Applied Statistical Analysis with Missing Data
Welcome and Introduction

Morten Frydenberg ©
Section of Biostatistics, Aarhus Univ, Denmark

The teachers, the programme and the participants

Participants

The birth weight data sets

The principles of statistical inference
- The likelihood method
- Estimates, standard errors,
  confidence intervals and tests
- Bias, coverage probabilities and efficiency

Morten Frydenberg Missing data - lecture 1 1

Why are data missing?

Inference ignoring the missing data problem

Different types of missingness

How to attack the missing problem

The Multiple Imputation procedure – an outline

A case study – from the last course

The drug study data

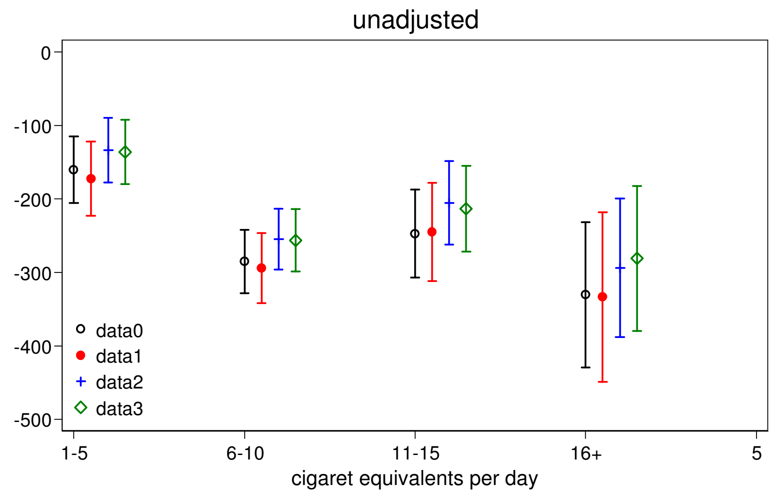
Morten Frydenberg Missing data - lecture 1 2

The birth weight data sets

Obs Variable	data set 0	1	2	3
id	10000	10000	10000	10000
sex	10000	10000	10000	10000
age	10000	8972	10000	10000
bweight	10000	9011	8914	10000
bmi4who	10000	8990	10000	10000
parity	10000	9065	10000	10000
cigs	10000	8979	10000	9468
nausea	10000	8929	10000	10000
alcohol	10000	9014	10000	10000
Model 0	10000	8090	8914	9468
Model 1	10000	4771	8914	9468

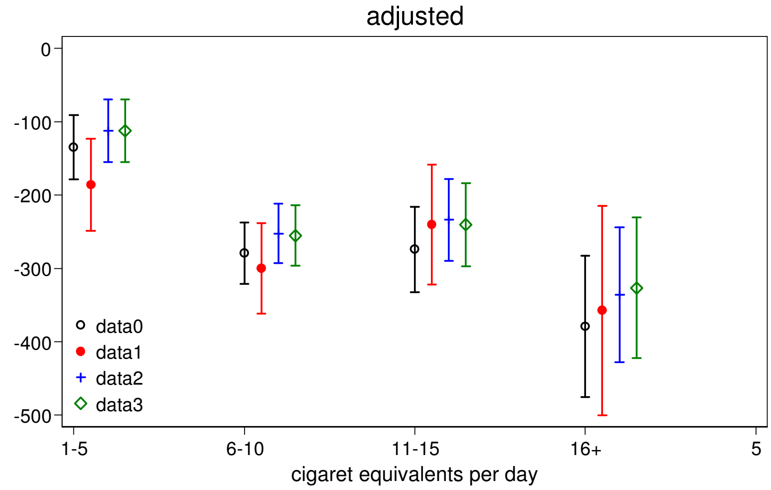
missing Variable	data set 0	1	2	3
id				
sex				
age		10%		
bweight		10%	11%	
bmi4who		10%		
parity		9%		
cigs		10%		5%
nausea		11%		
alcohol		10%		
Model 0		19%	11%	5%
Model 1		52%	11%	5%

Morten Frydenberg Missing data - lecture 1 3



Cigs	Data0	Data1	Data2	Data3
0				
1-5	-160 (-206;-115)	-172 (-223;-122)	-134 (-178;-90)	-136 (-180;-92)
6-10	-285 (-328;-242)	-294 (-342;-246)	-255 (-296;-213)	-256 (-299;-214)
11-15	-247 (-307;-187)	-245 (-312;-178)	-205 (-262;-148)	-213 (-272;-155)
16+	-330 (-429;-231)	-334 (-449;-218)	-294 (-388;-200)	-281 (-379;-182)
n	10000	8090	8914	9468

Morten Frydenberg Missing data - lecture 1 4



Cigs	Data0	Data1	Data2	Data3
0				
1-5	-135 (-179;-91)	-186 (-249;-123)	-113 (-155;-70)	-113 (-155;-70)
6-10	-280 (-321;-238)	-300 (-362;-238)	-252 (-293;-212)	-256 (-297;-214)
11-15	-274 (-332;-216)	-240 (-322;-159)	-234 (-290;-178)	-240 (-297;-184)
16+	-379 (-475;-283)	-358 (-500;-215)	-336 (-428;-244)	-327 (-422;-231)
n	10000	4771	8914	9468

Morten Frydenberg Missing data - lecture 1 5

The principles of statistical inference

In very general terms the purpose of statistical inference is to estimate the relationship between a response/outcome Y and a set of explanatory variables X.

In order to do that we specify a **statistical model** of the relationship between Y and X.

The model will typically contain a **systematic** and a **random** part with corresponding unknown constants – **parameters**.

Systematic part

Random part

$$y = f\left(x;\beta\right) \text{ "+" } e\left(\sigma\right)$$

Unknown Parameters

Unknown Parameters

Morten Frydenberg Missing data - lecture 1 6

The principles of statistical inference

$$y = f(x;\beta) \text{ "+" } e(\sigma)$$

We will in the course denote the **combined data** by $Z = (Y, X)$.
and combined set of parameters by $\theta = (\beta, \sigma)$.

That is, Z will contain both outcome and explanatory variables, and θ will contain parameters concerning the systematic and the random part of the model.

We will use the data Z **to make inference** concerning θ.

I.e. find
estimates, standard errors, confidence intervals
and calculate
test-statistics and p-values
for relevant hypotheses.

Morten Frydenberg Missing data - lecture 1 7

The maximum likelihood method

If we know θ, then we can calculate the probability of z

$$p_{Z;\theta}(z;\theta) = \text{Probability of } Z = z \text{ if parameter} = \theta$$

Most statistical inference is based on **the likelihood function**:

$$L(\theta \mid z) = p_{Z;\theta}(z;\theta)$$

I.e. we consider z fixed and then see how the probability varies as we vary θ.

The maximum likelihood estimate of θ is
the value of θ that maximizes the likelihood,
i.e.
the value of θ that maximizes the **probability** of observing the data, we actually did observe.

Morten Frydenberg Missing data - lecture 1 8

The maximum likelihood method

Most statistical inference is based on **the likelihood function**:

$$L(\theta \mid z) = p_{Z;\theta}(z;\theta)$$

Based on the likelihood function one can calculate:

- The **M**aximum **L**ikelihood **E**stimate (MLE) of θ
- Approximate **S**tandard **E**rror of the MLE
- Approximate **C**onfidence **I**ntervals (based on MLE and SE)
- Approximate (Wald) **tests** of hypotheses concerning θ

But the method requires that "you" can calculate the probability of the data (those that are on your hard disc!) given the parameters.

The principles of statistical inference
Bias, coverage probability, and efficiency

Let us denote the **true**, but unknown, parameter value by θ_T

And let $\hat{\theta}_n$ denote an estimator (it could be the MLE) based on n **independent** observations.

The estimate is said to be

unbiased if: Expected value of $\hat{\theta}_n = \theta_T$

asymptotically unbiased if: $\hat{\theta}_n \to \theta_T$ as $n \to \infty$

That is, the estimate gets ever closer to the true value as we get more observations.

Very few estimates are unbiased (notable exception: normal multiple regression), but MLE's are in general asymptotically unbiased.

The principles of statistical inference
Bias, coverage probability, and efficiency

The **coverage probability** of a Confidence Interval $(L_n; U_n)$ is

$$\Pr\left(L_n < \theta_T < U_n\right)$$

i.e. the probability that the true parameter is contained in the interval.

Ideally a 95% CI will have a coverage probability = 95%.

In practice very few methods will give exactly the stated coverage probability (normal multiple regression), and we will have to hope for **asymptotically correct** coverage probabilities (large data set – correct coverage).

Confidence intervals based on the MLE and the ML SE have in general asymptotically correct coverage probabilities.

Morten Frydenberg Missing data - lecture 1 11

The principles of statistical inference
Bias, coverage probability, and efficiency

The **width** of a Confidence Interval is $U_n - L_n$

When comparing two (asymptotically) **unbiased** methods of estimation, the method with the **smallest average width** of the confidence interval is said to be the **most efficient**.

In general we prefer :

- estimates that are (asympt.) **unbiased**
- CIs that have (asympt.) **correct coverage probabilities**
- methods that are (asympt.) the **most efficient**

Note:
High **efficiency** corresponds to higher statistical power.
High **efficiency** corresponds to small SE.

Morten Frydenberg Missing data - lecture 1 12

What do we observe/record when data are missing?

We do not fully observe Z, but rather:

Z_{Obs}

id	y	x_1	x_2	x_3
1	o	o	o	o
2	o	.	o	o
3	.	o	o	o
4	.	.	o	o
5	o	o	o	o
6	o	.	.	o

o observed

$R_{ecorded}$

id	y	x_1	x_2	x_3
1	1	1	1	1
2	1	0	1	1
3	1	1	1	1
4	0	0	1	1
5	1	1	1	1
6	1	0	0	1

Morten Frydenberg Missing data - lecture 1 13

Missing data

Avoid missing data!!!

If not, then collect as much information on the reason why the observation became missing:

- did the patient refuse to participate?
- is the patient dead?
  Is this missing data?
- did the patient not turn up?
- was the "measurement" never made?
- was the result not registered?

→ **Why???**

- was the value below the detection limit?
  Is this missing data?

Morten Frydenberg Missing data - lecture 1 14

Missing data – Solutions??

Complete case analysis - version one:
Ignore the problem and only analyse **patients** with information on all relevant variables!!!

Pros: Always possible – transparent model

Cons: Model often wrong
Estimate likely biased
Analysis likely inefficient

Complete case analysis - version two:
Ignore the problem and only use **variables** that are available for all patients!!!

Pros: Always possible - transparent model

Cons: Wrong/irrelevant model
Biased estimate!!

Different types of missingness

Three types of problems:

1 **The Easy:** MCAR
Complete case analysis version one will give an unbiased estimate of θ.

2 **The Tough:** MAR
It is possible (in theory) to get an unbiased estimate of θ by analysing the observed data correctly.

3 **The Unsolvable:** MNAR
It is impossible to get an unbiased of θ based solely on information in the observed data.

If the reason/mechanism behind the missingness is not known, then it is impossible to distinguish between situation 2 and 3.

How to attack the problem with missing data

First try to determine whether you are in situation 1, 2 or 3, by going through the possible missing data mechanisms.

If you are in **situation 1** then a complete case analysis will be valid, although not the most efficient.

If you are in **situation 3** then you **have to make additional** assumptions concerning the missing data in order to analyse the observed data. Specific modelling will be required and you will typically **not be able use standard programs**.

If you are in **situation 2** then you **might solve** the problem by **imputation** methods.

In all cases you should supplement your analysis with different types of **sensitivity analyses**.

Morten Frydenberg Missing data - lecture 1 17

The analyses of data with missing values by imputation

The analysis using imputation have 3 separate components:

1. **The complete data model**
   Specification of how to analyse the data, if it was without missing values.
2. **Imputing the missing values**
   Generate K complete data sets by generating K values of the missing data.
3. **The estimation**
   Find K estimates of θ - one for each of the K 'complete' datasets. The final, overall estimate of θ is found as the average of the K estimates. Calculate a suitable standard error.

Morten Frydenberg Missing data - lecture 1 18

A case study – from the last course*

Design

Patients treated with Percutaneous Coronary Intervention followed over three years by 8 questionnaires and in national registers.

Purpose

To estimates the Quality of Life after PCI.
To compare the QoL after PCI in predefined subgroups, given by sex, age, education…

Some data is missing!

Despite an initial response rate of 83% only 417 out of 1726 patients had **complete data** on all measure points and covariates.

*Joint work with Karin Biering and Niles Henrik Hjøllund

Tables:

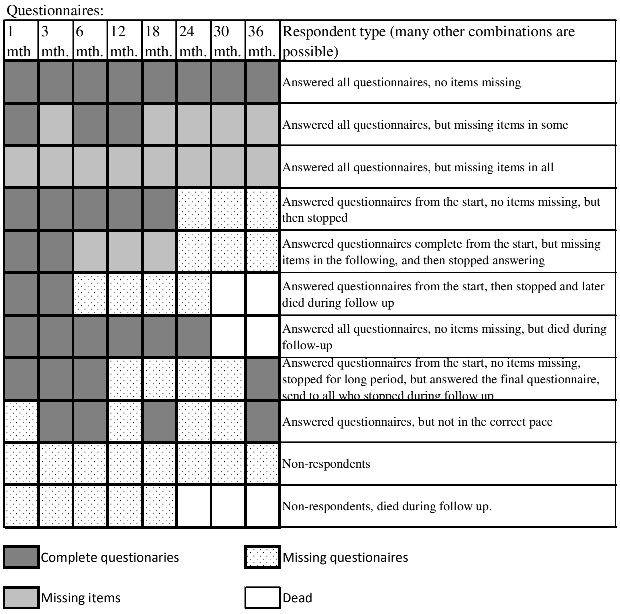
Table 1: Response patterns and attrition in a cohort of patients treated with PCI at Aarhus University Hospital, Skejby (N=1726)

	1 mth.	3 mth.	6 mth.	12 mth.	18 mth.	24 mth.	30 mth.	36 mth.
Overall mortality		5	5	9	15	14	14	12
Alive in current round	1726*	1721	1716	1707	1692	1678	1664	1652
From previous round	-	1323	1112	1057	1012	980	954	892
- Attrition #	262	211	55	45	32	26	62	39
= Available for next round	1323	1112	1057	1012	980	954	892	
- Intermittent missing questionnaire**	29	8	31	53	64	73	53	-
= Returned questionnaires	1294	1104	1026	959	916	881	839	853**
Responserate according to previous round	-	83.4%	92.2%	90.7%	90.5%	89.9%	87.9%	95.6%
SF-12 PCS/MCS								
Complete	1144	979	945	899	858	827	783	780**
Incomplete	150	125	81	60	58	54	56	73
Seattle Angina Questionnaire (frequency dimension)								
Complete	-	1046	1007	888	798	728	682	731**
Incomplete	-	58	19	71	118	153	157	122
Seattle Angina Questionnaire (stability dimension)								
Complete	-	1056	1015	891	805	738	690	736**
Incomplete	-	48	11	68	111	143	149	117

* 141 patients had hidden addresses and were not sent questionnaires.
\# Dead and non-respondents after 2 reminders
** Intermittent missing questionnaire in first round occurred when the first questionnaire was delayed from the patient until second time point 3 months after PCI. The following intermittent missing occurred because all patients who stopped answering during follow-up without any known reason was mailed a final questionnaire. This resulted in an increase in returned questionnaires in the final round.

Figure 1: Exemplified overview of some respondent types in a follow up study of PCI patients.

Questionnaires:



More Missing data

	All		Respondents						Non-respondents	
			Whole study course		With dropout		Returnees			
	N		N		N		N		N	
Total	1726	100%	761	100%	470	100%	92	100%	403	100%
Gender										
Male	1360	79%	612	80%	364	77%	76	83%	308	76%
Female	366	21%	149	20%	106	23%	16	17%	95	24%
Age										
-44 years	168	10%	28	4%	53	11%	12	13%	75	19%
45-54 years	476	28%	183	24%	139	30%	38	41%	116	29%
55-59 years	393	23%	176	23%	115	24%	24	26%	78	19%
60-67 years	689	40%	374	49%	163	35%	18	20%	134	33%
Indication										
Acute	557	32%	233	31%	157	33%	32	35%	135	33%
Elective	1169	68%	528	69%	313	67%	60	65%	268	67%
Comobidity										
Charlson Index 0	1010	59%	476	63%	259	55%	54	59%	221	55%
Charlson Index 1	393	23%	169	22%	106	23%	26	28%	92	23%
Charlson Index 2+	323	19%	116	15%	105	22%	12	13%	90	22%
Left Ventricular Ejection Fraction										
-34%	89	5%	30	4%	17	4%	1	1%	41	10%
35-54 %	612	35%	242	32%	197	42%	35	38%	138	34%
55+ %	895	52%	429	56%	226	48%	47	51%	193	48%
Missing	130	8%	60	8%	30	6%	9	10%	31	8%
Smoking										
Never	330	19%	186	24%	67	14%	16	17%	61	15%
Current	763	44%	272	36%	228	49%	40	43%	223	55%
Previous	597	35%	302	40%	164	35%	36	39%	95	24%
Missing	36	2%	1	0%	11	2%	0	0%	24	6%

	All		Respondents: Whole study course		Respondents: With dropout		Respondents: Returnees		Non-respondents	
	N		N		N		N		N	
Total	1726	100%	761	100%	470	100%	92	100%	403	100%
Body Mass Index										
-24.9 kg/m^2	485	28%	230	30%	126	27%	22	24%	107	27%
25-29.9 kg/m^2	774	45%	357	47%	207	44%	51	55%	159	39%
30+ kg/m^2	425	25%	173	23%	121	26%	19	21%	112	28%
Missing	42	2%	1	0%	16	3%	0	0%	25	6%
Physical activity										
<2 h/wks	96	6%	52	7%	39	8%	5	5%		0%
2-4 h/wks	402	23%	277	36%	91	19%	34	37%		0%
>4 h/wks, light	480	28%	352	46%	85	18%	43	47%		0%
>4 h/wks, heavy	82	5%	61	8%	14	3%	7	8%		0%
Missing	666	39%	19	2%	241	51%	3	3%	403	100%
Education level										
Low (<11 y)	253	15%	152	20%	66	14%	9	10%	26	6%
Intermediate (11-14 y)	742	43%	278	37%	205	44%	41	45%	218	54%
High (15+ y)	561	33%	304	40%	139	30%	41	45%	77	19%
Missing	170	10%	27	4%	60	13%	1	1%	82	20%

Why are data missing?

Several very different possible explanations!

Unknown address –"forskerbeskyttelse".

Related or unrelated to the focus of the study?

Did not return **a specific** questionnaire.

Related or unrelated to QoL?

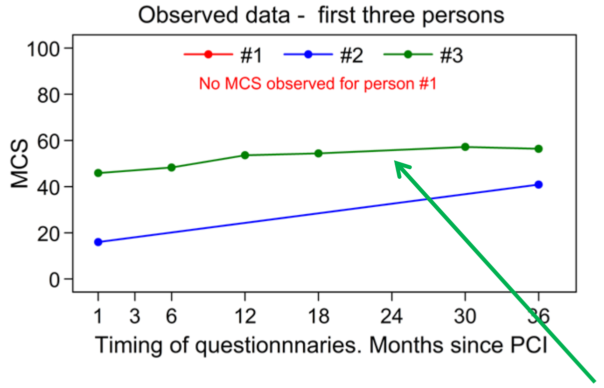
(in general or at that specific point in time?)

Related or unrelated to health?

(in general or at that specific point in time?)

Stopped returning questionnaires:

Related or unrelated to QoL, health, comorbidity, age, sex, education?



A missing Mental Component Score (SF-12 QoL) score at **24** months?

Missing **C**ompletely **A**t **R**andom?

No, we don't think so!

Could easily be related to health, general QoL and age.

But We know the age!
We have information on comorbidity at start!
We know the QoL at 1,3…, 18, 30 and 36 months.

So **maybe** - **given this information** – we have independency between QoL at 14 months and reporting it. That is: **M**issing **A**t **R**andom.

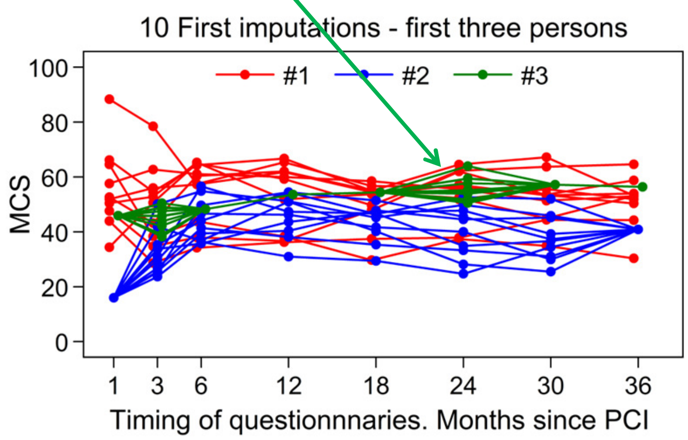
But if missingness is related to **current health status at 6 months**, then we do not have missing at random!

If this is the case we need information on current health status to obtain **MAR**!

In our study we collected **additional data** with information on social benefits on weekly basis (DREAM), with the sole purpose of making **MAR plausible.**

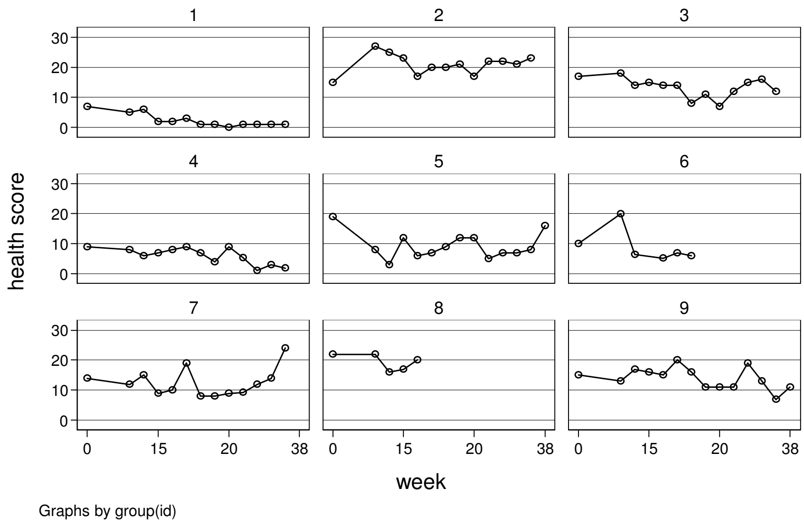
A possible model: $MCS_t : MCS_{t-1} + MCS_{t+1} + SAQs_{t-1} + SAQs_t + SAQf_{t-1} + SAQf_t + TPG_t + Age + Sex + LVEF + Indication + Comorbidity$

Simulated/imputed values = guesses of what the unobserved MCS values would have been if observed:

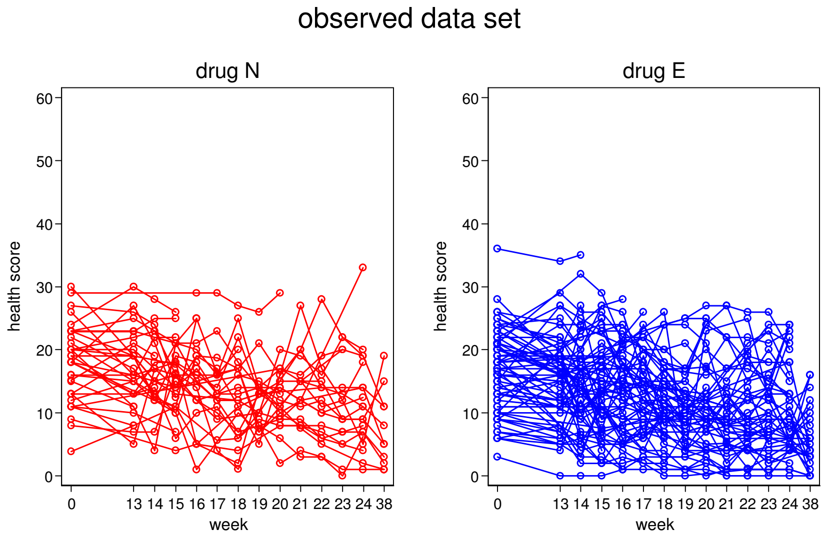


Morten Frydenberg Missing data - lecture 1 27

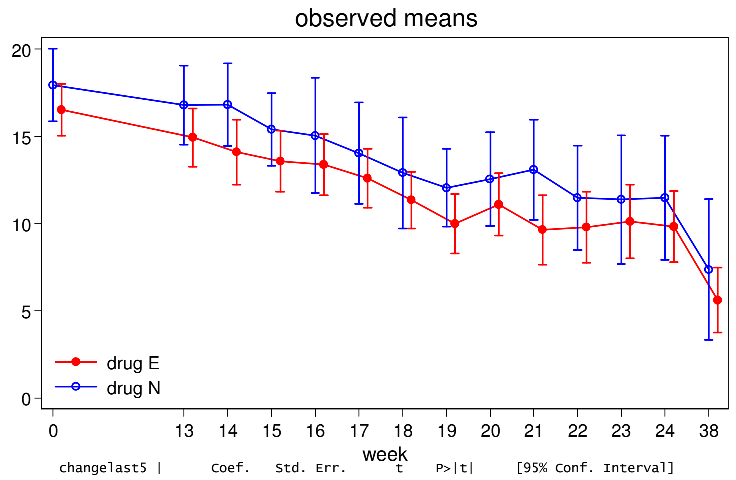
The drug study data



Morten Frydenberg Missing data - lecture 1 28



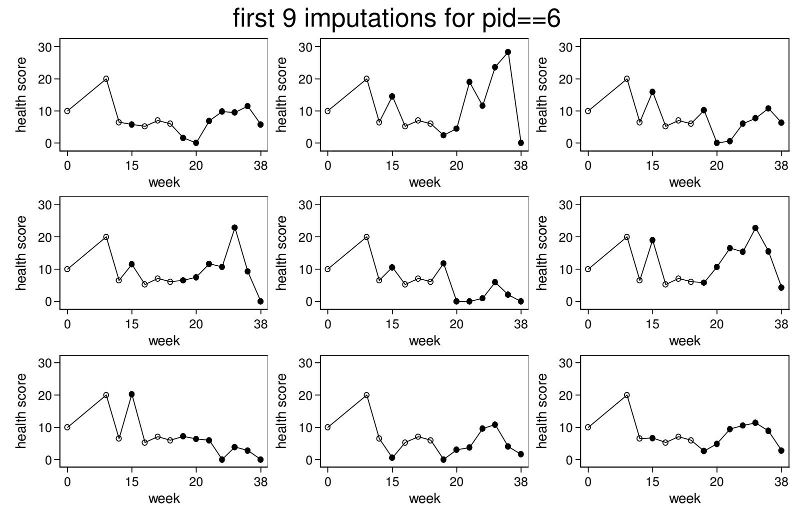
Why are the scores missing?



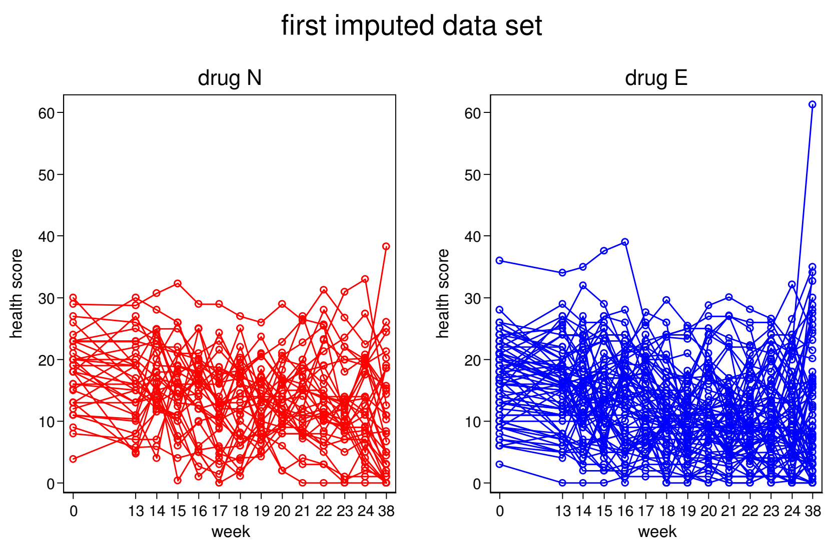
changelast5	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
drug 0	(base)					
1	**1.279789**	**2.810969**	**0.46**	**0.653**	**-4.49824**	**7.057818**
sex 1	(base)					
2	-2.159583	2.664323	-0.81	0.425	-7.636177	3.317012
duration	.0127123	.0766663	0.17	0.870	-.1448775	.1703021
_cons	-7.327105	3.429444	-2.14	0.042	-14.37643	-.2777814

N=30

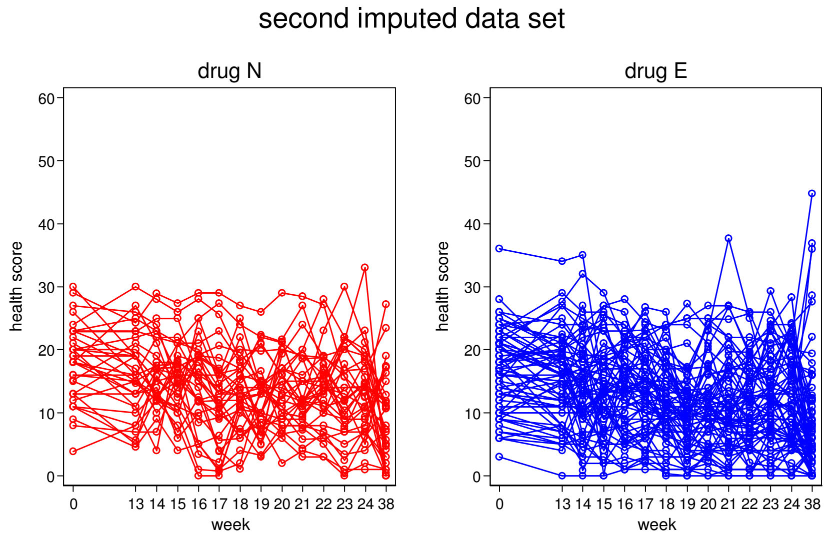
Can we impute (generate values) of the missing scores?



Morten Frydenberg Missing data - lecture 1 31



Morten Frydenberg Missing data - lecture 1 32



How can we combine the results based on the imputed scores?