Statistical modelling with missing data using multiple imputation

Lecture 2: Ad-hoc methods and introduction to multiple imputation

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

November 17, 2009

Overview

Lecture 2: outline

1. Critique of *ad-hoc* methods
   (a) Complete case (a.k.a. Completers) analysis
   (b) Imputation of simple mean
   (c) Imputation of regression mean
   (d) Creating an extra category
   (e) Last Observation Carried Forward (LOCF)
2. Introduction to Multiple Imputation
   (a) Motivation
   (b) Intuitive explanation
   (More detail this afternoon)
3. Example
4. Summary

2 / 34

Ad-hoc methods

Complete case analysis

	Variables	
Unit	1	2
1	3.4	5.67
2	3.9	4.81
3	2.6	4.93
4	1.9	6.21
5	2.2	6.83
6	3.3	5.61
7	1.7	5.45
8	2.4	4.94
9	2.8	5.73
10	3.6	.

- Complete case analysis deletes all units with incomplete data
- Under MCAR potentially inefficient, but unbiased
- Generally biased and inefficient
- Problematic in regression analyses

3 / 34

Simple mean imputation

	Variables	
Unit	1	2
1	3.4	5.67
2	3.9	4.81
3	2.6	4.93
4	1.9	6.21
5	2.2	6.83
6	3.3	5.61
7	1.7	5.45
8	2.4	4.94
9	2.8	5.73
10	3.6	**5.58**

- Missing observations replaced by completers' means *for that variable*
- Inappropriate for categorical variables
- Reduces associations in data (i.e. regression parameters biased to their null value)
- Variance underestimated

4 / 34

Regression mean imputation

	Variables	
Unit	1	2
1	3.4	5.67
2	3.9	4.81
3	2.6	4.93
4	1.9	6.21
5	2.2	6.83
6	3.3	5.61
7	1.7	5.45
8	2.4	4.94
9	2.8	5.73
10	3.6	**5.24**

- Use regression, here OLS: $V_2 = \alpha + \beta V_1 + e$
- Using units 1–9 we get: $\hat{V}_2 = 6.56 - 0.366 \times (V_1).$
- For unit 10 this gives $6.56 - 0.366 \times (3.6) = 5.24.$
- Now obtain unbiased estimators of means, associations, under MAR
- Variance still (often markedly) underestimated

5 / 34

Creating an extra category

	Variables		
Unit	1	2	3
1	1	3.4	5.67
2	1	3.9	4.81
3	1	2.6	4.93
4	1	1.9	6.21
5	$3 \leftarrow \cdot$	2.2	6.83
6	2	3.3	5.61
7	2	1.7	5.45
8	2	2.4	4.94
9	$3 \leftarrow \cdot$	2.8	5.73
10	$3 \leftarrow \cdot$	3.6	5.58

- Very dissimilar classes can be lumped into one group
- Severe bias can arise, in any direction
- Variable will not correctly adjust for confounding

- *Exception* is with missing baseline, where randomisation protects us as the chance of a missing baseline is the same in each arm. See [13].

6 / 34

LOCF — the idea

	Unit	
Time	1	2
1	2.1	3.4
2	3.8	3.9
3	$3.8 \leftarrow \cdot$	2.6
4	$3.8 \leftarrow \cdot$	1.9
5	$3.8 \leftarrow \cdot$	2.2
6	$3.8 \leftarrow \cdot$	$2.2 \leftarrow \cdot$
⋮	⋮	⋮

- Makes strong, implausible assumptions
- Imputes a degenerate distribution
- Means and variances wrong
- In general neither conservative or liberal; bias depends on *unknown* treatment effect!
- See [9, 5, 3]

7 / 34

Common defences

1. LOCF is conservative
   - No.
2. LOCF has correct size under the full null
   - So does rolling a 20-sided die.
3. No alternative available
   - No. Stick with the workshop...

LOCF is equivalent to analysing the last response we happened to measure

– This is not appropriate for evaluating exposure/treatment effects.

8 / 34

Principles...

- The hypotheses under investigation do not change when data are missing.
- However, extra assumptions have to be made to obtain an estimate of treatment.
- These should be as weak as possible, to maximise the chance of them being (approximately) true, so that the analysis is valid.
- For example, MAR implies that future behaviour for those who share the same past measurements and covariate values is identical *on average*, whether or not they dropout.

9 / 34

..vs LOCF shortfall

- we make the strong assumption that unseen observations equal the last observation seen;
- we get biased treatment estimates, with the direction of the bias depending on the (unknown) true treatment effect, and
- our confidence intervals do not have the correct coverage.

10 / 34

Summary

- Ad-hoc methods are an attempt to ‘solve’ the problem of missing data
- They avoid any serious thinking about the issues raised by missing data
- They do not utilize statistical principles
- Generally they result in misleading conclusions

11 / 34

Introduction to MI

12 / 34

Motivation for MI

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

In the first lecture, 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.

12 / 34

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].

13 / 34

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

Multiple imputation

Multiple imputation is an approximation to a full Bayesian analysis, which — at least for simpler problems — can be carried out in `WinBUGS`. See www.missingdata.org.uk under *example analyses*.

We next give an intuitive introduction.

Later, we will see that valid inference depends on the choice of the imputation model — this is where the effort needs to be spent.

In the imputation model, we may well want to include extra variables to increase the plausibility of the MAR assumption.

The imputation model also needs to be consistent with the model of interest, and ideally should correctly reflect the hierarchical structure of the data.

Additional variables — not themselves predictive of missingness — but predictive of the partially observed variables, should also be included in the imputation model.

14 / 34

MI: more details

15 / 34

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)

15 / 34

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.

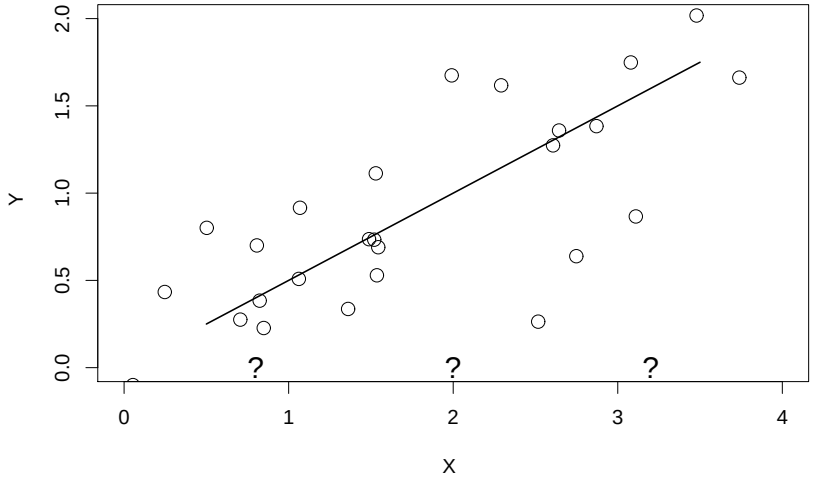
We combine the results using particular rules.

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

16 / 34

Intuition behind multiple imputation: 1

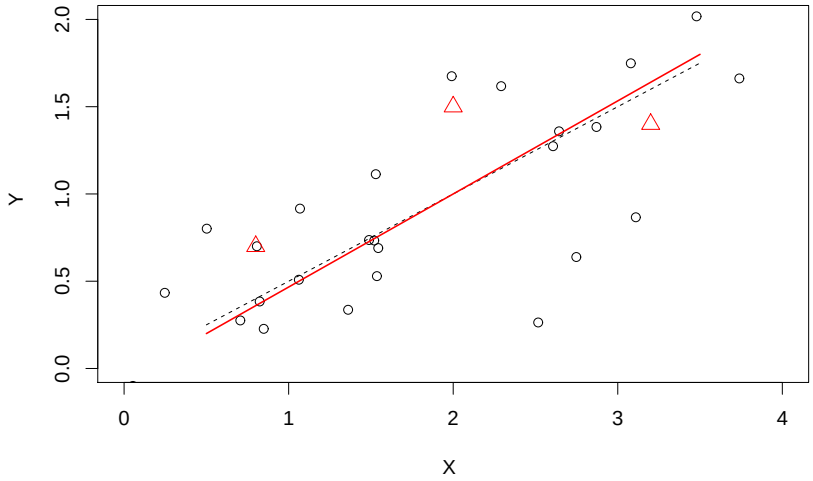
Model observed pairs, denoted (Y_O, X).



17 / 34

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.



18 / 34

Algortihm 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:

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

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

19 / 34

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

20 / 34

Intuition behind multiple imputation 3

Repeat step 2 a number of times:

	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

21 / 34

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.

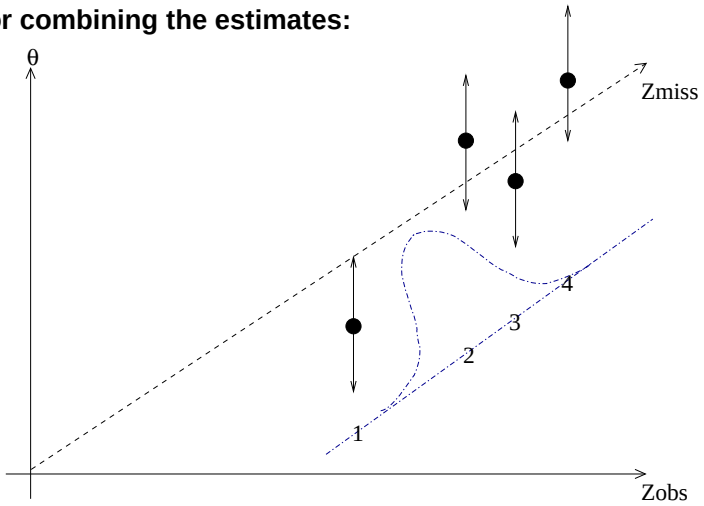
We combine these quantities to get our overall estimate and its variance using certain rules.

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. $Z_M|Z_O$.

22 / 34

Intuition for combining the estimates:



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

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

23 / 34

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 $\hat{\sigma}_{MI}$.

24 / 34

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

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

25 / 34

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

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

26 / 34

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.

27 / 34

Example

28 / 34

Complete case, extra category for missing data, and MI

We use data from the 'Class size project' [2], and consider a simple analysis to compare complete cases, creating and extra category for missing data, and MI.

Our model of interest regresses post-reception literacy, `nitpost`, on pre-reception maths, `nmatpre`, and special educational needs, `sen`:

$$\texttt{nlitpost}_i = \beta_0 + \beta_1 \texttt{nmatpre}_i + \beta_2 \texttt{sen}_i + \epsilon_i.$$

28 / 34

Analyses

We first fit the model to 4873 children with no missing data.

We then make some `sen` values missing, letting the missingness mechanism depend on `nlitpost` and `nmatpre`.

This leaves us with 2405 observations.

We then

1. fit the model of interest to these (analogous to a ‘Complete Cases’ analysis);
2. create an extra category, ‘missing `sen`’ fit the model, and
3. apply multiple imputation

29 / 34

MI for these data in Stata

```
. ice nlitpost nmatpre sen_m, saving(temp) m(20)

   #missing |
     values |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,405       49.35       49.35
          1 |      2,468       50.65      100.00
------------+-----------------------------------
      Total |      4,873      100.00

   Variable | Command | Prediction equation
------------+---------+-------------------------------------------------------
   nlitpost |         | [No missing data in estimation sample]
    nmatpre |         | [No missing data in estimation sample]
      sen_m | logit   | nlitpost nmatpre
------------------------------------------------------------------------------

Imputing
[Only 1 variable to be imputed, therefore no cycling needed]
1..2..3..4..5..6..7..8..9..10..11..12..13..14..15..16..17..18..19..20..file temp.dta saved

. use temp, clear

. mim: regress nlitpost nmatpre sen
```

30 / 34

Results

Method	Coefficients (SE) for		
	`nmatpre`	`sen`	Extra category `sen`
Full data ($n = 4873$)	0.585 (0.012)	-0.432 (0.043)	—
'Complete cases' ($n = 2405$)	0.434 (0.019)	-0.362 (0.047)	—
Extra category ($n = 4873$)	0.431 (0.014)	-0.365 (0.047)	0.584 (0.026)
MI (including individuals with partial data, $n = 4873$)	0.584 (0.012)	-0.432 (0.043)	—

MI has corrected the bias, with virtually no loss of information in this simple example.

31 / 34

Further reading and summary

32 / 34

Some MI references

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

Rubin (1987)[11] — Book bringing together the theory in a fairly accessible way.

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

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

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

Kenward & Carpenter (2007) [7]

32 / 34

Summary

- MI is most convenient under MAR.
  - To increase the chance that this is approximately true, we may 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 in trials.
- 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.
- A pattern mixture approach, using different imputation models for the different patterns, may be useful for sensitivity analysis. See [4], ch. 6; [8], and the second lecture this afternoon.
- This afternoon we discuss MI in more detail.

33 / 34

References

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

[2] P Blatchford, H Goldstein, C Martin, and W Browne. A study of class size effects in English school reception year classes. *British Educational Research Journal*, pages 169–185, 2002.

[3] J Carpenter, M Kenward, S Evans, and I White. Letter to the editor: Last observation carry forward and last observation analysis by J. Shao and B. Zhong, Statistics in Medicine, 2003, **22**, 2429–2441. *Statistics in Medicine*, 23:3241–3244, 2004.

[4] 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.

[5] R J Cook, L Zeng, and G Y Yi. Marginal analysis of incomplete longitudinal binary data; a cutionary note on locf imputation. *Biometrics*, pages 820–828, 2004.

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

[9] G Molenberghs, H Thijs, I Jansen, C Beunkens, M G Kenward, C Mallinkrodt, and R J Carroll. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5:445–464, 2004.

[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] Ian R. White and Simon G. Thompson. Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine*, 24:993–1007, 2005.