NTNU
Det skapende universitet

Analyse når noen observasjoner mangler ("missing values"). Problemstilling, mulige løsninger.

av Stian Lydersen

Forlesning i anvendt medisinsk statistikk, NTNU 11 mars 2009

Referanser:

Acock, A. C.: (2005) "Working With Missing Values". Journal of Marriage and Family", Nov 2005, Vol 67, 1012-1028

Schafer, J. L., Graham, J. W.: (2002). Missing Data: Our View of the State of the Art. Psychological methods, 2002, Vol 7, No 2, 147 – 177

Rosner, B. (2005) *Fundamentals of Biostatistics,* 6th ed, Wiley. Section13.13 (page 725 – 731) Missing data

Fayers, P. M., Machin, D. (2007) *Quality of Life. The Assessment, Analysis and Interpretation of Patient-reported Outcomes*. 2nd ed. Wiley. Chapter 15: (page 355-384) Missing Data.

Referanser (forts.)

Rässler, S., Rubin, D. B., & Zell, E. R. 2008, "Incomplete Data in Epidemiology and Medical Statistics," in Epidemiology and Medical Statistics (Handbook of Statistics vol 27)". C. R. Rao, J. P. Miller, & D. C. Rao, eds., Elsevier, pp. 569-601.

Carpenter, J. R. & Kenward, M. G. (2007) Missing data in randomised controlled trials - a practical guide. 186 pages. http://www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml

Little, R J A, Rubin, D B: (2002) *Statistical Analysis with Missing data*. 2nd ed. Wiley.

Schafer, J. L. (1997) Analysis of incomplete multivariate data. Chapman & Hall, London.

Manglende data:

- "Hull" i datamatrisen, som ideelt sett burde være komplett
- Vanligvis data som man hadde til hensikt å registrere, men som av en eller annen grunn ikke ble registrert.
- Det finnes en meningsfylt dataverdi som ikke er registrert.

Acock, A. C. (2005):
"Less than optimum strategies for missing values can produce biased estimated, distorted statistical power, and invalid conclusions. After reviewing traditional approaches (listwise, pairwise, and mean substitution), selected alternatives are covered including single imputation, multiple imputation, and full information maximum likelihood estimation. … When missing values cannot be avoided, multiple imputation and full information methods offer substantial improvements over traditional approaches."

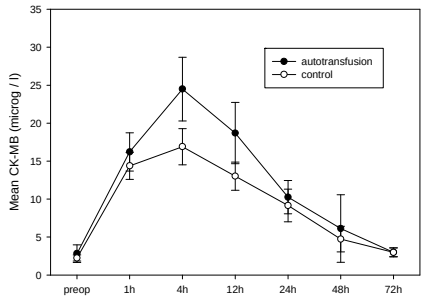
Example from:

Pleym H, Tjomsland O, Asberg A, Lydersen S, Wahba A, Bjella L, Dale O, Stenseth R. (2005)

Effects of autotransfusion of mediastinal shed blood on biochemical markers of myocardial damage in coronary surgery. Acta Anaesthesiol Scand. 2005 Oct;49(9):1248-54.

Randomised study, 23 autotransfusion and 24 control patients.

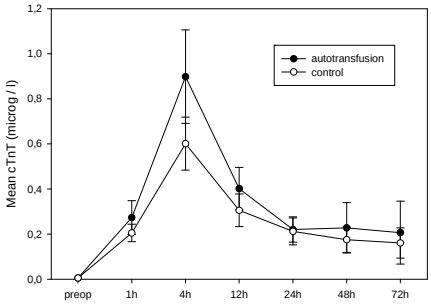
7



NTNU
Det skapende universitet

www.ntnu.no

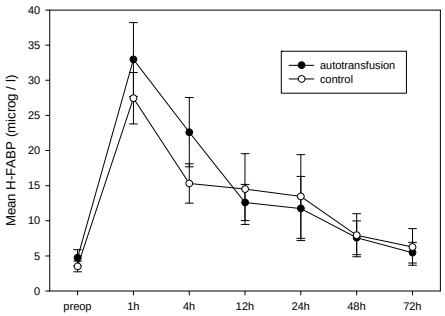
8



NTNU
Det skapende universitet

www.ntnu.no

9



NTNU
Det skapende universitet

www.ntnu.no

10

Missing data problem:

- 987 measurements (7 time points x47 patients x 3 substances)
- Missing data for 7 of 987 measurements
- This was 4 of the 47 patients!
- Repeated measurements ANOVA (as used in this study) requires complete data

NTNU
Det skapende universitet

www.ntnu.no

11

"A total of 7 out of 987 serum values were missing. Missing values were imputed using the EM algorithm with multivariate normal distribution on ln-transformed data. According to inspection of Q-Q plots, the ln-transformed data showed acceptable fit to the normal distribution, while the original data tended to be skewed. Repeated measurements ANOVA was used for joint analysis of the serum values of CK-MB, cTnT, and H-FABP, respectively, using the EM imputed ln-transformed values."

NTNU
Det skapende universitet

www.ntnu.no

12

Example from:

Hallan, S. I., Ritz, E., Lydersen, S., Romundstad, S., Kvenild, K., & Orth, S. R. Combination of estimated glomerular filtration rate and albuminuria provides best prediction of kidney failure: Results of the HUNT II study, Norway. In press, Journal of the American Society of Nephrology, 2009.

Cox proportional hazards regression with time to kidney failure (CKD stage 5) as dependent variable.

HUNT II (Helseundersøkelsen i Nord-Trøndelag), 1995-1997. Follow-up until 2007.

NTNU
Det skapende universitet

www.ntnu.no

13

92939 persons, 20 years and older, were invited.
65589 (70.6%) responded.
124 kidney failures.
8360 were hypertensive or had diabetes mellitus. These were asked to deliver urine samples, and 88.6% did so. In addition, a random 5% sample of non-diabetic non-hypertensive subjects (n=2,861) was also asked to deliver urine samples; 75.6% did so.

Hence: For 95% of the non-diabetic non-hypertensive subjects, urine samples were
Missing at random (MAR) by design.

14

Variable	n	% missing
Follow-up time	65589	0,0
Age	65589	0,0
Male sex	65589	0,0
Low education	61369	6,4
Depression	58423	10,9
Smoking	64395	1,8
Low physical activity	57881	11,8
Diabetes mellitus	64693	1,4
CVD	64624	1,5
BMI	64306	2,0
Waist circumference	64022	2,4

15

Variable	n	% missing
Systolic BP	64708	1,3
Diastolic BP	64708	1,3
Cholesterol	65158	0,7
HDL-Cholesterol	65155	0,7
GLUCOSE	65158	0,7
Triglycerides	65158	0,7
Creatinine	65158	0,7
eGFR [1)]	65158	0,7
ACR [2)]	9703	85,2

[1)] estimated glomerular filtration rate
[2)] Albumin creatinin ratio (from urine sample)
Not requested (Missing by design): 82,8 %
Requested, but not deliverd: 2,5%

16

Traditional classification of non response among survey methodologists.

- The sampled person does not participate (Unit nonresponse)
- Partial data are available for the sampled person. (Item, scale or form nonresponse)

The focus of this presentation

17

Ad hoc edits

- Tricks that apparently solve the missing data problem but redefine the parameter space:
  - Replacing the missing values by an arbitrary number and include a dummy indicator in regression.
  - For a nominal variable with categories 1, 2, …, k, let f.ex. k+1 represent missing.

18

Patterns of nonresponse:

Figure 1 Scafer & Graham p 150

19

The distribution of missingness

Let R denote what is known and what is missing.
For example with values 1 (0) if the corresponding value is observed (missing).
The probability distribution of R has been called
- response mechanism
- missingness mechanism
- probabilities of missingness
- distribution of missingness

NTNU Det skapende universitet

20

Types of missing data (Rubin, 1976)

- MCAR – Missing Completely at Random
- MAR – Missing at Random (ignorable nonresponse)
- MNAR – Missing Not at Random (nonignorable nonresponse)

NTNU Det skapende universitet

21

Types of missing data

$Y_{com} = (Y_{obs}, Y_{mis})$

MAR: The distribution of missingness does not depend on Y_{mis}:

$P(R|Y_{com}) = P(R|Y_{obs})$

MCAR: It does not depend on Y_{obs} either:

$P(R|Y_{com}) = P(R)$

NTNU Det skapende universitet

22

Repeated measurements with monotone dropout (attrition), Figure 1b

- MCAR: Y_j is missing with probability unrelated to any variables.
- MAR: It may be related only to $Y_1, \ldots, Y_{j-1}$ (only seen responses). *Noninformative* or *ignorable* dropout.
- MNAR: Related to $Y_1, \ldots, Y_p$ (seen and unseen responses). *Informative* dropout.

NTNU Det skapende universitet

23

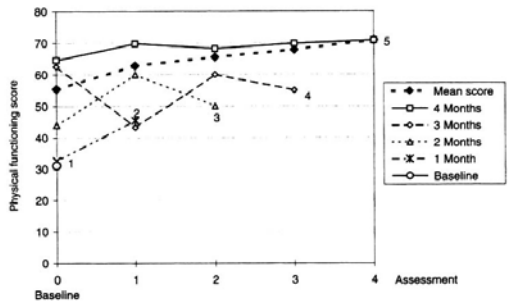
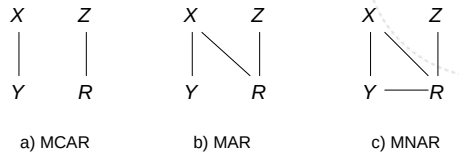


Figure 15.2 Mean physical functioning (PF) score stratified by time of dropout due to death or non-compliance in patients with metastatic breast cancer (Based on Curran *et al.*, 1998)

24



a) MCAR b) MAR c) MNAR

X is completely observed
Y is missing for some
Z denotes causes of missingness unrelated to X and Y
R represents missingness

NTNU Det skapende universitet

25

Plausibility and implications of MAR

- Planned missingness usually MCAR, sometimes MAR
  - Certain sequential designs
  - Multiple questionnaire forms
- MAR may be tested by obtaining follow-up data from non-respondents
- Else: NO WAY to test if MAR holds
- In some situations, erroneous assuming MAR has minor impact on results (refs in Schafer & Graham)

NTNU Det skapende universitet

26

Example – Systolic blood pressure at two time points.

Simulated data, Schafer and Graham Table 1

- $\mu_X = \mu_Y = 125$
- $\sigma_X = \sigma_Y = 25$
- $\rho(X, Y) = 0.6$

- 30 observations of X
- MCAR: 7 randomly selected Y
- MAR: Y observed if X > 140
- MNAR: Y observed if Y > 140

NTNU Det skapende universitet

27

Schafer & Graham
Table 1
page 153

NTNU Det skapende universitet

28

Methods for missing data (Schafer and Graham)

- Older methods
- ML estimation
- Multiple Imputation (MI)

impute to fill in data values (usually **missing data**) with values that are thought to be sensible ….
Simon Day: Dictionary for clinical trials, 2nd ed, Wiley, 2007.

NTNU Det skapende universitet

29

Older methods

- Case deletion / listwise deletion (LD) / complete-case analysis. Default in many computer programs.
- Available case (AC) analysis / pairwise deletion / pairwise inclusion
- Reweighting. Weights derived from estimated probabilities of nonresponse.
- Averaging available items
- Single imputation

NTNU Det skapende universitet

30

Bias in complete case analysis and in available case analysis:

P = proportion of cases (persons) with some missing data.
The combined mean, if we could observe it, is

$$\mu = (1-P)\mu_{\text{Responding}} + P\mu_{\text{NonResponding}}$$

The (unobservable) bias caused by complete case analysis is

$$B = \mu - \mu_{\text{Responding}} = (1-P)\mu_{\text{Responding}} + P\mu_{\text{NonResponding}} - \mu_{\text{Responding}}$$
$$= P(\mu_{\text{NonResponding}} - \mu_{\text{Responding}})$$

In clinical trials:

$$B_{\text{Difference}} = B_{\text{Active}} - B_{\text{Control}}$$

NTNU Det skapende universitet

31

Schafer & Graham
Table 2
page 156

32

Example – Quality of Life Questionnaire

European organization for research and treatment of cancer QLQ-C30 (EORTC QLQ-C30)

33

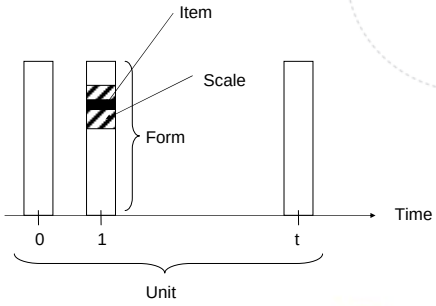
European organization for research and treatment of cancer QLQ-C30 (EORTC QLQ-C30)

	Not at all	A little	Quite a bit	Very much
21. Did you feel tense?	1✓	2	3	4
22. Did you worry?	1	2	3	4
23. Did you feel irritable?	1	2✓	3	4
24. Did you feel depressed?	1	2✓	3	4

Figure 15.3 The emotional functioning scale of the EORTC QLQ-C30

34

Example: Quality of Life questionnaires



35

Methods for missing items within a scale or form (Fayers & Machin, 2007)

- Treat the scale as missing
- Mean imputation: Compute the mean for the available items (if for example less than 50% are missing)
- Hierarchical scales (for example items 1-5 of EORTC QLQ-C30)
- Regression imputation: Predict the missing item from the remaining items

36

Beregning av gjennomsnitt i SPSS

Mean(q21, q22, q23, q24).
beregner hvis minst en av variablene er gitt

(q21+ q22 + q23 + q24)/4
beregner bare hvis alle variablene er gitt

Mean.2(q21, q22, q23, q24).
beregner hvis minst 2 av verdiene er gitt

37

Some methods for missing scales or forms

- Last Value Carried Forward (LVCF)
- Simple Mean Imputation (mean of those who completed the form at that time)
- Horizontal Mean Imputation – the mean of the patient's previous scores

Not OK

- Standardized Score Imputation
- Single imputatiton - EM (Expectation Maximation) algorithm *OK*
- Multiple Imputation
- Mixed models
- Generalized estimating equations (GEE)

OK

38

Warning:

Last Value Carried Forward (LVCF)
Same as Last Observation Carried Forward (LOCF)

Carpenter and Kenward (2007)

"As LOCF is neither valid under general assumptions nor based on statistical principles, it is not a sensible method, and should not be used."

"LOCF is not sensible, even when data are MCAR."

39

Standardized score imputation:
Two assessments, one patient fails to complete the second.
Standardized value at time 1:

$$Z_1 = (Q_1 - \bar{Q}_1) / SD_1$$

Assume the same standardized value at time 2 had it been observed:

$$Z_1 = (Q_2 - \bar{Q}_2) / SD_2$$

so Q_2 is computed as

$$Q_2 = \bar{Q}_2 + Z_1 SD_2$$

40

Mixed models / GEE for longitudinal data:

- Use data for the units and times where neither the response Y nor the covariates x are missing.
- Using Maximum Likelihood (ML), the estimates are consistent as long as data are missing MAR.

41

Single imputation (Schafer and Graham)

- Imputing unconditional means
- imputing from unconditional distributions (for example hot deck)
- imputing conditional means
- imputing from conditional distributions

42

Schafer and Graham
Figure 3
page 160

43

Schafer and Graham
Table 3
page 161

NTNU
Det skapende universitet
www.ntnu.no

44

Single imputation can be reasonable and better than casewise deletion.

Example:
Data set with 25 variables and 3% missing.
Casewise deletion discards ($1-0.97^{25}$) = 53% of the cases. Imputing once from a conditional distribution permits use of all data and with minor impact on estimates and uncertainty measures.

NTNU
Det skapende universitet
www.ntnu.no

45

ML estimation

Some theory …

$$P(Y_{obs};\theta) = \int P(Y_{com};\theta)dY_{mis}$$

is a correct
- probability distribution if MCAR
- likelihood if MAR or missing values are out of scope

$$P(Y_{obs}, R;\theta,\xi) = \int P(Y_{com};\theta)P(R \mid Y_{com};\xi)dY_{mis}$$

is a correct likelihood if MNAR (difficult)

NTNU
Det skapende universitet
www.ntnu.no

46

ML Estimation

When data are MAR, the observed data-likelihood is

$$L(\theta;Y_{obs}) = P(Y_{obs};\theta)$$

The ML estimate $\hat{\theta}$ that maximises the likelihood.

The log likelihood may be easier to calculate:

$$l(\theta;Y_{obs}) = \log L(\theta;Y_{obs})$$

NTNU
Det skapende universitet
www.ntnu.no

47

Basis for confidence intervals and testing hypotheses:

$\hat{\theta}$ is approximately normal distributed with

expectation θ

and covariance matrix

$$V(\hat{\theta}) \approx [-l''(\hat{\theta})]^{-1}$$

where $-l''(\hat{\theta})$ is the *observed information*.
The *expected information (Fisher information)* performs poorer for missing data.

NTNU
Det skapende universitet
www.ntnu.no

48

But the likelihood equation for missing data …

- can seldom be expressed in closed form
- hardly ever be solved explicitly for θ

NTNU
Det skapende universitet
www.ntnu.no

49

The EM Algorithm for ML estimation with missing data (Dempster & al 1977)

- Fill in missing data with a best guess
- Estimate the parameters for the complete data set
- Re-guess missing data with the estimated parameters
- Repeat until convergence
- May need many iterations
- Available (partly) in many statistical software packages

50

Schafer and Graham Table 4

51

MI (Multiple Imputation), Rubin (1987)

- Create m > 1 (for example m=5) data sets by single imputation from the conditional distribution
- Analyse each data set by a complete data method
- Combine the results using simple artihmetric to obtain overall estimates reflecting missing data uncertainty and finite-sample variations.

52

MI - advantages

- Retains the attractive of single imputation from conditional distribution
- A single imputed set may be randomly atypical
- Does not underestimate uncertainty
- Unlike other Monte Carlo methods, few repetitions are needed.

53

Rubin's (1987) rules for combining estimates and variances

Q = the population quantity of interest, $U = Var(\hat{Q})$

m estimates $\hat{Q}^{(j)}$, $U^{(j)}$, for j = 1, …, m

Estimate for Q:

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}^{(j)}$$

54

Average within-imputation variance

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U^{(j)}$$

Between-imputation variance

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left[\hat{Q}^{(j)} - \bar{Q}\right]^2$$

Total variance:

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

55

Student's t approximation for confidence intervals and tests for Q

$$\frac{\bar{Q}-Q}{\sqrt{T}} \sim t_{\upsilon}$$

where

$$\upsilon = (m-1)\left[1+\frac{\bar{U}}{(1+m^{-1})B}\right]^2$$

NTNU Det skapende universitet

www.ntnu.no

56

Proper MI reflecting the uncertainty in the model parameters

A single imputation is drawn from $P(Y_{mis} \mid Y_{obs}; \hat{\theta})$

MI:

simulate m plausible values $\theta^{(1)},...,\theta^{(m)}$
draw $Y_{mis}^{(t)}$ from $P[Y_{mis} \mid Y_{obs}; \theta^{(t)}]$ for t=1,…,m

Bayesian approach with a prior distribution for θ
is natural but not essential

NTNU Det skapende universitet

www.ntnu.no

57

Hallan & al (2009)
Statistical analyses were performed using Stata 10.0 (Stata Corp., TX, U.S.A.). In general, there were few missing data (<2% for most variables, see Table 1), but data on ACR were, by study design, available only in a subgroup. Multiple imputation is now considered the standard method for handling this type of data,(Clark & Altman 2003;Donders et al. 2006;Rassler et al. 2008;van Buuren et al 1999) whereas complete case analysis would yield too imprecise as well as biased results. The multiple imputation technique estimates the mean and uncertainty of the missing data using all information from the actually observed data in a proper way. In this way, unbiased estimates with the correct standard deviation and p-values are calculated.(Rassler et al, 2008).

NTNU Det skapende universitet

www.ntnu.no

58

Continued:
For most non-diabetic non-hypertensive subjects data were missing completely at random, and for those not returning urine samples as requested data were assumed to be missing at random, thus meeting the assumptions for the method. The analyses were carried out in the "ice" and "micombine" procedures for Stata,(Royston 2005) ACR was log-transformed and not used as predictor in the imputation of other missing variables,(van Buuren et al 1999b) study outcome variables were included in the imputation model,(Moons et al. 2006) and the time variable was log-transformed.(van Buuren, et al 1999a) Regression modelling revealed interactions between sex and both blood pressure and diabetes mellitus. Hence, these two interactions were included in the imputation model. We used m=20 imputations to achieve maximum accuracy.(Newgard & Haukoos 2007)

NTNU Det skapende universitet

www.ntnu.no

59

MI implementation in Stata

- ice – each variable with missing data is predicted from the other variables using the appropriate imputation regresson models (Linear, Binary logistic, or Ordinal logistic) and creates m imputed (complete) data sets.
- micombine – analyses each imputed data set using the relevant revant analysis model (for example Cox proportional hazards regression), and combines the results using Rubin's rules
- mim (2008) replaces micombine. Several new features.

NTNU Det skapende universitet

www.ntnu.no

60

Example Hallan et al 2009. Implementation in Stata (ice)

- Categorical variables must be coded 0,…k-1. For example female is coded 0 and 1
- Continuous variables are assumed normally distributed. Used ln(ACR) instead of ACR.
- Do not use a predictor with more than 50% missing. (Hence ln(ACR) used only as dependent variable)
- Include outcome variable as predictor. Here: follow-up time and event CKD.
- Use log transformed time variable as predictor (outcome variable in the Cox analysis model)
- Do not impute outcome if missing!
- Use an imputation model at least as rich as the analysis model. We included the interactions sex*bp and sex*diabetes.
- Used a high number of imputations (m=20) due to high proportion missing.

NTNU Det skapende universitet

www.ntnu.no

61

Needed to use the passive(), substitute(), and eq() options.

First, compute
indicator variables for categorical variables with >2 categories
Interaction terms
(is this necessary?)

62

Command:
ice acrmean ln_acrmean female alder95b mdrdnew royker currentsmoker formersmoker sbtm95 diab01 kol95 hdl95 ln_months ckd5 btmed female_sbt female_dbtm fys4_kat fys1 fys2 fys3 midje95 hofte95 utdanning bmi cvd hads_d8
using "M:\...\impute21sept07",
dryrun m(20)
passive(female_sbt:female* sbtm95\ female_dbtm:female*diab01\ formersmoker:royker==1\ currentsmoker:royker==2\ acrmean:exp(ln_acrmean)\ fys1:fys4_kat==1\ fys2:fys4_kat==2\ fys3:fys4_kat==3\)
substitute(royker: formersmoker currentsmoker\fys4_kat: fys1 fys2 fys3)
eq(ln_acrmean: female alder95b mdrdnew currentsmoker formersmoker sbtm95 diab01 kol95 hdl95 ln_months ckd5 btmed female_sbt female_dbtm fys1 fys2 fys3 midje95 hofte95 utdanning bmi cvd hads_d8 , royker: female alder95b mdrdnew sbtm95 diab01 kol95 hdl95 ln_months ckd5 btmed female_sbt female_dbtm fys1 fys2 fys3 midje95 hofte95 utdanning bmi cvd hads_d8 ….)

63

Continued:
A complete case analysis (n=7,832 and 43 kidney failure cases) would have introduced bias and imprecision: e.g. the adjusted HR (95% CI) for eGFR (not adjusted for ACR) would have changed from 10.3 (5.8-18.3) to 18.4 (6.8-49.9) for eGFR 45-59ml/min/1.73m2, from 42.5 (23.5-76.9) to 76.4 (28.6-204.3) for eGFR 30-44ml/min/1.73m2, and from 212.2 (115.5-389.7) to 253.2 (85.5-749.7) for eGFR 15-29 ml/min/1.73m2.

64

(results from 5 of the imputations in Hallan et al 2009)
Results of Cox Proportional hazard regression. Regression coefficient estimate (SE), p-value, and FMI (Fraction missing information). Results from Stata with commands ice and mim.

	Imputation number					Total, by Rubin's rules	p-value	FMI
	1	2	3	4	5			
Age, years	0.0707 (0.0067)	0.0705 (0.0067)	0.0701 (0.0067)	0.0701 (0.0067)	0.0707 (0.0067)	0.0704 (0.0067)	<0.001	0.002
Female sex	-0.612 (0.189)	-0.580 (0.190)	-0.570 (0.190)	-0.582 (0.190)	-0.589 (0.190)	-0.587 (0.191)	0.002	0.008
ACR	0.0276 (0.0013)	0.0282 (0.0013)	0.0285 (0.0013)	0.0283 (0.0013)	0.0280 (0.0013)	0.0281 (0.0014)	<0.001	0.082

65

MI example, Rosner (2005), p 725-731

- 2341 elderly patients in 1988-1989
- Predict death by 1991 by multiple logistic regression
- Covariates:
  - age (years)
  - male sex
  - physical performance (0 worst to 12 best)
  - self assessed health (1 excellent to 4 poor)
- Physical performance data missing for 550 patients unwilling or unable to perform the test.

66

Table 13.32 Descriptive statistics for 2341 older residents of East Boston interviewed in 1988–1989

	Completed the physical-performance evaluation			
	Yes		No	
Dead by 12/31/1991	No	Yes	No	Yes
n	1527	264	416	134
Age, median (IQR)*	77 (74–81)	78 (74–84)	78 (74–83)	85 (78–90)
Male, %	32.0	45.1	30.3	47.0
Physical performance median (IQR)	8 (5–10)	6 (2–8)	—	—
Self-assessed health median (IQR)	2 (2–3)	3 (2–3)	2 (2–3)	3 (2–3)

Notes: These are the 2341 residents of East Boston who participated in the 6-year follow-up evaluation of the EPESE, or Established Populations for Epidemiologic Studies of the Elderly (see, for example, Glynn et al. [35]). They ranged in age from 71 to 103 years at that time. Participants were asked to rate their health relative to others of their age as 1, excellent; 2, good; 3, fair; or 4, poor. Those who were able, also had objective evaluation of physical performance. This was based on brief tests of balance, gait, strength, and endurance, and summarized in an overall score ranging from 0 to 12, with higher scores indicting better function (Guralnik et al. [36]).
*Interquartile range.

67

Table 13.33 Comparison of effects of variables on the risk of death from alternative models; shown are logistic-regression parameters (standard errors)

	Analytic method			
	Complete case	No physical performance	Multiple imputation	Indicator method[a]
n (deaths)	1791 (264)	2341 (398)	2341 (398)	2341 (398)
Age	0.033 (0.013)	0.088 (0.009)	0.057 (0.011)	0.068 (0.01)
Male	0.92 (0.15)	0.82 (0.12)	1.00 (0.14)	0.92 (0.12)
Self-assessed health	0.38 (0.095)	0.60 (0.073)	0.39 (0.087)	0.46 (0.076)
Physical performance	−0.14 (0.023)	—	−0.15 (0.024)	−0.12(0.022)
Intercept	−4.76 (1.17)	−10.41 (0.79)	−6.60 (1.04)	−7.42 (0.92)
Indicator of missing performance				−0.47 (0.13)

[a]The indicator method assigns the average performance score (6.8) to those with missing values and includes an indicator variable for this group.

68

Table 13.34

Summary of linear-regression model predicting physical-performance score, and 5 draws of these regression parameters

		5 drawn values of parameters				
Variable	Estimate (SE)	1	2	3	4	5
Age (per year)	−0.22 (.013)	−0.23	−0.23	−0.23	−0.21	−0.22
Male gender	1.18 (0.15)	1.21	1.08	1.29	1.25	1.22
Self-assessed health	−1.38 (0.089)	−1.33	−1.43	−1.24	−1.43	−1.35
Dead by 1991	−1.30 (0.20)	−1.12	−1.47	−1.37	−1.60	−1.05
Intercept	27.2 (1.1)	27.6	27.7	27.5	26.3	27.2

Overall R^2 0.30; Root MSE 2.88

69

Table 13.35

70

Schafer and Graham Table 5

71

Software at NTNU

- Single imputation:
  - SPSS
  - R: MIXED, PAN
- Multiple imputation
  - Skal finnes noe i SPSS 17
  - Stata: ice, (micombine,) mim
  - R: MIXED, PAN
- Full ML / SEM:
  - AMOS (available through SPSS)
  - Stata and R / S-plus: gllamm
- Mixed models / GEE:
  - SPSS
  - Stata
  - R

72

ice – an implementation of the MICE algorithm (Multiple imputation using chained equations). Creates m imputed data sets

Mim (replaces Micombine) – combines the results from Ice using Rubin's rules

GLLAMM (Generalized latent and mixed model)

SEM – Structural equations modelling

73

Acock, A. C. (2005):
"The days that journals tolerate the absence of analysis of missing values and the use of traditional approaches to missing values should be numbered. … In general, multiple imputation and the approaches available in structural equation modelling software are the best that are currently available."

NTNU
Det skapende universitet
www.ntnu.no

74

Altman, D. G. (BMJ, 2007):
" … complete case analysis: … When only a very few observations are missing little harm will be done"

Schafer J. L. 1997, "Analysis of incomplete multivariate data" Chapman & Hall, London, page 1:
"When incomplete cases comprise only a small fraction of all cases (say, five percent or less) then case deletion may be a perfectly reasonable solution to the missing-data problem."

NTNU
Det skapende universitet
www.ntnu.no