가지 방법의 추정값들은 관측된 자료만을 이용한 경우인데 실제 추정값 $\theta=200.46$보다 모두 낮게 추정되었다. 또한, 추정(estimation)과 최대화(maximization)의 두 단계를 반복하는 계산인 EM 알고리즘(Dempster et. al., 1977)을 이용한 결측 대치법의 추정값은 179.15이었다. 본 논문에서 소개한 다중 대치법용 소프트웨어인 SOLAS를 이용하여 2장에 따라 실행하

면 다음과 같다.

표 3.1 Sweat Data

Individual	X_1 Sweat rate	X_2 Sodium	X_3 Potassium
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1*	55.5	9.7
6	4.6	36.1	7.9
7	2.4*	24.8*	14
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5*	58.8*	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Source : *Johnson and Wichern(1992),*

이와 같이 SOLAS를 이용하여 얻은 다대치의 추정값은 175.38이었다. 이 값은 EM 알고리즘을 이용한 결측 대치법의 값을 제외한 다른 방법보다는 실제 추정값에 가까운 값으로 근사함을 알 수 있다(표 3.2 참조).

표 3.2 대치 방법에 따른 추정값

방법	$\hat{\theta}$
실제값	$\theta = 200.46$
관측값만 이용	193.70
hot-deck	173.84
평균값 대치	170.40
EM	179.15
SOLAS	**175.38**

결측값 대치에 관한 상세한 설명은 김현정 외(2001)을 참고하기 바란다.
또한, 차후의 연구과제로 1장에서 언급했던 SOLAS이외의 다대치용 S/W인 CAT,

MIX, PAN등을 활용한 수치 예에 대한 논의와 앞에서 다룬 예 이외의 다른 데이터에의 적용도 유용할 것으로 생각된다.

참고문헌

1. Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum-likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series*, B39, 1-38.
2. Efron, B. (1994). Missing data, imputation, and bootstrap. *Journal of the American Statistical Association*, Vol 89, No 426, 463-479.
3. Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis 3rd Ed.*, Prentice Hall.
4. Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, John Wiley and Sons.
5. Orchard, T. and Woodbury, M. A. (1972). A missing information principle: theory and application, *Proc. 6th Berkely Symposium on Math. Statist. and Prob.* 1, 697-715.
6. Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall, Chapter 3.
7. Schafer, J.L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1), 3-15.
8. SOLAS for missing data analysis 2.0 (1999). *USER REFERENCE,* Statistical Solutions Ltd.
9. 김현정, 문승호, 신재경 (2001). "결측자료의 대치에 관한 방법적 비교", *한국통계학회 2001년 춘계 학술발표회 논문집*, 101-104.

[2003년 6월 접수, 2003년 8월 채택]