ANALYSIS OF INCOMPLETE DATA IN STATISTICAL SURVEYS

Ugo Guarnera [1]

[1]Italian National Institute of Statistics,
Italy
guarnera@istat.it

Jordan Twinning: Imputation - Amman, 6-13 Dec 2014

OUTLINE

OUTLINE

OUTLINE

OUTLINE

OUTLINE

NON-RESPONSE

One of the most important cause for data incompleteness is the non-response.

- total nonresponse (TNR):
  sample units where no information is available
- partial nonresponse(PNR):
  some questions are not answered
- intermediate cases
  some sections of the questonnaire empty, all questionnaire empty but info available from external (historical) source, ...

TOTAL NON-RESPONSE

Causes:

- non-contact
- refusal
- ...

typically (but not always) treated via weight adjustment

PARTIAL NON-RESPONSE

Causes:

- information not available to respondents
- ambiguous questions (questionnaire defect)
- "friction" along questionnaire
- ...

typically (but not always) treated via imputation of missing values

REMARK

Identification of PNRs may be not obvious. E.g., sometimes not reported values are to be interpreted as "zeros" rather than as missing values. This problem can be alleviated with suitable strategies in the data collection phase (for instance by using different codes for zero and missing). However, often some ad hoc procedure (possibly based on statistical modeling) for distinguishing zeros and missing values is necessary.

MISSING DERIVING FROM EDITING

Missing values can derive by dropping values that have been classified as erroneous in the editing phase. This is often the case when automatic procedures for random error localization are used. For instance, if a "balance edit" involving different numerical items is violated, one particular item is flagged as erroneous, it is canceled.

DATA INTEGRATION

Missing information can be originated by data integration from different sources. For instance, if a unique dataset is created based on data from two different surveys **A** and **B**, and some variables are not surveyed in both surveys.

$\mathbf{X}$: variables only in **A**

$\mathbf{Y}$: variables only in **B**

$\mathbf{Z}$: variables in both **A** and **B**
Interest is on the joint distribution of X, Y, Z

NONRESPONSE CLASSIFICATIONS

Two main aspects of non-response are:

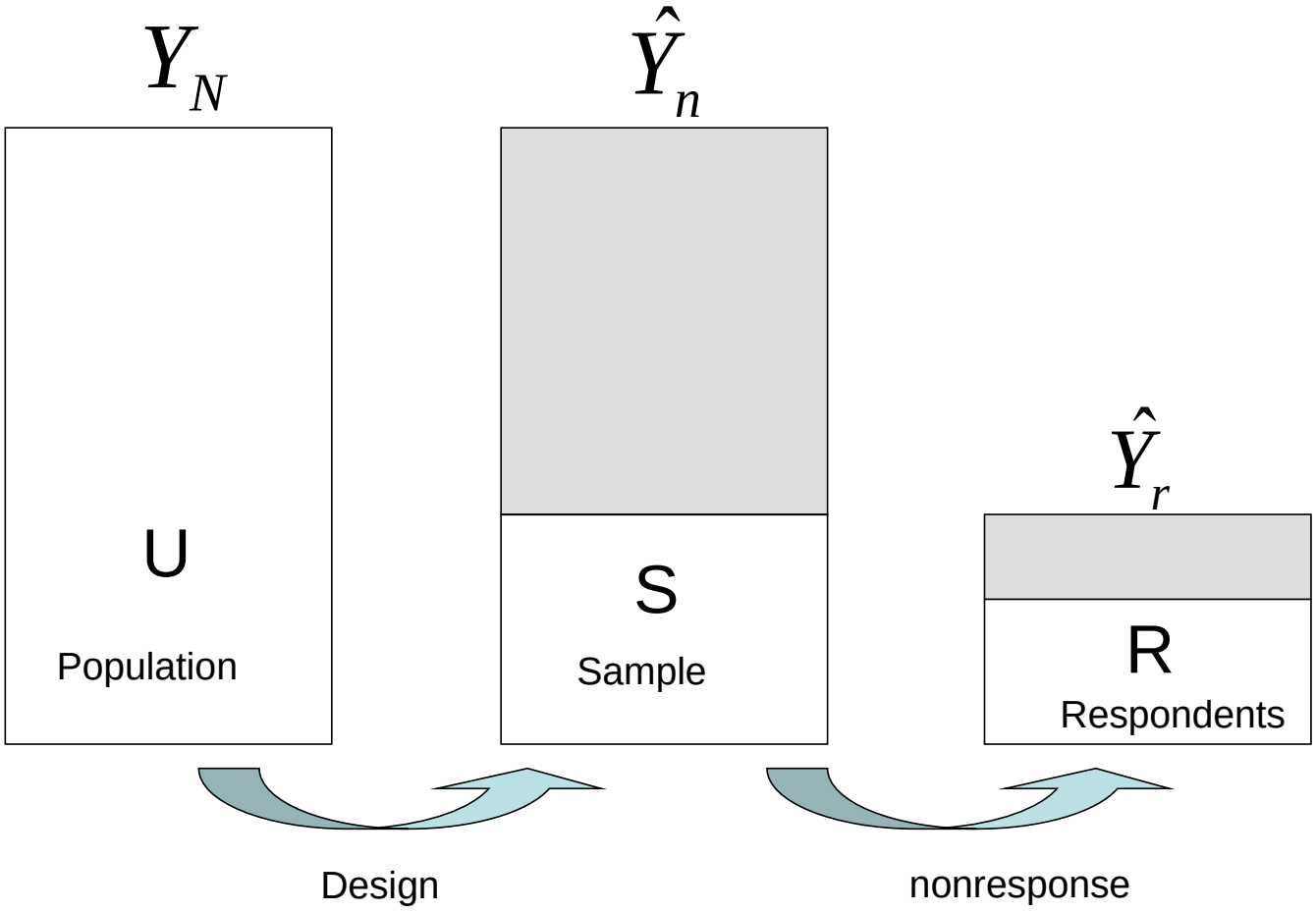
- mechanism
  analyzing (possibly modeling) non-response mechanism is important in order to have valid inferences.

- pattern
  sometimes pattern depends on mechanism. Applicability of different methods depend on nonreponse pattern .

QUASI RANDOMIZATION

- One possible approach to the inference on incomplete data is based on considering non-response as a second phase in sampling. However, the nonresponse mechanism is usually not under control of the researcher.
- Thus, validity of inferences based on respondents strongly depends on validity of assumptions made on non-response mechanism.

random mechanisms involved



EXAMPLE: SIMPLE RANDOM SAMPLING

Y = target variable for population of N units
$\hat{\overline{Y}}_n$ sample mean (estimator of population mean) on a sample of n units:

$$\hat{\overline{Y}} = \frac{\sum_{i=1}^{n} y_i}{n}$$

If there are r respondents we use the estimator $\hat{\overline{Y}}_r$ computed on r units instead of n.
is the estimator $\hat{\overline{Y}}_r$ unbiased?
is it possible to estimate its precision?

it depend on *the non-response mechanism*.

EXAMPLE: SIMPLE RANDOM SAMPLING (2)

If the set of respondents can be considered as a simple random sample of the units included in the sample the estimator $\hat{Y}_r$ is unbiased. It is a simple "expansion estimator" based on r units instead of n units. Of course the precision is lower since there is an increase in variance of $\sim n/r$.
If the hp of SRS for the (non)response is true the nonresponse mechanism is said to be *completely at random (MCAR - Missing Completely at Random)*

NOTATIONS

$\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ *data matrix* $n \times p$

$\mathbf{Y}_{obs}$ *observed data*

$\mathbf{Y}_{mis}$ *missing data*

$\mathbf{M}$ *matrix* $n \times p$ *of NR indicators:* $\mathbf{M}_{ij} = 0$ *if variable* $\mathbf{Y}_j$ *is observed* on the i-*esima unit* and 1 otherwise

NONRESPONSE MECHANISMS

- *MCAR* (Missing Completely At Random)

$$P(\mathbf{M}|\mathbf{Y}) = P(\mathbf{M}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{M})$$

- *MAR* (Missing At Random)

$$P(\mathbf{M}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{M}|\mathbf{Y}_{obs})$$

- *MNAR* (NMAR) (Missing Not At Random)

  $P(\mathbf{M}|\mathbf{Y})$ cannot be simplified as in the previous cases.

INTUITIVELY ..

MCAR Probability that $\mathbf{Y}_{mis}$ is missing does not depend on $\mathbf{Y}$ (the respondents are a (simple random) sub-sample of the original sample).

MAR Probability that $\mathbf{Y}_{mis}$ is missing depends only on $\mathbf{Y}_{obs}$. (not on the missing values)

NMAR Probability that $\mathbf{Y}_{mis}$ is missing depends on $\mathbf{Y}_{mis}$ ALSO upon conditioning on $\mathbf{Y}_{obs}$.

EXAMPLE

X=class of employees classes: 1-10; 11-50; 51-100; $>$ 100
Y= turnover
Assume that X is always observed, but Y is affected by nonresponse. If nonresponse probability $P(M)$ is independent of Y, then the mechanism is MCAR. If, *for each employee class*, nonresponse probability is constant, but there are different probabilities in different classes, then the mechanism is MAR. Finally, if even within each class, nonresponse prob. depends on turnover (Y), then we have NMAR.
If we wont to estimate the mean (or total) of Y, we can get unbiased estimators only in the firt two cases (MCAR and MAR)

EXAMPLE(CNTD)

n: sample size
n_j: size of employees classes $j \quad (j = 1, \ldots, 4)$
r: number of respondents
r_j: number of respondents in class j
y_i: turnover of ith unit(enterprise)

Let:

$$\hat{\overline{Y}}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \hat{\overline{Y}}_r = \frac{1}{r}\sum_{i=1}^{r} y_i \quad \hat{\overline{Y}}_{r_j} = \frac{1}{r_j}\sum_{i=1}^{r_j} y_i$$

all data mean, respondent mean, respondent mean in class j, respectively

EXAMPLE(CNTD)

Estimators:

- complete data: $\hat{\overline{Y}}_n$
- MCAR: $\hat{\overline{Y}}_r$
- MAR: $\frac{1}{n}\sum n_j \hat{\overline{Y}}_{r_j}$ (preferable also with MCAR)
- NMAR: ?

ADJUSTING FOR NONRESPONSE

In the MAR case we have re-weighted units separately in different adjustment cells (employee classes) to reduce bias. The idea is similar to that used in sampling (randomization): If the inclusion probability for the unit i in the sample is π_i, each unit "represents" $w_i = \pi_i^{-1}$ units in the target population. w_i = sampling weight.

Sampling weight are known because they are determined by the

sampling scheme (inverse of inclusion probabilities). Nonresponse can be viewed as a n additional selection process where, however probabilities are unknown (quasi randomization)

ADJUSTING FOR NONRESPONSE (CNTD)

If nonresponse probs p_{ri} were known for each sampled unit i, we could calculate the probability:

$\tilde{\pi}_i =$ Pr (i sampled and i respondent)= Pr(i sampled) $\times$
$\times$ Pr(i respondent | i sampled) $= \pi_i \times p_{ri}$.

Thus the adjusted weight would be:

$$\tilde{\pi}_i^{-1} = \frac{1}{\pi_i \times p_{ri}}$$

.

ADJUSTING FOR NONRESPONSE (CNTD)

Usually, response probabilities are unknown. Thus, in practice one to estimate them using the available auxuliary information. In the previous example we have made the assumption that the response probability depends only on the class of employees and we have estimated this probability with rthe response rate in each class:

$$p_{ri} = r_j/n_j$$

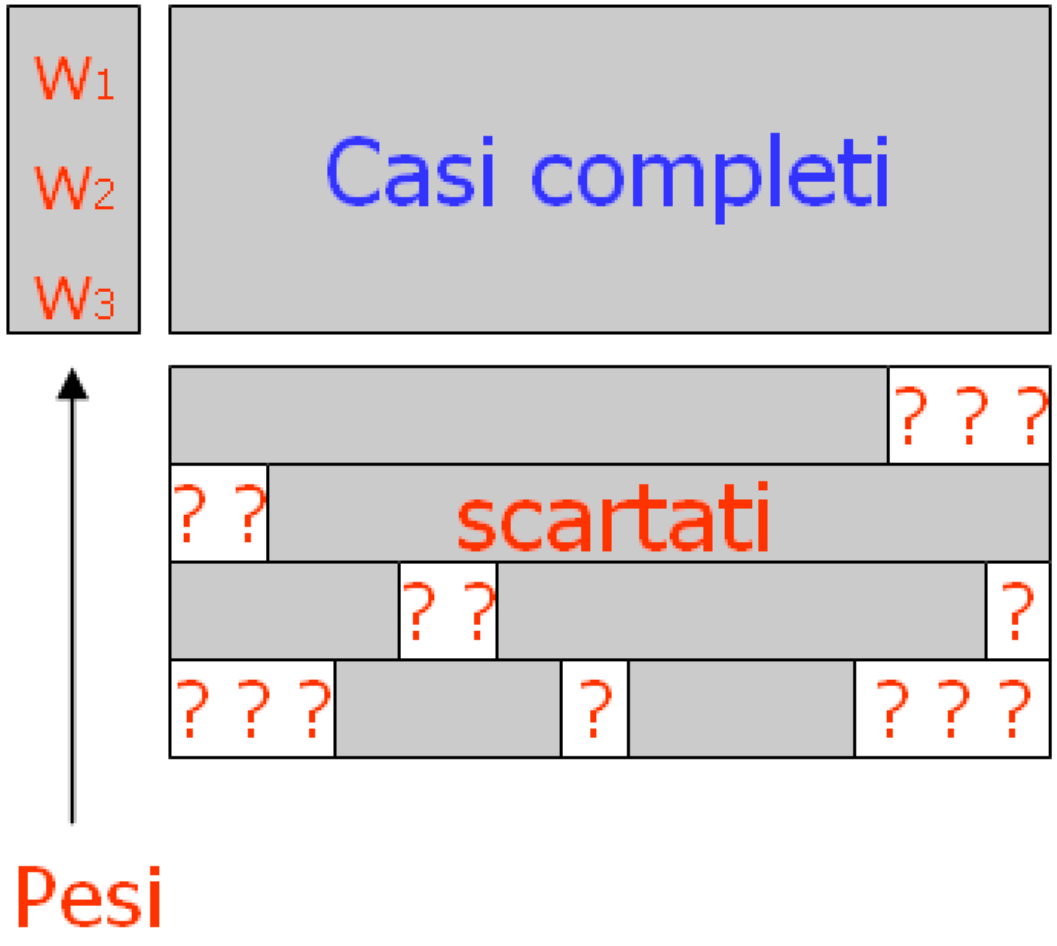
where j is the class of unit i. If all the sample weights are equal (w), this hypothesis leads to the estimator:

$$(\sum_{i}^{n} w)^{-1} \sum_{j}^{4} \sum_{i}^{r_j} w \frac{n_j}{r_j} y_i = \frac{1}{n} \sum n_j \hat{\overline{Y}}_{r_j}$$

COMPLETE CASE (CC) ANALYSIS

- In the example, in case MAR, we used the auxiliary variable "class of employees" to compensate for the difference between respondents and nonrespondents in terms of the target variable (turnover).
- If the mean turnover has been estimated through the mean on the whole population instead of separately for different classes we would have obtained a biased estimate. For Instance, if enterprises with low value of turnover have lower propensity to response, we over-estimate the turnover, because the respondents have higher values of turnover.
- in general, analysis based only on units where all variables are observed is said Complete Cases (CC) analysis.

CC ANALYSIS



ADVANTAGES OF CC

- simple (standard methods and software)
- no artificial data (imputation)
- ok in case of small number of missing values and MCAR
- univariate estimates are comparable, in fact they are computed on the same cases.

DRAWBACKS OF CC

- if nonresponse in not MCAR estimates are biased
- even with MCAR, some information (incomplete cases) is not used —> loss of precision in the estimates

EXAMPLE(1)

Let (X, Y) be Gaussian r.v.s. Assume that we have a sample of size n where X is always observed, Y has $n - r$ missing values (MCAR). We wont to estimate the expected value μ_y of Y. Consider the estimator $\bar{Y}_r = \sum_{i=1}^{r} y_i$, the mean of Y over the r respondents.

If data were complete, we would use the estimator $\bar{Y}_n = \sum_{i=1}^{n} y_i$. The ratio between the variances of the two estimators is: n/r, so for instance, if the nonresponse rate is 50%, missing values cause the first variance become twice the second one.

EXAMPLE(2)

If instead of discarding the incomplete records, we use all the available information (i.e., also the X-values), we can obtain a more efficient estimator $\bar{Y}_r$. For instance the MLE estimator of μ_y is:

$$\hat{\mu}_{y_{ML}} = \hat{\mu_y} = \bar{Y}_r + \hat{\beta}_r(\bar{X}_n - \bar{X}_r)$$

where $\hat{\beta}_r$ is the estimate of the regression coefficient of Y on X. In this case we have:

$$Var(\hat{\mu}_{y_{ML}})/Var(\bar{Y}_r) \approx 1 - \frac{n-r}{n}\rho^2$$

CONCLUSION

If $\rho^2 \to 0$, then the two variances are equal, while if $\rho^2 \to 1$ the variance ratio goes to r/n, i.e., we have the same gain in efficiency as using the estimator $\bar{Y}_n$ instead of $\bar{Y}_r$

This example shows how using auxiliary information can remarkably improve the precision of the estimate if it is related to the target population parameter (in the example ρ close to 1).

BIAS WITH CC

Consider the population mean μ. We can express it in terms of the respondent mean μ_r and the non-respondent mean μ_{nr} as:

$$\mu = \pi_r \mu_r + (1 - \pi_r)\mu_{nr}$$

thus the bias is

$$\mu_r - \mu = (1 - \pi_r)(\mu_r - \mu_{nr})$$

where π_r is the respondent proportion. Thus, the estimate is unbiased only with MCAR ($\mu_r = \mu_{nr}$)

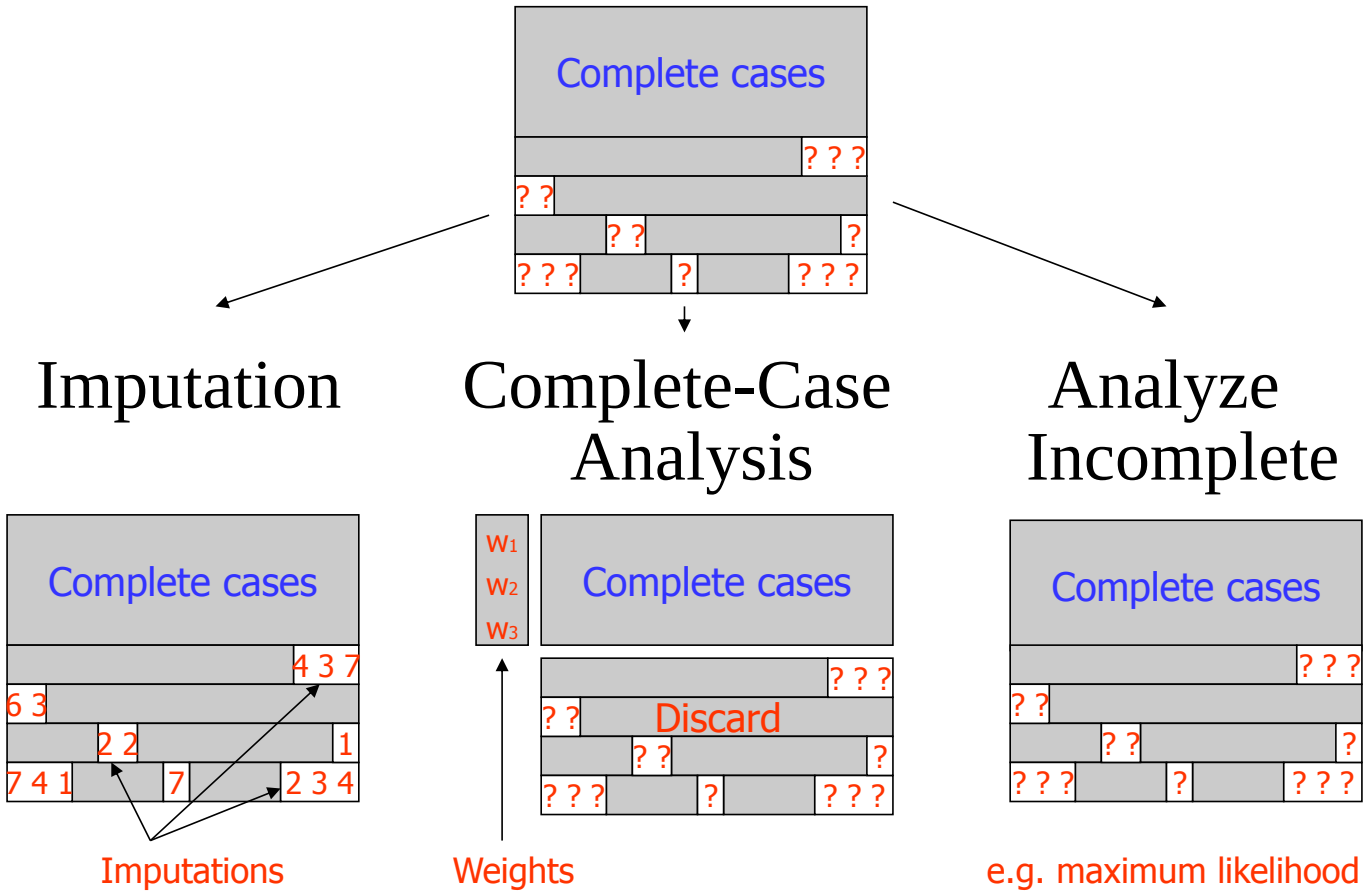
ALTERNATIVES

- analysis on "available cases" (AC)
- replacing missing values with artificial values (imputation)
- analysis using all data (complete and incomplete) (e.g. EM algorithm)

AVAILABLE CASES

- For each parameter of interest, one uses all the statistical units where the relevant variables are present. For example, we could use all the units where the variables X and Y are observed (but perhaps other variables are missing) to estimate the correlation coefficient ρ_{xy}.
- This method might produce inconsistent results: e.g., in each entry of the covariance matrix is estimated independently of other entries, the estimate could result in a non-positive definite matrix.

General Strategies



USING ALL DATA

this approach includes estimation methods for incomplete data and imputation.

- estimation with incomplete data
  - Under assumption of some parametric model it is sometimes possible to estimate (e.g. via ML) the parameters using also incomplete data. A popular technique is the EM algorithm.
- imputation
  - the dataset is "completed" to obtain a rectangular dataset ("without holes"), replacing missing values with plausible values (imputed).

MODELING

A parametric model has the advantages of being parsimonious of making the research assumptions explicit. On the other hand it is often difficult to specify a model that fits well data (e.g., zero inflation, non gaussian data, etc.). In the context of official

statistics one is usually interested in finite population quantities (e.g., totals or means), rather than in distribution parameters.

IMPUTATION

Missing values are imputed with "plausible" values so that standard methods and software can be applied to the completed dataset.

advantages: does not require special methodology for analysis.

drawbacks: an additional source of uncertainty is introduced. In the analysis phase, one should take into account the fact that imputed data are not really observed (incorporate in the estimation process the source of variability dure to nonresponse). Also, different imputation methods may be appropriate for different estimations objectives.

IMPUTATION METHODS

- **non parametric** (hot-deck, regression trees...)
- **parametric** (Regression, EM,...)
- **mixed** (Predictive Mean Matching,...)
- **semi-parametric** (mixture modeling,..)

SIMPLE METHODS (1)

- *mean imputation*. Missing values for each variable are imputed with the variable mean computed on the observed values. Often the population is partitioned into classes (*imputation cells*) and for each unit to be imputed means are computed within the class the unit belongs to. Within-cell imputation is similar to class-re-weighting within classes in sample theory.
- *regression-based imputation*. Imputed values are predictions from a regression model based on some set of covariates. Mean imputation is a particular case. One can add a random residual to the predictive mean.

SIMPLE METHODS (2)

- *hot-deck imputation*. The value to be imputed for an incomplete record is taken from a "similar" respondent in the same survey. This is a very common approach in the context of official Statistics. $\Longrightarrow$ flexibility but difficult to make assumptions explicit.
- *cold-deck imputation*. Similar to the previous method, but imputed values are taken from a different survey.

MEAN IMPUTATION(1)

May be appropriate if missing are MCAR (respondents are similar to non-respondents) and quantities to be estimated are means or totals (linear quantities). $\bar{y}_r$: mean over r respondents out of the n sampled units. with mean imputation:

$$\frac{1}{n}\left(\sum_{i}^{r} y_i + \sum_{r+1}^{n} \bar{y}_r\right) = \frac{1}{n}\left(r\bar{y}_r + (n-r)\bar{y}_r\right) = \bar{y}_r$$

since mechanism is MCAR $\bar{y}_r$ is an unbiased estimate of $\bar{y}_n$.

MEAN IMPUTATION(2)

If some non-linear parameter is to be estimated, e.g., the variance, mean imputation is not appropriate, in fact, using imputed data we would obtain:

$$s_r(n_r - 1)/(n - 1).$$

If s_r is a consistent estimate of the population variance, the estimate resulting from mean imputation is biased (under-estimated) by $(n_r - 1)/(n - 1)$. in fact, mean imputation makes distribution "flat".

CONDITIONAL MEAN IMPUTATION

Using auxiliary information can improve the precision of the estimates based on imputed data. If strongly predictive covariates are available one can impute with the conditional mean. Depending on whether auxiliary variables are categorical or numerical we obtain respectively:

- mean imputation within strata
- regression imputation without residual term

MEAN IMPUTATION WITHIN CELLS(1)

If population is divided into J cells (strata), and $\bar{y}_{jr}$ is the mean of variable Y over the r_j respondents in cell j, the estimate of the mean of Y is:

$$\bar{y}_{wc} = \frac{1}{n}\sum_{j=1}^{J}\left(\sum_{i=1}^{r_j} y_{ij} + \sum_{i=r_j+1}^{n_j} \bar{y}_{jr}\right) = \frac{1}{n}\sum_{j=1}^{J} n_j \bar{y}_{jr}$$

If (on average) $\bar{y}_{jr} = \bar{y}_j$ (MCAR) the estimator is unbiased. Note the similarity with stratification in sampling theory.

MEAN IMPUTATION WITHIN CELLS(2)

the gain deriving from imputing within cells is:

$$V(\bar{Y}) - V(\bar{Y}_{wc}) = n^2 \frac{1 - r/n}{r} \sum_{j=1}^{J} \frac{n_j}{n} (\bar{Y}_j - \bar{Y})^2$$

The difference is large if differences between the cell means and the overall mean is large.

IMPUTATION VIA REGRESSION WITHOUT RESIDUALS (1)

Missing are imputed with predictive means from (linear) regression.
Simple case, 2 variables (X, Y):

$$y_i = \hat{\alpha} + \hat{\beta} x_i$$

where $\hat{\alpha}$ and $\hat{\beta}$ are OLS estimates based on the units where both X and Y are observed.

In case where covariates are dummy variables identifying population strata, we have mean imputation within cells.

IMPUTATION VIA REGRESSION WITHOUT RESIDUALS (2)

Imputed values are (conditional) expected values, so although the resulting estimators of linear quantities are unbiased, the variance is under-estimated as in the case of mean imputation.
In fact, the regression variance is:

$$\sigma^2_{Y|X} = \sigma^2_Y - \frac{\sigma^2_{XY}}{\sigma^2_X}$$

in the imputed data Y the residual variance (non explained by the regression model) is not taken into account.

IMPUTATION VIA REGRESSION WITH RESIDUALS (1)

In order to preserve the data variability, one can add a residual term ϵ to the imputed (conditional) mean, e.g., drawing from a normal distribution with zero mean and variance equal to the residual regression variance estimated on complete data (both X and Y).

Estimates based on this imputed dataset are nearly consistent for all population parameter as far as the imputed values can be considered random draws from the predictive distribution of missing given observed data. In the following more complex missing patterns will be analyzed.

HOT-DECK (DONOR) IMPUTATION (1)

With *hot-deck* imputation values to be imputed for an incomplete record are taken from another unit (donor) where these values are available. Donor imputation includes random hot-deck and nearest neighbor donor. There are several versions: e.g., when many values are to be imputed they can all be taken from the same donor (*joint imputation*) or from different donors(*sequential imputation*).

HOT-DECK (DONOR) IMPUTATION (2)

Again, assume we have to estimate the mean of Y in a population with N units. Let us assume that from in simple random sample of n units there are only r respondents. If $r = n$ (no nonresponse)the (HT)estimator would be:

$$\bar{y}_n = \sum_{i=1}^{n} \frac{y_i}{n}$$

with variance:

$$Var(\bar{y}_n) = (n^{-1} - N^{-1})S_y^2$$

Now, we impute $n - r$ missing values via HD.

HOT-DECK (DONOR) IMPUTATION (3)

The mean on the imputed dataset can be expressed as:

$$\bar{y}_{HD} = \{r\bar{y}_r + (n-r)\bar{y}^*_{nr}\}/n$$

where $\bar{y}_r$ the mean on the observed values

$$\bar{y}^*_{nr} = \sum_{i=1}^{r} \frac{H_i y_i}{n-r}$$

and H_i is the number of times that y_i is selected as a donor $(\sum_{i=1}^{r} H_i = n-r)$.
The properties of the estimator depend on the procedure used to generate the numbers $\{H_1, \ldots, H_r\}$

HOT-DECK (DONOR) IMPUTATION (4)

Assume random hot-deck imputation (with repetition). In this case the random vector $\{H_1, \ldots, H_r\}$ can be considered as a multinomial r.v. with parameters $(n-r), (1/r, \ldots, 1/r)$ where $1/r$ is the selection probability for each unit. So the expected value of each H_j is $(n-r)/r$. It follows that the expected value of $\bar{y}_{HD}$, given the respondents is $\bar{y}_r$. We know that in case MCAR respondents can be considered as a SRS of the original sample. Thus the overall expectation (i.e., with respect to the sampling and nonresponse) is:

$$E(\bar{y}_{HD}) = \bar{Y}$$

that is the estimator is unbiased.

HOT-DECK (DONOR) IMPUTATION (5)

One can easily show that the variance is:

$$Var(\bar{y}_{HD}) = (r^{-1} - N^{-1})S_y^2 + (1 - r^{-1})(1 - r/n)S_y^2/n$$

Note that the first term is the variance of $\bar{Y}_r$ in case of SRS of size r. We conclude that imputation causes increase of the estimate variability.

ADVANTAGES

- Estimates concerning parameters of univariate distributions are approximately unbiased also for non-linear parameters, (e.g., the variance).
- hot-deck imputation is frequently used in practice in complex survey because it does not require model assumption (it is also applicable when data distribution is semi-continuous, e.g., zero-inflation)

DRAWBACKS

Hot-deck imputation causes decrease of associations among observed variables and missing variables, because observed values in the record to be imputed are not taken into account. One can alleviate the problem by using covariates:

- hot-deck within strata
- nearest neighbor donor (NND)

As far as the cell definition is concerned the problem is the same as in the case of mena imputation. Let us move on to NND.

NEAREST NEIGHBOR DONOR (NND)

Missing values in a incomplete record (recipient) are replaced with corresponding observed values in (one of) the most similar observation(s). Similarity is defined on the basis of a distance function that depends on suitably chosen covariates (matching variables). Common distance functions:

- l_p : $d_p(i,j) = \sum_k |x_{ik} - x_{jk}|^p$ ($p = 1$ Manhattan, $p = 2$ euclidean)
- max $d(i,j) = max_k|x_{ik} - x_{jk}|$
- Mahalanobis $d(i,j) = (x_i - x_j)^T S_{xx}^{-1}(x_i - x_j)$

REMARKS

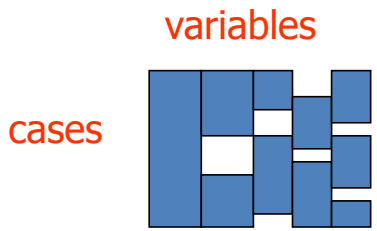
- NND is "asimptotic": large number of donors should be necessary (how many?)
- how to chose the matching variables?
- how to "weight"contributions from different variables to the distance function?
- variables should be standardized. How?
- in case of many variables to be imputed, joint imputation is preferable (rather than sequential)

MORE COMPLEX METHODS

Methods describes so far are applicable only with simple missing patterns (one variable with missing values and an "always observed variable"). Often data have irregular missing patterns, ("holes" randomly located). More general imputation methods are necessary.

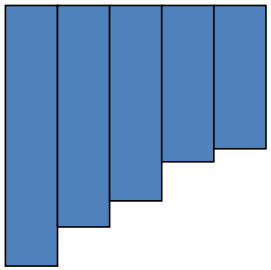
Patterns of Missing Data

- General Pattern

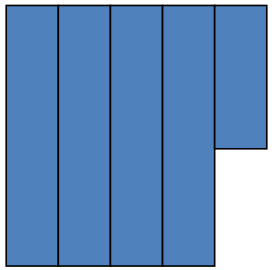


- Special Patterns

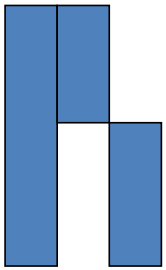
monotone



univariate



file matching



NND WITH GENERAL MISSING PATTERNS

Extending simple NND is simple: donor pool is composed of complete records, and the distance distance function is computed on the basis of the variables which are observed in the current recipient. For example, if we used the Euclidean distance with matching variables X_1, X_2, X_3, X_4 , and in the ith unit only variables X_1 and X_3 are observed, only these two variables will be used for computing the distance with respect to potential donor j:

$$d_2(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i3} - x_{j3})^2}$$

EXPLICIT MODELING (1)

We can use explicit parametric models in two ways:

- Regression: some variables are considered as *explanatory* and other as *responses*
- Joint distribution: what is modeled is the joint distribution of all variables simultaneously

In any case the objective is imputing missing values with plausible values derived from the assumed model.

EXPLICIT MODELING (2)

Methods based on regression are useful when the nonresponse pattern is univariate or when variables can be splitted into always observed variables ($\mathbf{X}$) and variables ($\mathbf{Y}$) which are never observed on a certain subset of sampled units. In this case multiple regression of $\mathbf{Y}$ on $\mathbf{X}$ can be used. For general patterns it is better to estimate joint distribution.

ESTIMATION OF JOINT DISTRIBUTION ON CC

in order to estimate parameters θ of the joint distribution of the data affected by missing values, we could use only complete records. However, this is not optimal for two reasons:

1. loss of precision (big variance)
2. risk of bias if nonresponse is not MCAR

ESTIMATION USING INCOMPLETE DATA

We would like to use all the available information, i.e., both complete and incomplete data. If for instance we use ML estimation, the involved probability distribution are:

$f(\mathbf{Y}|\theta)$: complete data distribution

$f(\mathbf{M}|\mathbf{Y},\psi)$: nonresponse ($\mathbf{M}$) distribution conditional on data

θ and ψ are distinct sets of parameters

MLE

We are interested in the estimates of parameters θ. ψ are "nuisance" parameters. Can we make inference on θ regardless of ψ-parameters (i.e., *ignoring* nonresponse mechanism? The answer is yes if the nonresponse mechanism is MAR:

$$f(\mathbf{M}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}; \psi) = f(\mathbf{M}|\mathbf{Y}_{obs}; \psi)$$

i.e., the nonresponse probability, ***conditional on observed data***, does not depend on the non-observed data.

ESTIMATION UNDER MAR

Using Bayes formula one can easily show that MAR hp is equivalent to assuming that the missing value distribution, conditional on observed data, is the same for respondents ($M = 0$) and non-respondents ($M = 1$):

$$f(\mathbf{Y}_{mis}|\mathbf{M}, \mathbf{Y}_{obs}) = \frac{f(\mathbf{M}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})}{f(\mathbf{M}|\mathbf{Y}_{obs})} = f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$$

OBSERVED DATA LIKELIHOOD (1)

In case MAR, the joint distribution of $\mathbf{Y}_{obs}$ and $\mathbf{M}$ can be factored as product of two functions depending on θ and ψ respectively:

$$f(\mathbf{Y}_{obs}, \mathbf{M}|\theta, \psi) = \int f(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}|\theta) f(\mathbf{M}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \psi) d\mathbf{Y}_{mis} =$$

$$= f(\mathbf{M}|\mathbf{Y}_{obs}, \psi) \int f(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}|\theta) d\mathbf{Y}_{mis} =$$

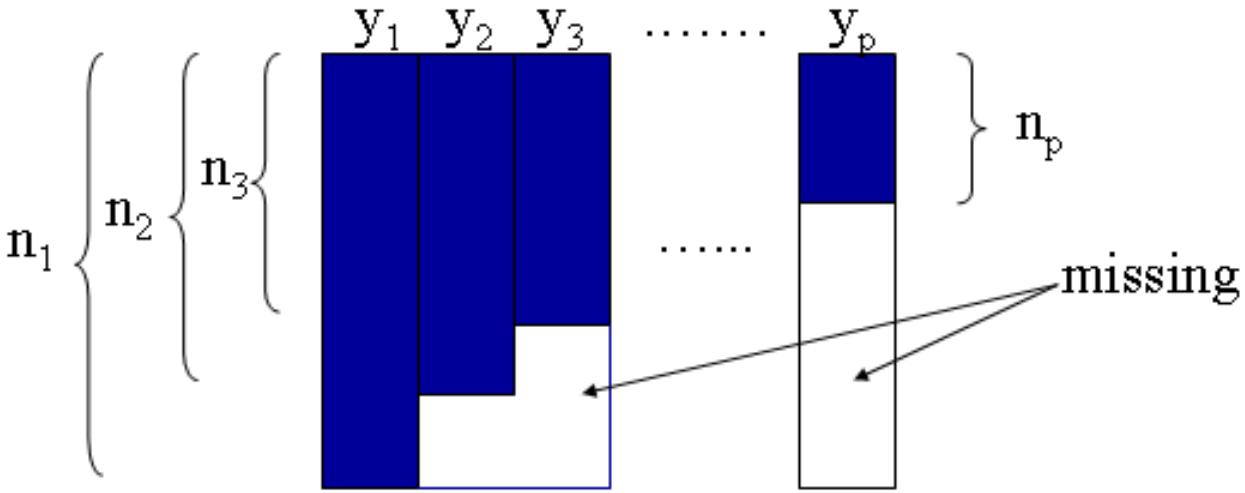
$$= f(\mathbf{M}|\mathbf{Y}_{obs}, \psi) f(\mathbf{Y}_{obs}|\theta) \propto L(\theta|\mathbf{Y}_{obs})$$

$L(\theta|\mathbf{Y}_{obs})$ *observed data likelihood*

OBSERVED DATA LIKELIHOOD (2)

Thus, MLEs can be computed by maximizing the observed likelihood $L(\theta|\mathbf{Y}_{obs})$. However, this is a complex function of the target parameters maximization is not easy. An important exception is when the missing pattern is "**monotone**", i.e. when variables $Y_1, \ldots, Y_p$ can be ordered so that for each unit, Y_k being observed implies Y_j is also observed $\forall j < k$. In fact, in this case the joint distribution can be factored via the "chain rule" as product of suitable conditional distributions whose parameters can be independently estimated.

Pattern monotono



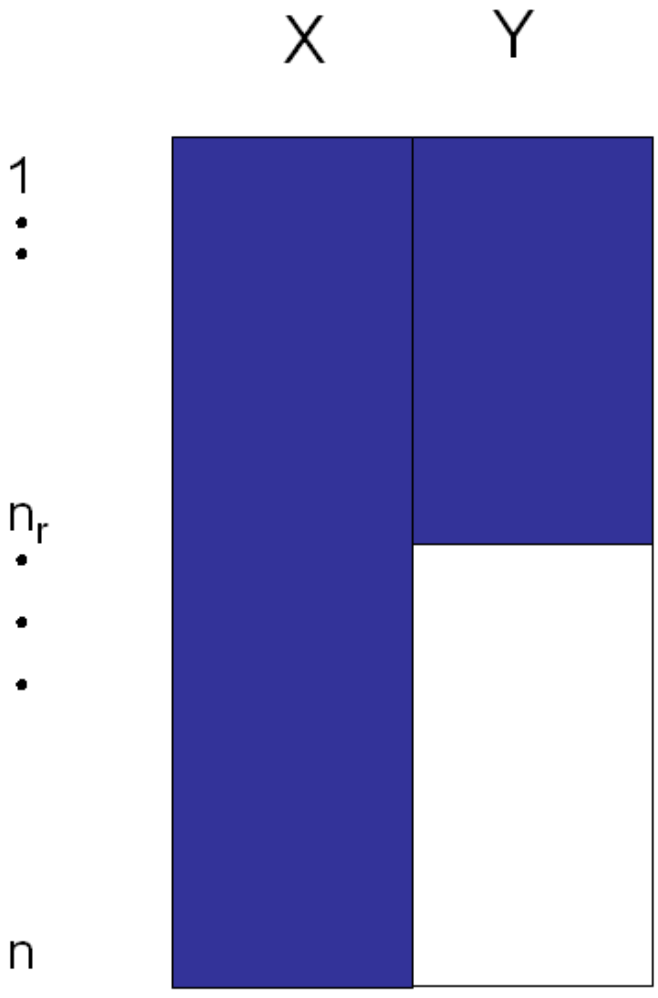
$$f(\mathbf{Y}_{obs}|\theta) = \prod_{i=1}^{n_1} f(y_{i1};\theta_1) \prod_{i=1}^{n_2} f(y_{i2}|y_{i1};\beta_{2.1}) \dots$$

$$\dots \prod_{i=1}^{n_p} f(y_{ip}|y_{i1}, y_{i2}, \dots, y_{i(p-1)};\beta_{p.1,\dots,p-1})$$

EXAMPLE: BIVARIATE NORMAL (1)

$\mathbf{V} = (X, Y)$ gaussian random vector. Distribution parameters: $\theta = (\mu_x, \sigma_x^2, \mu_y, \sigma_2^2, \sigma_{xy})$; X observed on n units; Y observed on $n_r < n$ units; We want to compute MLEs of the model parameters.

The task is easy because the pattern is monotone.



M_X	M_Y
0	0
0	0
	1
0	1

EXAMPLE: BIVARIATE NORMAL (2)

$$L(\theta|\mathbf{V}_{obs}) \propto |\Sigma|^{-n_r/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n_r}(\mathbf{v}_i-\mu)^t\Sigma^{-1}(\mathbf{v}_i-\mu)\right\} \times$$
$$\times\sigma_x^{(n-n_r)} \exp\left\{-\frac{1}{2\sigma_x^2}\sum_{i=n_r+1}^{n}(x_i-\mu_x)^2\right\}$$

re-parametrization: $\phi = (\mu_x, \sigma_x^2, \beta_0, \beta_1, \sigma_{y|x}^2)$

$$L(\theta|\mathbf{V}_{obs}) \propto \sigma_x^{-n} \exp\left\{-\frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(x_i-\mu_x)^2\right\} \times$$
$$\times\sigma_{y|x}^{-n_r} \exp\left\{-\frac{1}{2\sigma_{y|x}^2}\sum_{i=1}^{n_r}(y_i-\beta_0-\beta_1 x_i)^2\right\}$$

EXAMPLE: BIVARIATE NORMAL (3)

The two expressions are equivalent, but the second one can be more easily managed to obtain estimates of the parameters. In fact it is the product of the univariate lik. based on the units $1, \ldots, n$ and the lik. associated to the regression of Y on X based on the units $1, \ldots n_r$. Since parameters $\phi_1 = (\mu_x, \sigma_x^2)$ and $\phi_2 = (\beta_0, \beta_1, \sigma_{y|x}^2)$ are distinct, inferences can be independent. Also the first expression is the product of two likelihood functions, but the parameters in th etwo factors are not distinct (μ_x and σ_x^2 are in both factors).

EXAMPLE: BIVARIATE NORMAL (4)

MLEs can be easily derived: we use the n_r units with Y observed to estimate via OLS the parameters $\phi_2 = (\beta_0, \beta_1, \sigma^2_{y|x})$ and all the units to estimate $\phi_1 = (\mu_x, \sigma_x)$ $(\hat{\mu}_x = \bar{x}; \quad \hat{\sigma}^2_x = s^2_x;)$. Then, using re-parametrization formulas:

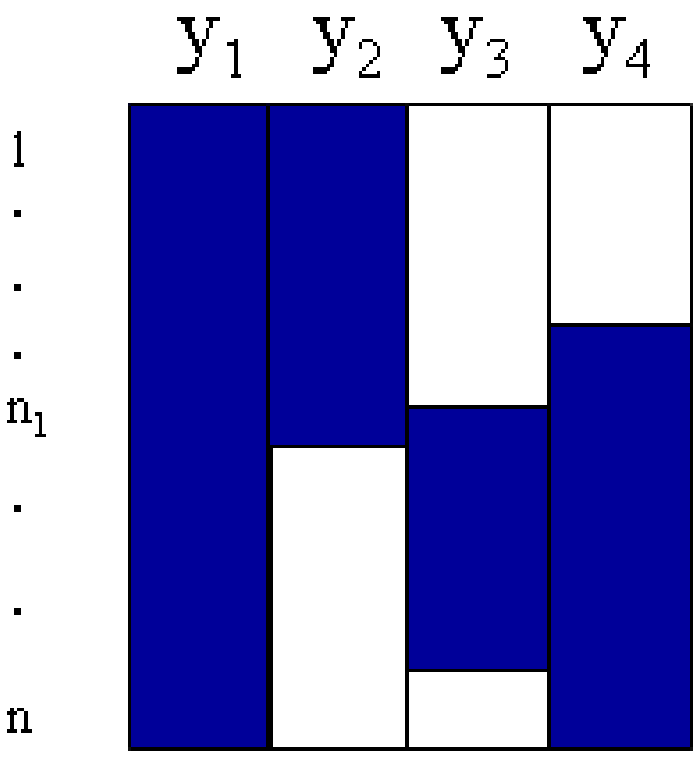
$\hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}_x; \qquad \hat{\sigma}^2_y = \hat{\sigma}^2_{y|x} + \hat{\beta}^2_1 \hat{\sigma}^2_x; \qquad \hat{\sigma}_{xy} = \hat{\beta}_1 \hat{\sigma}^2_x$

Replacing the appropriate estimators in the expression for $\hat{\mu}_y$ we obtain:

$\hat{\mu}_y = \bar{y}_r + \hat{\beta}_1 (\bar{x} - \bar{x}_r)$

where means are computed on respondents.

GENERAL PATTERNS



M

0	0	1	1
.	.	.	.
.	.	.	.
.	.		0
0	0	1	.
0	1	0	.
.	.	.	.
.	.	.	.
0	1	0	0
		1	

EM ALGORITHM

When analytic expression for MLEs are not available numerical routines have to be used. A popular method for maximization of the incomplete-data likelihood functions is the EM (Expectation-Maximization) algorithm. it allows one to obtain MLEs for incomplete data using (iteratively) standard techniques for complete data. Essentially, it is the formal version of the iterative procedure consisting of the iterative application of the following two steps:

1. replace missing values with values estimated (predicted) on the basis of the current estimates of the model parameters (E-step)
2. compute new estimates of parameters form the complete dataset imputed at the previous step (M-step)

EM

More exactly, the E-step at kth iteration is the computation of the expected value $Q(\theta|\theta^{(k)})$of the complete data log-likelihood $L(\theta|\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$. Expectation is with respect to the distribution of missing values conditional on observed data using the current parameters $\theta^{(k)}$:

$$Q(\theta|\theta^{(k)}) = \int l(\theta|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta^{(k)}) d\mathbf{Y}_{mis}$$

update of the estimates $\theta^{(k+1)}$ is obtained by maximizing $Q(\theta|\theta^{(k)})$ w.r.t. θ.
Under some regularity assumptions it can be shown that the sequence $\theta^{(n)}$ converges to the ML estimates of θ.

EXPONENTIAL FAMILY: E-STEP

For distribution of the exponential family, log-likeluhood based on n observations $\mathbf{y} = (y_1, \ldots, y_n)$ can be expressed as:
$l(\theta|\mathbf{Y}) = \eta(\theta)^t T(\mathbf{Y}) + ng(\theta) + c$
where $\eta(\theta) = (\eta_1(\theta)), \eta_2(\theta)), \ldots, \eta_s(\theta))^t$, is the natural parameter, $T(\mathbf{Y}) = (T_1(\mathbf{Y}), T_2(\mathbf{Y}), \ldots, T_s(\mathbf{Y}))^t$ is a s-dimensional vector of sufficient statistics and c is a constant.
Since $l(\theta|\mathbf{Y})$ is linear with respect to the sufficient statistics, the E-step reduces to replacing $T_j(\mathbf{Y})$ with $E(T_j(\mathbf{Y})|\mathbf{Y}_{obs}, \theta^{(t)})$. For the exponential family the resulting expression is simple.

EXPONENTIAL FAMILY: M-STEP

In the case of complete data, MLEs can be found as solution of the *moment equation*:

$$E(T(\mathbf{Y})|\theta) = t$$

where t is the realized value of $T(\mathbf{Y})$ the expected value is w.r.t. $f(\mathbf{Y}|\theta)$. In case of incomplete data the quantity on the r.h.s. has to be replaced by the output from the E-step. This allows explicit computation of M-step

EXAMPLE: UNIVARIATE GAUSSIAN (1)

Complete data ($N(\mu, \sigma^2)$):

$$T(\mathbf{Y}) = (T_1(\mathbf{Y}), T_2(\mathbf{Y}))^t = (\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} Y_i^2)^t$$

$E(T_1) = n\mu = t_1$, $E(T_2) = n\sigma^2 + n\mu^2 = t_2$ where (t_1, t_2) are realization from (T_1, T_2) :

$$(t_1, t_2) = (\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i^2)$$

$\hat{\mu} = \overline{y}$
$\hat{\sigma^2} = n^{-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$

EXAMPLE: UNIVARIATE GAUSSIAN (2)

Incomplete data: (n_1 respondents, $n_0 = n - n_1$ missing): E-step:

$$E(T_1|\mathbf{Y}_{obs},\theta) = E(\sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^{n} y_i) = \sum_{i=1}^{n_1} y_i + n_0\mu$$

$$E(T_2|\mathbf{Y}_{obs},\theta) = E(\sum_{i=1}^{n_1} y_i^2 + \sum_{i=n_1+1}^{n} y_i^2) = \sum_{i=1}^{n_1} y_i^2 + n_0(\sigma^2+\mu^2)$$

EXAMPLE: UNIVARIATE GAUSSIAN (3)

M-step:

$$\mu^{(t+1)} = n^{-1}\left[\sum_{i=1}^{n_1} y_i + n_0\mu^{(t)}\right]$$

$$\sigma^{2(t+1)} = n^{-1}\left[\sum_{i=1}^{n_1} y_i^2 + n_0\sigma^{2(t)} + n_0\mu^{2(t)}\right] - n^{-2}\left[\sum_{i=1}^{n_1} y_i + n_0\mu^{(t)}\right]^2$$

the above iteration converges to:

$\hat{\mu} = n_1^{-1}\sum_{i=1}^{n_1} y_i$
$\hat{\sigma}^2 = n_1^{-1}\sum_{i=1}^{n_1} y_i^2 - \hat{\mu}^2.$

PROBLEMS

- EM algorithm has to be initialized, i.e., it requires choosing (*starting points*)
- sometimes likelihood function has several local maxima. EM may converge to any of them (depending on the starting point)
- for some models (e.g. mixture of eteroshedastic Gaussian distributions) likelihood function is unbounded. In such cases EM may not converge.

IMPUTATION VIA EM(1)

For imputation purposes, EM is only an intermediate step used to estimate the distribution that generates data. In the next step, from the estimated joint distribution one has to derive the estimates of the relevant conditional distributions corresponding to various nonresponse patterns.

Example:

$$\mathbf{Y} = (Y_1, Y_2, Y_3) \sim N(\mu, \mathbf{\Sigma})$$

parameters:

$$(\mu_1, \mu_2, \mu_3 \; ; \sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{22}, \sigma_{23}, \sigma_{33})$$

IMPUTATION VIA EM(2)

For the pattern where only Y_2 is missing, the parameters of the distribution: $f(Y_2|Y_1, Y_3)$: $(\alpha_{2.13}, \beta_{2.13}, \sigma_{2.13})$ are to be estimated. These can be derived via standard formulas from the joint distribution parameters $(\mu, \mathbf{\Sigma})$:

$$\alpha_{2.13} = \mu_2 + \Sigma_{2(13)}\Sigma^{-1}_{(13)}\mu_{(13)}$$
$$\beta_{2.13} = \Sigma_{2(13)}\Sigma^{-1}_{(13)}$$
$$\sigma_{2.13} = \sigma_{22} - \Sigma_{2(13)}\Sigma^{-1}_{(13)}\Sigma_{(13)2}$$

where $\Sigma_{2(13)}$ is the row-vector $(\sigma_{21}, \sigma_{23})$, $\Sigma_{(13)}$ is the matrix (2×2) whose distinct elements are $\sigma_{11}, \sigma_{13}, \sigma_{33}$, $\mu_{(13)}$ is the column vector $(\mu_1, \mu_3)^t$ and $\Sigma_{(13)2} = \Sigma^t_{2(13)}$

ALTERNATIVE METHOD

More simply, one could use only complete data to estimate via standard linear all the relevant conditional distributions. However this is not optimal for two reason:

1. the parameters of the different conditional distributions are estimated separately rather than derived from the joint distribution parameters estimated only once
2. estimation process does not use all the available information

WITH OR WITHOUT RESIDUALS?

Once we have estimated the distribution $f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$ two imputation methods are possible:

1. using conditional means $E(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$
2. adding a random disturbance, i.e., obtaining values to be imputed by drawing from $f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$.

The first method produces more accurate estimates of linear quantities such as means and totals. However, if we wont to preserve the distributional characteristics, the second method is preferable.

MODELS

The model most commonly used for multivariate numerical data is the Gaussian model. In fact for a lot of method and software are for incomplete normal data are available.

However, often normal assumption is not realistic even after some preliminary transformation of data (e.g. logarithms, Box-Cox, etc.)

In these cases some other approach may be more appropriate.

PREDICTIVE MEAN MATCHING

Predictive Mean Matching (PMM) is more robust with respect to departure from normality assumption. With PMM, model is used only at an intermediate stage for computing expectations of missing given observed data. However, conditional means are not directly used for imputation. Instead, a suitable distance function is defined in terms of conditional means and used in NND imputation.

The function used for NND is not a proper distance function as a function of the observed variates. The method depends to some extent on the model and does not ensures the asimptotic properties of the NND imputation.

PMM can be useful when observations are not enough for NND being applied, and at the same time a full parametric approach is difficult (it is difficult to find an explicit model fitting adequately data)

PMM: ONE RESPONSE VARIABLE

In case of one response variable Y affected by nonresponse and p covariates $(X_1, \ldots, X_p)$ without missing values, PMM reduces to:

- for each unit u_i determine the conditional expectation $y_i^* = E(Y|x_{i1}, x_{i2}, \ldots, x_{ip})$ based on the estimated parameters of the regression of Y on $X_1, X_2, \ldots, X_p$
- for each u_i with Y missing, impute the value y_i taken from the donor u_j, whose predictive mean y_j^* is closest to y_i^*

If $p = 1$, PMM is the same as NND.

PMM: GENERAL CASE

In the general case of arbitrary response pattern, PMM consists of the following steps:

- estimate the joint distribution parameters via EM algorithm
- for each incomplete record u_i, compute the conditional mean $y^* = E(\mathbf{Y}_{mis,i}|\mathbf{y}_{obs,i})$
- compute the same conditional mean for all complete records (donors)
- impute u_i via NND using the Mahalanobis metrics based on the residual covariance matrix from the regression of $\mathbf{Y}_{mis}$ on $\mathbf{Y}_{obs}$.

TWO IMPORTANT REFERENCES

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data* 2nd edition, Wiley.

Shafer, J.L. (1997). *Analysis of Incomplete Multivariate Data* New York: CRC Press.