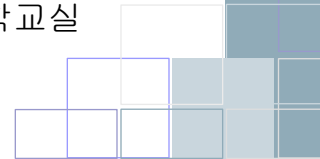


15차 SPSS OPEN HOUSE, 2008년 9월 24일



Introduction to Analysis of Missing Data

- 안 형 진, Ph.D.
- 고려대학교 의과대학
- 의학통계학교실



목차



1. Introduction
2. Pattern of Missing Data
3. Missing Data Mechanism
4. Example
5. Complete Case Analysis
6. Available Case Analysis
7. Maximum Likelihood Approach
8. Expectation and Maximization Algorithm
9. Single Imputation
10. Multiple Imputation
11. Sequential Regression Multivariate Imputation
12. Missing Data in Longitudinal Studies
13. SPSS/Missing Value Analysis
14. Conclusion

1. Introduction

SPSS

- 일반적으로 데이터는 사각행렬(rectangular matrix)의 구조를 가지고 있다
 - (행: 개체, 열: 변수)
- 일반적인 통계분석 방법은 이 데이터 행렬 안의 모든 값이 전부 관측된 경우를 가정한다.
- “결측자료(missing data or incomplete data)”
 - 데이터 행렬의 몇 몇 값이 관측되지 않은 경우

3

1. Introduction

SPSS

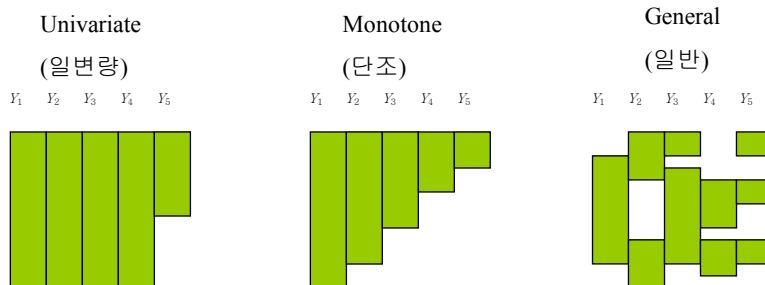
- 예
 - 표본조사(sample survey)
 - **개체 무응답(Unit non-response)**: 표본조사에서 한 개체(unit)가 응답거부 등 여러 가지 이유로 조사에 응하지 않아서 생긴 결측
 - **항목 무응답(Item non-response)**: 한 개체 내에서 몇 몇 항목의 값이 결측된 경우
 - 임상시험(Clinical trial to compare two treatment)
 - Randomize subjects to each treatment arm and take measurements at t -monthly clinical visits for each. But, some subjects drop out of the clinical trial after $t < T$ visits (중도탈락(**drop-out**) 또는 감소(**attrition**)) or miss some visits (간헐적 결측(**intermittent missing**)).

4

2. Pattern of Missing Data

SPSS

- The missing data **pattern**: 데이터 행렬에서 어떤 값이 관측되었는 지 아니면 결측인 지 나타내는 모양
- Example:



5

3. Missing Data Mechanism

SPSS

- 결측자료 메커니즘(Missing data **mechanism**): 결측과 데이터 행렬 안의 변수와의 관계.
- Let $Y = (y_{ij})$ denote an rectangular data set without missing values, with i th row where y_{ij} is the value of variable Y_j for subject i . With missing data, define the missing data indicator matrix $R = (r_{ij})$, such that

$$r_{ij} = \begin{cases} 1, & \text{if } y_{ij} \text{ is observed} \\ 0, & \text{if } y_{ij} \text{ is missing} \end{cases}$$

X_1	X_2	R
1	1	1
1	1	1
1	1	1
1	0	0
1	0	0
1	0	0

6

3. Missing Data Mechanism

SPSS

■ 완전 임의 결측(Missing Completely at Random: MCAR)

- 결측이 자료 안의 어떤 변수와도 관련이 없는 경우

$$f(R|Y, \phi) = f(R|\phi) \text{ for all } Y, \phi$$

- 매우 강한 가정
- 예: 수입에 관한 표본조사에서 수입이라는 항목에서 결측이 생겼다고 가정하자.
 - 이 때, 응답자와 무응답자 간에 어떤 본질적인 차이가 없다고 한다면 응답자의 수입에 관한 분포와 무응답자의 수입에 관한 분포가 같을 것이다.
 - 이런 경우, 완전 임의 결측이라고 한다.

7

3. Missing Data Mechanism

SPSS

■ 임의 결측(Missing at Random: MAR)

- 결측이 관측되어진 자료와 관련이 있고 결측된 자료와는 관련이 없다.

$$f(R|Y, \phi) = f(R|Y_{\text{obs}}, \phi) \text{ for all } Y_{\text{mis}}, \phi$$

- 완전 임의 결측보다는 덜 강한 가정
- 예: 수입에 관한 표본조사
 - 일반적으로 높은 소득자가 무응답률이 높을 것이다.
 - 하지만, 이제 모든 표본의 세금에 관한 정보가 있고 이 정보가 주어졌다면 소득에 관한 무응답은 임의적(random)이라고 가정하자. 즉, 같은 세금을 내는 사람들이 소득에 관해 무응답을 할 확률은 서로 같다.
 - 세금에 관한 정보가 주어진 경우, 소득에 관한 무응답은 소득과는 무관하므로 이런 경우를 임의 결측이라고 한다.

8

3. Missing Data Mechanism

SPSS

- 비임의 결측(Missing Not at Random: MNAR)
 - 결측 메커니즘이 MCAR도 아니고 MAR도 아닌 경우를 비임의 결측이라고 한다.
 - R의 분포가 결측값이 있는 변수에 의존한다.
 - 예: 수입에 관한 표본조사
 - 소득에 관한 무응답이 소득 자체와 관련이 있다.
 - 즉, 세금에 관한 정보가 주어졌더라도 소득이 높은 사람이 더 높은 무응답률을 보이는 경우.
 - 비임의 결측은 결측치 그 자체와 관련이 있다.
 - 유효한 통계적 추론을 위해서는 Y 와 R 의 결합모형이 필요하다.

9

3. Missing Data Mechanism

SPSS

- 결측 자료 분석은 이런 결측 메커니즘에 따라서 달라진다.
- 만일, 우리가 가진 결측 자료가 어떤 메커니즘인 지 안다면 좋겠으나 우리가 가진 자료로는 MCAR, MAR 또는 NMAR인 지 구분할 수 없다.
- MCAR과 MAR의 구분은 가능할 수도 있다.
- MNAR에서는 적절한 결측 자료 분석 모형을 알기가 어렵다.
- 이런 이유로 결측 자료 분석에서는 민감도 분석(sensitivity analysis)이 중요하다.
- 민감도 분석이란 다양한 메커니즘 가정과 모형 가정하에 분석결과가 어떻게 다른 지를 알아보는 방법이다.

10

4. Example

SPSS

- 해당하는 결측 메커니즘을 말해보라.
 - 임상시험에서 한 피험자가 미리 정해진 기준에 비추어 볼 때 상태가 좋지 않아서 시험에서 중도탈락하였다.
 - 동시에 한 변수에 대해 두 번 측정을 하였다. 이 측정값들의 차이가 미리 정해진 어떤 값보다 큰 경우는 다시 한 번 측정을 하였다. 이 때 세 번 째 측정된 값에는 결측치가 있는 데 이 결측은 어떤 메커니즘인가?
 - 여자가 남자보다 체중에 관해 이야기하지 않으려고 하는 경향이 있는 경우, 체중에서의 결측은 어떤 메커니즘인가? 이 때, 성별에 관한 정보는 있다.
 - 한 학교의 전체 학생에 대해 시험 1,2를 보고 다시 30%의 학생만 임의(random)로 추출하여 시험 3을 보았다. 이 경우 시험 3의 결과에는 결측이 발생한다.

11

5. Complete Case Analysis

SPSS

- Listwise deletion analysis라고도 한다.
- 모든 변수들이 관측된 개체들만 이용하여 분석
- 단 하나의 변수에서 결측값이 있어도 그 개체는 분석에서 제외
- 대부분의 통계프로그램에서 이 방법을 사용
- 장점
 - 간편성
 - 일변량 통계량들의 비교가능
 - MCAR 가정하에서 모수 추정치에 편향(bias)이 거의 발생하지 않음.
- 단점
 - 많은 표본수의 감소 → 정보의 손실 → 검정력의 약화
 - MCAR이 아닌 경우 편 발생

12

6. Available Case Analysis

SPSS

- Pairwise Deletion Analysis라고도 한다.
- 각 각의 분석 단계에서 사용 가능한 자료를 이용.
- 장점
 - 일반적으로 표본수는 complete case analysis보다 많다.
 - MCAR 가정하에서 모수 추정치에 편향이 거의 발생하지 않음.
- 단점
 - 표본의 기저(base)가 분석마다 변한다.
 - 모수 추정시 수학적 문제가 발생하기도 한다.
 - 이 방법은 실용적이지 못하며 추천되지 않는다.

13

7. Likelihood-based Inference

SPSS

- *Likelihood based estimation*
 - *Maximum likelihood estimation (MLE)*, such as implemented by the EM algorithm in the SPSS Missing Values option, applies MLE to the inference for missing data without recourse to the simulation.
 - MLE makes fewer demands of the data in terms of statistical assumptions and is generally considered superior to imputation by multiple regression.
 - In SPSS the user can specify the data distribution assumptions to be used by the EM algorithm: normal, mixed normal, and Student's t. For a mixed normal assumption, the user can specify the proportion and the standard deviation ratio. The user can also set the maximum number of iterations attempted in the iterative MLE process used by the EM algorithm (not recommended as the procedure may then stop even if estimates have not yet converged). – More on later

14

8. EM Algorithm

SPSS

- A method for finding MLEs
 - either in presence of missing data
 - or when the model can be simplified by adding ‘latent parameters’
- Selected References:
 - Hartley (1958) Biometrics, 14, 174-194 (available at JSTOR)
 - Demster, Laird, and Rubin. (1977) JRSS B, 39, 1-38(They used EM in the paper) (available at JSTOR)
 - McLachlan GJ and Krishnan T (1997) The EM algorithm and extensions, Wiley (This is a book on EM)

15

8. EM Algorithm

SPSS

- The EM algorithm works iteratively by alternatingly applying two steps: Expectation step (E-step) and the Maximization step (M-step).
 - E-step finds the conditional expectation of “missing data” given the observed data and current estimated parameters, and substitutes these expectations for “missing data”.
 - M-step updates the parameter estimates by maximizing the expected complete-data loglikelihood.

16

9. 단순 대치방법 (Single Imputation) SPSS

- 각 결측치를 하나의 값으로 대치하여 완전한 데이터 행렬로 만들고 이 대치된 값들을 실제로 관측한 값으로 여기고 분석을 한다.
- 이 때 통계적 모형을 이용하여 대치값(imputed value)을 구한다.

17

9. 단순 대치방법 (Single Imputation) SPSS

- Unconditional Mean Imputation
 - 각 결측값을 관측된 값들의 평균으로 대치한다.

y_1	y_2
0.419	0.556
1.235	2.282
0.756	1.102
0.422	0.480
1.909	1.867
-0.929	-0.572
-0.378	0.427
-1.321	-1.575
-0.074	?
0.905	?

$$\frac{\sum_{i=1}^8 y_{i1}}{8} = 0.571$$

$\tilde{y}_2 = 0.571$

18

9. 단순 대치방법 (Single Imputation) SPSS

■ Conditional Mean Imputation (Regression Mean Imputation)

- 결측값을 가진 변수를 다른 변수들과 회귀분석한 후 결측값의 예측값으로 대체한다.

y_1	y_2
0.419	0.556
1.235	2.282
0.756	1.102
0.422	0.480
1.909	1.867
-0.929	-0.572
-0.378	0.427
-1.321	-1.575
-0.074	?
0.905	?

$$\tilde{y}_{i2} = \tilde{\beta}_0 + \tilde{\beta}_1 y_{i1}$$

$$= 0.287 + 1.073 y_{i1}$$

$$0.208 = 0.287 + 1.073 \times (-0.074) \quad \tilde{y}_2 = 0.603$$

$$1.259 = 0.287 + 1.073 \times (0.905)$$

19

9. 단순 대치방법 (Single Imputation) SPSS

■ Conditional Mean Imputation (Stochastic Regression Imputation)

- 결측값을 회귀식의 예측값과 임의로 추출한 오차를 합하여 대체한다.
이렇게 함으로써 예측값의 불확실성을 고려해 줄 수 있다.

y_1	y_2
0.419	0.556
1.235	2.282
0.756	1.102
0.422	0.480
1.909	1.867
-0.929	-0.572
-0.378	0.427
-1.321	-1.575
-0.074	?
0.905	?

$$\tilde{y}_{i2} = \tilde{\beta}_0 + \tilde{\beta}_1 y_{i1}$$

$$= 0.287 + 1.073 y_{i1}$$

$$\tilde{\sigma}^2 = 0.214$$

$$0.355 = 0.287 + 1.073 \times (-0.074) + (0.147) \quad \tilde{y}_2 = 0.552$$

$$0.594 = 0.287 + 1.073 \times (0.905) + (-0.665)$$

20

9. 단순 대치방법 (Single Imputation)

■ Last Observation Carried Forward (LOCF)

- 경시적 자료(longitudinal data)에서 각 개체 내 결측값은 마지막으로 관측된 값으로 대치한다.

	Time					
Subject	1	2	3	4	5	6
1	2.3	3.2	4.5	? ← 4.5	? ← 4.5	? ← 4.5
2	1.3	1.5	2.4	1.5	? ← 1.5	? ← 1.5
3	2.1	2.0	3.3	3.5	2.9	3.5

- 평균과 공분산 구조가 매우 왜곡된다.
- 이 방법은 사용하지 않는 것이 좋다.

21

9. 단순 대치방법 (Single Imputation)

■ Hot Deck Imputation

- 주로 표본조사의 결측값을 대치하는 데 사용된다.
- 변수들이 비슷한 값을 가지는 다른 표본 개체를 임의로 골라 그 값들로 대치한다.

■ Cold Deck Imputation

- 이전의 표본조사에서 얻은 자료나 다른 자료에서 비슷한 값을 임의로 선택하여 결측값을 대치한다.

22

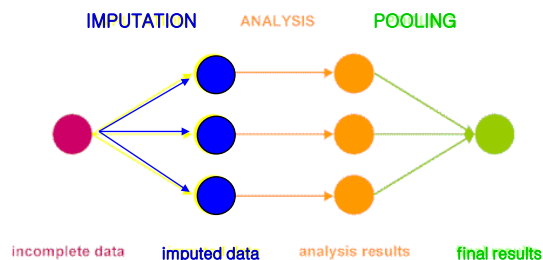
9. 단순 대치방법 (Single Imputation) SPSS

- 결측값이 모두 대치된 경우 일반적인 통계방법과 통계프로그램을 이용할 수 있기 때문에 대치법(imputation)은 많이 쓰이고 있고 또 매우 구미가 당기는 방법이다.
- 하지만 대치값을 만들기 위한 통계모형이나 결측 메커니즘에 의해 편향이 발생할 수도 있다.
- 또한, 단순 대치방법(single imputation)은 추정치의 표준편차가 과소 추정된다. 왜냐하면, 대치된 값을 실제로 관측된 값처럼 이용하였기 때문에 실제 표본수 보다 더 많은 표본수를 사용하였다.
- 다시 말해, 결측값 때문에 발생하는 불확실성(uncertainty)을 고려하지 않았기 때문이다.
- 이런 불확실성을 고려하는 방법으로는 최대우도(Maximum Likelihood)를 이용한 추정법이나 다중 대치방법(Multiple Imputation)등이 있다.

23

10. 다중 대치방법 (Multiple Imputation) SPSS

- 다중 대치법은 결측으로 인한 불확실성을 고려하기 위한 단순 대치법의 확장이다.
- 다중 대치법은 세가지 단계로 구성되어 있다.: **Imputation, Analysis, and Pooling**



24

10. 다중 대치방법 (Multiple Imputation)

■ Imputation

- 결측치를 가진 자료를 대치한다.
- 여러 번 반복하여 m 개의 대치된 dataset을 만든다.

■ Analysis

- 대치된 dataset들을 같은 통계적 방법을 이용하여 각 각 분석한다.
총 m 번 분석하게 된다.

■ Pooling

- m 개의 분석 결과를 Rubin이 제시한 결합방법을 이용하여 합치고
마지막 결과를 얻는다.

25

10. 다중 대치방법 (Multiple Imputation)

- 단순 대치방법에서의 한가지 중요한 문제는 결측으로 인한 불확실성을 고려할 수 있는 표준오차의 추정이다.
- 다중 대치방법을 이용하면 거의 모든 환경에서 결측으로 인한 불확실성을 고려할 수 있다.
- 이 방법은 표본조사론의 무응답 자료 분석을 위해 Rubin이 고안하였으나 지금은 일반적인 결측자료분석에 이용되고 있다.
- SPSS/Missing Value Analysis(MVA) version 17.0 (available on Fall 2008)

26

11. Sequential Regression Multivariate Imputation

SPSS

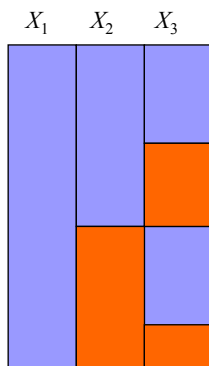
- 만일 일반패턴의 결측자료에서 변수들이 여러 형태(예: 연속형, 이항형, 순서형 등 등)인 경우는 어떨까?
- 여러 형태의 다중 변수들의 Joint Posterior 분포를 구하고 거기서 대체값을 얻으면 이상적인 결측값 분석이겠으나 문제는 형태가 다른 변수들의 joint 분포는 구할 수가 없다.
- 대신에 주변분포(marginal distribution)을 이용하여 joint 분포를 근사하는 방법이 바로 SRMI이다.

27

11. Sequential Regression Multivariate Imputation

SPSS

■ SRMI



Regress X_2 on X_1 and X_3 using a proper regression and impute missing values in X_2 .

Regress X_3 on X_1 and imputed X_2 using a proper regression and impute missing values in X_3 .

Repeat previous procedures until convergence achieved

28

11. Sequential Regression Multivariate Imputation



- IVEWare라는 프로그램을 이용하여 이 SRMI방법을 수행할 수 있다.
- 다음의 웹사이트에서 무료로 다운로드 받을 수 있다.
<http://www.isr.umich.edu/src/smp/ive/>

29

12. Missing Data in Longitudinal Studies



- LOCF – not recommended
- Mixed model for subject-specific trajectories
 - ☐ May assume linear trajectory
 - ☐ Use all available data
 - ☐ Valid under MAR assumption
 - ☐ In SPSS version 16.0, Analyze → Mixed Models → Linear
- Marginal model
 - ☐ Not individual effect, but only population effect
 - ☐ Fitted by GEE
 - ☐ Use all available data
 - ☐ Valid only under MCAR assumption
 - ☐ If the assumption is not believed to be MCAR, imputation or weighting methods should be used.
 - ☐ In SPSS version 16.0, Analyze → Generalized Linear Models → Generalized Estimating Equations

30

13. SPSS/MVA (V.16.0)



- **Create Higher-value Data and Build Better Models When You Estimate Missing Data**
- Use SPSS Missing Value Analysis to make better decisions about your data prior to analysis. With it, you can easily examine data using one of six diagnostic reports to uncover missing data patterns. You can then estimate summary statistics and impute missing values through regression or expectation maximization algorithms (EM algorithms). SPSS Missing Value Analysis helps you to:
 - Diagnose if you have a serious missing data imputation problem
 - Replace missing values with estimates, for example, impute your missing data with regression or EM algorithms
 - SPSS Missing Value Analysis, an SPSS add-on module, is a critical tool for anyone concerned about data validity, including survey researchers, social scientists, data miners, and market researchers.

31

13. SPSS/MVA (V.16.0)

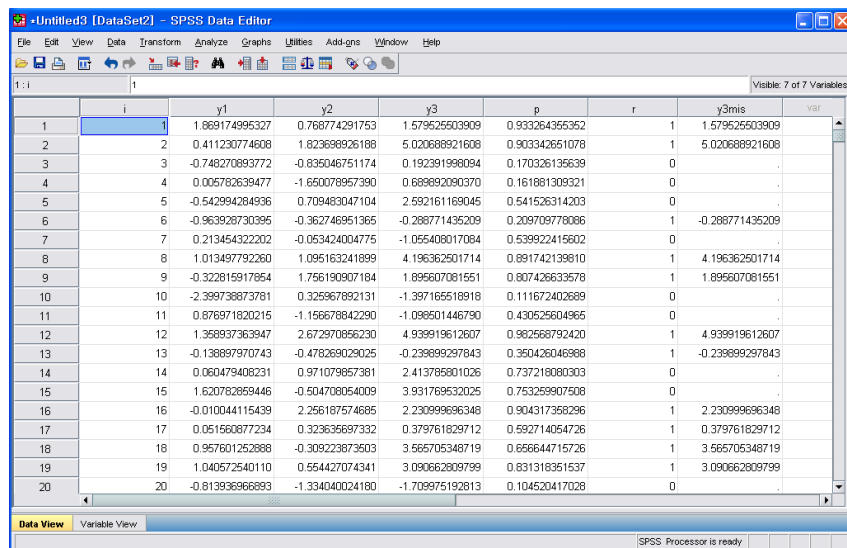


- **Simulated Data**
 - ☐ Y_1 and Y_2 are fully observed, but Y_3 is subject to missing.
 - ☐ $Y_1, Y_2 \sim N(0, 1)$ independently
 - ☐ $Y_3 \sim N(1+Y_1+Y_2, 1)$
 - ☐ $E(Y_3) = 1$
 - ☐ $R \sim \text{Bin}(1, \text{Pr}(R=1))$
 - ☐ $\text{Pr}(R=1) = \exp(Y_1+Y_2)/[1+\exp(Y_1+Y_2)]: \text{MAR}$
 - ☐ 50% of data are missing.
 - ☐ Generate 1,000 data

32

13. SPSS/MVA (V.16.0)

SPSS



SPSS Data Editor -Untitled3 [DataSet2] - SPSS Data Editor

Visible: 7 of 7 Variables

	i	y1	y2	y3	p	r	y3mis	var
1	1	1.869174995327	0.768774291753	1.579525503909	0.933264355352	1	1.579525503909	
2	2	0.411230774608	1.823698926188	5.020688921608	0.903342651078	1	5.020688921608	
3	3	-0.748270893772	-0.835046751174	0.192391998094	0.170326135639	0		
4	4	0.005782639477	-1.650078957390	0.689692090370	0.161881309321	0		
5	5	-0.542994284936	0.709483047104	2.592161169045	0.541526314203	0		
6	6	-0.963928730395	-0.362746951365	-0.288771435209	0.209709778086	1	-0.288771435209	
7	7	0.213454322202	-0.053424004775	-1.055408017084	0.539922415602	0		
8	8	1.013497792260	1.095163241899	4.196362501714	0.891742139810	1	4.196362501714	
9	9	-0.322815917854	1.756190907184	1.895607081551	0.807426633578	1	1.895607081551	
10	10	-2.399738873781	0.325967892131	-1.397165518918	0.111672402689	0		
11	11	0.768971820215	-1.156678842290	-1.098501446790	0.430525604965	0		
12	12	1.358937363947	2.672970866230	4.939919612607	0.982568792420	1	4.939919612607	
13	13	-0.138897970743	-0.478289029025	-0.239899297843	0.350426046968	1	-0.239899297843	
14	14	0.060479408231	0.971079857381	2.413785801026	0.737219080303	0		
15	15	1.620782959446	-0.504708054009	3.931789532025	0.753259907508	0		
16	16	-0.010044115439	2.256187574685	2.230999696348	0.904317358296	1	2.230999696348	
17	17	0.051560877234	0.323635687332	0.379761829712	0.592714054726	1	0.379761829712	
18	18	0.957601252888	-0.309223873503	3.565705348719	0.656644715726	1	3.565705348719	
19	19	1.040572540110	0.554427074341	3.090662809799	0.831318351537	1	3.090662809799	
20	20	-0.813936966893	-1.334040024180	-1.709975192813	0.104520417028	0		

Data View Variable View

SPSS Processor is ready

33

13. SPSS/MVA (V.16.0)

SPSS

Analysis using before deletion data (Y_1 , Y_2 , and Y_3)

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
y1	1000	-3.004	3.110	-0.020	0.984
y2	1000	-3.205	3.729	0.042	0.983
y3	1000	-4.410	6.700	1.063	1.783
Valid N (listwise)	1000				

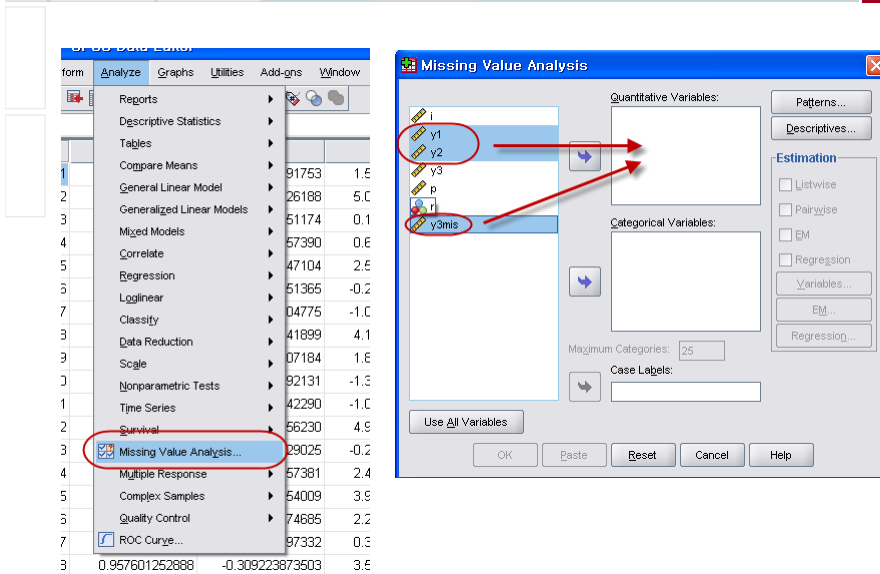
Correlations

		y1	y2	y3
y1	Pearson Correlation	1.000	.021	.553
	Sig. (2-tailed)		.501	.000
	N	1000.000	1000	1000
y2	Pearson Correlation	.021	1.000	.613
	Sig. (2-tailed)	.501		.000
	N	1000	1000.000	1000
y3	Pearson Correlation	.553	.613	1.000
	Sig. (2-tailed)	.000	.000	
	N	1000	1000	1000.000

34

13. SPSS/MVA (V.16.0)

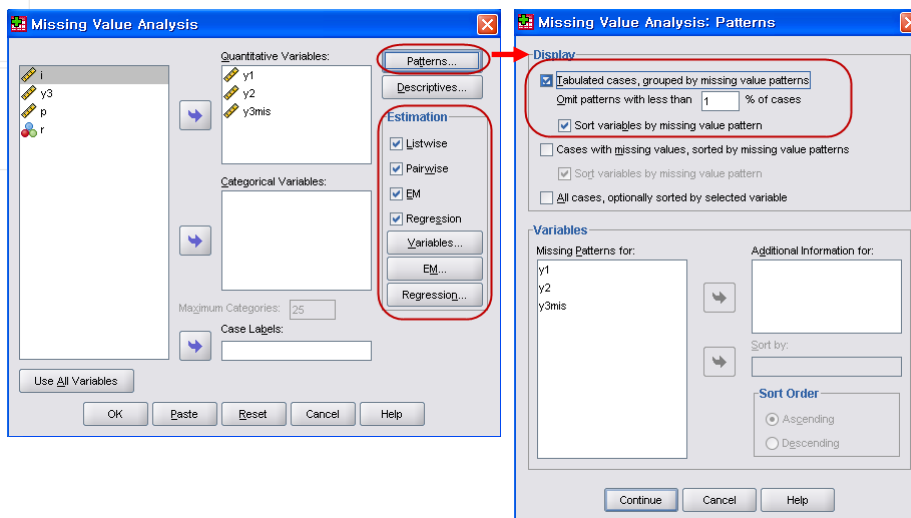
SPSS



35

13. SPSS/MVA (V.16.0)

SPSS



36

13. SPSS/MVA (V.16.0)



- The MVAR module in SPSS displays three types of 'indicator variable statistics' under its 'Descriptives' options, where the indicator variable is a dichotomous variable SPSS creates to flag whether or not the value of a given variable is present or missing.
- **Crosstabulations**: For each categorical variable, a table is generated showing for each category of the categorical variable the frequency and percent of nonmissing values for other variables.
- **Percent mismatch**: For each pair of variables, the percentage of cases in which one variable has a missing value and the other variable has a nonmissing value is displayed. This means also that the diagonal element in the table contains the percent of missing values for a single given variable.
- **t tests**: The means of the two groups formed by indicator variables are compared for each quantitative variable. The t statistic, degrees of freedom, counts of missing and nonmissing values, and means of the two groups are displayed. By default, indicator variables with fewer than 5% missing are not displayed, but the user may set this percentage under Descriptives. The probability of t tells the researcher whether the means of a given variable are different for the two groups formed by present vs missing values on a second variable. SPSS computes t and the corresponding df, but does not compute the p value.

37

13. SPSS/MVA (V.16.0)

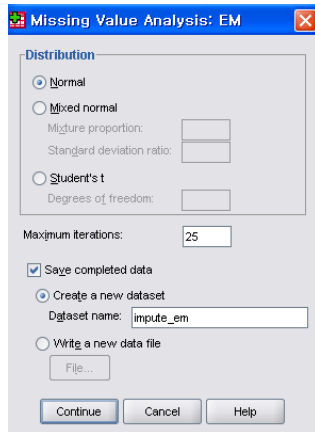


- **Note**
 - Listwise, pairwise, and regression estimation depend on the assumption of MCAR.
 - Therefore, all methods (including the EM method) for estimation give consistent and unbiased estimates of the correlations and covariances when the data are MCAR.
 - Violation of the MCAR assumption can lead to biased estimates produced by the listwise, pairwise, and regression methods.
 - If the data are not MCAR, you need to use EM estimation.
 - Data can be categorical or quantitative (scale or continuous). However, you can estimate statistics and impute missing data only for the quantitative variables.

38

13. SPSS/MVA (V.16.0)

SPSS



EM Estimation Options

Using an iterative process, the EM method estimates the means, the covariance matrix, and the correlation of quantitative (scale) variables with missing values.

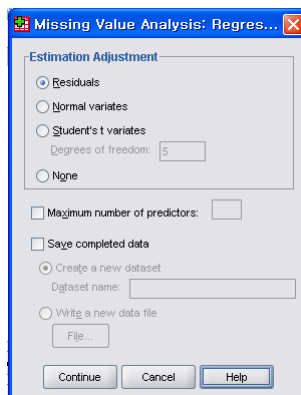
Distribution. EM makes inferences based on the likelihood under the specified distribution. By default, a normal distribution is assumed. If you know that the tails of the distribution are longer than those of a normal distribution, you can request that the procedure constructs the likelihood function from a Student's t distribution with n degrees of freedom. The mixed normal distribution also provides a distribution with longer tails. Specify the ratio of the standard deviations of the mixed normal distribution and the mixture proportion of the two distributions. The mixed normal distribution assumes that only the standard deviations of the distributions differ. The means must be the same.

Maximum iterations. Sets the maximum number of iterations to estimate the true covariance. The procedure stops when this number of iterations is reached, even if the estimates have not converged.

39

13. SPSS/MVA (V.16.0)

SPSS



- SPSS regression estimation is different from the regression imputation I explained before.

- SPSS regression estimation is only valid when missing is MCAR since the regression parameters are derived using pairwise deletion.

Regression Estimation Options

The regression method estimates missing values using multiple linear regression. The means, the covariance matrix, and the correlation matrix of the predicted variables are displayed.

Estimation Adjustment. The regression method can add a random component to regression estimates. You can select residuals, normal variates, Student's t variates, or no adjustment.

Maximum number of predictors. Sets a maximum limit on the number of predictor (independent) variables used in the estimation process.

40

13. SPSS/MVA (V.16.0)

SPSS

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes ^a	
				Count	Percent	Low	High
y1	1000	-2.042E-2	...	0	.0	3	2
y2	1000	0	.0	8	4
y3mis	500	1.7327E0	1.63983525E0	500	50.0	0	2

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Summary of Estimated Means

	y1	y2	y3mis
Listwise	1.7327E0
All Values	-2.042E-2	4.160E-2	1.7327E0
EM	-2.042E-2	4.160E-2	1.0169E0
Regression	-2.042E-2	4.160E-2	1.4893E0

Summary of Estimated Standard Deviations

	y1	y2	y3mis
Listwise	1.6398E0
All Values	9.838E-1	9.833E-1	1.6398E0
EM	9.838E-1	9.833E-1	1.8261E0
Regression	9.838E-1	9.833E-1	1.6369E0

Tabulated Patterns

	Missing Patterns			
	y1	y2	y3mis	Complete if ... ^b
500	y1	y2	y3mis	500
500			X	1000

Patterns with less than 1% cases (10 or fewer) are not displayed.

b. Number of complete cases if variables missing in that pattern (marked with X) are not used.

41

13. SPSS/MVA (V.16.0)

SPSS

Listwise and Pairwise Statistics

Listwise Means

Number of cases	y1	y2	y3mis
500	1.7327E0

Pairwise Frequencies

	y1	y2	y3mis
y1	1000		
y2	1000	1000	
y3mis	500	500	500

Pairwise Correlations

	y1	y2	y3mis
y1	1.000		
y2	.021	1.000	
y3mis	.475	.541	1.000

Listwise Covariances

	y1	y2	y3mis
y1	...		
y2	-1.277E-1	8.399E-1	
y3mis	7.093E-1	8.127E-1	2.6890E0

Pairwise Means

	y1	y2	y3mis
y1	-2.042E-2	4.160E-2	1.7327E0
y2	-2.042E-2	4.160E-2	1.7327E0
y3mis	3.039E-1	3.825E-1	1.7327E0

Mean of quantitative variable when other variable is present.

Listwise Correlations

	y1	y2	y3mis
y1	1.000		
y2	-.153	1.000	
y3mis	.475	.541	1.000

Pairwise Standard Deviations

	y1	y2	y3mis
y1	1.6398E0
y2	9.838E-1	9.833E-1	1.6398E0
y3mis	9.115E-1	9.164E-1	1.6398E0

Standard deviation of quantitative variable when other variable is present.

42

13. SPSS/MVA (V.16.0)

SPSS

Regression Estimated Statistics

Regression Means^a

y1	y2	y3mis
-2.042E-2	4.160E-2	1.4893E0

a. Random normal variate is added to each estimate.

Regression Covariances^a

	y1	y2	y3mis
y1	...		
y2	
y3mis	2.6794E0

a. Random normal variate is added to each estimate.

Regression Correlations^a

	y1	y2	y3mis
y1	1.000		
y2	.021	1.000	
y3mis	.466	.492	1.000

a. Random normal variate is added to each estimate.

43

13. SPSS/MVA (V.16.0)

SPSS

EM Estimated Statistics

EM Means^{a,b}

y1	y2	y3mis
-2.042E-2	4.160E-2	1.0169E0

a. Little's MCAR test: Chi-Square = 224.164, DF = 2, Sig. = .000

b. The EM algorithm failed to converge in 25 iterations.

EM Covariances^{a,b}

	y1	y2	y3mis
y1	...		
y2	
y3mis	1.0164E0	1.1075E0	3.3349E0

a. Little's MCAR test: Chi-Square = 224.164, DF = 2, Sig. = .000

b. The EM algorithm failed to converge in 25 iterations.

EM Correlations^{a,b}

	y1	y2	y3mis
y1	1.000		
y2	.021	1.000	
y3mis	.566	.617	1.000

a. Little's MCAR test: Chi-Square = 224.164, DF = 2, Sig. = .000

b. The EM algorithm failed to converge in 25 iterations.

44

13. SPSS/MVA (V.16.0)

SPSS

■ Data on Physical Fitness

- These measurements were made on men involved in a physical fitness course at N.C. State University. Only selected variables of Oxygen (oxygen intake, ml per kg body weight per minute), Runtime (time to run 1.5 miles in minutes), and RunPulse (heart rate while running) are used. Certain values were changed to missing for the analysis.

45

13. SPSS/MVA (V.16.0)

SPSS

	Oxygen	RunTime	RunPulse	var
1	44.609	11.37	178	.
2	45.313	10.07	185	.
3	54.297	8.65	156	.
4	59.571	.	.	.
5	49.874	9.22	.	.
6	44.811	11.63	176	.
7	.	11.95	176	.
8	.	10.85	.	.
9	39.442	13.08	174	.
10	60.055	8.63	170	.
11	50.541	.	.	.
12	37.388	14.03	186	.
13	44.754	11.12	176	.
14	47.273	.	.	.

46

13. SPSS/MVA (V.16.0)

SPSS
Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes ^a	
				Count	Percent	Low	High
Oxygen	28	47.11618	5.413047	3	9.7	0	2
RunTime	28	10.6882	1.37988	3	9.7	0	0
RunPulse	22	171.86	10.143	9	29.0	1	0

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Tabulated Patterns

Number of Cases	Missing Patterns			
	Oxygen	RunTime	RunPulse	Complete if ... ^a
21	Oxygen	RunTime	RunPulse	21
4			X	25
3		X	X	28
2	X		X	28
1	X			22

b. Number of complete cases if variables missing in that pattern (marked with X) are not used.

47

13. SPSS/MVA (V.16.0)

SPSS
Summary of Estimated Means

	Oxygen	RunTime	RunPulse
Listwise	46.35381	10.8095	171.67
All Values	47.11618	10.6882	171.86
EM	47.10418	10.5550	171.38
Regression	47.04135	10.5476	171.97

Summary of Estimated Standard Deviations

	Oxygen	RunTime	RunPulse
Listwise	5.477839	1.47277	10.351
All Values	5.413047	1.37988	10.143
EM	5.361870	1.44360	10.113
Regression	5.280996	1.43901	10.001

48

13. SPSS/MVA (V.16.0)

SPSS

■ Limitation

- ☐ Only single imputation (EM) → underestimated variance
- ☐ Ver. 17.0 incorporates multiple imputation
- ☐ Available Fall 2008

49

14. Conclusion

SPSS

- 본 강의에서는 결측분석에 관한 개론을 설명 – 이 강의에서 소개되지 않은 많은 다른 방법들도 존재함. (예: Weighting method)
- 결측의 비율이 자료에서 높은 경우 조심스러운 자료분석이 필요.
- 항상 더 많은 자료를 모으도록 노력함.
- 결측이 왜 생겼는 지 그 이유를 찾아보아 적절한 가정을 함.
- 맹목적으로 한 가지 방법만 사용하지 않음.
- 민감도 분석의 수행

50