예방의학회지 제37권 제3호(2004년 8월)
J Prev Med Public Health 2004;37(3):209-211

특강원고

결측값의 대치법
Imputation of Missing Values

윤 성 철
Sung-Cheol Yun

서울대학교 자연과학대학 통계학과
Department of statistics, Seoul National University

소 개

결측/무응답은 사회과학 및 자연과학 등의 전반적 분야의 관측 또는 실험되어 얻어지는 자료에 종종 나타나는 현상이다. 이런 결측/무응답을 가진 자료는 복잡한 통계적 분석 기법을 요한다. 일반적으로 결측/무응답을 가진 자료를 분석 할 때는 다음 세가지점이 고려된다.

첫째, 효율성(efficiency) 문제

둘째, 자료 처리 및 분석의 복잡성 문제

셋째, 관측된 자료와 결측된 자료간의 차이에서 기인하는 편이(bias) 문제

많은 연구자들은 보편적으로 결측/무응답을 가진 자료를 분석할 때는 적절한 가정 하에 (Missing Completely at Random)[1] 불완전한 자료는 무시하고 완전하게 관측 또는 실험 된 자료만으로 표준적 통계기법으로 분석을 하게 되는데, 이는 분석의 간편성을 가지는 장점이 있는 반면, 효율성의 문제와 가정이 맞지 않았을 경우에 통계적 추론이 틀릴 수 있다는 문제가 있다. 결측/무응답을 가지는 자료를 분석 할 수 있는 모형은 최근 10여년 사이에 많은 연구가 되었다[2,3,4]. 특히 무시 할 수 없는 결측을 가진 자료 분석을 위한 통계 모형들은 여전히 많은 통계학자들에 의해 현재 활발히 연구가 진행 중에 있다. 본서에서는 결측/무응답을 가진 자료를 무시하지 않고 분석 할 수 있는 통계 방법론의 하나인 대치법(imputation)과 간단한 문제에 대하여 일반 연구자들이 쉽게 분석 할 수 있는 SAS 프로시져를 소개한다.

단순 대치법 (Single Imputation)

1. Completes Analysis

불완전 자료는 모두 무시하고 완전하게 관측된 자료만으로 표준적 통계기법에 의해 분석하는 방법을 말한다. 분석이 쉽다는 장점이 있지만 부분적 관측된 자료를 무시하므로 생기는 효율성 상실과 통계적 추론의 타당성 문제가 있다.

2. 평균 대치법(Mean Imputation)

이 방법은 관측 또는 실험되어 얻어진 자료의 적절한 평균값으로 결측값을 대치해서 불완전한 자료를 완전한 자료로 만든 후, 완전한 자료를 마치 관측 또는 실험되어 얻어진 자료라 생각하고 분석하는 방법을 말한다. 대표적 방법으로 비조건부 평균 대치법과 조건부 평균 대치법이 있다. 조건부 평균 대치법의 확장으로써 Buck's[5] 방법이 일반적으로 많이 알려져 있다.

예1) 비조건부 평균 대치법

10	?	15	19	12	18	?	?	16
10	*15*	15	19	12	18	*15*	*15*	16

sol> 관측값의 평균 : $15 = \frac{10+15+19+12+18+16}{6}$
결측값 대치 : ?= 15

평균 대치법은 사용하기가 간단하고 Completes Analysis에 비해 효율성이 향상된다. 그러나 관측된 자료를 토대로 한 추정값으로 결측값을 대치함으로써 통계량의 표준오차가 과소 추정되는 문제가 있다.

예2) 조건부 평균 대치법(Regression Imputation)

Y_1	Y_2	Y_3	Y_4
10	15	20	20
12	25	30	30
15	35	40	40
25	48	57	57
30	49	60	60
35	55	65	65
37	47	70	70
40	60	?$_1$	*76.89*
42	65	?$_2$	*81.67*
50	70	?$_3$	*92.39*

sol> $Y_{i4}=\beta_0+\beta_1 Y_{i0}+\beta_2 Y_{i4}+\epsilon_i,\ i=1,\cdots,7$
$\rightarrow \beta_0=3.69, \beta_1=0.099, \beta_2=0.56$
?$_1$=3.69+40 · 0.99+60 · 0.56=76.89

3. 단순 확률 대치법 (Single Stochastic Imputation)

이 방법은 평균대치법에서 추정량 표준오차의 과소 추정문제를 보완하고자 고안된 방법으로 Hot-deck[2]방법, Nearest-Neighbour[2] 방법 등이 있다. 기본적 아이디어는 평균대치법에서 관측된 자료를 토대로 추정된 통계량으로 결측값을 대치할 때 어떤 적절한 확률값을 부여 한 후 대치하는 방법이다. 이 방법은 추정량의 표준오차가 과소 추정되는 문제는 보완 되지만, 간단 문제를 제외한 대부분의 경우에 추정량의 표준오차 계산자체가 어려운 문제가 있다.

다중 대치법 (Multiple Imputation)

단순 대치법 (Single Imputation)은 결측치를 가진 자료 분석에 사용하기가 용이하고, 통계적 추론에 사용된 통계량의 효율

2004년도 예방의학회 하계 워크숍에서 발표된 내용을 특강의 형태로 정리하였음.
책임저자 : 윤성철 (서울시 관악구 신림동 산56-1, 전화 : 02-3010-5269, 팩스 : 02-477-2898, E-mail : ysch@statcom.snu.ac.kr)

성 및 일치성 등의 문제를 부분적 보완을 해 준다. 그러나 추정량 표준오차의 과소추정 또는 계산의 난해성의 문제를 여전히 가지고 있다. 이 장에서는 단순 대치법의 문제를 보완 할 수 있는 다중 대치법에 관하여 소개한다. 다중 대치법은 단순 대치법을 한번 하지 않고 m번의 대치를 통한 m개의 가상적 완전한 자료를 만들어서 분석하는 방법으로 다음과 같이 3가지 단계로 구성되어있다.

1 단계 : 대치 (Imputations step)
2 단계 : 분석 (Analysis step)
3 단계 : 결합 (Combination step)

1. 대치 단계(Imputation Step)

결측/무응답을 가진 자료 분석을 할 때 일반적 결측 메카니즘(Missing Mechanism)에 대해서는 MAR (Missing at Random)[1]을 가정을 하게 된다. 이 가정 하에서 보편적으로 사용되는 대치법들의 통계적 추론이 유효하다. 본서에도 MAR이라는 가정 하에 결측을 가진 자료를 분석 할 수 있는 대치법을 소개한다.

대치 단계에서 일반적으로 많이 사용되는 대치 모형은 다양하다. 결측 되어지는 자료의 형태가 Monotone할 때 모수적 모형으로는 회귀 방법(Regression Method), 비모수적 모형으로는 Propensity[6] 방법 등이 있고, Non-monotone한 경우에는 일반적으로 MCMC(Markov Chain Monte Carlo)[4] 방법이 많이 사용되어 진다. 이러한 대치 방법을 이용해서 가상의 완전한 자료를 m개를 생성한다. 가상 데이터 set의 수 m을 많이 생성하면 할수록 보다 바람직한 결과를 도출 할 수 있지만, 분석에 있어서 너무 많은 시간이 소요된다. 일반 적으로 m은 3개에서 5개정도 충분하다고 알려져 있다. m의 가상 데이터 수에 기저한 대치 추정량의 효율성은 식1 로 알려져 있다. 여기서 r은 결측정보비(Proportion Missing Information)로써 3.절에서 자세히 소개하도록 하겠다.

<식1> $(1+\frac{r}{m})^{-1}$

표1 로부터 만약 결측정보비가 0.3이하일 경우 m이 3개에서 5개 정도만 되어도 m=∞에 대해 효율성이 크게 저하되지 않음 을 알 수있다.

m	r				
	0.1	0.3	0.5	0.7	0.9
3	97%	91%	86%	81%	77%
5	98%	94%	91%	88%	85%
10	99%	97%	95%	93%	92%
20	100%	99%	98%	97%	96%

표1 : 결측정보비 r와 가상데이터 수 m에 기저한 다중 대치법의 효율성

2. 분석 단계(Analysis Step)

이 단계에서는 대치단계에서 만든 m개의 완전한 가상 자료 각각을 표준적 통계분석을 통하여 관심이 있는 추정량 Θ_i와 분산 V_i을 계산한다 (i=1,…,m). 만약 원 자료가 결측/무응답이 없는 자료일 때 분석하는 표준적 통계분석법과 같은 기법을 사용하는 것이다.

3. 결합 단계(Combination Step)

이 단계에서는 분석단계에서 생성된 m개의 추정량과 분산의 결합통한 통계적 추론를 한다. 추론 방법 및 결합은 다음과 같다.

-다중 대치 추정량 : $\Theta_{MI}=\frac{1}{m}\sum_{i=1}^{m}\Theta_i$

-다중 대치 추정량의 분산 :

$T=V+(1+m^{-1})B$

대치내 분산(Within imputation variance) :

$V=\frac{1}{m}\sum V_i$

대치간 분산(Between imputation variance) :

$B=\frac{1}{m-1}\sum_{i=1}^{m}(\Theta_i-\Theta_{MI})^2$

-통계적 추론 : MI의 (1-α) · 100% 신뢰구간

$\Theta_{MI}\sim T$분포 : $\Theta_{MI}+t_\alpha(df)\sqrt{T}$,

$df=(m-1)[1+\frac{V}{(1+m^{-1})B}]^2$

-결측정보비(Proportion Missing Information)

$r=\frac{R+\frac{2}{(df+3)}}{R+1}$ 여기서 $R=\frac{(1+m^{-1})B}{V}$

SAS 프로시져

다중 대치법을 이용할 수 있는 소프트웨어 패키지는 여러 종류가 있다. 예를 들면 SOLAS, SAS, S-plus, MICE 등이 있으며, 이러한 패키지들은 나름대로의 장단점들을 가지고 있다. 본서에서는 일반 사용자들이 많이 사용하고 있는 SAS프로그램에서의 다중 대치법 사용방법을 간략히 설명한다. SAS에서는 현재까지 자료들의 값이 연속형인 경우에 한에서 만 대치할 수 있다.

- 대치 단계 : Proc MI
  MI 프로시저에서의 대치 모형으로는 Regression, Propensity, MCMC 방법을 제공해준다.
- 분석 단계 : Proc GLM, REG, Logistic, etc.
  분석 단계에서는 일반 사용자가 흔히 사용하는 프로시저를 토대로 분석을 시행한다.
- 결합 단계 : Proc MIANALYZE
  분석 단계에서 생성된 m개의 추정량에 대한 결합 및 통계적 추론을 제공해준다.

기본적 프로그램은 다음과 같고 기타 자세한 내용은 SAS 메뉴얼을 참고하기 바란다.

참고문헌

1. Rubin, D.B. (1976). Inference and Missing Data, Biometrika, 63, 581-590.
2. Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys, New York: Wiely.
3. Rubin, D.B. (1996). Multiple Imputation 18+Years, Journal of the American Statistical Association, 91, 473-489.
4. Schafer, J.L. (1999). Multiple Imputation: A primer, Statistical Method in Medical Resealch, 8, 3-15.
5. Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer, Journal of the Royal Statistical Society, Series B, 22, 302-306.
6. Rosenbaum, P.R. and Rubin, D.B. (1983). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome, Journal of the Royal Statistical Society, Series B, 45, 212-218.

Data Step

```
DATA YUN;
Input Y X1 X2 @@;
Datalins;
44.609  11.37  178  45.313  10.07  185  54.294  8.65  156
59.571  .  .  49.874  9.22  .  44.811  11.63  176
.  11.95  176  .  10.85  .  39.442  13.08  174
60.055  8.63  170  50.541  .  .  37.388  14.3  186
44.754  11.12  176  47.273  .  .  51.855  10.33  166
49.156  8.95  180  40.836  10.95  168  46.672  10.00  .
46.774  10.25  .  50.388  10.08  168  39.407  12.63  174
46.080  11.17  156  45.441  9.63  164  .  8.92  .
45.118  11.08  .  39.203  12.88  168  45.790  10.47  186
50.545  9.93  148  48.673  9.40  186  47.920  11.50  170
47.467  10.50  170
;RUN;
```

Imputation Step

```
PROC MI data=YUN seed=1234 noprint out=miout;
MCMC; VAR Y X1 X2; RUN;
```

Analysis Step

```
PROC REG data=miout outest=outreg covout noprint;
MODEL Y=X1 X2;
by _Imputation_; run;
```

Combination Step

```
PROC MIANALYZE data=outreg;
VAR intercept X1 X2; run;
```