# Imputation of Missing values

Seoul National University

Sungcheol Yun

2004. 7. 7.

## Contents

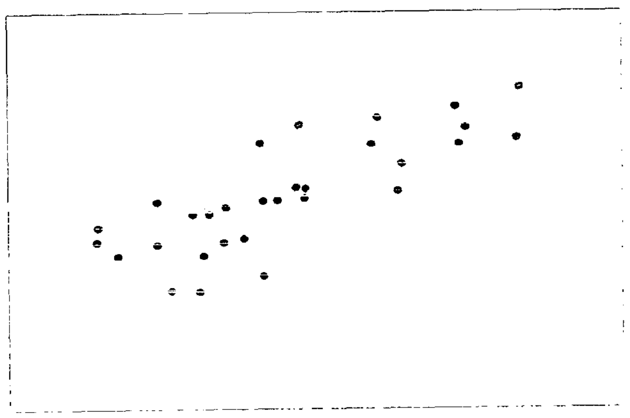
# 1 Introduction

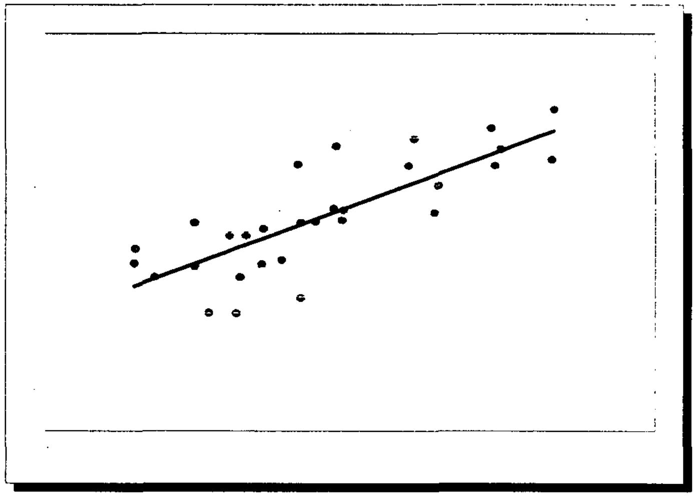
## 1.1 A Simple Continuous Example

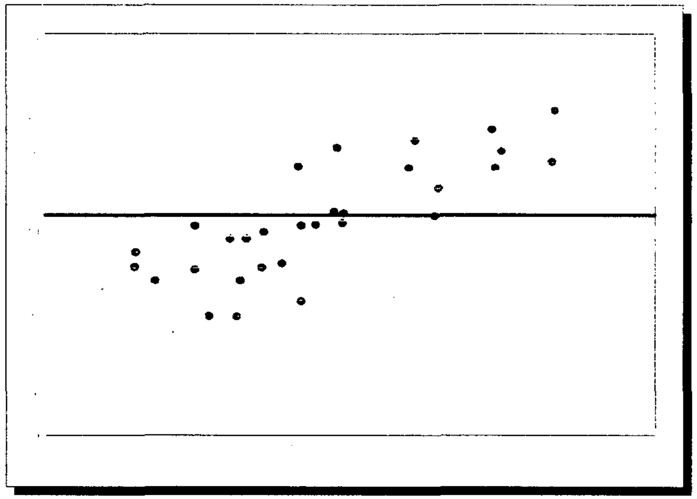
Consider a simple linear regression setting:

$$E(Y_i) = \theta_0 + \theta_1 x_i = \mathbf{x}_i^T\theta, \quad i = 1, \ldots, n.$$

$$Y_i \sim N(0, \sigma^2), \quad \text{independently.}$$





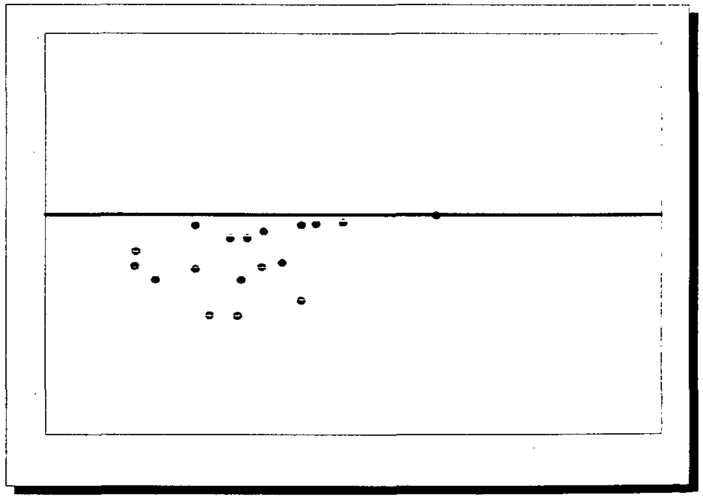


Suppose now that some **response** observations are missing.

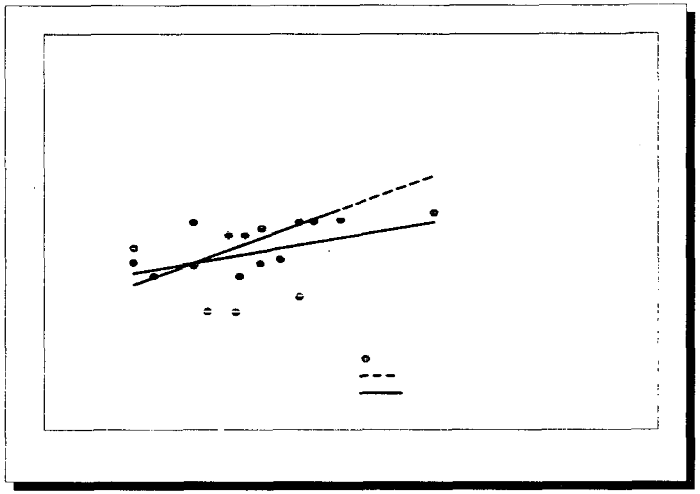
What are the implications?

→Possible bias and effects on precision.

Suppose that all observations in the example greater than 13 are unobservable.

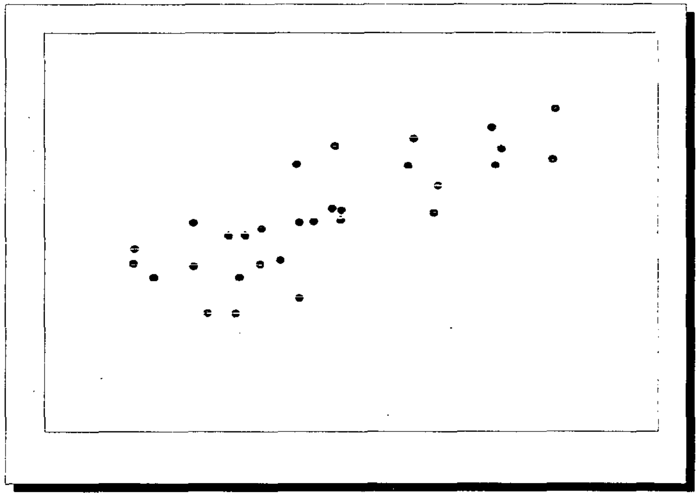


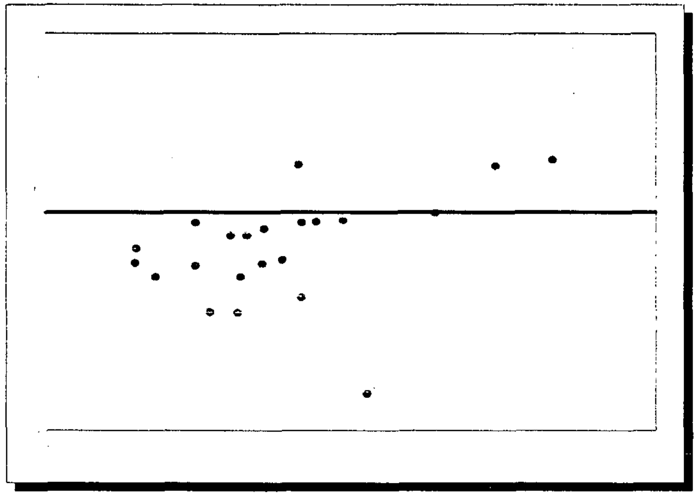
The "completers" regression line is now biased (and inconsistent)



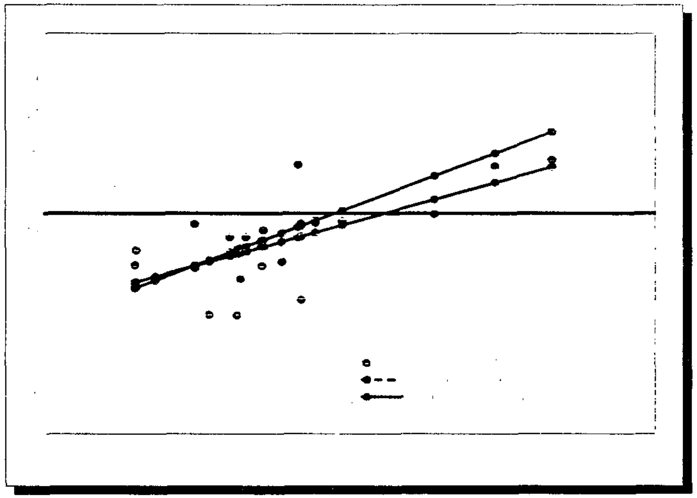
To make it a little more realistic:

suppose now that an observation greater than 13 has a probability of 0.25 of being observed.





The completers line is still inconsistent:



### 1.2 Some Notation and Definitions

*All response/outcome data*, whether observed or not:

$$\mathbf{Y} = \{\mathbf{Y}_o, \mathbf{Y}_m\}.$$

$\mathbf{Y}_o$: observed, $\mathbf{Y}_m$: missing.

*Covariates:*

$$\mathbf{X} = \{\mathbf{X}_o, \mathbf{X}_m\}.$$

$\mathbf{X}_o$: observed, $\mathbf{X}_m$: missing.

$\longrightarrow$ $Z^m = (\mathbf{Y}_m, \mathbf{X}_m)$ and $Z^o = (\mathbf{Y}_o, \mathbf{X}_o)$

*Response indicator:*

Let **Z** be partially observed, where **R** is a set of response indicators :

$$R_j = \begin{cases} 1 & \text{if the } j\text{th element of } Z \text{ observed} \\ 0 & \text{if the } j\text{th element of } Z \text{ missing} \end{cases}$$

**Goal:** we want to estimate a model for **Z** using only the observed data:

$$\{\mathbf{Z}^o, \mathbf{R}\}.$$

**Question:**

Under what circumstances are "simple" analyses "valid"?

Where

1. "simple" means that we act as though the missing data are missing by design;

2. and "valid" means inferences have the conventional justification, *e.g.* effects and their SE's consistently estimated, tests have the correct size and so on.

The answer depends on the **missing value mechanism.**

References:

Rubin DB (1976) Inference and missing data. *Biometrika*, 63, 581-592.

Little RJA and Rubin DB (2002) Statistical Analysis with Missing Data. Second Edition. Chichester: Wiley and Sons.

Look at this in the simpler case of missing response only.

It is natural (but not necessary) to formulate the model in terms of the marginal distribution of **Y**:

$$f(\mathbf{Z}^o, \mathbf{Z}^m, \mathbf{R}) = f(\mathbf{Z}^o, \mathbf{Z}^m \mid \theta) P(\mathbf{R} \mid \mathbf{Z}^o, \mathbf{Z}^m; \beta).$$

Often called the **selection model** representation.

The answer to the question depends on the **missing value mechanism:**

$$P(\mathbf{R} \mid \mathbf{Z}^o, \mathbf{Z}^m; \beta)$$

"simple" analyses are "valid" when the missing data are **Missing Completely at Random (MCAR).**

This is true when:

$$P(\mathbf{R} \mid \mathbf{Z}^o, \mathbf{Z}^m) \quad = \quad P(\mathbf{R})$$

A second **question:**

What are the most general conditions under which a "valid" analysis can be done using only the observed data, and **no** information on the functional form of the missing mechanism?

Answer: **Missing At Random (MAR)**:

$$P(\mathbf{R} \mid \mathbf{Z}^o, \mathbf{Z}^m) \quad = \quad P(\mathbf{R} \mid \mathbf{Z}^o) \qquad (1)$$

**Why?**

The likelihood for the observed data partitions

$$f(\mathbf{Z}^o, \mathbf{R}) = f(\mathbf{Z}^o) f(\mathbf{R} \mid \mathbf{Z}_o).$$

Provided the two parts depend on different parameters, likelihood inference about the response can be based on the first part alone

When neither MCAR nor MAR hold, we have:

**Missing not at Random (MNAR)**

Also called **non-ignorable**.

Even accounting for all available observed information, the missing value mechanism still depends on the unobserved data.

A "valid" analysis must take some account of the missing value mechanism.

But we will (almost) never know what this is in the MNAR case.

## 2 "Simple" Methods : Single Imputation

Nearly all such methods are based on the principle of producing a dataset that can be analysed by "complete data" methods.

2.1 Completers analysis

2.2 Mean imputation

2.3 Single stochastic imputation
Very important in a sample survey analysis.

2.4 Last Observation Carried Forward (LOCF)/Last Observation Analysed (LOA)
Commonly used in clinical trials with repeated measurements.
Currently under much criticism.

## 2.1 Completers analysis

- Delete all units with incomplete sets of observations (response and/or covariate).
- Inefficient.
- Problematic when covariate values are missing and models with several sets of explanatory variables need to be compared.

  Keep changing the underlying dataset or use a (potentially small) common subset?
- Inconsistent unless MCAR, or unless inferences wanted about the "completers population".

## 2.2 Mean Imputation

- Replace each missing value by an "appropriately" chosen mean value.
- *Simple unconditional group means (based on completers)*

  Reproduces the completers means.

- *e.g.*) Unconditional mean imputation

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 10 | ? | 15 | 19 | 12 | 18 | ? | ? | 16 |
| 10 | 15 | 15 | 19 | 12 | 18 | 15 | 15 | 16 |

← Missing value $y_{ij}$ are imputed by $\overline{y_o} = 15$

*Buck's method (conditional mean imputation.)*

Reference : Buck SF (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.

Replace each missing value by a predicted mean based on available information on that unit.

Estimate the predictive models from completers.

*e.g.* multivariate Gaussian:

1. To estimate the mean vector $\mu$ and the covariance matrix $\Sigma$ from the complete cases.

   → This step builds on the assumption that $\mathbf{Y} \sim N(\mu, \Sigma)$.
2. For a subject with missing components, the regression of the missing components $\mathbf{Y}_i^m$ on the observed ones $y_i^o$ is

$$\mathbf{Y}_i^m | \mathbf{y}_i^o \sim N(\mu^m + \Sigma^{mo}(\Sigma^{oo})^{-1}(\mathbf{y}_i^o - \mu_i^o), \\ \Sigma^{mm} - \Sigma^{mo}(\Sigma^{oo})^{-1}\Sigma^{om}).$$

*e.g.*) Regression imputation

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $Y_1$ | 10 | 12 | 15 | 25 | 30 | 35 | 37 | 40 | 42 | 50 |
| $Y_2$ | 15 | 25 | 35 | 48 | 49 | 55 | 47 | 60 | 65 | 70 |
| $Y_3$ | 20 | 30 | 40 | 57 | 60 | 65 | 70 | $?_1$ | $?_2$ | $?_3$ |

$Y_{3i} = \beta_0 + \beta_1 Y_{1i} + \beta_2 Y_{2i} + \varepsilon_i,\ i = 1, 2, \cdots, 7$

$\mapsto \widehat{\beta_0} = 3.69,\ \widehat{\beta_1} = 0.99.\ \widehat{\beta_2} = 0.56$

Then,

$\widehat{?_1} = 3.69 + 40 \cdot 0.99 + 60 \cdot 0.56 = 76.89$

$\widehat{?_2} = 3.69 + 42 \cdot 0.99 + 65 \cdot 0.56 = 81.67$

$\widehat{?_3} = 3.69 + 50 \cdot 0.99 + 70 \cdot 0.56 = 92.39$

In this way, "vertical" information (estimates for $\mu$ and $\Sigma$) is combined with "horizontal" information ($\mathbf{y}_i^?$).

Buck (1960) showed that under mild regularity conditions, the method provides consistent estimators under MCAR mechanisms.

Little and Rubin (1987) added that the method provides consistent estimators under certain types of MAR mechanism.

Even though the distribution of the observed components is allowed to differ between complete and incomplete observations, it is very important that the regression of the missing components on the observed ones is constant *across* missingness patterns

General problem with mean imputation:

Even if subsequent means are consistently estimated the precision will be wrong.

Variability will be underestimated.

## 2.3 Single stochastic imputation

- Broadly, single stochastic imputation methods seek to provide draws from the appropriate distribution.
- Such methods are very widely used in survey analysis, where, in particular, many variations on **hot-deck** methods exist
- These methods involve replacing missing values in one unit from another unit (donor) that is in some sense similar.
- Implicitly they use non-parametric estimates of the distribution of the missing data.

- *Simple hot-deck*
  A unit is chosen randomly from those that match the incomplete unit on the *observed values*.
  Problem : may be few or none to chose from.
- *Hot-deck within classes*
  Imputation classes are formed of "homogeneous" units.
  A random draw of a completer from the same class is used as the donor
- *Nearest-Neighbour*
  Some distance function is set up between units, based on fully observed measurements, and the nearest complete unit is used as donor.

Obviously we can't use the usual complete data estimators of precision.

For each particular class of estimator (*e.g.* mean, ratio, percentile) each type of imputation has an associated variance estimator that may be

- design based (*i.e.* using the sampling structure of the survey)
- model based, or model assisted (*i.e.* using some additional modelling assumptions)

These variance estimators can be *very* **complicated** and are **not convenient** for generalization.

*e.g.*, Särndal C-E, Swensson B and Wretman J (1992) *Model Assisted Survey Sampling*. New York: Springer.

Skinner C and Rao JNK (2002) Jackknife variance estimation for multivariate statistics under hot deck imputation from common donors. *Journal of Statistical Planning and Inference*, **102**, 421-436.

## 2.4 Last Observation Carried Forward (LOCF)/Last Observation Analysed (LOA)

Specific to repeated measurements problems, largely *restricted to* clinical trials.

Missing values are replaced by the last observed measurement.

Under MCAR, LOCF typically produces bias of which the direction and magnitude depend on the true but *unknown* treatment effects.

CC also produces bias under MAR that can *go in either* direction.

The LOCF estimator is not conservative in general.

Their remains the important issue of appropriate *precision* for LOCF *estimators*. For full repeated measurements analyses this is clearly disastrous: means and covariance structure are seriously distorted.

# 3 Multiple Imputation

- Imputation – fill the holes in the data
  - usually with *best possible estimate*
  - followed by standard analysis
  - overestimate sample size, thus underestimates error

- Multiple Imputation (MI) – *do this m times*
  - with randomly chosen *estimate* from distribution of possible
  - followed by *m standard analyses*
  - the *m* outcomes are then combined
  - the variation of *m* imputations restores the error

Steps for Multiple Imputation

1. **Imputations**
   : Generate a set of $m > 1$ plausible values for $\mathbf{Z}^m = (\mathbf{Y}_m, \mathbf{X}_m)$

2. **Analysis**
   : Analyze the *m* datasets using complete-case methods.

3. **Combination**
   : Combine the results from the *m* analyses.

### 3.1. Imputation Step

◇ The MAR assumption is key to the validity of MI

- To generate imputations $(Z^{(1)}, Z^{(2)}, \cdots, Z^{(m)})$ from the distribution $f(\mathbf{Z}^m | \mathbf{Z}^o)$

◇ There are a variety of imputation models that have been used.

- Regression Method
- Propensity Score Method
- MCMC method
- etc.

◇ How Many ?

- An estimator based on $m < \infty$ imputations has efficiency

$$(1 + \frac{\gamma}{m})^{-1}$$

where $\gamma$ is proportion missing information

Answer : 3-5 is enough!

| | $\gamma$ | | | | |
|---|---|---|---|---|---|
| $m$ | 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 |

### 3.2. Analysis

◇ Standard complete data analysis techniques

◇ Obtain $m$ sets of point estimates $\theta_i$ and variances $V_i$

◇ Combine $m$ results into single outcome

### 3.3. Combination

◇ Simply compute mean of $m$ estimates

$$\bar{\theta}_{MI} = \frac{1}{m} \sum \hat{\theta}_i$$

◇ Standard Errors

- Within imputation variance ( i.e. mean of $m$ variances)

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

- Between imputation variance ( i.e. variance of point estimates )

$$B = \frac{1}{m-1} \sum (\hat{\theta}_i - \bar{\theta}_{MI})^2$$

→ Total error variance

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

◇ MI confidence interval:

$$\overline{\theta}_{MI} \pm t_{df}\sqrt{T} \quad \text{where} \quad df = (m-1)\left[1 + \frac{\bar{V}}{(1+m^{-1})B}\right]^2$$

◇ The relative increase in variance due to missing values is estimated by

$$R = \frac{(1+m^{-1})B}{\bar{V}}$$

◇ The rate of missing information is estimated by

$$\gamma = \frac{R + \frac{2}{df \cdot 3}}{R + 1}$$

## 4 Example : Using SAS procedure

◇ The imputation step is carried out by PROC MI

- regression, propensity, MCMC method

◇ Once PROC MI has been run, use of complete data methods

- PROC GLM, PHREG, LOGISTIC, etc.
- BY statement to repeated these methods for each value of the variable _Imputation_

◇ The results are combined using PROC MIANALYZE

- which provides a clear summary of results.