Missing Data: Multipele Imputatie

Mark Huisman

Rijks*universiteit* **Groningen**

Statistiek in de Praktijk

30 maart 2006

Inhoud

1. Omgaan met ontbrekende scores:
   Imputeren
2. Procedures voor *Single Imputation*:
   (un)conditional means and distributions
3. Procedures voor *Multiple Imputation*:
   multivariate normale verdeling
   NORM (Schafer, 1997, 2000)

1. Omgaan met ontbrekende scores

Verschillende missing data procedures:

- Analyse van geobserveerde data:
  complete case, *available case*
- Procedures gebaseerd op *(her)wegen* van volledig geobserveerde cases
- Procedures gebaseerd op modelleren van de geobserveerde data: *EM-algoritme*, *selection models*, *pattern-mixture models*
- Procedures voor imputeren:
  Single Imputation, *Multiple Imputation*

Imputeren

Invullen van plausibele waarden voor de ontbrekende scores

Voordelen:

- Efficiënter dan analyse op complete cases
- Gebruik maken van informatie over ontbrekende scores in de geobserveerde data
- Opgevulde dataset kan worden geanalyseerd met standaard methoden en software
- Eenmalig imputeren zorgt er voor dat de dataset voor alle vervolganalyses hetzelfde blijft

Nadelen:

- Soms moeilijk te implementeren, m.n. multivariate gevallen
- Sommige (ad hoc) procedures vertekenen verdelingen en relaties

Dempster & Rubin (1983):

The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.

2. Single Imputation

(Schafer & Graham, 2002)

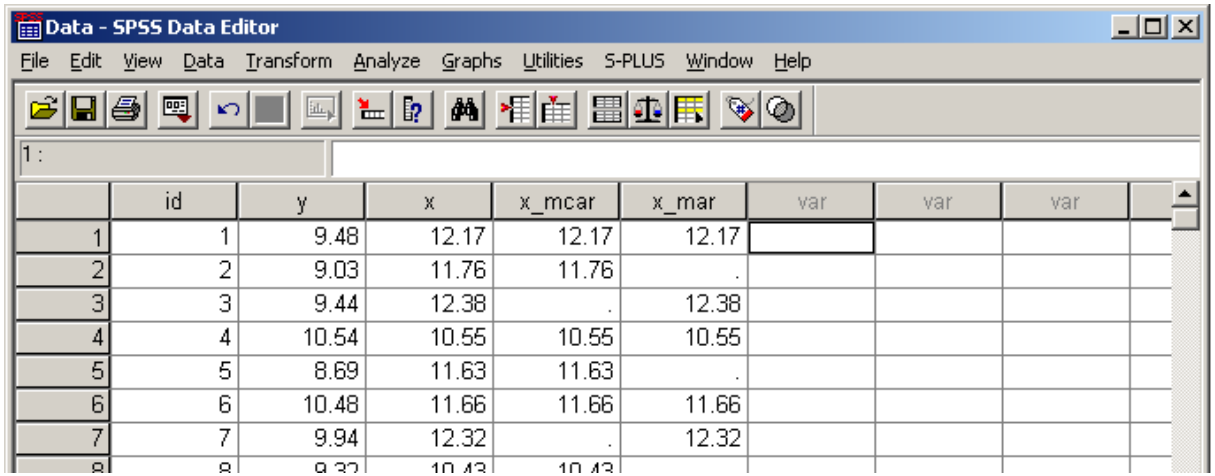
- Imputeren van *unconditional means*: invullen gemiddelden
- Imputeren met *unconditional distributions*: trekking uit geobserveerde scores $\Rightarrow$ intact houden van verdelingen
  hot deck imputation
- Imputeren van *conditional means*: invullen voorspellingen (bijv. met regressiemodel)
- Imputeren met *conditional distributions*: trekking uit de verdeling van Y_{mis} gegeven Y_{obs}
  invullen van (regressie) voorspellingen plus random error

Voorbeeld

y en x bivariaat normaal verdeeld: $\mu = \begin{pmatrix} 10 \\ 12 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$

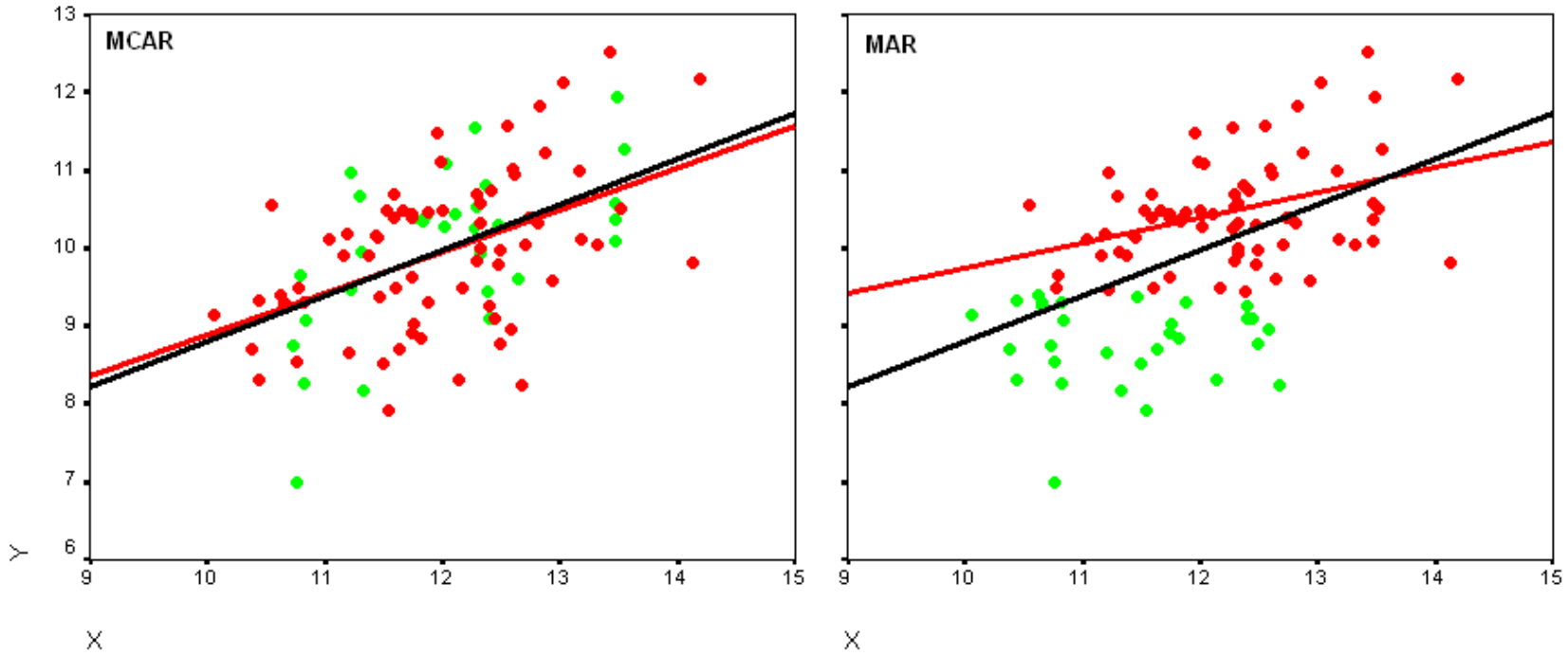
Twee keer missing data voor x: 30% ontbrekende scores

1. *Missing Completely at Random* (MCAR): willekeurig
2. *Missing at Random* (MAR): x ontbreekt als $y < 9.4$



	id	y	x	x_mcar	x_mar	var	var	var
1	1	9.48	12.17	12.17	12.17			
2	2	9.03	11.76	11.76	.			
3	3	9.44	12.38	.	12.38			
4	4	10.54	10.55	10.55	10.55			
5	5	8.69	11.63	11.63	.			
6	6	10.48	11.66	11.66	11.66			
7	7	9.94	12.32	.	12.32			
8	8	9.32	10.43	10.43				

Voorbeeld: ontbrekende scores zijn groen

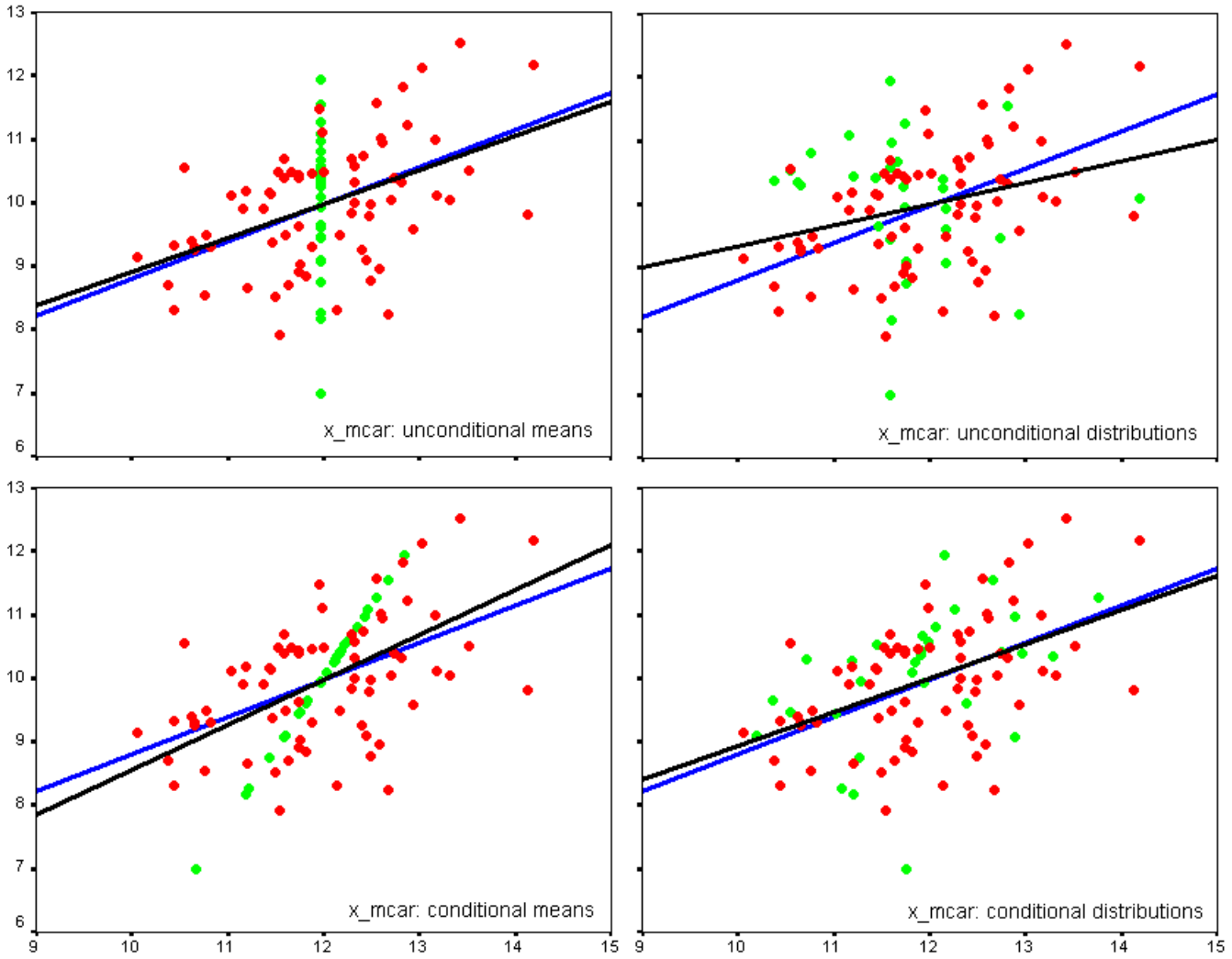


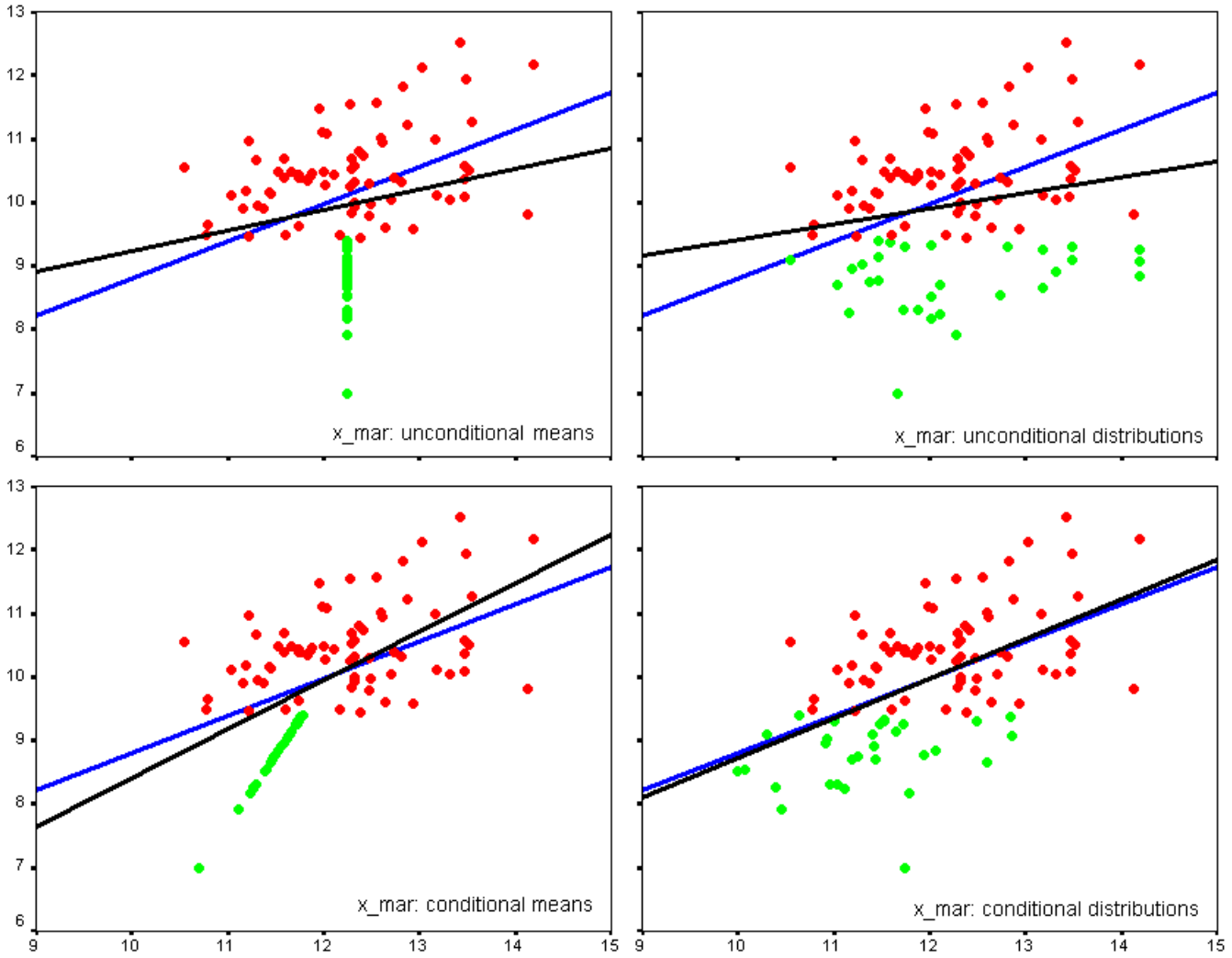
	$\bar{x}$	$SD(x)$	b_1	$SE(b_1)$
Compleet	11.98	0.892	0.583	0.098
MCAR	11.97	0.902	0.536	0.117
MAR	12.25	0.807	0.322	0.096

Voorbeeld: resultaten *Single Imputation* (MCAR)

1. *unconditional means*: $\widehat{x}_{mis} = 11.97$
   $\Rightarrow$ onzuivere schattingen varianties en covarianties
2. *unconditional distribution*: $\widehat{x}_{mis}$ random uit geobserveerde scores x
   $\Rightarrow$ onzuivere schattingen covarianties
3. *conditional means*: $\widehat{x}_{mis} = 7.59 + 0.44\, y$
   $\Rightarrow$ onzuivere schattingen covarianties
4. *conditional distributions*: $\widehat{x}_{mis} = 7.59 + 0.44\, y + 0.79\, z$, met $z \sim N(0, 1)$
   $\Rightarrow$ zuivere schattingen onder MAR

(Alle berekeningen met SPSS behalve 2, hot deck)


x_mcar: unconditional means
x_mcar: unconditional distributions
x_mcar: conditional means
x_mcar: conditional distributions


x_mar: unconditional means
x_mar: unconditional distributions
x_mar: conditional means
x_mar: conditional distributions

Single Imputation

Tekortkomingen:

- *onzuivere schattingen*: gemiddelden en/of varianties en/of covarianties
  $\Rightarrow$ *gezamenlijke verdeling* van de variabelen 'verstoord'
- SE's, p-waarden, en andere *maten voor onzekerheid* zijn misleidend omdat ze de extra onzekerheid veroorzaakt door missing data niet weergeven
- Bovendien worden imputaties behandeld als observaties
  $\Rightarrow$ steekproefgrootte is niet gelijk aan n

Oplossing: *Multiple Imputation*

3. Multiple Imputation

Herhaal het imputatieproces

Parameterschattingen variëren door het *random karakter* van de imputaties en deze variatie kan worden gebruikt voor de *correctie van de varianties* en SE's

MCAR	Mean		Regr.			Reg+error		
$\bar{x}$	$SE(\bar{x})$	$r(y,x)$	$\bar{x}$	$SE(\bar{x})$	$r(y,x)$	$\bar{x}$	$SE(\bar{x})$	$r(y,x)$
11.97	0.075	0.400	11.98	0.079	0.559	11.93	0.089	0.474
						11.95	0.090	0.416
						12.02	0.088	0.516
						11.98	0.090	0.552
						11.96	0.097	0.519
11.98	0.089	0.516	11.98	0.089	0.516	11.98	0.089	0.516

Multiple Imputation

Herhaal het imputatieproces m keer $\Rightarrow$ m geïmputeerde datasets

Analyseer de m datasets met een *standaard techniek*
en vat de resultaten samen (Rubin, 1987):

schatting $\bar{Q}_m = \frac{1}{m}\sum_i^m \hat{Q}_i$

variantie van schatting $T_m = \bar{U}_m + (1 + \frac{1}{m})B_m,$

met variantie *binnen datasets* $\bar{U}_m = \frac{1}{m}\sum_i^m U_i$

en variantie *tussen datasets* $B_m = \frac{1}{m-1}\sum_i^m (\hat{Q}_i - \bar{Q}_m)^2$

MCAR	Mean		Regr.			Reg+error		
$\bar{x}$	$SE(\bar{x})$	$r(y,x)$	$\bar{x}$	$SE(\bar{x})$	$r(y,x)$	$\bar{x}$	$SE(\bar{x})$	$r(y,x)$
11.97	0.075	0.400	11.98	0.079	0.559	11.93	0.089	0.474
						11.95	0.090	0.416
						12.02	0.088	0.516
						11.98	0.090	0.552
						11.96	0.097	0.519
						12.07	0.090	0.451
						11.97	0.086	0.501
						11.96	0.091	0.546
						12.03	0.092	0.480
						11.92	0.093	0.473
11.97	0.075	0.400	11.98	0.079	0.559	11.98	0.103	0.493
11.98	0.089	0.516	11.98	0.089	0.516	11.98	0.089	0.516

schatting $\bar{Q}_m = 11.98$ met

variantie $T_m = (0.103)^2 = 0.0082 + 1.1 \times 0.0022$

(Samenvattingen met Excel spreadsheets)

Genereren van multipele imputaties

Imputeren met (*conditional distributions*; SPSS):

$$\widehat{x}_i = b_0 + b_1\, y_i + s_e\, z_i, \text{ met } z_i \sim N(0,1)$$
$$b_0 = 7.590, b_1 = 0.440, s_e = 0.795$$

Op deze manier wordt de *kansverdeling van de ontbrekende waarden* gegenereerd: $P(X_{mis}|Y_{obs})$
$\Rightarrow$ Imputaties zijn trekkingen uit deze verdeling

Probleem:
b_0, b_1 en s_e worden beschouwd als 'ware' populatieparameters, terwijl het steekproefschattingen zijn

De populatiewaarden zijn onbekend, maar voor zgn. *proper multiple imputations* (Rubin, 1987) moet iedere geïmputeerde dataset gebaseerd zijn op verschillende waarden (trekkingen) van b_0, b_1 en s_e

Proper Multiple Imputations geven twee maten van onzekerheid weer:

1. onzekerheid over de kansverdeling van de missing data
   $\Rightarrow$ trekking uit verdeling ontbrekende waarden $P(X_{mis}|Y_{obs})$
2. onzekerheid over de onbekende modelparameters
   $\Rightarrow$ trekking uit verdeling van de parameters

Gebruik zgn. *Bayesian posterior distributions*

Is het nodig beide maten van onzekerheid weer te geven?

De eerste (trekking uit verdeling missing data): *ja*
$\Rightarrow$ is vrij eenvoudig (zelf) te doen in SPSS

De tweede (trekking parameters): *in veel gevallen wel*
Als de dataset groot is en het percentage missing data laag, dan zullen verschillen niet zo groot zijn (Allison, 2001)
$\Rightarrow$ niet eenvoudig zelf te doen

Twee algoritmes:

- Data augmentation (*NORM*, Solas, SAS, Lisrel)
- Sampling importance/resampling (Amelia, SAS)

Multivariate Normale Model

Nodig voor multipele imputaties: *imputatiemodel*

Tot nu toe (in voorbeeld): regressiemodel

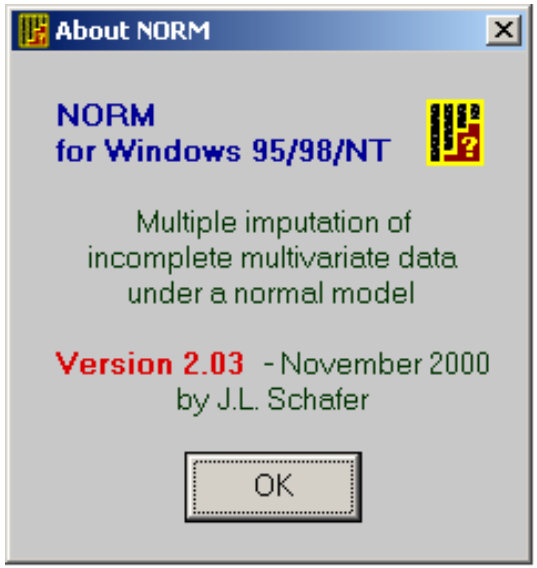
Meest gebruikte model: *multivariate normale verdeling*

- alle variabelen univariaat normaal verdeeld
- elke variabele te schrijven als lineaire functie van de andere

Ook voor *niet-normale data*:

- transformaties kunnen variabelen 'normaler' maken
- niet-normale variabelen zijn volledig geobserveerd
- multipele imputatie (waarschijnlijk) robuust tegen schendingen van het imputatiemodel als percentage missing data (informatie) laag is (Schafer, 1997)

Multivariate Normale Model



NORM (Schafer, 1997, 2000)
Gebaseerd op

- multivariate normale verdeling
- *data augmentation*

Vrij verkrijgbaar op internet

Voorbeeld

y en x bivariaat normaal verdeeld: $\mu = \begin{pmatrix} 10 \\ 12 \end{pmatrix}$,

missing data voor x: 30% scores MCAR en MAR

Voorbeeld: Bivariate Normale model

Tot nu toe imputeren met (MCAR situatie):

$$\widehat{x}_i = b_0 + b_1\, y_i + s_e\, z_i, \text{ met } z_i \sim N(0,1),$$

waarbij $b_0 = 7.590$, $b_1 = 0.440$ en $s_e = 0.795$ worden vastgelegd

Dit imputatiemodel gebruikt NORM ook: *bivariate normale model*

Voor *proper multiple imputations*:
trek b_0, b_1 en s_e uit hun *posterior distribution*

$\Rightarrow$ *Data Augmentation*:
techniek om die verdeling te simuleren en daaruit trekkingen te genereren (MCMC: Markov Chain Monte Carlo)

Data Augmentation

Iteratief proces om verdelingen te simuleren
Twee stappen die afwisselend worden uitgevoerd:

I-stap: *Imputeer de missing data* door trekkingen uit hun (posteriori) verdeling, gegeven de geobserveerde data en de huidige waarden van de parameters: $P(X_{mis}|Y_{obs}, \theta)$
$\Rightarrow$ regressiemodel

P-stap: *Simuleer nieuwe waarden voor de parameters* door deze te trekken uit hun (posteriori) verdeling, gegeven de geobserveerde data en de geïmputeerde missing data

Dit proces geeft (simuleert) de gezamenlijke verdeling van de missing data en parameters

Voorbeeld: Data Augmentation in NORM

1. Bepaal *startwaarden voor de paramaters* van het bivariate normale model: μ en $\Sigma \Rightarrow$ hieruit volgen b_0, b_1 en s_e

   EM levert goede startwaarden (voor MCAR situatie):

   $$\widehat{\mu} = \begin{pmatrix} 9.96 \\ 11.98 \end{pmatrix}, \widehat{\Sigma} = \begin{pmatrix} 1.006 & 0.4432 \\ 0.4432 & 0.8081 \end{pmatrix}$$

2. Gebruik de huidige waarden van de parameters μ en Σ om de *regressiecoëfficiënten te berekenen*:
   $\widehat{x}_i = 7.589 + 0.441\, y_i + 0.780\, z_i$, met $z_i \sim N(0, 1)$

3. **I-stap**: *Imputeerde de missing data* met het regressiemodel (trekking uit posteriori verdeling)

Voorbeeld: Data Augmentation in NORM

4. **P-stap**: Gegeven de geobserveerde data en de geïmputeerde data, *trek nieuwe waarden voor de parameters* μ en Σ uit hun posteriori verdeling
5. Ga terug naar stap 2, gebruik de nieuwe parameterschattingen (trekkingen) om het regressiemodel te berekenen

   Herhaal dit proces tot convergentie

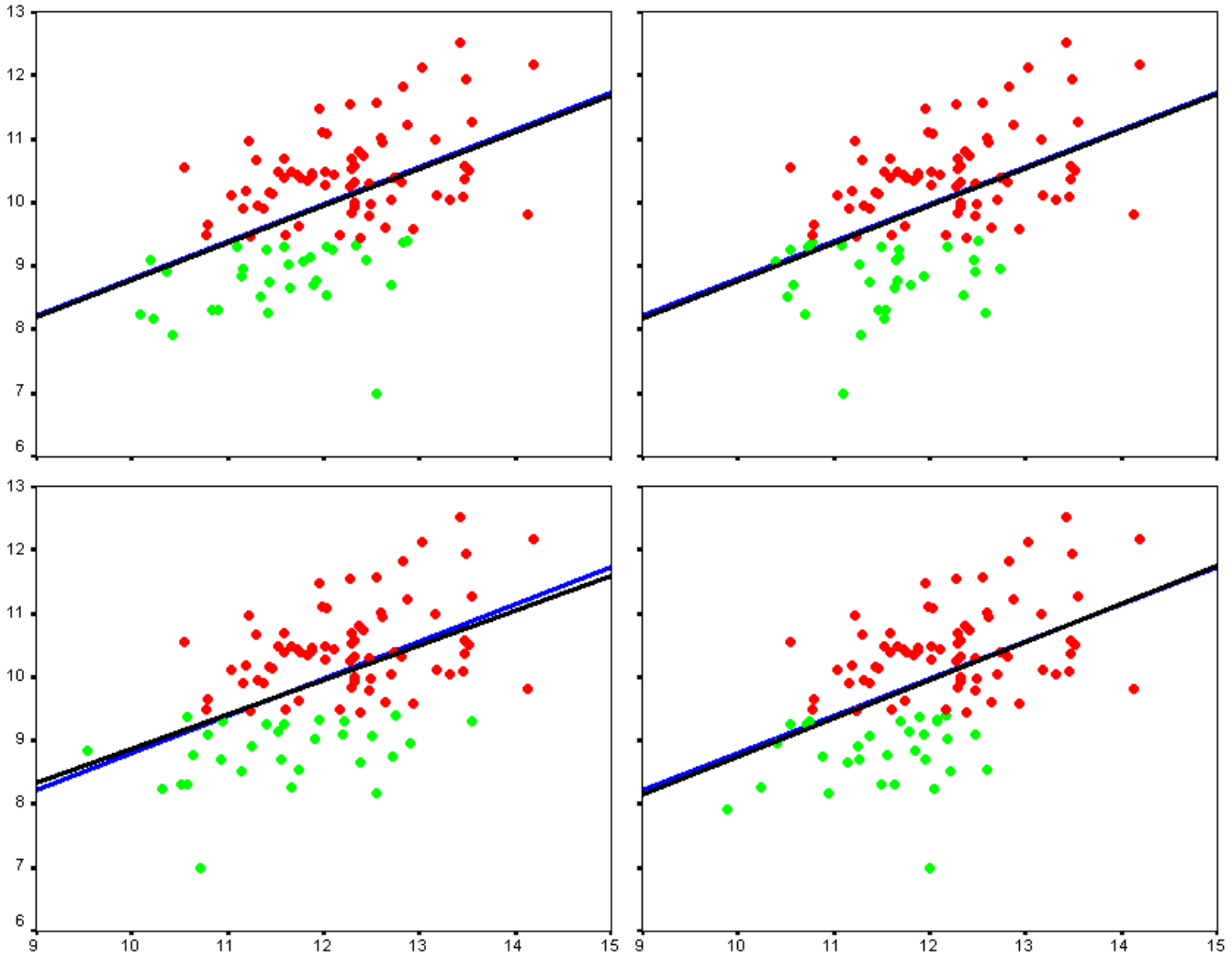
In stap 3 worden de parameters opgevat als 'ware' waarden
In stap 4 worden imputaties beschouwd als 'ware' observaties
$\Rightarrow$ Daarom is de procedure iteratief:
convergeert naar gezamenlijke verdeling data en parameters
Controle op convergentie is belangrijk

Voorbeeld: Resultaten

NORM: 30% data MAR, $m = 10$ en $n = 100$

$\bar{x}$	$SE(\bar{x})$	$r(y,x)$	$b_1(y,x)$	$SE(b_1)$
12.04	0.086	0.497	0.580	0.102
12.03	0.084	0.491	0.591	0.106
12.01	0.096	0.517	0.543	0.091
12.03	0.085	0.502	0.599	0.104
12.17	0.083	0.309	0.374	0.117
11.98	0.091	0.555	0.613	0.093
12.09	0.085	0.494	0.586	0.104
11.96	0.094	0.588	0.631	0.088
11.97	0.092	0.563	0.617	0.091
12.06	0.086	0.453	0.531	0.106
12.03	0.110	0.497	0.566	0.127
11.98	0.089	0.516	0.583	0.098

(Samenvattingen met Excel spreadsheets)


13
12
11
10
9
8
7
6
9
10
11
12
13
14
15

Voorbeeld: Resultaten

Regressie: $\widehat{y}_i = b_0 + b_1\, x_i$

$b_1 = \sum_i^m b_{1i} = 0.566$

$SE(b_1) = \sqrt{\bar{U}_m + (1 + \frac{1}{m})B_m} = \sqrt{0.0101 + 1.1 \times 0.0056} = 0.127$

variantie binnen datasets $\bar{U}_m = 0.0101$

variantie tussen datasets $B_m = 0.0056$

$\Rightarrow$ relative toename in variantie veroorzaakt door missing data:

$$r_m = \frac{(1 + \frac{1}{m})B_m}{\bar{U}_m} = \frac{1.1 \times 0.0056}{0.0101} = 0.61$$

Voorbeeld: Inferenties

Inferenties worden gebaseerd op de *t-verdeling*:
de betreffende gestandaardiseerde parameter heeft bij benadering een t-verdeling met $\nu = (m-1)(1+\frac{1}{r_m})^2$ *vrijheidsgraden* (Rubin, 1987)
Voor de regressiecoëfficiënt b_1 geldt een t-verdeling met $\nu = 63$ vrijheidsgraden

Toets H_0: $\beta_1 = 0$

$$t = \frac{0.566}{0.127} = 4.44,\ p = 0.00004$$

95% Bhi voor β_1

$$0.566 \pm 1.998 \times 0.127 = (0.312, 0.821)$$

Voorbeeld: *Missing Information*

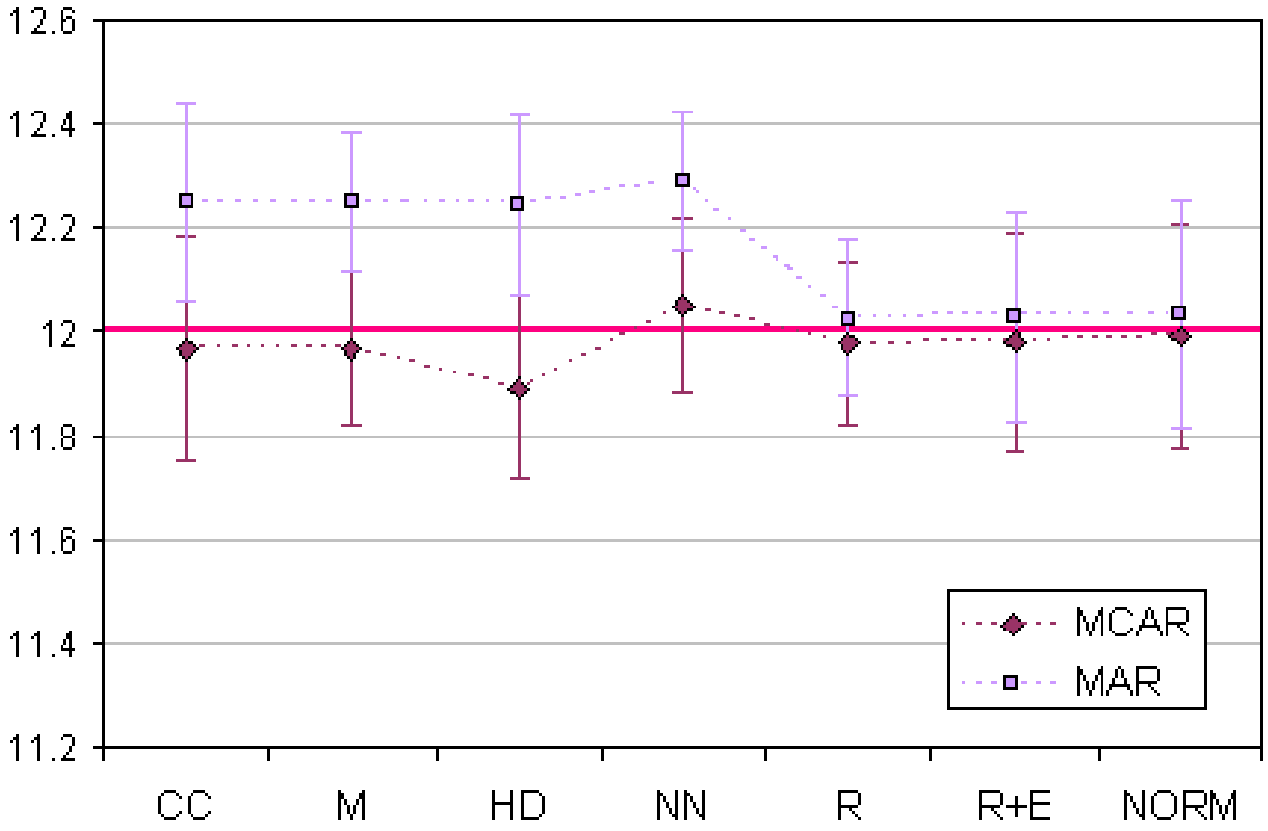
Schatting van *hoeveelheid informatie over een parameter die verloren is gegaan* door missing data (Rubin, 1987)

Gebaseerd op varianties: $\widehat{\gamma}_m = \frac{r_m + 2/(\nu+3)}{r_m+1}$

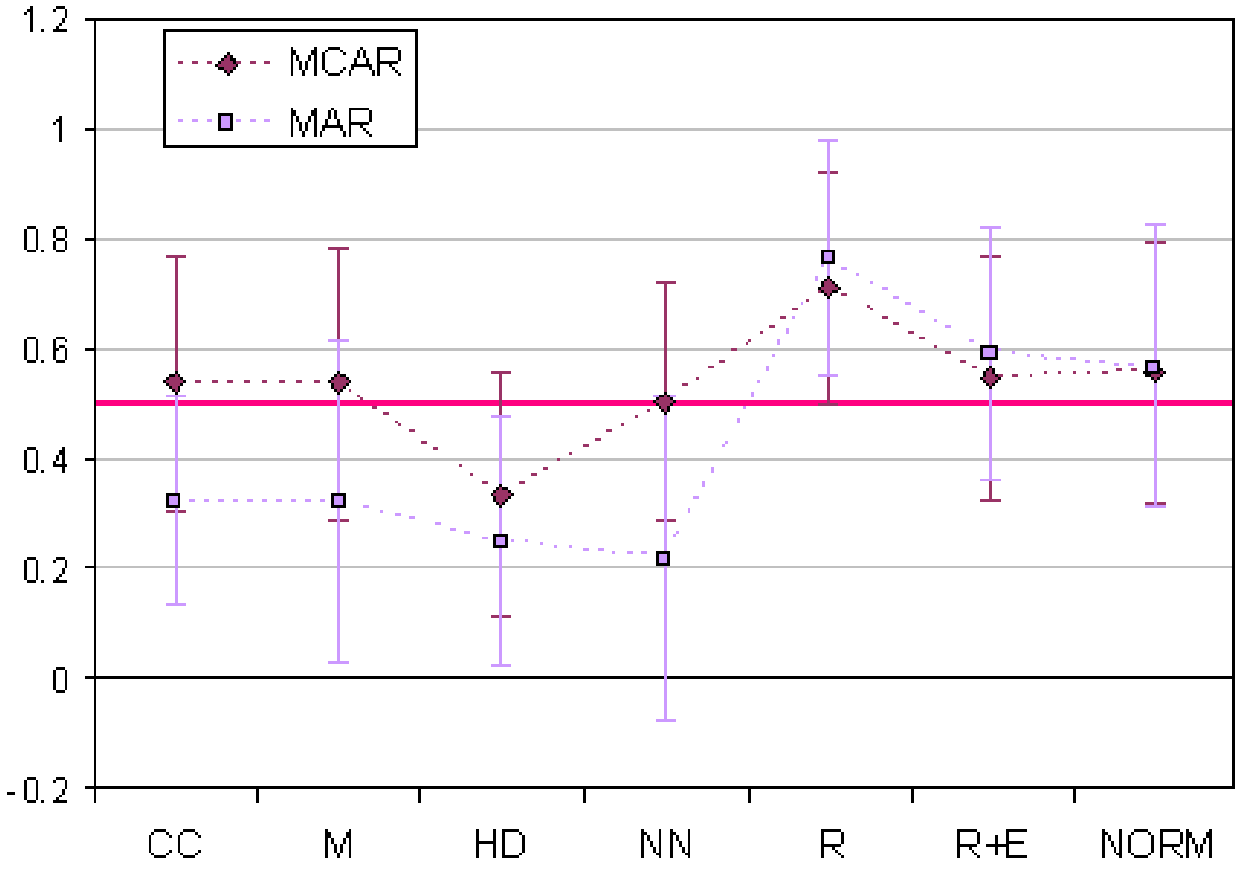
Gebaseerd op de variaties tussen en binnen datasets
Veel variatie is tussen dataset is een indicatie voor een groot verlies van informatie

	$\bar{x}$	$r(y,x)$	$b_1(y,x)$
MCAR	0.30	0.40	0.33
MAR	0.37	0.57	0.39

Voorbeeld: gemiddelde van x



Voorbeeld: regressiecoëfficiënt b_1



Multiple Imputation

Voordelen:

- Er worden complete dataset gegenereerd, die kunnen worden geanalyseerd met *standaard technieken*
- Informatie uit het dataverzamelingsproces kan worden gebruikt bij het imputeren
  $\Rightarrow$ *Missing data mechanisme*: MAR, MNAR?

Nadelen:

- Het imputeren van de missing data kan veel werk zijn en/of erg moeilijk zijn
- Het analyseren van de datasets is (veel) meer werk: analyses per dataset en het samenvatten van de resultaten

Discussie

- *Multivariaat model*: interacties en niet-lineaire verbanden?
- *Categorische data*: multinomiaal model (Schafer, 1997)
  Software: S-Plus procedures CAT en MIX (Schafer, 1997)
- *Nonparametrische methoden*: verbanden tussen variabelen?
- *Longitudinale data*

Referenties

Allison, P.D. (2001). *Missing Data* (Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-136). Thousand Oaks: Sage.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Rubin, D.B. (1987). *Multiple Imputation for Nonreponse in Surveys*. New York: Wiley.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J.L. (2000). *NORM. Version 2.03*. `http://www.stat.psu.edu/~jls/`.

Schafer, J.L. & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.