Statistical modelling with missing data using multiple imputation

Lecture 3: More on multiple imputation

James Carpenter
London School of Hygiene & Tropical Medicine
Email: james.carpenter@lshtm.ac.uk
www.missingdata.org.uk

Support from MRC, ESRC & German Research Foundation

November 17, 2009

Overview

Lecture 3: outline

1. Motivation: two examples when MI makes a difference
2. More formal justification for MI
3. Further comments
4. Algorithms for MI
5. Likely difference between MI and CC analyses
6. Example: NCDS data
7. Survival data
8. Discussion

2 / 49

Two examples

Relative survival in cancer

A key focus in cancer epidemiology is estimating the relative survival of cancer patients, and exploring how this varies with covariates (not least country).

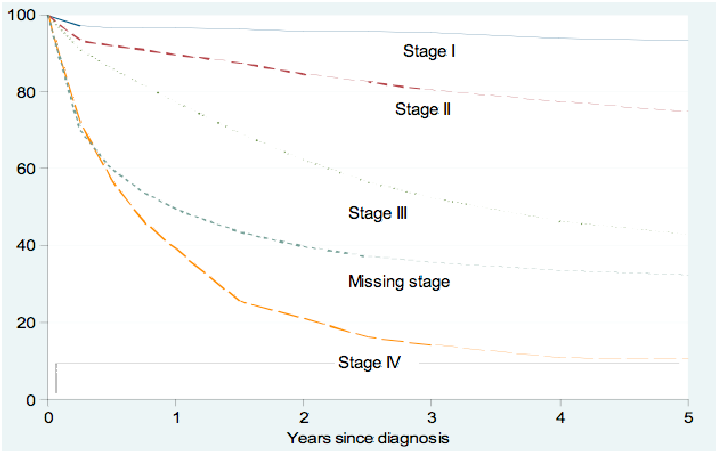
The outcome is thus time from diagnosis to death, the latter usually extracted from registry data.

A key predictor is the ‘stage’ of the cancer at diagnosis, which is an ordinal variable taking values 1,...4. Unfortunately, this is often not recorded/observed. Further, the suspicion is this is related to the severity of the cancer at diagnosis.

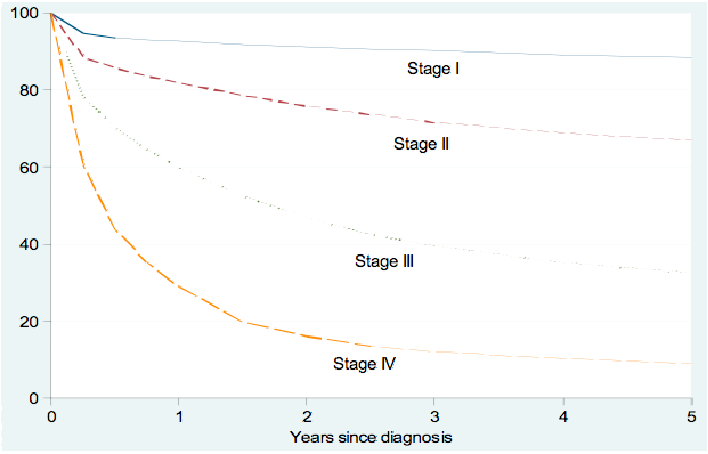
We applied MI to the analysis of survival in 29,563 colorectal cancer patients who were diagnosed between 1997 and 2004 and registered in the North West Cancer Intelligence Service [9].

Incomplete information, mostly on stage, meant that only 55% could be included in a complete case analysis.

3 / 49



4 / 49



5 / 49

Multiple Imputation for Costs

Burton et al [3] used data from a randomised controlled trial to compare the cost-effectiveness of chemotherapy with that of standard palliative care in patients with advanced non-small cell lung cancer. Resource usage data were obtained for a subset of 115 patients, but were complete for only 82 patients.

Patient and tumour characteristics were stated to be comparable in those with complete and incomplete data, but the effect of treatment on survival was stated to differ.

The authors used the MI. Variables included in the imputation models were listed. Five imputed datasets were created. Log and logit transformations were used to deal with non-normality, and a two-stage procedure was used to deal with variables with a high proportion of zero values (semi-continuous distributions). Complete data were transformed back to their original scales prior to analyses being performed.

6 / 49

Results

The complete case analysis resulted in a higher mean cost for chemotherapy compared with palliative care (£2804, 95% CI £1236 to £4290) than did the analyses using multiple imputation (£2384, 95% CI £833 to £3594).

The complete case analyses implied that chemotherapy was not cost-effective (mean net monetary benefit -£3346, but the MI analyses implied that it was cost-effective (mean net monetary benefit £1186), although confidence intervals were wide.

7 / 49

More on MI

8 / 49

Intuition for MI

Recall from this morning, that we divide the data into the 'observed' and 'missing' parts, Z_O, Z_M.

We then proceed as follows:

1. assume that the missing data are MAR (given the observed data);
2. model the data, so that all the partially observed variables are responses;
3. impute the missing data from this model multiple times, taking full account of the variability (i.e. including the uncertainty in estimating the parameters of the imputation model), and
4. fit the model of interest to each 'completed' data set, and combine the results using Rubin's rules.

A more formal intuition for MI is given in some slides at the end of this session.

8 / 49

Further comments

9 / 49

Frequently asked questions

- How many imputations?
  - With 50% missing information, an estimate based on 5 imputations has SD 5% wider than one with an infinite number of imputations. But this isn't the whole story...
- What if not MAR?
  - Most software assumes MAR, but MAR is not necessary for MI.
- Why not compute just one imputation?
  - Underestimates variance, as can't estimate $\hat{\sigma}_b^2$.
- What if I am interested in more than one parameter?
  - Imputation proceeds in the same way, as does finding the overall estimate of θ. However, estimating the covariance matrix can be tricky. Typically more imputations will be needed. See Schafer (1997)[12] for a discussion.

9 / 49

Why 'multiple' imputation?

One of the main problems with the single stochastic imputation methods is the need to develop appropriate variance formulae for each different setting.

Multiple imputation attempts to provide a procedure that can get the appropriate measures of precision relatively simply in (almost) any setting.

Once we choose the imputation model, it proceeds automatically (although an appropriate choice may not always be straightforward).

10 / 49

Structuring the imputation model

In order to do multiple imputation, it suffices to fit a model where partially observed variables are responses, and fully observed covariates.

This is tricky in general!

Thus, people have started with the assumption of multivariate normality, and tried to build out from that. Implicit in that the regression of any one variable on the others is linear.

Skew variables can be transformed to (approximate) normality before imputation and then back transformed afterwards.

With an unstructured multivariate normal distribution, it doesn't matter whether we condition on fully observed variables or have them as additional responses: so most software treat them as responses.

11 / 49

Software taxonomy: methods derived from multivariate normal

	Complexity		
Response type	Normal		Mixed response
Data structure	Independent	Multilevel	Multilevel
Package			
Standalone	NORM	PAN	REALCOM
SAS	NORM-port	—	—
Stata	NORM-port	—	—
R/S+	NORM-port	—	—
MLwiN	MCMC algorithm emulates PAN		+ 1–2 binary

All methods: General missingness pattern; fitting by Markov Chain Monte Carlo (MCMC) or data augmentation algorithm (see references on later slides).

Relationships essentially normal/linear (except MLwiN).

Interactions must be handled by imputing separately in each group.

Schafer has a general location model package, relatively little used.

12 / 49

Chained equations

The 'chained equation' AKA 'full conditional specification' approach approximates a joint model by a series of conditionals, in the manner of the Gibbs sampler.

In this case, though, no unique joint distribution may exist!

However, this can handle non-monotone data, and non-normal data, relatively easily.

Multilevel data problematic, as is data with irregular follow-up, but this is not such an issue for trials data.

13 / 49

Algorithm

1. To get started, for each variable in turn fill in missing values with randomly chosen observed values.
2. 'Filled-in' values in the first variable are discarded leaving the original missing values. These missing values are then imputed using regression imputation on all other variables.
3. The 'filled-in' values in the second variable are discarded. These missing values are then imputed using regression imputation on all other variables.
4. This process is repeated for each variable in turn. Once each variable has been imputed using the regression method we have completed one 'cycle'.
5. The process is continued for several cycles, typically ~ 10.

Comments

The attraction of this approach is that linear regression models can be replaced with GLMs etc. for non-normal responses.

Software

SAS — IVEware

R — mice, mi

Stata — ice

There has been little theory for this approach, but recent work with Ian White and Rachael Hughes (Bristol) has found a condition for ICE to work, and suggests when this condition doesn't hold, errors are likely to be very small.

Need to be careful about interactions and non-linearities: are all our models consistent?

Joint modelling versus chained equations

Over the last few months there have been a number of simulations studies comparing the two approaches, which are currently working their way through the publication process.
These indicate that the joint normal model does very well for binary and ordinal data, especially when

- ordinal variables are transformed to approximate normality before imputation
- adaptive rounding is used to convert imputed non-integer values to integer values

The attraction of chained equations is that categorical data can be handled more readily.

However, simulations suggest that chained equations suffers from bias — most likely due to overfitting — with large numbers of variables. Joint normal modelling seems considerably more robust to this.

Joint modelling can naturally handle multilevel (and other) structures.

Some MI references

Allison (2000)[1] — a cautionary tale!

Allison (2002)[2] — nice monograph, very cheap!

Horton and Lipsitz (2001)[6] — Comparison of software packages.

Kenward & Carpenter (2007) [7] — up-to-date review.

Rubin (1987)[11] — The original souce; this book brings together the theory in a 'fairly accessible' way.

Rubin(1996)[10] — review of the use of MI after $\sim$ 18 years.

Schafer (1997)[12] — Key book giving details of data augmentation, MCMC algorithms and MI methods in many models.

Schafer & Graham (2002)[13] — nice overview

17 / 49

What we haven't covered

1. If we can formulate our imputation model in terms of our parameters of interest, then Maximum Likelihood and MI will approximately agree.

   Thus, many longitudinal trials with missing data can be analysed without using MI. See [5], chs 3, 4

   - In general, Generalised Estimating Equations (GEEs) only give valid inference if data are MCAR (as they are moment based estimators). This is more of an issue for discrete data. Options in [5], ch 5.
2. So far done multiple imputation under MAR: multiple imputation under MNAR this afternoon.

18 / 49

MI vs CC analysis

19 / 49

Correcting bias: missing response values

Consider a regression of Y on two covariates X, Z

Suppose only Y has missing data

CC (Complete Cases) will be unbiased when:

- Y MCAR
- Y MAR given X, Z.
- Y MAR given some W, but W independent of $[Y, X, Z]$.

CC biased when

- Y MAR given W, and W dependent on $[Y, X, Z]$.
- Y MNAR

Implication: Variables predictive of Y being missing, and associated with variables in the analysis, should be included in the imputation model.

19 / 49

Correcting bias - missing covariate values

Consider a regression of Y on two covariates X, Z

Suppose only X has missing data

CC will be unbiased when:

- X is MCAR
- X is MAR given Z (but not Y)
- X is MAR given some W, but W independent of $[Y, X, Z]$.
- X is MNAR (dependent on X, possibly Z, but not Y)

20 / 49

Missing covariate values (ctd)

CC biased when

- X MAR, and mechanism depends on Y
- X is MAR, and mechanism depends on some W, and W not independent of $[Y, X, Z]$.

Implication: Variables predictive of X being missing, and associated with variables in the model, should be included in the imputation model.

Warning: If covariates MNAR (mechanism unrelated to response), then MI may be biased (since it requires MAR to be unbiased) while CC would not be.

More discussion in White & Carlin (2009) (under review with Statistics in Medicine)

21 / 49

When is bias correction most likely with MI?

We assume that we have variables such that data are MAR.

In general the simpler the model of interest, the more likely that we have omitted a variable predictive of missingness, and correlated with response and covariates. Thus the more likely the CC analysis is biased.

The simplest 'model' is the sample mean, sample variance etc.

Example
In clinical trials with partially observed longitudinal follow-up, marginal means are often very biased.

Suppose now the response is MAR given treatment, baseline response and baseline age.

As we bring these terms into the model we reduce the bias.

22 / 49

Recovering information

Even if the CC analysis is approximately unbiased, MI can recover information.

Given the cost of collecting the data, versus the cost of MI, this alone is sufficient to justify its use.

With MI, broadly speaking, information is recovered through two routes:

1. bring cases with response and almost all variables observed into analysis, and
2. bring in information on missing values through additional variables correlated with them.

Implication: Include variables predictive of partially observed variables in the imputation model (even if they are not predictive of missingness).

Warning: If the principal missing data patterns have a missing response, information only comes in by route (2) above.

23 / 49

Example: NCDS data

24 / 49

NCDS data: model of interest

With MI, to get valid answers we need to think hard about the imputation model. This needs to be at least as general as the model of interest.

Consider the following model for the 1956 NCDS birth cohort:

$$\begin{aligned}\texttt{logit}\{\Pr(\text{child has no educational qualifications at 23 years})\} =& \beta_0 + \beta_1 (\text{in care before age 7}) \\ &+ \beta_2 (\text{in social housing before age 7}) \\ &+ \beta_3 (\text{inverse birthweight}) \\ &+ \beta_4 (\text{mother's age}) \\ &+ \beta_5 (\text{mother's age})^2 \\ &+ \beta_6 (\text{mother's age}) \times (\text{social housing}) \\ &+ \beta_7 (\text{mother's age})^2 \times (\text{social housing})\end{aligned}$$

24 / 49

Analyses

Besides the complete case analysis, we will consider the following MI analyses, fitted using chained equations:

- MI I: imputation model without non-linear or interaction terms;
- MI II: imputation model with non-linear terms as far as possible, together with additional mid-life variables (behavioural score and number of family moves)
- MI III: separate impuation in the two social housing categories (to handle interaction)

25 / 49

Complete cases: 10,279 from 17,631

Explanatory variable	Complete Cases	
In care	1·07	(0·16)
In social housing	0·98	(0·058)
Inverse birthweight	123	(14·2)
Mo' age at birth	−0·029	(0·0065)
Mo' age squared	0·0035	(0·00081)
Mo' age × housing	0·024	(0·0090)
Mo' age sq'd × housing	−0·0015	(0·0011)
Constant	−2·6	(0·13)

26 / 49

Patterns of missing data

```
. mvpatterns care soch7 invbwt mo_age noqual2
Variable     | type     obs   mv   variable label
-------------+-------------------------------------------------------------
care         | byte   14360 3271   in care before age 7
soch7        | byte   14232 3399   social housing
invbwt       | double 16783  848   reciprocal birth weight (oz)
mo_age       | byte   17402  229   mother's age at birth (centered around 28)
noqual2      | byte   12044 5587   no qualifications at age 23
-------------------------------------------------------------------------------

Patterns of missing values

  +------------------------+
  | _pattern   _mv   _freq |
  |------------------------|
  |    +++++     0   10279 |
  |    ++++.     1    3324 |
  |    ..++.     3    1886 |
  |    ..+++     2    1153 |
  |    ++.++     1     349 |
  |------------------------|
  |    ...+.     4     124 |
  |    ++.+.     2     116 |
  |    ++..+     2     109 |
  |    +.++.     2      64 |
  |    +.+++     1      59 |
```


MI I: no interactions or non-linearities

```
ice care soch7 invbwt mo_age noqual2, m(15) cycles(10) saving(imp1) replace dryrun

   Variable | Command | Prediction equation
------------+---------+-----------------------------------------------------
       care | logit   | soch7 invbwt mo_age noqual2
      soch7 | logit   | care invbwt mo_age noqual2
     invbwt | regress | care soch7 mo_age noqual2
     mo_age | regress | care soch7 invbwt noqual2
    noqual2 | logit   | care soch7 invbwt mo_age
-------------------------------------------------------------------------------

set seed 1389
ice care soch7 invbwt mo_age noqual2, m(15) cycles(10) saving(imp1) replace
```

Analysis of imputed data

```
* load the imputed data
use imp1, clear

drop mo_agesq
drop agehous
drop agesqhous

gen mo_agesq=mo_age*mo_age
gen agehous=mo_age*soch7
gen agesqhous=mo_agesq*soch7

* Now we fit the model of interest to each of the imputed data sets
* and combine the results:

mim:  logit  noqual2 care soch7 invbwt mo_age mo_agesq agehous agesqhous
```

29 / 49

Results with MI I:

Explanatory variable	Complete Cases		Multiple imputation using MI I		MI II	MI III
In care	1·07	(0·16)	1·00	(0·14)		
In social housing	0·98	(0·058)	0·98	(0·053)		
Inverse birthweight	123	(14·2)	116	(15·8)		
Mo' age at birth	−0·029	(0·0065)	−0·021	(0·0073)		
Mo' age squared	0·0035	(0·00081)	0·0021	(0·00081)		
Mo' age × housing	0·024	(0·0090)	0·015	(0·0085)		
Mo' age sq'd × housing	−0·0015	(0·0011)	−0·00099	(0·0011)		
Constant	−2·6	(0·13)	−2·5	0·15		

30 / 49

MI II: with non-linearity and itermediate variables

```
ice care soch7 invbwt mo_age noqual2 mo_agesq
     sqbsag gfammove gfammove1 gfammove2,
    passive (mo_agesq: mo_age*mo_age \ gfammove1 : gfammove==1 \ gfammove2: gfammove==2)
            substitute (gfammove: gfammove1 gfammove2)
            cmd(gfammove: ologit)
            seed(1389) m(50) cycles(20) saving(imp_soch7_ordinal) replace dryrun
```

31 / 49

gives

```
   Variable | Command | Prediction equation
------------+---------+-------------------------------------------------------
       care | logit   | soch7 invbwt mo_age noqual2 mo_agesq sqbsag gfammove1
            |         | gfammove2
      soch7 | logit   | care invbwt mo_age noqual2 mo_agesq sqbsag gfammove1
            |         | gfammove2
     invbwt | regress | care soch7 mo_age noqual2 mo_agesq sqbsag gfammove1
            |         | gfammove2
     mo_age | regress | care soch7 invbwt noqual2 sqbsag gfammove1 gfammove2
    noqual2 | logit   | care soch7 invbwt mo_age mo_agesq sqbsag gfammove1
            |         | gfammove2
   mo_agesq |         | [Passively imputed from mo_age*mo_age]
     sqbsag | regress | care soch7 invbwt mo_age noqual2 mo_agesq gfammove1
            |         | gfammove2
   gfammove | ologit  | care soch7 invbwt mo_age noqual2 mo_agesq sqbsag
  gfammove1 |         | [Passively imputed from gfammove==1]
  gfammove2 |         | [Passively imputed from gfammove==2]
-------------------------------------------------------------------------------

* Then use mim
mim:  logit  noqual2 care soch7 invbwt mo_age mo_agesq agehous agesqhous
```

32 / 49

Results with MI II

Explanatory variable	Complete Cases		Multiple imputation using MI I		MI II		MI III
In care	1·07	(0·16)	1·00	(0·14)	1·13	(0·15)	
In social housing	0·98	(0·058)	0·98	(0·053)	0·95	(0·061)	
Inverse birthweight	123	(14·2)	116	(15·8)	115	(14·2)	
Mo' age at birth	−0·029	(0·0065)	−0·021	(0·0073)	−0·026	(0·0062)	
Mo' age squared	0·0035	(0·00081)	0·0021	(0·00081)	0·0032	(0·00077)	
Mo' age × housing	0·024	(0·0090)	0·015	(0·0085)	0·017	(0·0086)	
Mo' age sq'd × housing	−0·0015	(0·0011)	−0·00099	(0·0011)	−0·00102	(0·0011)	
Constant	−2·6	(0·13)	−2·5	0·15	−2·50	(0·13)	

MI III

```
use NCDS, clear

* Work on data with soch7==1
drop if soch7==.
drop if soch7==0

gen sqbsag= sqrt(bsag)

gen gfammove=0
replace gfammove=1 if fammove >1 & fammove <=5
replace gfammove=2 if fammove >5
replace gfammove=. if fammove==.

gen gfammove1=(gfammove==1)
replace gfammove1=. if(gfammove==.)

gen gfammove2=(gfammove==2)
replace gfammove2=. if(gfammove==.)
```

MI III, slide 2

```
ice care  invbwt mo_age noqual2 mo_agesq
     sqbsag gfammove gfammove1 gfammove2,
    passive (mo_agesq: mo_age*mo_age \
            gfammove1 : gfammove==1 \ gfammove2: gfammove==2)
            substitute (gfammove: gfammove1 gfammove2)
            cmd(gfammove: ologit)
            seed(1389) m(15) cycles(10) saving(imp_soch7_1) replace
```

35 / 49

MI III, slide 3

Repeat above, in subset where soch7 is zero, to create imp_soch7_0. Then:

```
use imp_soch7_1, clear
mim: append using imp_soch7_0

* Generate interactions in combined dataset
replace agehous=soch7*mo_age
replace agesqhous=soch7*mo_age*mo_age

* Analyse imputed data
mim: logit noqual2 care soch7 invbwt mo_age mo_agesq agehous agesqhous
```

36 / 49

Final Results

Explanatory variable	Complete Cases		Multiple imputation using MI I		MI II		MI III	
In care	1·07	(0·16)	1·00	(0·14)	1·13	(0·15)	1·15	(0·17)
In social housing	0·98	(0·058)	0·98	(0·053)	0·95	(0·061)	0·97	(0·058)
Inverse birthweight	123	(14·2)	116	(15·8)	115	(14·2)	122	(15·3)
Mo' age at birth	−0·029	(0·0065)	−0·021	(0·0073)	−0·026	(0·0062)	−0·030	(0·0067)
Mo' age squared	0·0035	(0·00081)	0·0021	(0·00081)	0·0032	(0·00077)	0·0034	(0·00077)
Mo' age × housing	0·024	(0·0090)	0·015	(0·0085)	0·017	(0·0086)	0·021	(0·0098)
Mo' age sq'd × housing	−0·0015	(0·0011)	−0·00099	(0·0011)	−0·00102	(0·0011)	−0·0012	(0·0011)
Constant	−2·6	(0·13)	−2·5	0·15	−2·50	(0·13)	−2·60	(0·14)

37 / 49

Discussion

- Complete Case analysis: potential loss of information, and if data are MAR potential bias too.
- MI I: Analysis without non-linearities, interactions or auxiliary variables: imputed data does not have structure we are investigating with model of interest; further we are not making use of information on variables on the causal path.
- MI II: Analysis with non-linearities and auxiliary variables: this better maintains non-linear relationship, but some issues remain; information from auxiliary variables limited as it turns out they are often missing when response, `noqual2` is missing—possibly a design flaw? Interactions still not included in imputation model.
- MI III: Having noted that `soch7`'s missingness mechanism depends on covariates, no bias if restrict to complete cases on `soch7` and impute separately in the two `soch7` groups to preserve any interaction.

Conclude: MI confirms the interaction is present under MAR, and (if you do some more imputations) gains some information.

The moral of the story is, think carefully about the data and imputation model before you start. If you apply MI, make sure the imputation model is compatible as possible with the model of interest. See [4] for further details.

38 / 49

Survival data

Survival data

Survival data raises some issues, and was investigated by Patrick Royston and Ian White [17]. They concluded that when using the full conditional specification (ice) approach, with mostly missing covariates, the preferred method is:

1. always include the censoring indicator as a binary variable in all chained equations;
2. bring in survival through the baseline hazard, $H_0(t)$ as follows:
   In each imputation cycle, as well as updating each incomplete variable in turn, we also update $H_0(t)$ by fitting the Cox model.
3. use the current estimate of $H_0(t)$ as a covariate in the other regressions in the full conditional specification.

Note that simpler approximations, such as including $\log(T)$ may work pretty well in practice.

39 / 49

Some further references

Use of multiple imputation in the epidemiologic literature: Klebanoff and Cole, 2008, [8].

Multiple imputation for missing data in epidemiological and clinical research—potential and pitfalls: Sterne *et al*, 2009, [15].

Strategies for multiple imputation in longitudinal studies, Spratt *et al*, 2009, [14].

Note also a recent paper on a new approach for handling interactions in imputation, which may work better (but appears to require that missing data are approximately MCAR), by Paul von Hippel [16].

40 / 49

Discussion

Reporting analyses with missing data

For any analysis potentially affected by missing data:

1. Report the number of missing values for each variable of interest. Give reasons for missing values if possible, and indicate how many individuals were excluded because of missing data when reporting the flow of participants through the study. If possible, describe reasons for missing data in terms of other variables.
2. Clarify whether there are important differences between individuals with complete and incomplete data.
3. For analyses that account for missing data, describe the nature of the analysis (e.g. multiple imputation), and the assumptions that were made (e.g. missing at random).

41 / 49

For analyses based on multiple imputation:

1. Provide details of the imputation modelling: software, number of imputations, variables in imputation model, use of interactions, transformations.
2. If a large fraction of the data is imputed, give a comparison of observed and imputed values. Marked differences cast doubt on the imputation procedure, unless they can be explained.
3. Where possible, provide results from analyses restricted to complete cases, for comparison with results based on multiple imputation. If there are important differences between the results, suggest explanations, bearing in mind that analyses of complete cases may suffer more chance variation, and that under the MAR assumption multiple imputation should correct biases that may arise in complete-cases analyses.
4. Discuss whether the variables included in the imputation model make the missing at random assumption plausible.

It is also desirable to investigate the robustness of key inferences to possible departures from the MAR assumption.

42 / 49

Summary

- Taken another look at the justification for MI
- Compared joint and full conditional specification imputation algorithms (note both fit a joint model; the latter implicitly)
- Reviewed available software
- Discussed the use of preliminary analysis to
    - help identify key additional (a.k.a. auxiliary) variables to include in the imputation model, and
    - help identify likely differences between the complete case and multiple imputation analysis
- Stressed that imputation model and model of interest need to be consistent (a.k.a. congenial)
- Illustrated with data from ALSPAC study
- Suggested reporting guidelines

43 / 49

How do we draw $Z_M|Z_O$?

This morning, we described a regression method for drawing $Z_M|Z_O$. This should work reasonably if the data set is large, as it is then an approximation to a Bayesian rule:

Let η be the parameter vector for a model for Z_0 (keep in mind the regression of observed Y's on X's this morning). This model must be such that all the missing data are missing responses; in other words only fully observed variables can be conditioned on.

The posterior distribution for η is $[\eta|Z_O] \propto [Z_O|\eta][\eta]$. (1)
(We can approximate this by drawing from the distribution of the regression parameters estimated by maximum likelihood — as this morning).

Then $[Z_M, \eta|Z_O] = [Z_M|\eta, Z_O][\eta|Z_O]$, where $[\eta|Z_O]$ is from (1).
(Having drawn a regression line, we draw the missing data about that line).

Discard unwanted η's. Calculate $\theta(Z_M, Z_O)$, estimating the posterior distribution of our parameter of interest, θ.

45 / 49

More formal Intuition for MI

Assuming MAR (i.e. ignore the dropout mechanism).

The posterior is

$$\begin{aligned}
[\theta, Z_M|Z_O] &= \int [\theta, \eta, Z_M|Z_O]\, d\eta \\
&= \int [\theta|\eta, Z_M, Z_O][Z_M|\eta, Z_O][\eta|Z_O]\, d\eta \\
&= [\theta|Z_M, Z_O] \int [Z_M|\eta, Z_O][\eta|Z_O]\, d\eta \\
&\qquad \text{(as } [\theta|\eta, Z_M, Z_O] = [\theta|Z_M, Z_O]\text{)} \\
&= [\theta|Z_M, Z_O][Z_M|Z_O].
\end{aligned}$$

46 / 49

Mean estimator

It follows that

$$\begin{aligned}[\theta|Z_O] &= \int [\theta, Z_M|Z_O]\, dZ_M \\ &= \int [\theta|Z_M, Z_O][Z_M|Z_O]\, dZ_M \\ &= \mathbf{E}_{Z_M|Z_O}[\theta|Z_M, Z_O] \\ \text{Thus } \mathbf{E}[\theta|Z_O] &= \mathbf{E}_{Z_M|Z_O}\mathbf{E}_\theta[\theta|Z_M, Z_O].\end{aligned}$$

Suppose draw $Z_M^1, \ldots, Z_M^K$ from $[Z_M|Z_O]$, and $\theta(Z_M^k, Z_O)$ estimates $\mathbf{E}_\theta[\theta|Z_M^k, Z_O]$. Then

$$\mathbf{E}[\theta|Z_O] \approx \frac{1}{K}\sum_{k=1}^{K} \theta(Z_M^k, Z_O) = \hat{\theta}_{MAR}$$

47 / 49

Variance estimator

Recall $[\theta|Z_O] = \mathbf{E}_{Z_M|Z_O}[\theta|Z_M, Z_O]$.

Thus

$$\mathbf{V}[\theta|Z_O] = \mathbf{E}_{Z_M|Z_O} V_\theta[\theta|Z_M, Z_O] + \mathbf{V}_{Z_M|Z_O}\mathbf{E}_\theta[\theta|Z_M, Z_O].$$

Suppose draw $Z_M^1, \ldots, Z_M^K$ from $[Z_M|Z_O]$, and $\sigma^2(Z_M^k, Z_O)$ estimates $V_\theta[\theta|Z_M^k, Z_O]$. Then

$$\begin{aligned}\mathbf{V}[\theta|Z_O] \approx &\frac{1}{K}\sum_{k=1}^{K} \sigma^2(Z_M^k, Z_O) \\ &+ \frac{1}{K-1}\sum_{k=1}^{K} \left(\hat{\theta}(Z_M^k, Z_O) - \hat{\theta}_{MAR}\right)^2.\end{aligned}$$

48 / 49

References

[1] P D Allison. Multiple imputation for missing data: a cautionary tale. *Sociological methods and Research*, 28:301–309, 2000.

[2] P D Allison. *Missing Data*. Thousand Oaks, CA: Sage, 2002.

[3] A Burton, L J Billingham, and S Bryan. Cost-effectiveness in clinical trials: using multiple imputation to deal with incomplete cost data. *Clinical Trials*, 4:154–161, 2007.

[4] James Carpenter and Ian Plewis. Coming to terms with non-response in longitudinal studies. Under revision for *The SAGE handbook of Mehodological Innovation, editors Malcolm Williams and Paul Vogt*, London: SAGE, 2009.

[5] James R Carpenter and Michael G Kenward. *Missing data in clinical trials — a practical guide*. Birmingham: National Health Service Co-ordinating Centre for Research Methodology. Freely downloadable from http://www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml, accessed 28 May 2009, 2008.

[6] N J Horton and S R Lipsitz. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, pages 244–254, 2001.

[7] Michael G Kenward and James R Carpenter. Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, 16:199–218, 2007.

[8] M A Klebanoff and S R Cole. Use of multiple imputation in the epidemiologic literature. *American Journal of Epidemiology*, 168:355–357, 2008.

[9] Ula Nur, Lorraine G Shack, Bernard Rachet, James R Carpenter, and Michel P Coleman. Modelling relative survival in the presence of incomplete data: a tutorial. *International Journal of Epidemiology*, –:in press, 2009.

[10] D Rubin. Multiple imputation after 18 years. *Journal of the American Statistical Association*, 91:473–490, 1996.

[11] D B Rubin. *Multiple imputation for nonresponse in surveys*. New York: Wiley, 1987.

[12] J L Schafer. *Analysis of incomplete multivariate data*. London: Chapman and Hall, 1997.

[13] J L Schafer and J W Graham. Missing data: our view of the state of the art. *Psychological Methods*, 7:147–177, 2002.

[14] M Spratt, J A C Sterne, K Tilling, J R Carpenter, and J B Carlin. Strategies for Multiple Imputation in Longitudinal Studies. Revision under review, 2009.

[15] J A C Sterne, I R White, J B Carlin, M Spratt, P Royston, M G Kenward, A M Wood, and J R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, 339:157–160, 2009.

[16] Paul T von Hippel. How to impute interactions, squares and other transformed variables. *Sociological Methodology*, 39:265–291, 2009.

[17] I R White and P Royston. Imputing missing covariate values for the cox model. *Statistics in Medicine*, pages 1982–1998, 2009.