Missing data patterns and a first introduction to imputation

Henrik Støvring
(stovring@biostat.au.dk)

Department of Biostatistics
SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

November 28, 2013

Outline

Overview

- Imputing missing values
  - Single value imputation – variants and shortcomings
  - Multiple imputations – a first example
  - Rubin's rule/formula
- Missing data patterns
  - One variable only
  - Monotone
  - General patterns

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Similar value imputation

- Consider subject i and i'
- i and i' are equal, except...
- ... i has a missing value, i' has not
- If MAR, we might as well have observed i as i'
- Idea: Substitute i's value for the missing value of i
- An example of (single) imputation

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Mean value imputation

- Consider two groups (for each variable)
  - Subjects with an observed value
  - Subjects with a missing value
- If MAR, the missing values are similar to observed values
- Idea: Substitute average observed value for missing values
- Often done stratified on other covariates (think MAR and similar value imputation)

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Missing as separate category

- Warning: Popular, but has no real rational motivation
- Only relevant when variable with missing values is
  - categorical
  - covariate
- Idea: Add a new category corresponding to the missing values
- Estimate as usual with the categorical variable

Example: Missing smoking status (m3_bw.dta)

- Of 10,000 sampled, smoking status is missing for 529
- Three approaches for replacing missing values:
  - Find birth weight closest to child with missing smoking status, impute this smoking status for the missing value
  - Compute “mean” smoking status, i.e. probability of smoking, and impute this
  - Replace missing smoking status with a new value, indicating a “new” category

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Example: Stata code, similar value

```
. use m3_bw.dta
. sort bweight
. generate tmpgrp = sum(!missing(cigs[_n]) & missing(cigs[_n+1]))
. bysort tmpgrp (bweight): replace cigs = cigs[1] ///
                                    if missing(cigs)

. regress bweight cigs
```

Example: Stata code, mean value

```
. use m3_bw.dta
. summarize cigs, mean
. replace cigs = r(mean) if missing(cigs)

. regress bweight cigs
```

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Example: Stata code, missing as category

```
. use m3_bw.dta
. recode cigs (. = 10), gen(cigscat)
. local cigslab : value label cigs
. label define `cigslab' 10 "Missing", add
. label values cigscat `cigslab'

. regress bweight i.cigscat
```

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Single value imputations: Pro's and Con's

- Pro's:
  - Simple to implement
  - Makes use of all observed data
- Con's:
  - Uses data more than once
  - Overconfident in imputation step
  - Underestimates uncertainty due to missing observations
- Should never be used as general solution

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

The fundamental idea

- What does MAR mean?
- Consider variable Z_1 with a missing value, $z_{1k} = ?$
- Assume distribution of Z_1 depends on Z_2 and a parameter θ

$$P(Z_1 < z) = F(z; Z_2, \theta)$$

- The concealed value of z_{1k} also has distribution

$$P(Z_{1k} < z) = F(z; Z_2, \theta)$$

  because under MAR, it does not depend on the value being missing when we know Z_2
- Assume we can estimate shape of $F(z; Z_2, \theta) \equiv F_\theta(z)$

The fundamental idea (II)

- If we know $F_\theta(z)$, we can sample from it:
  1. Sample value, replace missing value with it
  2. Keep observed variables
  3. Save *completed* dataset j
  4. Repeat m times to create m complete datasets
- Analyze each completed dataset j to obtain an estimate β_j

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

The fundamental idea (III)

- Now have *m* different estimates β_j of the same quantity β
- Rubin's formula

$$\widehat{\beta} \equiv \frac{1}{m} \sum_j \widehat{\beta}_j$$

- Uncertainty estimate is a sum of
  - The mean of the standard errors of $\widehat{\beta}_j$
  - Variability between $\widehat{\beta}_j$'s across imputations

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

The fundamental idea (III)

- Formula for standard error

$$s.e.(\widehat{\beta}) = \sqrt{\mathrm{E}\left(s.e.(\widehat{\beta}_j)^2\right) + \left(1 + \frac{1}{m}\right)\frac{\sum_j(\widehat{\beta}_j - \mathrm{E}(\widehat{\beta}_j))^2}{m-1}}$$

- Used in t-distribution with the following degrees of freedom:

$$df = (m-1)\left(1 + \frac{m \cdot \mathrm{E}\left(s.e.(\widehat{\beta}_j)^2\right)}{(m+1)\frac{\sum_j(\widehat{\beta}_j - \mathrm{E}(\widehat{\beta}_j))^2}{m-1}}\right)$$

Combining Estimates (general – for statisticians) I

- For each imputed dataset, estimate $\widehat{Q}_j$ and variance estimate U_j (think of U_j as $s.e.(\widehat{\beta}_j^{\,2})$)
- Parameter estimate

$$\overline{Q} = \frac{1}{m}\sum_{j=1}^{m}\widehat{Q}_j$$

- Within-imputation variance

$$\overline{U} = \frac{1}{m}\sum_{j=1}^{m}U_j$$

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Combining Estimates (general – for statisticians) II

- Between-imputation variance

$$B = \frac{1}{m-1} \sum_{j=1}^{m} (\widehat{Q_j} - \overline{Q})^2$$

- Total variance is estimated as

$$T = \overline{U} + \left(1 + \frac{1}{m}\right) B$$

- Use $\sqrt{T}$ for standard error and *t*-distribution for tests and confidence intervals, where degrees of freedom are

$$df = (m-1)\left(1 + \frac{m\overline{U}}{(m+1)B}\right)^2$$

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Combining Estimates (general – for statisticians) III

- Estimated rate of missing information is

$$\gamma = \frac{r + 2(df + 3)^{-1}}{r + 1}$$

with

$$r = \frac{(1 + m^{-1})B}{\overline{U}}$$

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Conditions for validity

- Imputations must be *proper*, i.e.
  1. Estimates from imputed datasets asymptotically follow a normal distribution
  2. Variance of estimates is a consistent estimate of true within-imputation variance, and smaller asymptotically than variance of estimate
- Hard to verify in practice
- Rule of thumb:
  Whenever complete case analysis is OK asymptotically, and imputations are not degenerate: It works

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Why are only a few imputed datasets needed?

- Assume rate of missing information is not too large ($\gamma < 0.3$)
- Relative efficency of MI is $> 94\%$ with 5 imputations
- Details in Rubin (1987, p. 114)
- Approximate formula for efficiency is

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

Why you often need many imputed datasets

- Assume you
  1. estimate “many” parameters
  2. do combined tests of parameters
- Then your m should be relatively large
  for B to be well estimated with respect to correlation of estimates
  where B is between-imputation covariance matrix

Multiple imputation in practice

- Implemented in statistical packages:

SAS	`MI` and `MIANALYZE`
Stata	`-ice-` (add-on) and `-mi-`
R	`mi` and `mice` (add-ons)
SPSS	`MULTIPLE IMPUTATION`

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Gaussian response: Temperature data I

- Assume temperature is normally distributed for each gender j ($N(\mu_j, \sigma^2)$)
- $\rightarrow$ Temperature can be predicted from gender assuming MAR

Gaussian response: Temperature data II

Algorithm

1. Estimate μ and σ from observed data
```
xi:  regress tempC i.sex
```
2. Predict probabilities
```
predict mu if tempC == .
predict sdf if tempC == ., stdf
```
3. Impute (guess) missing values
```
gen outcome = tempC if tempC != .
replace outcome = rnormal(mu, sdf)
                         if tempC == .
```

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

“Non-parametric” response

- Situation: Want to apply non-parametric analysis
- No obvious distribution for missing values
- Alternatives to consider
  - Similar value imputation
  - Transform to normality
  - Impute based on parametric distribution (Normal?) $\rightarrow$ analyze non-parametrically

Monotone patterns

- Assume order exist for $x_{[1]}, \ldots, x_{[k]}$ such that
  1. At least one covariate $x_{[1]}$ has no missings
  2. If $x_{[j]}$ has missings then $x_{[1]}, \ldots, x_{[j-1]}$ are not missing for these observations
- Can be inspected with `-misstable-` in Stata
- If MAR can be assumed, then a sequential approach is possible for predicting missing values:
  - Let $j = 1, \ldots, k$:
  - Predict $x_{[j]}$ from $x_{[1]}, \ldots, x_{[j-1]}$

Non-monotone patterns

- The most common situation:
  - Two or more covariates missing for the same individual(s)
- Problem in predicting unobserved values under MAR
  - Want to predict x_1 from observed values in x_2, x_3
  - Want to predict x_2 from observed values in x_1, x_3
  - When x_2 is missing, observed value in x_3 is dropped in estimation of relationship between x_1 and x_2, x_3
  - When x_1 is missing, observed value in x_3 is dropped in estimation of relationship between x_2 and x_3, x_3

Binary or categorical covariate

- Covariate can only assume *r* different values
- Strategy
  - Estimate probability for each category (logistic, ordered logistic, multinomial)
  - Impute category for missing values based on estimated probabilities
- Requires assessing assumptions for modeled relationship:
  - Linearity in log-odds (logistic regression)
  - Collinearity/perfect prediction
  - Proportional odds (ordered logistic regression)
  - ...

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Continuous covariate

- Ordinary regression: No assumed distribution of covariate
- Need distribution for imputing missing values
- Standard assumption: Normal or transformed-normal
- Use linear regression for prediction
- Requires assessing assumptions for modeled relationship:
  - Linearity
  - Homo-schedasticity (constant standard deviation)
  - Normality of residuals

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Two variables: one with missings, one without

- Consider *m3_bw* and the variables:

  cigs: Cigarette equivalents per day, has missings
  bweight: Birth weight in grams, has no missings

- Imputation strategy is straightforward:
  1. Model relationship between smoking and birthweight among observed values
  2. Assume identical relationship for unobserved smoking
  3. Impute values of smoking when missing

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Three variables: two with missings, one without

- Consider *m4_bw* and the variables:

  cigs: Cigarette equivalents per day, has missings
  alko: Alcohol units per week, has missings
  but only if cigs is missing
  bweight: Birth weight in grams, has no missings

- Example of monotone missing data pattern

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Imputation in monotone patterns

- Assume
  - Birth weight predicts alcohol intake
  - Alcohol intake (and birth weight?) predicts smoking
- Imputation strategy becomes sequential
  1. Model relation between birth weight and alcohol,
     $\rightarrow$ impute alcohol
  2. Model relation between alcohol and smoking including imputed alcohol values
     $\rightarrow$ impute smoking
- Implemented in Stata with `-mi impute monotone-`

Comments on monotone imputation

- Is transparent
- Is computationally efficient
- Does not allow for feedback:
  Based on an imputed smoking value, one might want to re-impute alcohol value
- Only useful if imputation model follows missing data pattern

More variables, not monotone

- Two strategies
  1. Impute several variables jointly (`-mi impute mvn-`)
  2. Impute iteratively in round-robin fashion (`-mi impute chained-`

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

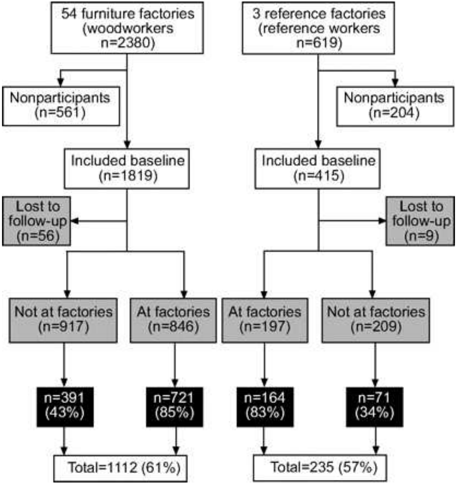
Wood dust and Lung Function

- Paper by Jacobsen et al, “Longitudinal lung function decline and wood dust exposure in the furniture industry”, 2008, Eur Respir J
- Included woodworkers from furniture factories and reference workers from reference factories
- Follow-up study

  Inclusion: Workers at factories in Viborg County, 1997-98

  Follow-up: First through factories, then mail, 2003-5

Flow chart of study population



SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Selection process

1. Invited: $n = 2,999$
2. Non-participation (75.5% retained)
3. Lost to follow-up (44.9% retained)
4. Non-response for individual items (measurements, questions) xx% ??
5. Study population: $n = 1,347$

Selection process

Using a *complete case analysis*

- Study population included: $n = 1,347$
- Table 1: "#: valid cases vary between variables"
- Table 2: $n = 1,335$
- Table 3: $n = 1,260$
- Table 4: $n = 1,199$
- Table 5: $n = 1,230$ or $n = 1,212$
- Table 6: $n = 1,190$
- Minimum participation rate: 39.7%
- Maximum non-response rate among participants: 11.7%

Wood dust and lung function

- Use dataset underlying the paper – thanks to V Bælum, Public Health, AU
- Study population included: $n = 1,347$
- Maximum non-response rate among participants: 11.7%
- Objective I: Do analysis with all 1,347 included
- Objective II: Maintain original hypothesis and models (outcome, exposure, other covariates)

Table of missingness pattern

Stata commands:

- `-misstable summarize-`
- `-misstable patterns-`

Missing as outcome

- Idea:
  1. Create a missing indicator for each variable of interest
  2. Use this indicator variable as outcome (logistic regression, say)
- Example: smoking status and wood dust exposure
- Stata commands:

```
gen miss_wooddust = missing(wooddustgrp)
generate miss_smoke = missing(packryg)
```

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

Missing as outcome

- Findings
  1. No statistical significant relations for smoking being missing
  2. Statistical significant association between wood dust exposure being missing and smoking status
- Can rule out MCAR for wood dust
- May still be MNAR for both wood dust and smoking!

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

A first example “by hand”

- Outcome: Annual change in FEV1
- Exposure: Wood dust (in 4 categories)
- No missings in outcome, 84 missing values in exposure
- Imputation model: Predict wood dust exposure from outcome with polytomous logistic regression (`-mlogit-`)
- See details in do- and log-file

Thank you for your attention!

Slides prepared with LaTeX and Beamer

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY