Handling Missing Data for Indicators

Susanne Rässler

Institute for Employment Research & Federal Employment Agency
Nürnberg, Germany

First Workshop on Indicators in the Knowledge Economy, Tübingen, 3-4 March 2005

Agenda

- **Missing data**
- **Multiple imputation principle**
- **Indicators are missing**
- **Multivariate MI approach**
- **Alternative approaches via flexible chained equations**
- **Conclusions**

Missing Data - “everybody has them, nobody wants them”

Unit no.	Gender	Age	Education	Health state	Personal Net-Income	...
1	female	40-45	high	good	?	...
2	male	30-35	middle	poor	4500-5000	...
3	female	>60	?	poor	4000-4500	...
4	male	20-25	high	?	?	...
5	male	20-25	low	?	1500-2000	...
6	female	30-35	low	good	1500-2000	...
...	...	...				...

Case Deletion

Unit no.	Gender	Age	Education	Health state	Personal Net-Income	...
2	male	30-35	middle	poor	4500-5000	...
6	female	30-35	low	good	1500-2000	...
...	...	...				...

Missingness may be either

- **MCAR (missing completely at random),**
- **MAR (missing at random), or**
- **MNAR (missing not at random)**

(Rubin and Little 1987, 2002)

$\Rightarrow$ In multivariate analysis often 30% to 40% of the data are lost with case deletion assuming MCAR!

Handling missing data

- Procedures based on the available cases only, i.e., only those cases that are completely recorded for the variables of interest

- Weighting procedures such as Horvitz-Thompson type estimators or raking estimators that adjust for nonresponse

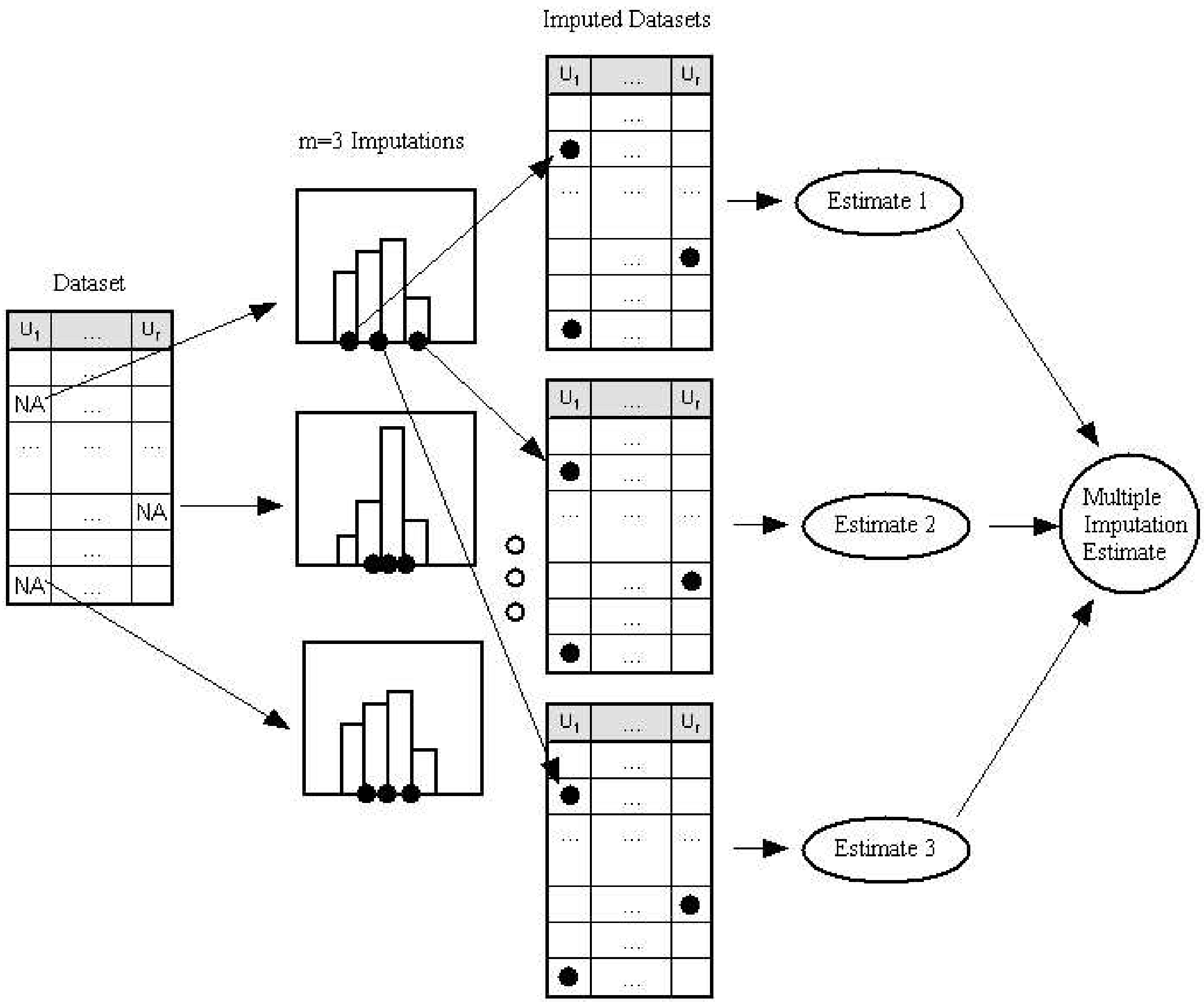
- Single imputation and correction of the variance estimates to account for imputation uncertainty

- Multiple imputation (MI) according to Rubin (1978, 1987) and standard complete-case analysis

- Model-based corrections of parameter estimates such as the expectation-maximization (EM) algorithm

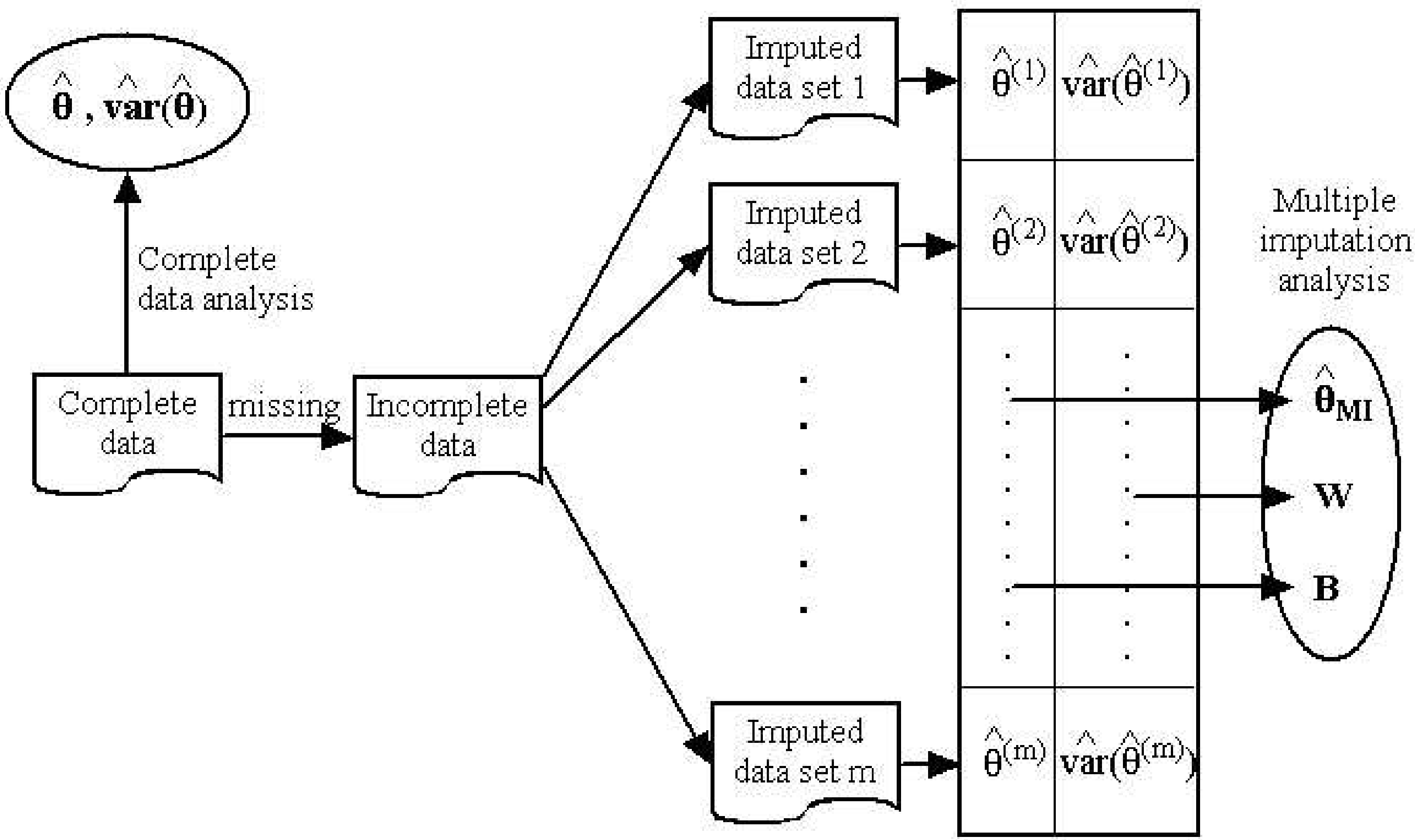
$\Rightarrow$ **We regard multiple imputation as most flexible for multipurpose complex surveys such as KEI**

The Multiple Imputation Principle (1)



⇒ **MI reflects uncertainty about which value to impute**

The Multiple Imputation Principle (2)



$\Rightarrow$ **Correct MI analysis is based on an analysis of variance**

The Multiple Imputation Principle (3)

- **Estimates: with complete data, tests and intervals based on the normal approximation should be appropriate (Rubin 1978, 1987, or t-approximation, Barnard and Rubin 1999); i.e.,**

$$(\widehat{\theta}-\theta)/\sqrt{\widehat{var}(\widehat{\theta})} \sim N(0,1)$$

- **Produce m completed data sets and calculate $\widehat{\theta}^{(j)}$ and $\widehat{var}(\widehat{\theta}^{(j)})$, $j = 1, 2, \ldots, m$**

- **Multiple imputation estimate $\widehat{\theta}_{MI} = \frac{1}{m}\sum_{j=1}^{m} \widehat{\theta}^{(j)}$**

- **Its estimated total variance is $T = W + (1+\frac{1}{m})B$ with "within-imputation" variance $W = \frac{1}{m}\sum_{j=1}^{m} \widehat{var}(\widehat{\theta}^{(j)})$ and "between-imputation" variance $B = \frac{1}{m-1}\sum_{j=1}^{m}(\widehat{\theta}^{(j)} - \widehat{\theta}_{MI})^2$**

$\Rightarrow$ **Tests can be based on $(\widehat{\theta}_{MI}-\theta)/\sqrt{T} \sim t_v$ with $v = (m-1)\left(1+\frac{W}{(1+m^{-1})B}\right)^2$**

Basic Principle of Multiple Imputation Procedures

- **Create m independent random draws of the missing data according to their posterior predictive distribution**

$$f_{Y_{mis}|Y_{obs}}(y_{mis}|y_{obs}) = \int f_{Y_{mis}|Y_{obs},\Theta}(y_{mis}|y_{obs},\theta) f_{\Theta|Y_{obs}}(\theta|y_{obs}) d\theta$$

- **Realization either by**
  (1) random draws of the parameters Θ according to their observed-data posterior distribution $f_{\Theta|Y_{obs}}$ as well as
  (2) random draws of Y_{mis} according to their conditional predictive distribution $f_{Y_{mis}|Y_{obs},\Theta}$ for actual draws of Θ.

- **or realization iteratively (MCMC, data augmentation) by**
  (1) random draws of the parameters Θ according to their complete-data posterior distribution $f_{\Theta|Y_{obs},Y_{mis}}$ for actual draws of Y_{mis} as well as
  (2) random draws of Y_{mis} according to their conditional predictive distribution $f_{Y_{mis}|Y_{obs},\Theta}$ for actual draws of Θ.

Indicators are missing

- **Countries: EU + USA + Japan**
- **Time period: 1995 … 2002/03, early *estimates* for 2003/04**
- **Indicators:**

GERD	**Gross domestic expenditure for R & D per capita (POP)**
PhD	**Total new science and technology PhDs per capita**
FTE	**Total researchers (FTE) per capita**
GFCF	**Total gross fixed capital formation (excl. building) per capita**
EGov	**E-government**
TEE	**Total education expenditure per capita**
LLL	**Life-long learning (per population aged 25-64 years participating in education and training; POP1)**

$\Rightarrow$ **Some indicators are missing at the most recent point of time**

Where to Go Intermediate: The multivariate model for KEI

- **Data augmentation algorithm using the multivariate linear mixed-effects model (Schafer & Yucel 2002)**

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i\,, \quad i = 1, 2, \ldots, n$$

$$\begin{array}{lcl}
Y_i & = & (T \times r) \textbf{ matrix of indicators} \\
X_i & = & (T \times p) \textbf{ matrix of covariates} \\
Z_i & = & (T \times q) \textbf{ matrix of covariates with } Z_i \textbf{ basically} \in X_i \\
\beta & = & (p \times r) \textbf{ matrix of fixed effects} \\
\textbf{vec}(b_i) & \sim & N_{qr}(0, \Psi) \textbf{ vector of random effects} \\
\textbf{vec}(\epsilon_i) & \sim & N_{Tr}(0, \Sigma \otimes I_T) \textbf{ vector of random errors} \\
\Psi^{-1} & \sim & Wishart_{qr}(a, B)\textbf{, } a, B \textbf{ hyperparameter} \\
\Sigma^{-1} & \sim & Wishart_{r}(c, D)\textbf{, } c, D \textbf{ hyperparameter}
\end{array}$$

- **times of measurement t incorporated into X_i and possibly Z_i.**

- **allows unequal spacing, time-varying covariates, unbalanced panels for T_i, correlation between indicators.**

Univariate Multiple Imputation Models for Complex Data

Simple case with 3 variables A, B and C each with missing data (Rubin 2003, applied in the NMES):

- **"Begin by arbitrarily filling in all missing B and C.**
- **Fit a model of $A|B, C$ using those units where A is observed and impute the missing A values.**
- **Toss the imputed B values and fit a model of $B|A, C$ using those units where B is observed and impute the missing B values.**
- **Toss the imputed C values and fit a model of $C|A, B$ using those units where C is observed and impute the missing C values.**
- **Iterate..."**

$\Rightarrow$ **Great flexibility due to the possible conditional specifications!**

Univariate KEI-Imputations Based on PAN (1)

- **Assume indicators are missing at random (MAR)**

- **Fit univariate mixed-effects model for each KEI indicator separately (SPLUS library pan by Schafer 1997):**

$$y_i = X_i\beta + Z_ib_i + \epsilon_i\,, \quad i = 1, 2, \ldots, n$$

$$\begin{array}{lcl}
y_i & = & (y_{i1}, y_{i2}, \ldots, y_{iT}) \text{ KEI indicator} \\
X_i & = & \text{(intercept, time)} \\
Z_i & = & \text{(intercept)} \\
\beta_0, \beta_1 & = & \text{fixed effects of intercept and time} \\
b_i & \sim & N(0, \psi) \text{ random effect for country } i \\
\epsilon_i & \sim & N_T(0, \sigma^2 I_T) \text{ random errors}
\end{array}$$

- **Leads to model** $y_i \sim N_T(X_i\beta, \psi + \sigma^2 I_T)$ **for** $i = 1, 2, \ldots, n$ **such that**

$$Cov(y_{it}, y_{js}) = \begin{cases} \psi, & t, s = 1, 2, \ldots, T, t \neq s, i = j \\ \psi + \sigma^2, & t, s = 1, 2, \ldots, T, t = s, i = j \\ 0, & \text{else, i.e. for all } i \neq j \end{cases}$$

First Results of KEI-Imputations Based on PAN (2)

- **Generate $m = 10$ imputations after a burn-in period of 1000 Gibbs cycles.**
- **ACF's of ψ, σ^2 and β suggest quick convergence**
- **Lags of 100 between each imputation are used**
- **To Do:**
  - **allow correlation between indicators $\Rightarrow$ Pan for KEI according to Schafer & Yucel (2002)**
  - **allow for heteroscedasticity $\Rightarrow$ possibly with approach Schafer & Yucel (2002)**
  - **allow for flexible serial autocorrelation $\Rightarrow$ future research**
  - **allow for spacial autocorrelation $\Rightarrow$ future research**

Conclusions

- MI is in general applicable when the complete-data estimates are asymptotically normal (like ML estimates are) or t distributed.
- The regression switching approach seems to be quite promising in large data sets and for high amounts of missing values.
- Even in the context of “mass imputation”, such as split questionnaire survey designs and data fusion we find good frequentist properties.
- In the U.S. applied for MI in the NHANES (split project) and NMES.
- The basic routines are already implemented in MICE (SPLUS and R version) and IVEware, Raghunathan’s SAS callable application.

$\Rightarrow$ Multiple imputation displays nonresponse uncertainty while using standard complete-case analysis!

References

- **Barnard, J., Rubin, D.B. (1999), Small-Sample Degrees of Freedom with Multiple Imputation, Biometrika, 86, 948-955.**
- **Rubin, D.B. (1978), Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse, (with discussion and reply), Proceedings of the Survey Research Methods Section of the American Statistical Association, 20-34.**
- **Rubin, D.B. (1987), Multiple Imputation for Nonresponse in Surveys, Wiley, New York.**
- **Little , R.J.A., Rubin, D.B. (1987, 2002), Statistical Analysis with Missing Data, Wiley, New York.**
- **Schafer, J.L. (1997), Analysis of Incomplete Multivariate Data, Chapman and Hall, London.**
- **Schafer, J.L., Yucel, R. (2002) Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values, Journal of Computational and Graphical Statistics, 11, 437-457.**