Lecture 5
Incomplete data

Ziad Taib
Biostatistics, AZ

May 8, 2009

Outline of the problem

- Missing values in longitudinal trials are a big issue
  - First aim should be to reduce proportion
  - Ethics dictate that it can't be avoided
  - Information lost can't be conjured up
  - There is no magic method to fix it
- Magnitude of problem varies across areas
  - 8-week depression trial: 25%−50% may drop out by final visit
  - 12-week asthma trial: maybe only 5%−10%

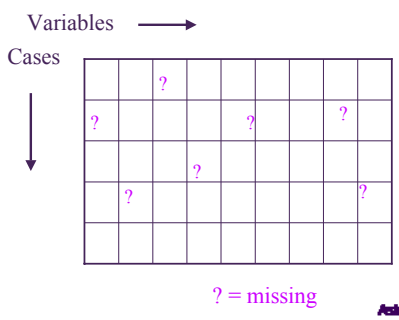
Outline of the lecture

Part I: Missing data

Part II: Multiple imputation

Part I: Missing data

1. Introduction to missing data



What is missing data?

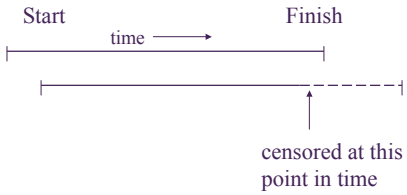
- The missingness hides a real value that is useful for analysis purposes.

Survey questions:

1. What is your total annual income for FY 2008?
2. Who are you voting for in the 2009 election for the European parlament???

What is missing data?

Clinical trials:



Missingness

- *It matters why data are missing*. Suppose you are modelling weight (Y) as a function of sex (X). Some respondents wouldn't disclose their weight, so you are missing some values for Y. There are three possible mechanisms for the nondisclosure:

1. There may be no particular reason why some respondents told you their weights and others didn't. That is, the probability that Y is missing may has no relationship to X or Y. In this case our data is *missing completely at random*
2. One sex may be less likely to disclose its weight. That is, the probability that Y is missing depends only on the value of X. Such data are *missing at random*
3. Heavy (or light) people may be less likely to disclose their weight. That is, the probability that Y is missing depends on the unobserved value of Y itself. Such data are *not* missing at random

Example: The analgesic trial

- single-arm trial with 530 patients recruited (491 selected for analysis)
- analgesic treatment for pain caused by chronic nonmalignant disease
- treatment was to be administered for 12 months
- we will focus on Global Satisfaction Assessment (GSA)
- GSA scale goes from 1=very good to 5=very bad
- GSA was rated by each subject 4 times during the trial, at months 3, 6, 9, and 12.

- Research questions:
  - Evolution over time
  - Relation with baseline covariates: age, sex, duration of the pain, type of pain, disease progression, Pain Control Assessment (PCA), ...
  - Investigation of dropout

GSA	Month 3		Month 6		Month 9		Month 12	
1	55	14.3%	38	12.6%	40	17.6%	30	13.5%
2	112	29.1%	84	27.8%	67	29.5%	66	29.6%
3	151	39.2%	115	38.1%	76	33.5%	97	43.5%
4	52	13.5%	51	16.9%	33	14.5%	27	12.1%
5	15	3.9%	14	4.6%	11	4.9%	3	1.4%
Tot	385		302		227		223	

Missing data patterns & mechanisms

- **Pattern:** Which values are missing?
- **Mechanism:** Is missingness related to the response?

(Y_i, R_i) = Data matrix, with COMPLETE DATA
R_{ij} = Missing data indicator matrix

$$R_{ij} = \begin{cases} 1, & Y_{ij} \text{ missing} \\ 0, & Y_{ij} \text{ observed} \end{cases}$$

Y_i^0 = Observed part of Y
Y_i^m = Missing part of Y

Missing data patterns & mechanisms

"Pattern" concerns the distribution of R
"Mechanism" concerns the distribution of R given Y
Rubin (Biometrika 1976) distinguishes between:

- Missing Completely at Random (MCAR)

  $P(R|Y) = P(R)$ for all Y
- Missing at Random (MAR)

  $P(R|Y) = P(R|Y^0)$ for all Y^m
- Not Missing at Random (NMAR)

  $P(R|Y)$ depends on Y^m

Missing At Random (MAR)

- What are the most general conditions under which a valid analysis can be done using only the observed data, and no information about the missingness value mechanism.

$$P(R/Y^o, Y^m)$$

- The answer to this is when, *given the observed data, the missingness mechanism does not depend on the unobserved data*. Mathematically,

$$P(R/Y^o, Y^m) = P(R/Y^o)$$

- This is termed ***Missing At Random***, and is equivalent to saying that the behaviour of two units who share *observed values* have the same statistical behaviour on the other observations, whether observed or not.

Example

	Variables					
Unit	1	2	3	4	5	6
1	1	3	4.3	3.5	1	4.6
2	1	3	?	3.5	?	?

- As units 1 and 2 have the same values where both are observed, given these observed values, under MAR, variables 3, 5 and 6 from unit 2 have the same distribution (NB not the same value!) as variables 3, 5 and 6 from unit 1.
- Note that under MAR the probability of a value being missing will generally depend on observed values, so it does not correspond to the intuitive notion of 'random'. The important idea is that the missing value mechanism can expressed solely in terms of *observations that are observed.*
- Unfortunately, this can rarely be definitively determined from the data at hand!

- *If data are MCAR or MAR*, you can ignore the missing data mechanism and use multiple imputation and maximum likelihood.
- *If data are NMAR*, you can't ignore the missing data mechanism; two approaches to NMAR data are **selection models** and **pattern mixture**.

- Suppose Y is weight in pounds; if someone has a heavy weight, they may be less inclined to report it. So the value of Y affects whether Y is missing; the data are NMAR. Two possible approaches for such data are selection models and pattern mixture.
- ***Selection models***. In a selection model, you simultaneously model Y and the probability that Y is missing. Unfortunately, a number of practical difficulties are often encountered in estimating selection models.
- ***Pattern mixture*** (Rubin 1987). When data is NMAR, an alternative to selection models is multiple imputation with pattern mixture. In this approach, you perform multiple imputations under a variety of assumptions about the missing data mechanism. In ordinary multiple imputation, you assume that those people who report their weights are similar to those who don't. In a pattern-mixture model, you may assume that people who don't report their weights are an average of 20 pounds heavier. This is of course an arbitrary assumption; the idea of pattern mixture is to try out a variety of plausible assumptions and see how much they affect your results. Pattern mixture is a more natural, flexible, and interpretable approach.

Simple analysis strategies

(1) Complete Case (CC) analysis

Advantages:

Easy

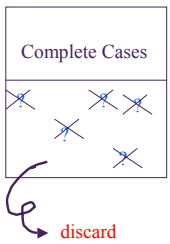
Does not invent data

Disadvantages:

Inefficient

Discarding data is bad

CC are often biased samples



Analysis strategies

(2) Analyze as incomplete (summary measures, GEE, …)

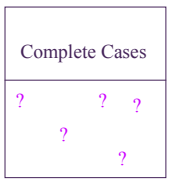
Advantages:

Does not invent data

Disadvantages

Restricted in what you can infer

Maximum likelihood methods may be computationally intensive or not feasible for certain types of models.



Analysis strategies

(3) Analysis after single imputation

Advantages:

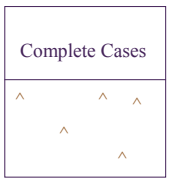
Rectangular file

Good for multiple users

Disadvantages:

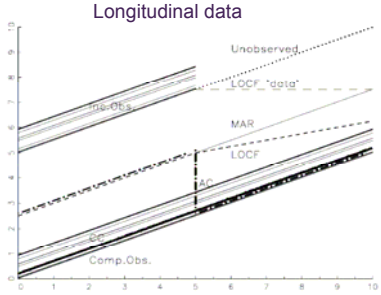
Naïve imputations not good

Invents data- inference is distorted by treating imputations as the truth



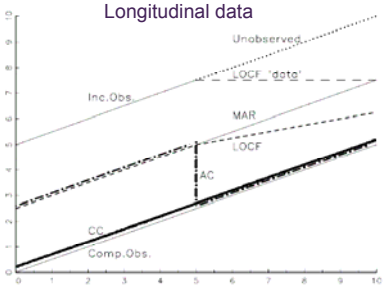
Data and modelling strategies

Longitudinal data



Modelling strategies

Longitudinal data



Simple methods of analysis of incomplete data

MCAR

cc	locf
Complete case analysis:	Last observation carried forward:
⇒ **delete** incomplete subjects	⇒ **impute** missing values
• Standard statistical software	• Standard statistical software
• Loss of information	• Increase of information
• Impact on precision and power	• Constant profile after dropout: unrealistic
• Missingness ≠ MCAR ⇒ bias	• Usually bias

Various strategies

MCAR	→ MAR	→ MNAR
CC ?	direct likelihood !	joint model !?
LOCF ?	expectation-maximization (EM).	sensitivity analysis ?!
imputation ?	multiple imputation (MI).	
⋮	(weighted) GEE !	

Notation

- Subject i at occasion (time) $j = 1, \ldots, n_i$
- **Measurement** Y_{ij}
- **Missingness indicator** $R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$
- Group Y_{ij} into a vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})' = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$

  $\boldsymbol{Y}_i^o$ contains Y_{ij} for which $R_{ij} = 1$,

  $\boldsymbol{Y}_i^m$ contains Y_{ij} for which $R_{ij} = 0$.
- Group R_{ij} into a vector $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{in_i})'$
- D_i: time of dropout: $D_i = 1 + \Sigma_{j=1}^{n_i} R_{ij}$

Ignorability

- Let us decide to use likelihood based estimation.
- The full data likelihood contribution for subject i:

$$L^*(\boldsymbol{\theta}, \boldsymbol{\psi} | \boldsymbol{Y}_i, D_i) \propto f(\boldsymbol{Y}_i, D_i | \boldsymbol{\theta}, \boldsymbol{\psi}).$$

- Base inference on the observed data:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi} | \boldsymbol{Y}_i, D_i) \propto f(\boldsymbol{Y}_i^o, D_i | \boldsymbol{\theta}, \boldsymbol{\psi})$$

with

$$\begin{aligned} f(\boldsymbol{Y}_i^o, D_i | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{Y}_i, D_i | \boldsymbol{\theta}, \boldsymbol{\psi}) d\boldsymbol{Y}_i^m \\ &= \int f(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m | \boldsymbol{\theta}) f(D_i | \boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, \boldsymbol{\psi}) d\boldsymbol{Y}_i^m. \end{aligned}$$

In a likelihood setting the term *ignorable* is often used to refer to MAR mechanism. It is the mechanism which is ignorable - not the missing data!

- Under a MAR process:

$$\begin{aligned} f(\boldsymbol{Y}_i^o, D_i | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m | \boldsymbol{\theta}) f(D_i | \boldsymbol{Y}_i^o, \boldsymbol{\psi}) d\boldsymbol{Y}_i^m \\ &= f(\boldsymbol{Y}_i^o | \boldsymbol{\theta}) f(D_i | \boldsymbol{Y}_i^o, \boldsymbol{\psi}). \end{aligned}$$

- The likelihood factorizes into two components.

Direct likelihood maximisation

MAR: $f(\boldsymbol{Y}_i^o | \boldsymbol{\theta}) \cancel{f(D_i | \boldsymbol{Y}_i^o, \boldsymbol{\psi})}$

Mechanism is MAR
$\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ distinct
Interest in $\boldsymbol{\theta}$
Use observed information matrix
$\Longrightarrow$ Likelihood inference is valid

Outcome type	Modeling strategy	Software
Gaussian	Linear mixed model	SAS proc MIXED
Non-Gaussian	Generalized linear mixed model	SAS proc GLIMMIX, NLMIXED

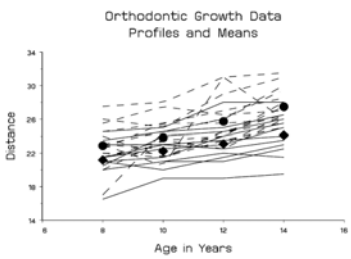
Example 1: Growth data

- Taken from Potthoff and Roy, Biometrika (1964)
- Research question:

> Is dental growth related to gender ?

- The distance from the center of the pituitary to the maxillary fissure was recorded at ages 8, 10, 12, and 14, for 11 girls and 16 boys

- Individual profiles:
  - ▷ Much variability between girls / boys
  - ▷ Considerable variability within girls / boys
  - ▷ Fixed number of measurements per subject
  - ▷ Measurements taken at fixed time points



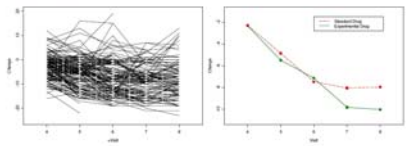
Growth data

- Data
  - ▷ Complete cases
  - ▷ LOCF imputed data
  - ▷ All available data
- Analysis methods
  - ▷ Direct likelihood
    - ◦ ML
    - ◦ REML
  - ▷ MANOVA
  - ▷ ANOVA per time point
- Model
  - ▷ Unstructured group by time mean
  - ▷ Unstructured covariance matrix

Principle	Method	Boys at Age 8	Boys at Age 10
Original	*Direct likelihood, ML*	*22.88 (0.56)*	*23.81 (0.49)*
	Direct likelihood, REML = MANOVA	*22.88 (0.58)*	*23.81 (0.51)*
	ANOVA per time point	*22.88 (0.61)*	*23.81 (0.53)*
Direct Lik.	Direct likelihood, ML	22.88 (0.56)	23.17 (0.68)
	Direct likelihood, REML	22.88 (0.58)	23.17 (0.71)
	MANOVA	24.00 (0.48)	24.14 (0.66)
	ANOVA per time point	22.88 (0.61)	24.14 (0.74)
CC	Direct likelihood, ML	24.00 (0.45)	24.14 (0.62)
	Direct likelihood, REML = MANOVA	24.00 (0.48)	24.14 (0.66)
	ANOVA per time point	24.00 (0.51)	24.14 (0.74)
LOCF	Direct likelihood, ML	22.88 (0.56)	22.97 (0.65)
	Direct likelihood, REML = MANOVA	22.88 (0.58)	22.97 (0.68)
	ANOVA per time point	22.88 (0.61)	22.97 (0.72)

Example: The depression trial

- Clinical trial: experimental drug *versus* standard drug
- 170 patients
- Response: change versus baseline in $HAMD_{17}$ score
- 5 post-baseline visits: 4–8



Patients are evaluated both pretreatment and posttreatment with the 17-item Hamilton Rating Scale for Depression (Ham-D-17),

The depression trial

- Complete case analysis:

```
%cc(data=depression, id=patient, time=visit, response=change, out=cc);
```

⇒ performs analysis on CC data set

- LOCF analysis:

```
%locf(data=depression, id=patient, time=visit, response=change, out=locf);
```

⇒ performs analysis on LOCF data

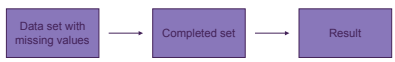
- Direct-likelihood analysis: ⇒ fit linear mixed model to incomplete data

- Treatment effect at visit 8 (last follow-up measurement):

Method	Estimate	(s.e.)	p-value
CC	-1.94	(1.17)	0.0995
LOCF	-1.63	(1.08)	0.1322
MAR	-2.38	(1.16)	0.0419

Observe the slightly significant p-value under the MAR model

5. Part II: Multiple imputation



Single Imputation

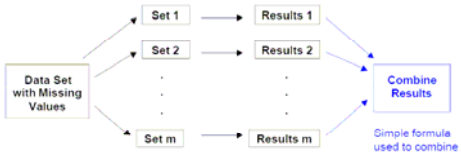
Substitute a value for each missing value. Some of the ways to choose this value:

- Mean Estimation
  - Replace missing data with the mean of non-missing values.
  - Standard deviation and standard errors are underestimated (no variation in the imputed values).
- Hot-deck Imputation
  - Stratify and sort by key covariates, replace missing data from another record in the same strata.
  - Underestimation of standard errors can be a problem.
- Predict missing values from Regression
  - Impute each independent variable on the basis of other independent variables in model.
  - Produces biased estimates.

Disadvantage:
In general, Single Imputation results in the sample size being over-estimated with the variance and standard errors being underestimated.

Multiple Imputation

- Requires Missing At Random (MAR) or Missing Completely At Random (MCAR) Assumption.
- Combine results from repeated single imputations.



Impute a complete data set:

- Fill in missing values with a value "drawn" from the distribution of your data.
- Each drawn value will be different for the missing item.

Analyze the imputed data set.

General principles

- Valid under MAR
- An alternative to direct likelihood ~~and WGEE~~
- Three steps:
  1. The missing values are filled in M times $\Longrightarrow$ M complete data sets
  2. The M complete data sets are analyzed by using standard procedures
  3. The results from the M analyses are combined into a single inference
- Rubin (1987), Rubin and Schenker (1986), Little and Rubin (1987)

Informal justification

- We need to estimate $\boldsymbol{\theta}$ from the data (e.g., from the complete cases)
- Plug in the estimated $\hat{\boldsymbol{\theta}}$ and use

$$f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\hat{\boldsymbol{\theta}})$$

to impute the missing data.

- We need to acknowledge that $\hat{\boldsymbol{\theta}}$ is a random variable; its uncertainty needs to be included in the imputation process
- Given this distribution we:
  - ▷ draw a random $\boldsymbol{\theta}^*$ from the distribution of $\hat{\boldsymbol{\theta}}$
  - ▷ put this $\boldsymbol{\theta}^*$ in to draw a random $\boldsymbol{Y}_i^m$ from

$$f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{\theta}^*).$$

The algorithm

1. Draw $\boldsymbol{\theta}^*$ from its posterior distribution
2. Draw $\boldsymbol{Y}_i^{m*}$ from $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{\theta}^*)$.
3. To estimate $\boldsymbol{\beta}$, then calculate the estimate of the parameter of interest, and it estimated variance, using the completed data, $(\boldsymbol{Y}^o,\boldsymbol{Y}^{m*})$:

$$\boldsymbol{\beta}=\boldsymbol{\beta}(\boldsymbol{Y})=\boldsymbol{\beta}(\boldsymbol{Y}^o,\boldsymbol{Y}^{m*})$$

The *within* imputation variance is

$$U=\hat{\mathrm{Var}}(\boldsymbol{\beta})$$

4. Repeat steps 1, 2 and 3 a number of M times

$$\Rightarrow \quad \hat{\boldsymbol{\beta}}^m \quad \& \quad U^m \qquad (m=1,\ldots,M)$$

Pooling information

- With M imputations, the estimate of $\boldsymbol{\beta}$ is

$$\bar{\boldsymbol{\beta}}^*=\frac{\sum_{m=1}^{M}\hat{\boldsymbol{\beta}}^m}{M}.$$

- Further, one can make normally based inferences for $\boldsymbol{\beta}$ with

$$(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}^*)\sim N(\mathbf{0},V),$$

where

$$V=W+\left(\frac{M+1}{M}\right)B$$

Hypothesis testing

- Two "sample sizes":
  - ▷ N: The sample size of the data set
  - ▷ M: The number of imputations
- Both play a role in the asymptotic distribution (Li, Raghunathan, and Rubin 1991)

$$H_0:\boldsymbol{\theta}=\boldsymbol{\theta}_0$$

$$\downarrow$$

$$p=P(F_{k,w}>F)$$

k : length of the parameter vector θ

$$F_{k,w} \sim F$$

$$F = \frac{(\theta^* - \theta_0)' W^{-1} (\theta^* - \theta_0)}{k(1+r)}$$

$$w = 4 + (\tau - 4)\left[1 + \frac{(1-2\tau^{-1})}{r}\right]^2$$

$$r = \frac{1}{k}\left(1 + \frac{1}{M}\right)\mathrm{tr}(BW^{-1})$$

$$\tau = k(M-1)$$

- Limiting behavior:

$$F \xrightarrow{M\to\infty} F_{k,\infty} = \chi^2/k$$

MI in practice

- Many analyses of the same incomplete set of data
- A combination of missing outcomes and missing covariates
- MI in SAS:

Imputation Task: PROC MI

↓

Analysis Task: PROC "MYFAVORITE"

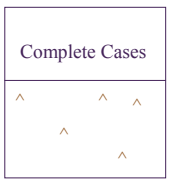
↓

Inference Task: PROC MIANALYZE

MI in practice

A simulation-based approach to missing data

1. Generate $M > 1$ plausible versions of Y_i^m.
2. Analyze each of the M datasets by standard complete-data methods.
3. Combine the results across the M datasets (M =3-5 is usually OK).



^ = imputation for M^{th} dataset

MI in practice... Step 1

Generate $M > 1$ plausible versions of Y_i^m via software, i.e. obtain M different datasets.

- An assumption we make: the data are MCAR or MAR, i.e. the missing data mechanism is ignorable.
- Should use as much information is available in order to achieve the best imputation.
- If the percentage of missing data is high, we need to increase M.

How many datasets to create?

The efficiency of an estimator based on M imputations is $(1+ \gamma/ M)^{-1}$, where γ is the fraction of missing information.

Efficiency of multiple imputation (%)

	γ				
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

MI in practice... Step 2

Analyze each of the M datasets by standard complete-data methods.

- Let β be the parameter of interest.
- $\hat{\beta}_j$ is the estimate of β from the complete-data analysis of the j^{th} dataset. ($j = 1 \ldots M$)
- U_j is the variance of $\hat{\beta}_j$ from the analysis of the j^{th} dataset.

MI in practice... Step 3

Combine the results across the M datasets.

- $\bar{\beta} = M^{-1} \Sigma_j \hat{\beta}_j$ is the combined inference for β.
- Variance for $\bar{\beta}$ is $M^{-1} \Sigma U_j + (1 + M^{-1})\Theta$.
- $\Theta = 1/(M-1) \Sigma (\beta_j - \bar{\beta})^2$
- $(1 + M^{-1})\Theta$ incorporates added uncertainty from imputation.

within dataset variance

between dataset variance

Software

2. SAS software (experimental)
It is part of SAS/STAT version 8.02

SAS institute paper on multiple imputation, gives an example and SAS code:
http://www.sas.com/rnd/app/papers/multipleimputation.pdf

SAS documentation on PROC MI
http://www.sas.com/rnd/app/papers/miv802.pdf

SAS documentation on PROC MIANALYZE
http://www.sas.com/rnd/app/papers/mianalyzev802.pdf

Software

1. Joe Schafer's software from his web site. ($0)
http://www.stat.psu.edu/%7Ejls/misoftwa.html#top

Schafer has written publicly available software primarily for S-plus. There is a stand-alone Windows package for data that is multivariate normal.

This web site contains much useful information regarding multiple imputation.

Software

3. SOLAS version 3.0 ($1K)
http://www.statsol.ie/solas/solas.htm

Windows based software that performs different types of imputation:

- Hot-deck imputation
- Predictive OLS/discriminant regression
- Nonparametric based on propensity scores
- Last value carried forward

Will also combine parameter results across the M analyses.

MI Analysis of the Orthodontic Growth Data

- The same Model 1 as before
- Focus on boys at ages 8 and 10
- Results

	Boys at Age 8	Boys at Age 10
Original Data	22.88	23.81
Multiple Imputation	22.88	22.69

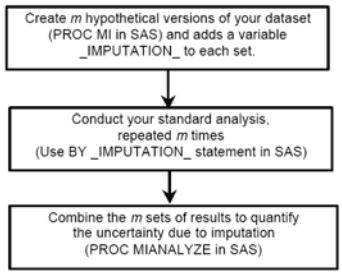
- Confidence interval for Boys at age 10: [21.08,24.29]

SAS PROCEDURES FOR DOING MI

1. PROC MI creates Multiple Imputed data sets.
2. PROC MIANALYZE combines results after analysis.

- SAS released these two new procedures in experimental form in Release 8.1 of the SAS system.
- Implement's Rubin's method for multiple imputation.
- Documentation for PROC MI and PROC MIANALYZE is located at: http://statweb.unc.edu/onldoc.htm

Flowchart for MI Using SAS Procedures



IMPORTANT PROC MI OPTIONS AND STATEMENTS:

PROC MI

Specify data sets:

DATA = *input data set*. *Contains all variables in the analysis set & other related variables.*

OUTPUT = *output data set.* *This will be used for standard analysis.*

Specify imputation options:

NIMPU = *number of imputations.* *This is the number of data sets that will be created.*

SEED = *number to start random sequence*

MAXIMUM = *maximum value of each variable in imputation model*

MINIMUM = *minimum value of each variable*

} *Sets plausible range of values.*

ROUND = *units to round variables in imputation;* *Use mainly for categorical outcome variables.*

VAR *variables in imputation model;* *Include variables in analysts' model & other related variables.*

BY *variables;* *Useful for imputing for different sub-populations, such as males and females.*

IMPORTANT PROC MIANALYZE OPTIONS AND STATEMENTS.:

PROC MIANALYZE

Specify input data sets using statements from only one of the following three lines:

DATA = *COV, CORR, EST type data set output from SAS analysis procedure*

PARMS = *matrix of parameter estimates*, COVB = *Covariance matrices*

PARMS = *matrix of parameter estimates*, XPXI = $(X'X)^{-1}$ *matrices*

Specify statistical analysis with these statements:

EDF = *complete data degrees of freedom from analysis*

ALPHA = *level used to construct confidence-limits;*

VAR *covariates in model;*

REVIEW OF STEPS FOR MULTIPLE IMPUTATION:

1. Determine if your data meets the multivariate normal and MAR or MCAR missing data assumptions.
2. Determine the set of variables to be used in the imputation model.
3. Use PROC MI to create several imputed data sets.
4. Check convergence of the imputation process.
5. Analyze each imputed data set with technique of your choice.
6. Use PROC MIANALYZE to combine the results.

You are now ready to interpret your results. Be sure to include information about the imputation model and the percent of the data that needed to be imputed when you report the findings.

Overview

- Ignore drop-out
  - CC (complete-case analysis)
- Single imputation of missing values
  - LOCF (last observation carried forward)
- Generate small samples from estimated distributions
  - MI (multiple imputation)
- Fit model for response at all time-points
  - GEE (generalized estimating equations)
  - (MNLM (multivariate normal linear model; also referred to as MMRM, or mixed-model repeated measures))
- Model drop-out as well as response
  - SM (selection models)
  - PMM (pattern-mixture models)

Properties of methods

- MCAR: drop-out independent of response
  - CC is valid, though it ignores information
  - LOCF is valid if there are no trends with time
- MAR: drop-out depends only on observations
  - CC, LOCF, GEE invalid
  - MI, MNLM, weighted GEE valid
- MNAR: drop-out depends also on unobserved
  - CC, LOCF, GEE, MI, MNLM invalid
  - SM, PMM valid if (uncheckable) assumptions true

References

- Allison, P. (2002). *Missing data*. Thousand Oaks, CA: Sage [greenback].
- Horton, NJ & Lipsitz, SR. (2001) Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician* **55**(3): 244-254.
- Little, R.J.A. (1992) Regression with missing X's: A review. *Journal of the American Statistical Association* **87**(420):1227-1237.
- Roderick J. A. Little and Donald B. Rubin (2002) Statistical Analysis with Missing Data, 2nd edition April 2002, Applications of Modern Missing Data Methods, by Roderick J. A. Little.
- by Joseph L. Schafer Joe Schafer's (1997) Analysis of Incomplete Multivariate Data, web site: http://www.stat.psu.edu/%7Ejls.
- Anderson, T.W. (1956) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing.

Further References

- Little, RL & Rubin, DB. (1st ed. 1990, 2nd ed. 2002). *Statistical analysis with missing data*. New York: Wiley.
- Rubin, DB. (1987). *Multiple imputation for survey nonresponse*. New York: Wiley.
- Mallinckrodt et al. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry* 53, 754–760.
- Gueorguieva & Krystal (2004) Move Over ANOVA. *Archives of General Psychiatry* 61, 310–317.
- Mallinckrodt et al. (2004). Choice of the primary analysis in longitudinal clinical trials. *Pharmaceutical Statistics* 3, 161–169.
- Molenberghs et al. (2004). Analyzing incomplete longitudinal clinical trial data (with discussion). *Biostatistics* 5, 445–464.
- Cook, Zeng & Yi (2004). Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics* 60, 820-828.