Statistical modelling with missing data using multiple imputation

Session 2: Multiple Imputation

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

March 23, 2010

Overview 2 / 54

Session 2: multiple imputation

- Intuitive introduction
- Example: cancer epidemiology
- Pitfalls
- Sensitivity analysis
- Publishing analyses that use multiple imputation
- Discussion

2 / 54

Introduction to MI 3 / 54

Motivation for MI

Suppose our data set has variables X, Y with some Y values MAR given X.

In the first session, we saw that using only subjects with both observed we can get valid estimates of the regression of Y on X.

However, inference based on observed values of Y alone (eg sample mean, variance) is typically biased.

This suggests the following idea

1. Fit the regression of Y on X
2. Use this to impute the missing Y
3. With this completed data set, calculate our statistic of interest (eg sample mean, variance, regression of X on Y).

As we can only ever know the *distribution* of missing data (given observed), steps 2 & 3 have to be repeated, and the results averaged in some way.

3 / 54

Why and when MI?

Why?

1. MI is attractive, because once we have imputed the missing data, we can analyse the completed data sets as we would have done if no data were missing.
2. MI is particularly attractive when we have missing covariates, when other options are relatively tricky.

When?

1. MI is not often needed if only responses are missing and our model of interest is a regression, *and* we are prepared to assume MAR given the covariates in the model — for then we get valid estimates from the observed data
2. Thus MI not as frequently used in trials as elsewhere, as in trials usually outcomes data are missing[a].

4 / 54

[a] Note missing baseline can be treated as an outcome in the analysis

Simple illustration

For simplicity, suppose we have only two variables in our data set.

Suppose one of them is observed on every unit. Call this X.

Suppose one is only observed on some units. Call this Y and write $Y = (Y_M, Y_O)$ (missing, observed)

5 / 54

The key idea

The key idea is to use the data from units where both Y and X are observed, together with the rest of the X's — i.e. (Y_O, X) — to learn about the relationship between Y and X.

Then, if $\tilde{X}$ represents the vector of X values from individuals with missing Y's, we use this relationship to complete the data set by drawing the missing observations from $Y_M|\tilde{X}$.

We do this K (typically $>>$ 5) times, giving rise to K complete data sets.

We analyse each of these data sets in the usual way.

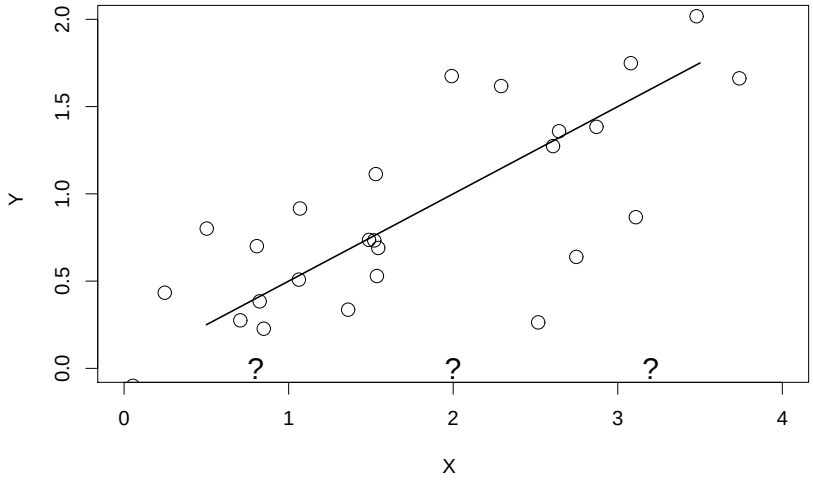
We combine the results using particular rules.

Suppose the analysis of interest is calculating the marginal mean of Y, or regressing X on Y.

6 / 54

Intuition behind multiple imputation: 1

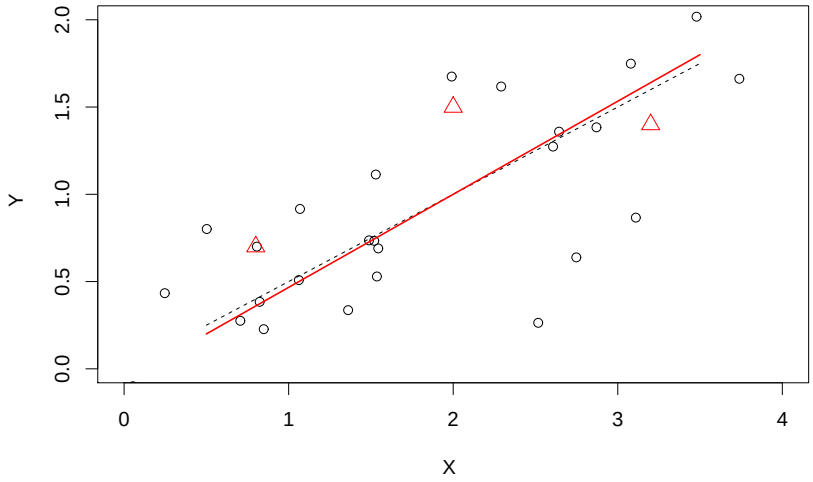
Model observed pairs, denoted (Y_O, X).



7 / 54

Intuition behind multiple imputation: 2

Draw Y_M by (i) drawing from distribution of regression line (this gives us the red line below) (ii) then drawing from variability about that line.



8 / 54

Algorithm for this simple example

Let n_0 be the number of fully observed individuals. Let W be the design matrix, consisting of two columns; one of '1's and the other of the n_0 X's (where Y is observed).

The sampling distributions of the estimators are:

$$\hat{\sigma}^2 \sim \frac{\sigma^2 \chi^2_{n_0-2}}{(n_0-2)},$$

$$\begin{pmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{pmatrix} \sim N\left\{\begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}, \sigma^2(W^TW)^{-1}\right\}$$

9 / 54

Continued...

We then

1. Draw a $\tilde{\sigma}^2$ from $\hat{\sigma}^2(n_0-2)/\chi^2_{n_0-2}$.
2. Draw $(\tilde{\beta}_0, \tilde{\beta}_1)$ from
$$N\left\{\begin{pmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{pmatrix}, \tilde{\sigma}^2(W^TW)^{-1}\right\}$$
3. For each missing Y, draw a $\tilde{\epsilon} \sim N(0, \tilde{\sigma}^2)$.
4. Create an imputed data set by, for each missing Y imputing using the appropriate X
$$\tilde{\beta}_0 + \tilde{\beta}_1 X + \tilde{\epsilon}.$$

Repeat the whole to create the second, third,... imputations

10 / 54

Intuition behind multiple imputation 3

Results of multiple imputation:

	Data		Imputation 1		Imputation 2		Imputation 3		Imputation 4	
Unit	Y	X	Y	X	Y	X	Y	X	Y	X
1	1.1	3.4	1.1	3.4	1.1	3.4	1.1	3.4	1.1	3.4
2	1.5	3.9	1.5	3.9	1.5	3.9	1.5	3.9	1.5	3.9
3	2.3	2.6	2.3	2.6	2.3	2.6	2.3	2.6	2.3	2.6
4	3.6	1.9	3.6	1.9	3.6	1.9	3.6	1.9	3.6	1.9
5	0.8	2.2	0.8	2.2	0.8	2.2	0.8	2.2	0.8	2.2
6	3.6	3.3	3.6	3.3	3.6	3.3	3.6	3.3	3.6	3.3
7	3.8	1.7	3.8	1.7	3.8	1.7	3.8	1.7	3.8	1.7
8	**?**	0.8	**0.2**	0.8	**0.8**	0.8	**0.3**	0.8	**2.3**	0.8
9	**?**	2.0	**1.7**	2.0	**2.4**	2.0	**1.8**	2.0	**3.5**	2.0
10	**?**	3.2	**2.7**	3.2	**2.5**	3.2	**1.0**	3.2	**1.7**	3.2

Using the imputed data sets

Notation for analyses of imputed data sets

As described above, we have imputed K 'complete' data sets.

Analysing each of them in the usual way (i.e. using the model intended for the complete data) gives us K estimates of the original quantity of interest, say θ. Denote these estimates $\hat{\theta}_1, \ldots, \hat{\theta}_K$.

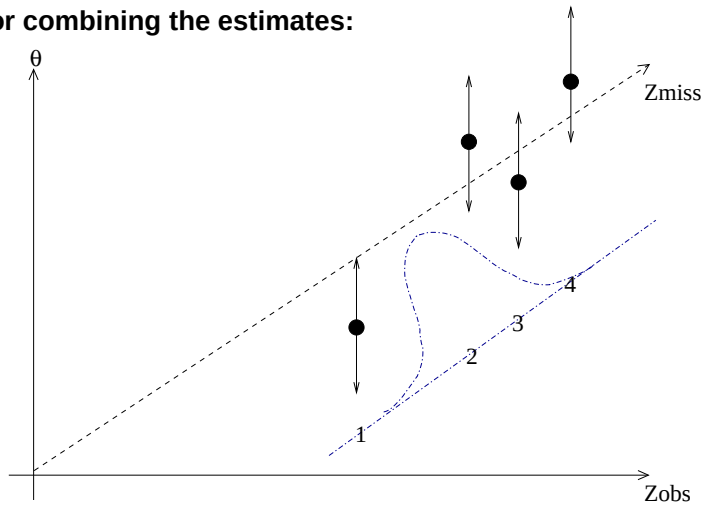
The analysis of each imputed data set will also give an estimate of the variance of the estimate $\hat{\theta}_k$, say $\hat{\sigma}_k^2$. Again, this is the usual variance estimate from the model.

We combine these quantities to get our overall estimate and its variance using certain rules.

Write $Z_M = Y_M$ (the set of missing data) and $Z_O = (Y_O, X)$ (the set of observed data).

For inference, we need to average over the distribution of the missing given observed data, i.e. $Z_M | Z_O$.

Intuition for combining the estimates:



$\hat{\theta}_{MI} = \mathbf{E}_{Z_M|Z_O}\mathbf{E}[\theta(Z_O, Z_M)]$.

$\mathbf{V}[\hat{\theta}_{MI}] = \mathbf{E}_{Z_M|Z_O}\mathbf{V}[\theta(Z_O, Z_M)] + \mathbf{V}_{Z_M|Z_O}\mathbf{E}[\theta(Z_O, Z_M)]$.

13 / 54

MI rules

14 / 54

Combining the estimates

Let the multiple imputation estimate of θ be $\hat{\theta}_{MI}$. Then

$$\hat{\theta}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k.$$

Further define the within imputation and between imputation components of variance by

$$\hat{\sigma}_w^2 = \frac{1}{K}\sum_{k=1}^{K}\hat{\sigma}_k^2, \quad \text{and} \quad \hat{\sigma}_b^2 = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta}_{MI})^2,$$

Then

$$\hat{\sigma}_{MI}^2 = \left(1 + \frac{1}{K}\right)\hat{\sigma}_b^2 + \hat{\sigma}_w^2,$$

so the estimated standard error of $\hat{\theta}_{MI}$ is $\hat{\sigma}_{MI}$.

14 / 54

Inference for θ

To test the null hypothesis $\theta = \theta_0$, compare

$$\frac{\hat{\theta}_{MI} - \theta_0}{\hat{\sigma}_{MI}} \quad \text{to} \quad t_\nu,$$

where

$$\nu = (K-1)\left[1 + \frac{\hat{\sigma}_w^2}{(1+1/K)\hat{\sigma}_b^2}\right]^2.$$

Thus, if $t_{\nu,0.975}$ is the 97.5% point of the t distribution with ν degrees of freedom, the 95% confidence interval is

$$(\hat{\theta}_{MI} - \hat{\sigma}_{MI} t_{\nu,0.975}, \quad \hat{\theta}_{MI} + \hat{\sigma}_{MI} t_{\nu,0.975})$$

The rate of missing information

If there were no missing data, and we used MI, we should find that $(1+1/K)\hat{\sigma}_b^2 = 0$.
Thus the relative increase in variance due to the missing data is

$$r = \frac{(1+1/K)\hat{\sigma}_b^2}{\hat{\sigma}_w^2}.$$

Alternatively, the 'rate of missing information' is

$$\frac{(1+1/K)\hat{\sigma}_b^2}{\hat{\sigma}_w^2 + (1+1/K)\hat{\sigma}_b^2} = \frac{r}{1+r}.$$

It turns out a better estimate of this quantity is

$$\frac{r + 2/(\nu+3)}{1+r}.$$

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.

The key is thus appropriate choice of the imputation model. This should

1. be consistent (also known as congenial) with the scientific model of interest, and
2. appropriately incorporate relevant auxiliary variables.

Summary

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.

18 / 54

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)[15] for a discussion.

19 / 54

Example

20 / 54

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.

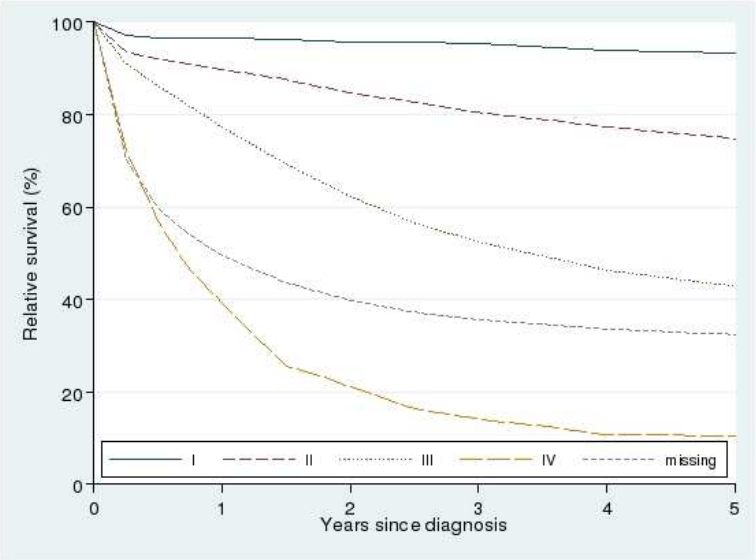
A key predictor is the 'stage' of the cancer at diagnosis, which is an ordinal variable taking values 1,...4. Unfortunately, this is often not recorded/observed. Further, the suspicion is that this is related to the severity of the cancer at diagnosis.

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 (Nur *et al*, 2009[10]).

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

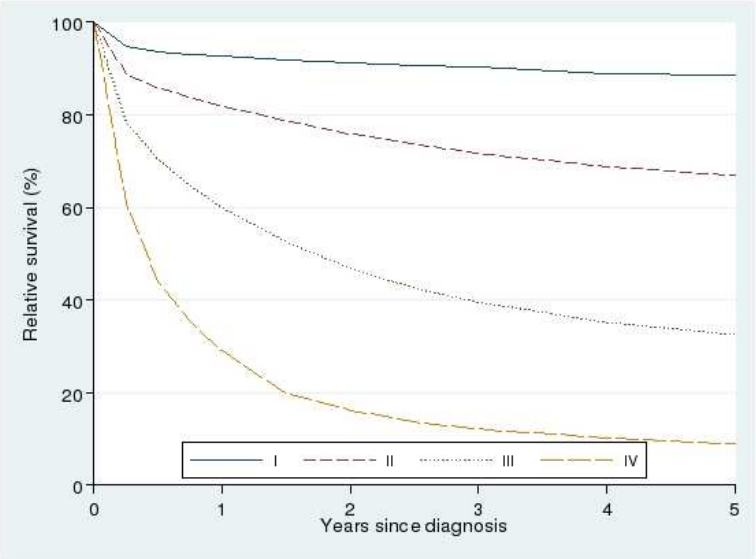
20 / 54

Results: extra category for missing stage



21 / 54

Results: multiple imputation assuming MAR



22 / 54

Algorithms for MI

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.

24 / 54

Software taxonomy: methods derived from multivariate normal

	Complexity		
Response type	Normal		Mixed response
Data structure	Independent	Multilevel	Multilevel
Package			
Standalone	NORM	PAN	REALCOM[a]
SAS	NORM-port	—	—
Stata	NORM-port	—	—
R/S+	NORM-port	—	—
MLwiN	MCMC algorithm emulates PAN[a]		+ 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, REALCOM).

Interactions must be handled by imputing separately in each group.

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

25 / 54

[a] free from www.missingdata.org.uk

Chained equations/full conditional specification

An alternative to specifying an explicit joint distribution the partially observed variables is to specify a sequence of conditional models. This leads to the chained equations (also known as full conditional specification) 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.

26 / 54

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 (unpublished) 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.

Multilevel and more complex data structures are more naturally handled using joint modelling.

27 / 54

Pitfalls 28 / 54

Likely pitfalls

We have already noted that once the imputation model is specified, MI is essentially automatic.

It follows that most problems arise because of inappropriate choice of imputation model.

Likely pitfalls[17]:

1. omitting variables in the model of interest (e.g. response) or handling them incorrectly (e.g. with time to event data);
2. handling non-normal variables incorrectly;
3. having an imputation model which is inconsistent (uncongenial) with the model of interest;
4. an inappropriate MAR assumption, and
5. convergence problems with the MI model.

Point (3) usually arises when the model of interest contains non-linear relationships and/or interactions not in the imputation model.

28 / 54

Survival analysis

Suppose the model of interest is a survival analysis with partially observed covariates.

When setting up the chained equations for the imputation, we need to remember to include the censoring indicator as a covariate in all imputation models, together with a suitable measure of survival.

A paper by White and Royston (2009)[19] suggests the cumulative hazard is preferable.

29 / 54

Case study: QRISK

A recent BMJ article reported the development of the QRISK tool for cardiovascular risk prediction, based on a large general practice research database[6].

The researchers used multiple imputation to handle the missing data in their analysis.

In their published prediction model, cardiovascular risk was found to be unrelated to cholesterol (coded as the ratio of total to HDL cholesterol), which was surprising.

30 / 54

continued...

The authors have subsequently clarified that when their analysis was performed in the usual way by restricting to individuals with complete information (no missing data) there was a clear association between cholesterol and cardiovascular risk.

Furthermore, a similar result was obtained after using a revised, improved, imputation procedure .

LIKELY EXPLANATION: The censoring indicator was omitted from the imputation model. As most healthy individuals were censored, this meant the imputation model could not properly distinguish the distributions of key risk factors between healthy and unhealthy individuals.

31 / 54

Pitfall 2: non-normal data

The second pitfall is to unthinkingly treat continuous, ordinal and binary data as if it were normal in the imputation model.

For example, if a biochemical factor had a highly skewed distribution but was implicitly assumed to be normally distributed, then imputation procedure could impute some implausibly low or even negative values.

A pragmatic approach here is to transform such variables to approximate normality before imputation, and then transform the imputed values back to the original scale.

Different problems arise when data are missing in binary or categorical variables. The chained equation procedure handles these naturally through logistic/ordinal/multinomial models.

If data are binary/ordinal, a normal approximation may be acceptable[2], but this is less plausible the closer probabilities get to 1 or 0, and implausible for unordered categorical data.

32 / 54

Pitfall 3: inconsistent (uncongenial) imputation model

This is arguably the most difficult pitfall to avoid. We describe pitfall in general, and then illustrate with a case study.

The problem arises if the imputation model does not contain all the structure present in the model of interest. For instance, the model of interest may contain an interaction — say between mothers age and whether she is in social housing. However, the imputation model may not include this interaction.

Thus, all the imputed data will not have this interaction present. Fitting the model of interest to the resulting ‘completed’ data will result in a point estimate that is incorrectly pulled towards the null.

33 / 54

Case study: UK 1958 Birth Cohort

The UK National Childhood Development Study consists of a target sample of 17634 children born in the UK in 1958.

Our analysis here focuses on how well the probability of having any educational qualifications at age 23 can be predicted by four variables measured at birth/early childhood: birth weight, mothers age, living in social housing and spending a period in care.

By age 23, the target sample had reduced as a result of death and permanent emigration, to 15885.

24% of these were missing at age 23 — a non-trivial percentage (see Plewis et al., 2004[11]).

34 / 54

Model of interest

$$\begin{aligned}&\texttt{logit}\{\Pr(\text{child has no educational qualifications at 23 years})\}\\&= \beta_0 + \beta_1\texttt{care} + \beta_2\texttt{soch7} + \beta_3\texttt{invbwt} +\\&\beta_4\texttt{mo\_age} + \beta_5\texttt{mo\_agesq} + \beta_6\texttt{agehous} + \beta_7\texttt{agesqhous}\end{aligned}$$

We focus particularly on the variable `agehous` — the interaction between mother's age when the child was born and whether they were living in social housing before the child was 7 years old.

Also of interest are the non-linear terms `mo_agesq` and `agesqhous`

35 / 54

Naive multiple imputation

We illustrate the Stata `ice` command by Royston (2007)[12], for imputation using chained equations:

```
. 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
-----------------------------------------------------
End of dry run. No imputations were done, no files were created.
```

After imputation, we load the imputed data and use the `mim` prefix to fit the model of interest to each imputed data set and combine the results for final inference using Rubin's rules.

36 / 54

More careful multiple imputation

A more careful look at the data suggests the following:

1. The reason for social housing being missing depends on the other covariates, and *given these* not the response. So we can drop those with missing social housing without inducing bias.
   We can then impute separately in the two social housing groups, thus allowing interactions, in particular with mother's age.
2. There are additional variables on the causal path — a behavioural score `bsag` and the number of family moves, `fammove`. These auxiliary variables appear to contain information and be predictors of missing data, and should be included (after transformation if necessary).
3. We should allow for the non-linear relationship with mother's age — as best we can.

37 / 54

Code

We impute separately in the two social housing groups; the code for each group is as follows:

```
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
```

38 / 54

continued...

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

End of dry run. No imputations were done, no files were created.
```

39 / 54

Results

Explanatory variable	Complete Cases		Multiple imputation			
			Naive		With interaction	
In care	1·07	(0·16)	1·00	(0·14)	1·15	(0·17)
In social housing	0·98	(0·058)	0·98	(0·053)	0·97	(0·058)
Inverse birthweight	123	(14·2)	116	(15·8)	122	(15·3)
Mo' age at birth	−0·029	(0·0065)	−0·021	(0·0073)	−0·030	(0·0067)
Mo' age squared	0·0035	(0·00081)	0·0021	(0·00081)	0·0034	(0·00077)
Mo' age × housing	0·024	(0·0090)	0·015	(0·0085)	0·021	(0·0098)
Mo' age sq'd × housing	−0·0015	(0·0011)	−0·00099	(0·0011)	−0·0012	(0·0011)
Constant	−2·6	(0·13)	−2·5	(0·15)	−2·60	(0·14)

Take home messages

MI is a powerful tool, but do not be deceived by how easy it is to use the software:

- Think about your model of interest and your imputation model: they need to be consistent/congenial.
  Be careful with interactions and non-linearities, also 'follow-on' questions, where a second question is only relevant if the first has a particular answer (e.g. Do you smoke (Y/N), if 'Yes' how many a day?)
- Be careful with survival data: have you included the censoring indicator and time-to-event appropriately
- Be careful with non-normal data
- Be aware of the risks of over-fitting with binary/categorical data

Email for a practical work sheet going through this example using Stata.

Sensitivity analysis 42 / 54

Sensitivity analysis

So far we have talked about imputation under MAR.

If data are MNAR a useful approach is the following:

1. Impute under MAR
2. Modify the imputations to reflect departures from MAR
3. Fit the model of interest to the imputed data, and combine the results using Rubin's rules for inference.

For example, in the cancer epidemiology example, we may want to explore the robustness of the conclusions to the assumption that cancer stage is MAR.

If we believe the probability of cancer stage IV is greater than predicted by MAR, we can initially impute under MAR, then move to MNAR by further increasing the probability that missing stage data are stage IV.

42 / 54

Some references for sensitivity analysis

Multiple imputation based approaches to sensitivity analysis with missing outcome data, derived in a clinical trials context, are described by Little and Yau (1996)[9] and in Chapter 6 of Carpenter & Kenward (2008)[3]. See also Carpenter *et al* (2009)[5] and White *et al* (2007)[18].

An alternative approach based on re-weighting estimates after multiple imputation under MAR, is described by Carpenter and Kenward (2007)[4].

Email for example analyses in Stata.

43 / 54

Reporting MI analyses 44 / 54

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).

44 / 54

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 need a careful explanation.
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.

45 / 54

Discussion

46 / 54

Summary

- MI is most convenient under MAR. The assumptions it relies on are usually more plausible than a complete case analysis. If these assumptions are correct, MI give unbiased, efficient inference.
- To increase the plausibility of the MAR assumption, we will often wish to include several predictors of missingness that we do not want to adjust for in the final analysis. This is potentially a key advantage.
- Multiple imputation is particularly useful for missing covariates, especially in:
  - survey settings where there is a separate imputer and analyst;
  - large and messy problems, where a full likelihood or Bayesian analysis is impractical.

46 / 54

Summary II

- Beware the pitfalls:
  - omitting key variables;
  - handling non-normal variables incorrectly;
  - having an inconsistent/uncongenial imputation model
  - making an inappropriate MI assumption
  - convergence problems with the imputation model.
- Remember that all analyses of partially observed data are based on inherently untestable assumptions. Hence sensitivity analysis is important.
- A pattern mixture approach, using different imputation models for the different patterns, may be useful for sensitivity analysis.

47 / 54

Further reading and summary

Some MI references

Schafer (1997)[15] — Key book giving details of data augmentation and MI methods in many models.

Rubin (1987)[13] — Book bringing together the theory in a fairly accessible way.

Rubin (1996)[14] — review of the use of MI after $\sim$ 18 years.

Horton and Kleinman (2007)[7] — Comparison of software packages.

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

Kenward & Carpenter (2007) [8] — up to date overview.

Spratt *et al* (2009)[16] — Strategies for multiple imputation in longitudinal studies,

Sterne *et al* (2009)[17] — Potential and pitfalls of multiple imputation.

48 / 54

More details on MI

How do we draw $Z_M|Z_O$?

In the pictures above, 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. This model must be such that all the missing data are missing responses; in other words only fully observed variables can be conditioned on (keep in mind the regression of observed Y's on X's above).

The posterior 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 max. like. — as above).

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.

50 / 54

More formal Intuition for MI

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

The posterior is

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

51 / 54

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

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

52 / 54

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

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

53 / 54

References

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

[2] C A Bernaards, T R Belin, and J L Schafer. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26(6):1368–1382, mar 2007.

[3] 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 `www.missingdata.org.uk`, accessed 15 December 2009, 2008.

[4] James R Carpenter, Michael G Kenward, and Ian R White. Sensitivity analysis after multiple imputation under missing at random — a weighting approach. *Statistical Methods in Medical Research*, 16:259–275, 2007.

[5] James R Carpenter, James H Roger, and Mike G Kenward. Relevant, Accessible Sensitivity Analyses Using Multiple Imputation *revision in preparation*, 2009.

[6] J Hippisley-Cox, C Coupland, Y Vinogradova, J Robson, M May, and P Brindle. Derivation and validation of qrisk, a new cardiovascular disease risk score for the united kingdom: prospective open cohort study. *British Medical Journal*, 335:7611–7623, 2007.

[7] N J Horton and K P Kleinman. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *American Statistician*, 61(1):79–90, 2007.

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

[9] R J A Little and L Yau. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, 52:471–483, 1996.

[10] 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, pre-print online, doi:10.1093/ije/dyp309*, 2009.

[11] I Plewis, L Calderwood, D Hawkes, and G Nathan. *National Child Development Study and 1970 British Cohort Study Technical Report: Changes in the NCDS and BCS70 Populations and Samples over Time (1st ed.)*. London: Institute of Education, University of London, 2004.

[12] P Royston. Multiple imputation of missing values: further update of ice with emphasis on interval censoring. *The Stata Journal*, 7:445–464, 2007.

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

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

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

[16] 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.

[17] 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.

[18] I White, J Carpenter, Stephen Evans, and Sara Schroter. Eliciting and using expert opinions about non-response bias in randomised controlled trials. *Clinical Trials*, 4:125–139, 2007.

[19] I R White and P Royston. Imputing missing covariate values for the cox model. *Statistics in Medicine*, 28:1982–1998, 2009.