

결측(缺測) 데이터와 그 대책

Chanseok Park (박찬석)

Applied Statistics Laboratory
Department of Industrial Engineering
Pusan National University

August 5, 2020

Hosted by SEC



부산대학교
PUSAN NATIONAL UNIVERSITY

1 Missing · Incomplete Data

- Types of Missing and Incomplete
- Illustration of Missing Mechanism
- Ad-hoc methods
- Ad-hoc methods (deletion)
- Ad-hoc methods (single imputation)
- Which method can be used?

2 Multiple Imputation

- MLE
- EM algorithm
- MI algorithm

1 Missing · Incomplete Data

- Types of Missing and Incomplete
- Illustration of Missing Mechanism
- Ad-hoc methods
- Ad-hoc methods (deletion)
- Ad-hoc methods (single imputation)
- Which method can be used?

2 Multiple Imputation

- MLE
- EM algorithm
- MI algorithm

1. Missing · Incomplete data

Types of Missing and Incomplete

- Missing data: no value is observed. (Little and Rubin, 2002)
 - **MCAR**: Missing Completely AT Random
if missingness does not depends on the data, $Y = (Y_{\text{obs}}, Y_{\text{mis}})$.
 $f_{\theta}(M|Y) = f_{\theta}(M)$, where M is a missing indicator.
 - **MAR**: Missing At Random
if missingness depends only on the observed data Y_{obs} .
 $f_{\theta}(M|Y) = f_{\theta}(M|Y_{\text{obs}})$.
 - **MNAR**: Missing Not At Random
if missingness depends on the data $Y = (Y_{\text{obs}}, Y_{\text{mis}})$.
 $f_{\theta}(M|Y) = f_{\theta}(M|Y_{\text{obs}}, Y_{\text{mis}})$
- Incomplete data: value is partially observed.
 - Truncation
 - Censoring
 - Grouping
 - Masking

1. Missing · Incomplete data: Illustration

Illustration A: Missing Mechanism (IQ is asked and GPA is known.)

Complete Data		MCAR		MAR		MNAR	
GPA	IQ	GPA	IQ	GPA	IQ	GPA	IQ
2.0	93	2.0	?	2.0	?	2.0	?
2.2	115	2.2	115	2.2	?	2.2	115
2.4	96	2.4	96	2.4	?	2.4	?
2.6	116	2.6	?	2.6	?	2.6	116
2.8	94	2.8	94	2.8	?	2.8	?
3.0	106	3.0	106	3.0	106	3.0	106
3.2	98	3.2	?	3.2	98	3.2	?
3.4	103	3.4	103	3.4	103	3.4	103
3.6	95	3.6	95	3.6	95	3.6	?
3.8	112	3.8	?	3.8	112	3.8	112
4.0	100	4.0	100	4.0	100	4.0	100
4.2	120	4.2	?	4.2	120	4.2	120

1. Missing · Incomplete data: Ad-hoc methods

Deletion – Terminology

- Complete-case analysis (listwise deletion, casewise deletion)
- Available-case analysis (pairwise deletion)

Single imputation

- Mean substitution
- Regression imputation
- Hot-deck, Cold-deck

In the next illustration, we assume that GPA, IQ, and Hours of study in a week are asked. We are interested in relation between **GPA** and **Hours**.

1. Missing · Incomplete data: Ad-hoc methods (deletion)

Illustration B: Complete and Available cases (GPA, IQ, Hr are asked)

Original Data			Complete case			Available case		
GPA	IQ	Hours	GPA	IQ	Hours	GPA	IQ	Hours
2.0	93	NA	2.4	96	26	2.4	96	26
2.2	115	NA	2.6	116	28	2.6	116	28
2.4	96	26	3.8	112	40	3.2	NA	34
2.6	116	28	4.0	100	42	3.8	112	40
NA	NA	30	4.2	120	44	4.0	100	42
NA	NA	32				4.2	120	44
3.2	NA	34				2.0	93	NA
NA	103	36				2.2	115	NA
3.6	95	NA				NA	NA	30
3.8	112	40				NA	NA	32
4.0	100	42				NA	103	36
4.2	120	44				3.6	95	NA

Available case depends on an estimate. (Here, $\text{Cov}(X_1, X_3)$ is assumed.)

1. Missing · Incomplete data: Ad-hoc methods (deletion)

Complete case (listwise/casewise deletion)

- MCAR: unbiased.
- MAR: biased.
- Popular in regression data.
- Loss in power. (small sample size).

Available case (pairwise deletion)

- MCAR: biased.
- MAR: biased.
- Correlation estimate (small value case is OK).

Refer to `Talk-R.r` at [▶ 2020/Talk-R](#).

For available case, the correlation can be **even greater than 1.0**.
(note: complete case (**MCAR**) is OK).

Illustration for **available case** continues on the next page.

1. Missing · Incomplete data: Example (Available Case)

mean/sd GPA (X_1)	mean/sd Hours (X_3)	Covariance between X_1 & X_3	
		GPA (X_1)	Hours (X_3)
2.0	26	2.4	26
2.2	28	2.6	28
2.4	30	3.2	34
2.6	32	3.8	40
3.2	34	4.0	42
3.6	36	4.2	44
3.8	40	$\text{Cov}(X_1, X_3) = 5.67$	
4.0	42		
4.2	44		
$\bar{X}_1 = 3.11$	$\bar{X}_3 = 34.67$		
$S_1 = 0.83$	$S_3 = 6.32$		

Thus, $r_{13} = \frac{\text{Cov}(X_1, X_3)}{S_1 \cdot S_3} = \frac{5.67}{0.83 \times 6.32} = \mathbf{1.08 > 1}$, which does not make sense at all. Refer to Talk-R.r at

► 2020/Talk-R

1. Missing · Incomplete data: Ad-hoc (single imputation)

Single imputation

- **Mean substitution** (location estimate)

Median, mode, HL, etc. are also OK instead of mean.

Distorted variance/covariance. Easily biased.

- **Regression imputation**

Distorted variance/covariance. can be unbiased.

A random error can be added to avoid distortion of variance/covariance. Note: if a dummy variable is used for a predictor, it can include mean substitution.

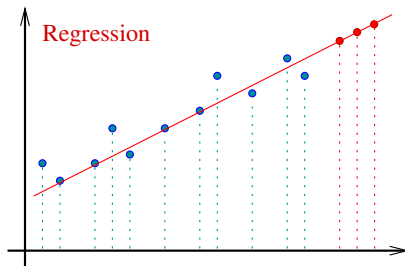
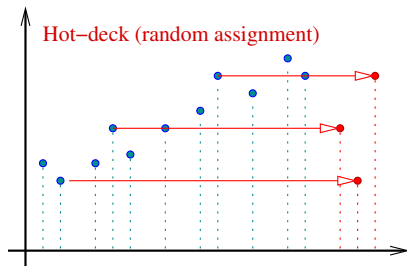
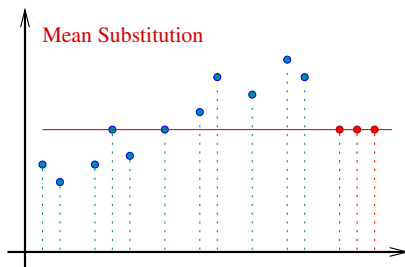
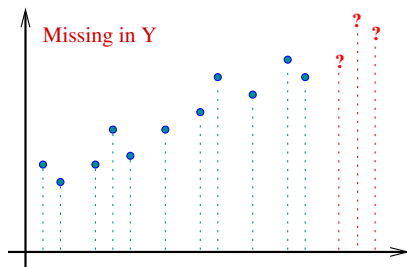
- **Hot-deck / Cold-deck** (similar to bootstrap/jackknife).

Hot-deck: random sample from similar hot responding values. (here, “hot” means current source.)

Cold-deck: random sample from similar cold responding values. (here, “cold” means previous/external source.)

Both can be easily biased.

1. Missing · Incomplete data: Ad-hoc (single imputation)



1. Missing · Incomplete data: Which method can be used?

Recall “Complete case” versus “Available case”

Complete case (listwise/casewise deletion)

- MCAR: unbiased \Leftarrow It looks OK although wasteful with deletion.
- MAR: biased.

Available case (pairwise deletion)

- MCAR: biased.
- MAR: biased.

It is possible to test MCAR vs. MAR, but impossible to test MNAR.

Section 2.2.4 of van Buuren (2018) stated: several tests have been proposed to test MCAR vs. MAR. These tests are not widely used, and their practical value is unclear. ... It is not possible to test MAR versus MNAR since the information that is needed for such a test is missing.

1. Missing · Incomplete data: Which method can be used?

Which method can be used?

- Listwise deletion (complete case) is **OK** (in a sense of unbiasedness) under **MCAR** although it is **wasteful**.

Note: listwise deletion is **not** robust to violations of the MCAR assumption. Also, the MCAR (very strong condition) is often **unrealistic** in practice.

- There are several *ad-hoc* methods under MAR, but these are **not** robust to violation of MAR assumption.
- We can think of imputation.
However **single** imputation also has several **drawbacks**.

Multiple Imputation (MI) can handle both MAR and MNAR.

Some research papers show that listwise deletion method can outperform the MI method, but it is extremely rare in practice.

2. Multiple Imputation: MLE

Multiple imputation has a very similar mechanism as EM algorithm. We look at **MLE** and **EM** first, and then **multiple imputation** later.

Likelihood and log-likelihood

$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^n f(x_i) \quad \text{and} \quad \ell(\boldsymbol{\theta}|\mathbf{x}) = \log L(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^n \log f(x_i),$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)$.

MLE (maximum likelihood estimate/estimator)

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{x}) \quad \text{or} \quad \hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\mathbf{x}).$$

In many practical cases, the MLE is obtained in a closed form.

2. Multiple Imputation: EM algorithm

We hope to have full observations, $\mathbf{x} = (x_1, x_2, \dots, x_m, x_{m+1}, \dots, x_n)$, but \mathbf{x} can be split into complete observations (\mathbf{y}) and incomplete observations (\mathbf{z}) so that $\mathbf{x} = (\mathbf{y}, \mathbf{z})$.

That is, what if $\mathbf{y} = (x_1, x_2, \dots, x_m)$ are complete and $\mathbf{z} = (x_{m+1}, x_{m+2}, \dots, x_n)$ are **incomplete**. Say, $a_j \leq x_j \leq b_j$, where $j = m+1, m+2, \dots, n$. Then we have

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \prod_{i=1}^m f(x_i) \prod_{j=m+1}^n \{F(b_j) - F(a_j)\}$$
$$\ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \sum_{i=1}^m \log f(x_i) + \sum_{j=m+1}^n \log\{F(b_j) - F(a_j)\}.$$

In general, the MLE can **not** be obtained in a closed form.

2. Multiple Imputation: EM algorithm

Treat incomplete part as random variable with an appropriate distribution. In this case, we can set up $\mathbf{z} = (z_{m+1}, z_{m+2}, \dots, z_n)$ where z_j has the pdf $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$. Then the complete likelihood is given by

$$L^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \prod_{i=1}^m f(x_i) \prod_{j=m+1}^n f(z_j),$$

where $\ell^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \log L^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$, z_j has a distribution whose value is between $a_j < z_j < b_j$ (we need to take an expectation w.r.t. z_j since z_j is like an random variable).

Recall: compare the above complete likelihood with the previous original likelihood function:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \prod_{i=1}^m f(x_i) \prod_{j=m+1}^n \{F(b_j) - F(a_j)\}$$

2. Multiple Imputation: EM algorithm

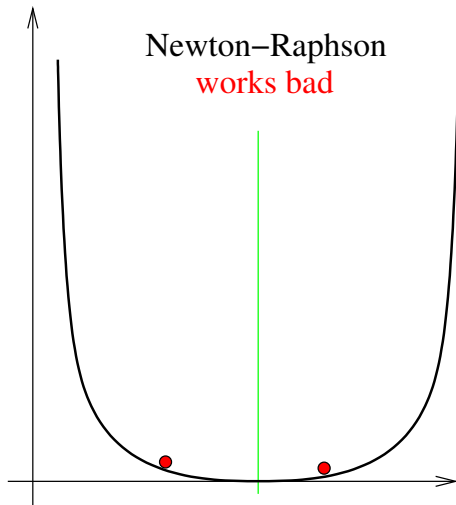
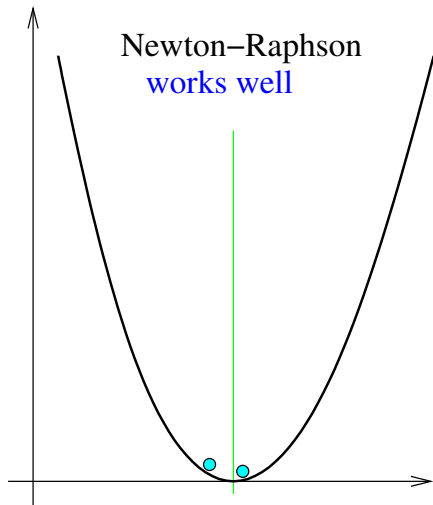
Start with an initial value (say, $t = 0$). Repeat E-step and M-step below.
(Note: we take an expectation w.r.t. \mathbf{z}_j).

EM Algorithm

- **E-step:** $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int \log L^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(t)}) d\mathbf{z}$
- **M-step:** $\boldsymbol{\theta}^{(t+1)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$

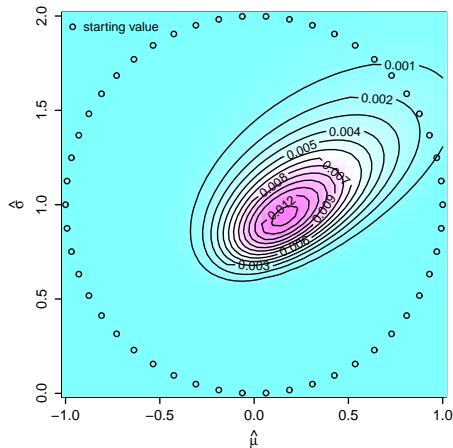
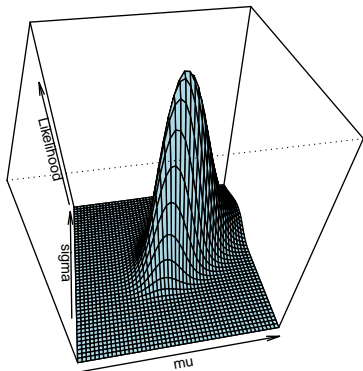
- In the M-step, we need to obtain the **closed-form** maximizer.
(in general, using EM, this M-step helps one to obtain the closed-form).
- This iterative method provides clear benefit over the Newton-Raphson method. (If the likelihood is unimodal, finding the MLE is guaranteed).
- The issue is how to obtain the closed-form maximizer in the M-step. To relax this, **MC-EM** and **Q-EM** are developed.

2. Multiple Imputation: EM algorithm



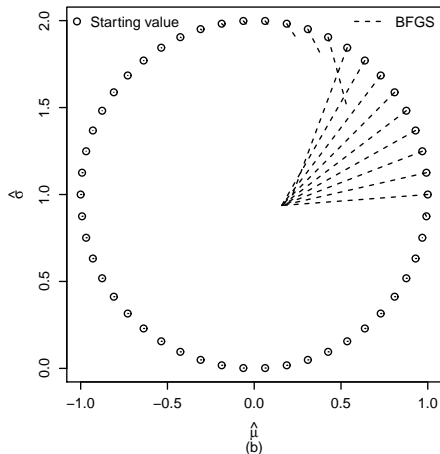
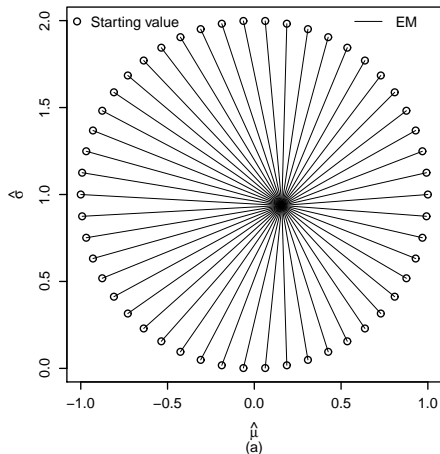
2. Multiple Imputation: EM algorithm

Illustration: Likelihood function with two parameters. Lognormal example from (Park, 2013).



2. Multiple Imputation: EM algorithm

Illustration: Convergence of BFGS (improved Newton-Raphson) method (Park, 2013)



2. Multiple Imputation: MI algorithm

MI is similar to EM which solves an **incomplete-data** problem by repeatedly solving the **complete-data** version. In MI, the unknown missing data Y_{mis} are replaced by simulated values $Y_{\text{mis}}^{(1)}, Y_{\text{mis}}^{(2)}, \dots, Y_{\text{mis}}^{(m)}$.

Multiple Imputation (MI) Algorithm

- **I-step:** $Y_{\text{mis}}^{(t+1)} \sim p(Y_{\text{mis}} | Y_{\text{obs}}, \theta^{(t)})$
- **P-step:** $\theta^{(t+1)} \sim p(\theta | Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)})$

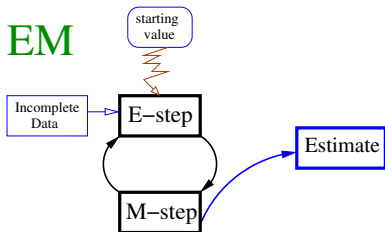
Thus, we obtain m completed data sets. With each of data sets, we analyze it by a **standard method** with a completed data set. We will have m different results. By pooling them (summarizing them), we can obtain a result along with the uncertainty due to missing.

R Packages for MI

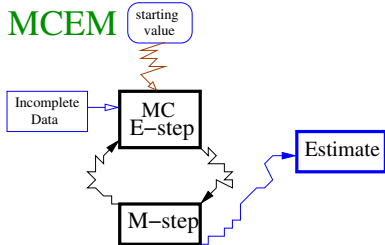
- NORM Package: <https://cran.r-project.org/web/packages/norm>
- MICE Package by van Buuren and Groothuis-Oudshoorn (2011)
<https://cran.r-project.org/web/packages/mice>

2. Multiple Imputation: Algorithms

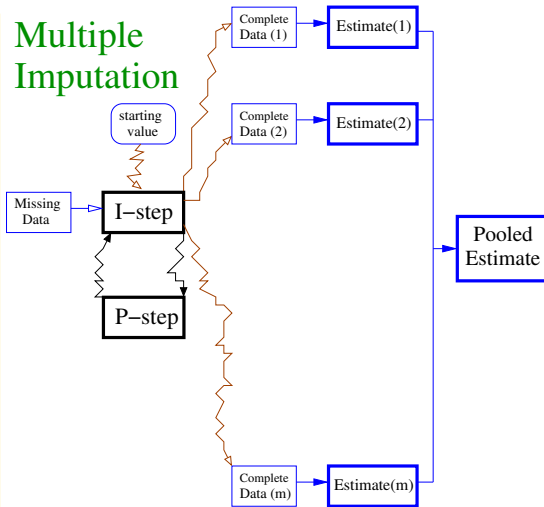
EM



MCEM



Multiple Imputation



2. Multiple Imputation: Summary

Recall

- MCAR: **unbiased** with listwise deletion (but, wasteful).
- MAR: **biased** with either listwise or pairwise deletion.
- MNAR: **biased**. Impossible to test MNAR.

When does MI work well?

- MAR with distinctness: **unbiased**.

The parameters θ and ψ are distinct if $g(\theta, \psi) = g_1(\theta) \cdot g_2(\psi)$.
See Definition 6.4 of Little and Rubin (2002).

- MI method is very **robust** to MNAR.
See Section 6.2 of van Buuren (2018).

NOTE: with MI, we can generate complete-data sets (not parameter estimate).

- Little, R. J. A. and Rubin, D. B. (2002). Statistical Analysis with Missing Data. John Wiley & Sons, New York, 2nd edition.
- Park, C. (2013). Parameter estimation from load-sharing system data using the expectation-maximization algorithm. IIE Transactions, 45:147–163.
- Park, C. (2018). A quantile variant of the Expectation-Maximization algorithm and its application to parameter estimation with interval data. Journal of Algorithms & Computational Technology, 12:253–272.
- Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall, Boca Raton, FL.
- van Buuren, S. (2018). Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton, second edition.
- van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3):1–67.

- Wei, G. C. G. and Tanner, M. A. (1990a). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. Journal of the American Statistical Association, 85:699–704.
- Wei, G. C. G. and Tanner, M. A. (1990b). Posterior computations for censored regression data. Journal of the American Statistical Association, 85:829–839.