Overview

The afternoon consists of two 105-minute sessions, discussing issues raised by missing data and practical approaches for analyzing the resulting partially observed data. Specifically,

- we will review a range of methodological approaches, with a focus on multiple imputation, and
- the aim is that participants gain an overview of this research area, alongside the knowledge and confidence to apply these techniques in their own work.

Session 1

- 14:05–14:25: Rod
  - Introduction and issues raised by missing data;
  - Rubin’s framework
  - Complete records analyses
- 14:25–14:55: James
  - Introduction to Multiple Imputation (MI)
  - Algorithms
  - Planning, performing and reporting MI analyses
  - Example: youth cohort study
- 14:55–15:25: Rod
  - Likelihood based approaches
- 15:25–15:45: James
  - MI with non-linear relationships
Session 2

16.15–16.45: Rod
- Inverse probability weighting and extensions

16:45–17:15: James
- MI for hierarchical data

17:15–17:35: Rod
- Introduction to sensitivity analysis
- Pattern mixture, selection and shared parameter models
- Issues raised by clinical trials

17:35–17:50: James
- Examples
  - Trial to improve the quality of peer review
  - Asthma trial

17:50–18:00: Discussion

SESSION 1, part I

Example 1. Survey nonresponse

- National Health and Nutrition Examination Survey (NHANES) III
- Public Use Files subject to:
  - Unit nonresponse
    - noncontact
    - refusal
  - Partial response
    - questionnaire interview complete
    - health examination missing
  - Item nonresponse

Overview

- Introduction and issues raised by missing data
- Rubin’s framework
- Complete case/complete records analysis
Example 1 contd. Survey nonresponse

- Issue: different users of public use file get different answers to the same question because missing data are handled in different, sometimes naïve, ways
- Methods
  - weighting for unit nonresponse
  - multiple imputation for partial and item nonresponse

Example 2. Administrative censoring in Longitudinal Studies

- Censoring by termination of study

Example 3. Attrition in Longitudinal studies

- Longitudinal studies often have drop-outs
  - Individuals relocate outside study area
  - In drug trials, individuals drop out because of side-effects of drugs, ineffective or effective treatments
- Common analyses have problems:
  - complete case analysis is biased, particularly if drop-out is differential across treatment arms
  - Naïve imputation (e.g. last observation carried forward) involves unrealistic assumptions
- We discuss better alternatives

Example 4. Measurement error with a calibration sample as missing data

(a) External Calibration design

(b) Internal Calibration design

\[ D \] measured in calibration sample

\[ X \]: covariate of interest but unobserved
\[ Y \]: observed error-prone measurement related to \( X \)
\[ D \]: response variable, interest in regression of \( D \) on \( X \)
(more generally \( D \) can include other covariates).
Guo and Little (2013)
What are missing data?

- Always assume missingness hides a meaningful value for analysis.

Examples:
- Missing data from missed clinical visit (✓)
- No-show for a preventive intervention (✗)
- In a longitudinal study of blood pressure medications:
  - Losses to follow-up (✓)
  - Deaths (✗)

Pattern versus mechanism

- Pattern: Which values are missing?
- Mechanism: Why? Reasons related to the study variables?

\[ Y = \text{data matrix, if no data were missing} \]
\[ R = \text{response indicator matrix} \]
\[ (i,j) \text{th element indicates whether } (i,j) \text{th element of } Y \]
\[ \text{is observed ("responds", 1) or missing (0)} \]
- Pattern concerns distribution of \( R \)
- Mechanism concerns distribution of \( R \) given \( Y \)

Patterns of Missing Data

- Some methods work for a general pattern
- Other methods apply only to special patterns

Rubin’s (1976) framework

- Data are:
  - Missing completely at random (MCAR) if missingness independent of \( Y \):
    \[ p(R \mid Y) = p(R) \text{ for all } Y \]
  - Missing at random (MAR) if missingness only depends on observed components \( Y_{obs} \) of \( Y \):
    \[ p(R \mid Y) = p(R \mid Y_{obs}) \text{ for all } Y \]
  - Missing not at random (MNAR) if missingness depends on missing (as well as perhaps on observed) components of \( Y \)
MAR for univariate nonresponse

\( X_j \) = complete covariates
\( Y \) = incomplete variable
\( R = 1, Y \) observed
\( 0, Y \) missing

MAR: missingness independent of \( Y \) given \( X_1, \ldots, X_k \)
That is, \( R \) can depend on \( X \)'s …
but not on \( Y \) given \( X \)'s

A non-monotone example

Mechanism is MAR if
\[
\Pr(Y_2 \text{ missing}) = g(Y_1)
\]
\[
\Pr(Y_1 \text{ missing}) = f(Y_2)
\]
\[
\Pr(\text{complete}) = 1 - f(Y_2) - g(Y_1)
\]

MAR for monotone missing data

MAR if dropout depends on values recorded prior to drop-out
MNAR if dropout depends on values that are missing (that is, after drop-out)
Censoring by end of study: plausibly MCAR
Drop-out from side effect: MAR if side effect is measured and included in analysis

Example of MAR: planned missing data: matrix sampling

- Raghunathan and Grizzle (1995)
- Extension of double sampling
- To reduce burden, split questionnaire into core plus modules 1, ..., \( K \)
- Split sample into subsamples 1, ..., \( K \)
- Subsample \( k \) receives the common core plus module \( k \)
- Selection of subsamples could be MCAR or MAR (conditional on common core characteristics)
- Other patterns of designed missing data are possible
MAR for unplanned missing data

- An assumption: can't be checked from observed data without additional assumptions
- More realistic after adjustment for observed covariates that are predictive of the missing values

Properties of a good missing-data method

- Makes use of partial information on incomplete cases, for reduced bias, increased efficiency
- Frequency valid (“calibrated”) inferences under plausible model for missing data (e.g. confidence intervals have nominal coverage)
- Propagates missing-data uncertainty, both within and between imputation models
- Favor likelihood based approaches
  - Maximum Likelihood (ML) for large samples
  - Multiple Imputation/Bayes for small samples

General Strategies

Imputation → Complete case Analysis → Analyze Incomplete

- Complete cases
- Discard
- Weights
- e.g. maximum likelihood

Complete-case analysis

- Complete cases
- Discard
Complete case (CC) analysis

- Easy (default in many computer packages)
- Simple and may be good enough with small amounts of missing data
  - but defining "small" is problematic; depends on
    - fraction of incomplete cases
    - recorded information in these cases
    - parameter being estimated

Limitations of CC analysis

- Assumes the complete cases are "representative" of all the cases
- Loss of information in incomplete cases, which has two aspects:
  - Increased variance of estimates
  - Bias when complete cases differ systematically from incomplete cases
    - restriction to complete cases requires that the complete cases are representative of all the cases for the analysis in question – for many analyses this implies MCAR (but not all: e.g. CC valid for regression where missingness depends on covariates)
    - this assumption is often questionable!

Increased variance from CC analysis

Inefficiency of CC Analysis depends on how much information is contained in the discarded incomplete cases

In a likelihood analysis, this is measured by the information matrices for the complete cases ($I_{CC}$) and all the observed data ($I_{CC} + I_{IC}$)

Example: univariate missing data

Suppose $X_1, ..., X_p$ are strong predictors of $Y$

$I_{IC}$ is substantial for unconditional mean of $Y$

$I_{IC} = 0$ for conditional mean of $Y$ given $X_1, ..., X_p$!
Weighted CC analysis

- Weight respondents differentially to reduce nonresponse bias – e.g. mean becomes weighted mean
- Common for unit nonresponse in surveys
- More on this later

Next: multiple imputation

**Imputation**: unlike complete-case analysis, retains information in incomplete cases

**Multiple imputation**: propagates imputation uncertainty, yielding valid inferences

SESSION 1, part II

Introduction to Multiple Imputation
Motivation for MI

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

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 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 can 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\(^1\). However, this is beginning to change with the use of MI in sensitivity analysis (final topic of Session 2).

\(^1\) Note missing baseline can be treated as an outcome in the analysis.
Motivating 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)

The key idea

The key idea is to use the observed data to estimate the distribution of the missing data given the observed data (typically assuming MAR).

We then impute the missing data from this distribution.

In this setting we therefore use data from units where both $Y$ and $X$ are observed, to learn about the distribution of $Y$ given $X$.

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.

Intuition behind multiple imputation: 1

Model observed pairs:

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

We combine the results using Rubin’s rules.

Suppose the analysis of interest is calculating the marginal mean of $Y$, or regressing $X$ on $Y$. 
Intuition behind multiple imputation: 2

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

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:

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

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

Continued...

We then

- Draw a $\tilde{\sigma}^2$ from $\tilde{\sigma}^2(n_0 - 2) / \chi^2_{n_0-2}$.
- Draw $(\tilde{\beta}_0, \tilde{\beta}_1)$ from

$$N\left(\left(\tilde{\beta}_0, \tilde{\beta}_1\right), \tilde{\sigma}^2(W^TW)^{-1}\right)$$

- For each missing $Y$, draw a $\tilde{\epsilon} \sim N(0, \tilde{\sigma}^2)$.
- Create an imputed data set by, for each missing $Y$ imputing using the appropriate $X$

$$\tilde{Y} = \tilde{\beta}_0 + \tilde{\beta}_1 X + \tilde{\epsilon}.$$  

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

Intuition of multiple imputation: 3

Results of multiple imputation:

<table>
<thead>
<tr>
<th>Unit</th>
<th>Data</th>
<th>Imputation 1</th>
<th>Imputation 2</th>
<th>Imputation 3</th>
<th>Imputation 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>4</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
</tr>
<tr>
<td>7</td>
<td>3.8</td>
<td>3.8</td>
<td>3.8</td>
<td>3.8</td>
<td>3.8</td>
</tr>
<tr>
<td>8</td>
<td>?</td>
<td>0.8</td>
<td>0.8</td>
<td>0.3</td>
<td>2.3</td>
</tr>
<tr>
<td>9</td>
<td>?</td>
<td>2.0</td>
<td>2.4</td>
<td>1.8</td>
<td>3.5</td>
</tr>
<tr>
<td>10</td>
<td>?</td>
<td>3.2</td>
<td>2.5</td>
<td>1.0</td>
<td>1.7</td>
</tr>
</tbody>
</table>
**Notation for analyses of $K$ imputed datasets**

Analysing each imputed (i.e. 'completed') dataset the usual way (i.e. using the model intended if there were no missing data) gives us $K$ estimates of the original quantity of interest, say $\theta$. 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.

**Intuition for combining the estimates:**

$$
\hat{\theta}_{MI} = E_{Z_M|Z_O}[E_{\Theta|Z_M,Z_O}[\Theta]] .
$$

$$
$$

We combine these quantities to get our overall estimate and its variance using Rubin’s rules.

These assume that, in the absence of missing data, the posterior distribution of $\theta$ is normal. Let $Z$ denote the data. The posterior mean is therefore the maximum likelihood estimate,

$$
\hat{\theta} = E_{\Theta|Z}[\Theta].
$$

Now let $Z_M$ be the set of missing data ($= Y_M$ in our example) and $Z_O$ be the set of observed data ($= (Y_O, X)$).

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

**Combining the estimates: Rubin’s rules**

Let the multiple imputation estimate of $\theta$ 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

$$
\hat{\sigma}_w^2 = \frac{1}{K} \sum_{k=1}^{K} \hat{\sigma}_k^2,
$$

and

$$
\hat{\sigma}_b^2 = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta}_{MI})^2 .
$$

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 $\hat{\sigma}_{MI}$.
Inference for \( \theta \)

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 \( \nu \) degrees of freedom, the 95% confidence interval is

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

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 (‘bespoke’) 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

- be consistent with the scientific model of interest, and
- appropriately incorporate relevant auxiliary variables.
Introduction to MI: 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.

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. See C&K, §2.6
- **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 $\hat{\sigma}^2_u$.
- **What if I am interested in more than one parameter?**
  - Imputation proceeds in the same way, as does finding the overall estimate of $\theta$. However, estimating the covariance matrix can be tricky. Typically more imputations will be needed. See Schafer (1997) for a discussion.

Do the MI rules always work?

MI’s key attraction is the variance formula, which gives a relatively simple way of calculating the variance in a wide range of settings.

Critics of MI have tended to home in on the variance formula. In general it will be biased if

- imputations are not approximate draws from a Bayesian posterior: $\rightarrow$ downward bias, usually serious
- the imputer assumes more than the analyst (so called super-efficiency) (Meng, 1994): $\rightarrow$ upward bias, usually mild
- survey weights, based on the inverse of selection probabilities, are used in calculating quantities of interest (see Kim et al. 2006). See also Rubin (1987): $\rightarrow$ upward bias, often mild.

See C&K p.64–70, p.73, Ch.11 and references therein.

Example: Youth Cohort Study (YCS)

The YCS is an ongoing UK government funded representative survey of pupils in England and Wales at school leaving age.


Our analysis is illustrative, and relates GCSE score (range 0–84) to parental occupation, ethnicity, cohort and sex.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cohort</td>
<td>year of data collection: 1990, 93, 95, 97, 99</td>
</tr>
<tr>
<td>boy</td>
<td>indicator variable for boys</td>
</tr>
<tr>
<td>occupation</td>
<td>parental occupation, categorised as managerial, intermediate or working</td>
</tr>
<tr>
<td>ethnicity</td>
<td>categorised as Bangladeshi, Black, Indian, other Asian, Other, Pakistani or White</td>
</tr>
</tbody>
</table>
Missing data pattern in YCS

<table>
<thead>
<tr>
<th>Pattern</th>
<th>GCSE score</th>
<th>occupation</th>
<th>ethnicity</th>
<th>No.</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>55145</td>
<td>87%</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>6821</td>
<td>11%</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>687</td>
<td>1%</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>592</td>
<td>1%</td>
</tr>
</tbody>
</table>

Predictors of complete records

<table>
<thead>
<tr>
<th>Variable</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>cohort '93</td>
<td>-0.085</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
</tr>
<tr>
<td>cohort '95</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>(0.038)</td>
</tr>
<tr>
<td>cohort '97</td>
<td>0.178</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
</tr>
<tr>
<td>cohort '99</td>
<td>0.135</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
</tr>
<tr>
<td>boy</td>
<td>-0.053</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
</tr>
<tr>
<td>GCSE score</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
</tr>
<tr>
<td>Non-white</td>
<td>-1.723</td>
</tr>
<tr>
<td></td>
<td>(0.0288)</td>
</tr>
<tr>
<td>ROC</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Implications

We see that parental occupation (reported by the pupil) is the variable with the greatest proportion of missing data.

We see that the response in the substantive model, GCSE score, is a key predictor of missing parental occupation, so a complete records analysis will be biased.

Next we observe that ethnicity is the key covariate in the substantive model which predicts missing parental occupation. The above analysis leads us to expect that MI under the MAR assumption will have the greatest effect on the coefficients for ethnicity.

Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete Records</th>
<th>Multiple imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 54872$</td>
<td>$n = 62578$</td>
</tr>
<tr>
<td>Cohort90</td>
<td>reference</td>
<td></td>
</tr>
<tr>
<td>Cohort93</td>
<td>5.66 (0.20)</td>
<td>5.44 (0.20)</td>
</tr>
<tr>
<td>Cohort95</td>
<td>9.42 (0.22)</td>
<td>9.21 (0.20)</td>
</tr>
<tr>
<td>Cohort97</td>
<td>8.09 (0.21)</td>
<td>8.03 (0.20)</td>
</tr>
<tr>
<td>Cohort99</td>
<td>12.70 (0.22)</td>
<td>12.91 (0.21)</td>
</tr>
<tr>
<td>Boys</td>
<td>-3.44 (0.13)</td>
<td>-3.35 (0.13)</td>
</tr>
<tr>
<td>Managerial</td>
<td>reference</td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>-7.42 (0.15)</td>
<td>-7.75 (0.16)</td>
</tr>
<tr>
<td>Working</td>
<td>-13.74 (0.17)</td>
<td>-14.32 (0.17)</td>
</tr>
<tr>
<td>White</td>
<td>reference</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>-5.61 (0.57)</td>
<td>-7.16 (0.51)</td>
</tr>
<tr>
<td>Indian</td>
<td>3.58 (0.44)</td>
<td>2.97 (0.42)</td>
</tr>
<tr>
<td>Pakistani</td>
<td>-2.03 (0.58)</td>
<td>-3.63 (0.47)</td>
</tr>
<tr>
<td>Bangladeshi</td>
<td>0.27 (1.04)</td>
<td>-3.20 (0.74)</td>
</tr>
<tr>
<td>Other asian</td>
<td>5.52 (0.68)</td>
<td>4.49 (0.63)</td>
</tr>
<tr>
<td>Other</td>
<td>-0.25 (0.70)</td>
<td>-1.32 (0.66)</td>
</tr>
<tr>
<td>Constant</td>
<td>39.66 (0.19)</td>
<td>39.09 (0.18)</td>
</tr>
</tbody>
</table>
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. This assumes that the regression of any one variable on the others is linear.

Skew variables can be transformed to hopefully better approximate multivariate 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 treats them as responses.

Software taxonomy

Methods derived from multivariate normal

<table>
<thead>
<tr>
<th>Response type</th>
<th>Complexity</th>
<th>Data structure</th>
<th>Package</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Independent</td>
<td>Multilevel</td>
<td>Standalone</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>Multilevel</td>
<td>SAS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Stata</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R/S+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MLwiN</td>
</tr>
</tbody>
</table>

SAS — IVEware; R — mice, mi; Stata — ice

Hughes et al (2014) show that for an unstructured multivariate normal, and for saturated log-linear models FCS and joint modelling are equivalent (with appropriate priors). For a mix of binary and continuous data they are formally not equivalent, but differences are unlikely to be practically important. See C&K p. 86–87, 103–104, 122–3.

Full Conditional Specification (FCS)
An alternative to specifying an explicit joint distribution the partially observed variables is to specify a sequence of conditional models. This is the FCS (chained equations) 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

Hughes et al (2014) show that for an unstructured multivariate normal, and for saturated log-linear models FCS and joint modelling are equivalent (with appropriate priors). For a mix of binary and continuous data they are formally not equivalent, but differences are unlikely to be practically important. See C&K p. 86–87, 103–104, 122–3.
Comments

Lee & Carlin (2010) found similar performance to multivariate normal imputation, even for ordinal variables, provided appropriate rounding is used.

They also raised concerns about bias with FCS due to over-specification of models for unordered data, and difficulties with large numbers of variables.

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

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 are (Sterne et al 2009):

1. omitting variables in the substantive model (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 substantive model;
4. an inappropriate MAR assumption, and
5. convergence problems with the MI model.

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

Survival analysis

Suppose the substantive model 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.

White and Royston (2009) suggest the cumulative hazard is preferable.

Nevertheless, this is only approximate, and if a non-trivial proportion of data are missing and the variance of partially observed covariates is large this approximation may be poor.

For a full discussion, see C&K Ch.8.

Case study: QRISK

A 2007 BMJ article (Hippisley-Cox et al) reported the development of the QRISK tool for cardiovascular risk prediction, based on a large general practice research database.

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.
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 I
(thanks to Tim Morris, MRC-CTU)

The authors stated:
We imputed total serum cholesterol and HDL separately in the original imputation model. We then calculated the ratio by dividing total serum cholesterol by HDL.

Such ‘passive’ imputation is best avoided, but here was particularly poor. Imputed HDL values had relatively high variance, so that many were close to zero, massively inflating the total/HDL ratio and removing the association with heart disease.

Likely explanation II

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.

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 substantive model and your imputation model: they need to be consistent/congenial.
- Be clear about your assumptions: is MAR plausible?
- 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
Reporting analyses based on multiple imputation

- Provide details of the imputation modelling: software, number of imputations, variables in imputation model, use of interactions, transformations.
- If a large fraction of the data is imputed, give a comparison of observed and imputed values. Marked differences need a careful explanation.
- Where possible, provide results from analyses restricted to complete records for comparison. If there are important differences between the results, suggest explanations.
- 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 contextually relevant departures from the MAR assumption.

Some MI references

Schafer (1997) — Details of data augmentation, EM and joint MI algorithms for MVN and general location model.
Rubin (1987) — Sets out MI in the survey setting.
Rubin (1996) — review of the use of MI after ~ 18 years.
Carpenter & Kenward (2013) — reviews theoretical basis of MI and illustrates its use in a wide range of situations.

Overview

- Likelihood-based methods – maximum likelihood, Bayes
- Little and Rubin (2002, chapter 6)
Likelihood methods

- Statistical model + data $\Rightarrow$ Likelihood
- Two general approaches based on likelihood
  - maximum likelihood inference for large samples
  - Bayesian inference for small samples:
    $\log(\text{likelihood}) + \log(\text{prior}) = \log(\text{posterior})$
- Methods can be applied to incomplete data
  - do not require rectangular data sets
- First review main ideas for complete data

Definition of Likelihood

- Data $Y$
- Statistical model yields probability density $f(Y \mid \theta)$
  for $Y$ with unknown parameters $\theta$
- Likelihood function is then the density as a function of $\theta$
  $L(\theta \mid Y) = \text{const} \times f(Y \mid \theta)$
- Loglikelihood is often easier to work with:
  $l(\theta \mid Y) = \log L(\theta \mid Y) = \text{const} + \log\{f(Y \mid \theta)\}$
  Constants can depend on data but not on parameter $\theta$
Example: Normal sample

- $Y = (y_1, \ldots, y_n)$ univariate iid normal sample

$$\theta = (\mu, \sigma^2)$$

$$f(Y | \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right)$$

$$l(\mu, \sigma^2 | Y) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2$$

Example: Multinomial sample

- $Y = (y_1, \ldots, y_n)$ univariate $K$-category multinomial sample

$n_j$ number of $y_i$ equal to $j$ ($j=1, \ldots, K$)

$$\theta = (\pi_1, \ldots, \pi_{K-1}); \quad \pi_K = 1 - \pi_1 - \cdots - \pi_{K-1}$$

$$f(Y | \pi_1, \ldots, \pi_{K-1}) = \frac{n!}{n_1! \cdots n_K!} \left(\prod_{j=1}^{K-1} \pi_j^{n_j}\right) (1 - \pi_1 - \cdots - \pi_{K-1})^{n_K}$$

$$l(\pi_1, \ldots, \pi_{K-1} | Y) = \left(\sum_{j=1}^{K-1} n_j \log \pi_j\right) + n_K \log(1 - \pi_1 - \cdots - \pi_{K-1})$$

Maximum Likelihood Estimate

- The maximum likelihood (ML) estimate $\hat{\theta}$ of $\theta$ maximizes the likelihood, or equivalently the log-likelihood

$$L(\hat{\theta} | Y) \geq L(\theta | Y) \quad \text{for all } \theta,$$ or

$$\log L(\hat{\theta} | Y) \geq \log L(\theta | Y) \quad \text{for all } \theta$$

- The ML estimate is the “value of the parameter that makes the data most likely”

- The ML estimate is not necessarily unique, but is for many regular problems given enough data

Computing the ML estimate

- In regular problems, the ML estimate can be found by solving the **likelihood equation**

$$S(\theta | Y) = 0$$

where $S$ is the **score function**, defined as the first derivative of the loglikelihood:

$$S(\theta | Y) = \frac{\partial \log L(\theta | Y)}{\partial \theta}$$

- In simple examples, explicit solutions available

- More generally, iterative algorithms like Newton-Raphson or scoring are needed
Basic Examples

- Univariate normal sample \( Y = (y_1, \ldots, y_n) \quad \theta = (\mu, \sigma^2) \)
  \[ \hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \]
  (Note the lack of a correction for degrees of freedom)
- Multivariate Normal sample
  \[ \hat{\mu} = \bar{y}, \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T / n \]
- Normal Linear Regression (possibly weighted)
  \( (y_i \mid x_i, \ldots, x_q) \sim N(\beta_0 + \sum_{j=1}^{p} \beta_j x_i, \sigma^2 / w_i) \)
  \[ \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \text{weighted least squares estimates} \]
  \[ \hat{\sigma}^2 = (\text{weighted residual sum of squares})/n \]
- Multinomial sample \( \hat{x}_j = n_j / n, \quad j = 1, \ldots, K \)

Properties of ML estimates

- Under assumed model, ML estimate is:
  - Consistent (not necessarily unbiased)
  - Efficient for large samples
  - not necessarily the best for small samples
- ML estimate is transformation invariant
  - If \( \hat{\theta} \) is the ML estimate of \( \theta \) then \( \phi(\hat{\theta}) \) is the ML estimate of \( \phi(\theta) \)
  - Property is exploited in factored likelihood methods for monotone missing data

More general models

- Generalized linear regression (normal, logit, Poisson, etc)
- Normal repeated measures with various mean and covariance structures (e.g. PROC MIXED in SAS).
  Unbalanced data can be viewed as a missing data problem
- Generalized linear mixed models
- Factor analysis, latent class models
- Loglinear models for contingency tables
- Fitting these often requires iterative algorithms

Large-sample ML Inference

- Basic large-sample approximation:
  For regular problems,
  \[ \hat{\theta} - \theta \sim N(0, C) \]
  where \( C \) is a covariance matrix estimated from the sample
  - Frequentist treats \( \hat{\theta} \) as random, \( \theta \) as fixed; equation defines the sampling distribution of \( \hat{\theta} \)
  - Bayesian treats \( \theta \) as random, \( \hat{\theta} \) as fixed; equation defines posterior distribution of \( \hat{\theta} \)
**Forms of precision matrix**

- The precision of the ML estimate is measured $C^{-1}$ by
  - Observed information (recommended)
    \[ C^{-1} = I(\hat{\theta} | Y) = \left( \frac{\partial^2 \log L(\theta | Y)}{\partial \theta \partial \theta'} \right)_{\theta = \hat{\theta}} \]
  - Expected information (not as good, may be simpler)
    \[ C^{-1} = J(\hat{\theta}) = E \left( I(\theta | Y, \theta) \right)_{\theta = \hat{\theta}} \]
  - Some other approximation to curvature of loglikelihood in the neighborhood of the ML estimate (e.g. sandwich estimate)

**Interval estimation**

- 95% (confidence, probability) interval for scalar $\theta$ is:
  \[ \hat{\theta} \pm 1.96 \; C^{1/2}, \text{where} \; 1.96 \text{ is 97.5 ptile of normal distribution} \]
- Example: univariate normal sample
  \[ I = J = \begin{bmatrix} n/\hat{\sigma}^2 & 0 \\ 0 & n/(2\hat{\sigma}^4) \end{bmatrix} \Rightarrow C = \begin{bmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 2\hat{\sigma}^4/n \end{bmatrix} \]

Hence some 95% intervals are:

\[ y \pm 1.96 \frac{s}{\sqrt{n}} \text{ for } \mu \]
\[ s^2 \pm 1.96 \frac{s^2}{\sqrt{n}/2} \text{ for } \sigma^2 \]
\[ \ln(s) \pm 1.96 \frac{\sqrt{2/n}}{\ln(\sigma)} \text{ for } \ln(\sigma) \]

**Significance Tests**

Tests based on likelihood ratio (LR) or Wald (W) statistics:

- $\theta = (\theta_{(1)}, \theta_{(2)}); \hat{\theta}_{(0)} = \text{null value of } \theta_{(1)}; \theta_{(2)} = \text{other parameters}$
- $\hat{\theta} = \text{unrestricted ML estimate}$
- $\hat{\theta} = (\hat{\theta}_{(1)}, \hat{\theta}_{(2)}); \hat{\theta}_{(2)} = \text{ML estimate of } \theta_{(2)} \text{ given } \theta_{(1)} = \theta_{(0)}$

**LR statistic:**

\[ LR(\hat{\theta}, \hat{\theta}) = 2 \left[ \ell(\hat{\theta} | Y) - \ell(\hat{\theta} | Y) \right] \]

**Wald statistic:**

\[ W(\hat{\theta}, \hat{\theta}) = (\hat{\theta}_{(0)} - \hat{\theta}_{(1)})^T C^{-1} (\hat{\theta}_{(0)} - \hat{\theta}_{(1)}) \]

yield P-values

\[ P = pr \left( \chi^2_q > D(\hat{\theta}, \hat{\theta}) \right) \]

- $D = \text{LR or Wald statistic}; q = \text{dimension of } \theta_0$
- $\chi^2_q = \text{Chi-squared distribution with } q \text{ degrees of freedom}$

**Bayes inference**

- Given a prior distribution $p(\theta)$ for the parameters, inference can be based on the posterior distribution using Bayes' theorem:
  \[ p(\theta | Y) = \text{const.} \times p(\theta) \times L(\theta | Y) \]

- For small samples, Bayes' inference based on the posterior distribution is often better than the large sample ML approximation.
  - In important standard problems with non-informative priors, Bayes yields inference comparable to small-sample frequentist inference
Example: linear regression
The normal linear regression model:
\[ (y_i | x_{i1}, ..., x_{ip}) \sim N(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2) \]
with non-informative “Jeffreys’” prior:
\[ p(\beta_0, ..., \beta_p, \log \sigma^2) = \text{const.} \]
yields the posterior distribution of \((\beta_0, ..., \beta_p)\) as multivariate
\(T\) with mean given by the least squares estimates \((\hat{\beta}_0, ..., \hat{\beta}_p)\)
covariance matrix \((X^T X)^{-1}s_i^2\) where \(X\) is the design matrix,
and degrees of freedom \(n - p - 1\).
Resulting posterior probability intervals are equivalent to
standard \(t\) conﬁdence intervals.

Simulating Draws from Posterior Distribution
- With problems with high-dimensional \(\theta\), it is often
easier to draw values from the posterior distribution,
and base inferences on these draws
- For example, if
\[ (\hat{\theta}_i^{(d)} : d = 1, ..., D) \]
is a set of draws from the posterior distribution for a
scalar parameter \(\theta\), then
\[ \tilde{\theta}_i = D^{-1} \sum_{d=1}^{D} \hat{\theta}_i^{(d)} \] approximates posterior mean
\[ s_{\theta}^2 = (D - 1)^{-1} \sum_{d=1}^{D} (\hat{\theta}_i^{(d)} - \tilde{\theta}_i)^2 \] approximates posterior variance
\( \tilde{\theta}_i \pm 1.96s_{\theta} \) or 2.5th to 97.5th percentiles of draws
approximates 95% posterior credibility interval for \(\theta\)

Example: Posterior Draws for Normal Linear Regression
\((\hat{\beta}, s^2) = \text{ls estimates of slopes and resid variance}\)
\[ \sigma^{(d)^2} = (n - p - 1)s^2 / \chi^2_{n-p-1} \]
\[ \beta^{(d)} = \hat{\beta} + A^T z \sigma^{(d)} \]
\[ \chi^2_{n-p-1} = \text{chi-squared deviate with } n - p - 1 \text{ df} \]
\[ z = (z_1, ..., z_{p+1})^T, z_i \sim N(0,1) \]
\[ A = \text{upper triangular Cholesky factor of } (X^T X)^{-1} : \]
\[ A^T A = (X^T X)^{-1} \]

Likelihood methods with incomplete data
- Statistical model + incomplete data -> Likelihood
- Statistical models needed for:
  - data without missing values
  - missing-data mechanism
- Model for mechanism not needed if it is ignorable
  (to be deﬁned later)
- With likelihood, proceed as before:
  - ML estimates, large sample standard errors
  - Bayes posterior distribution
  - Little and Rubin (2002, chapter 6)
The Observed Data

\[ Y = (y_{ij})_{n \times k} = (Y_{\text{obs}}, Y_{\text{mis}}) \]

\[ R = (r_{ij})_{n \times k} \]

\[ Y = (y_{ij})_{n \times k} = (Y_{\text{obs}}, Y_{\text{mis}}) \]

\[ R = (r_{ij})_{n \times k} \]

\[ r_{ij} = \begin{cases} 1, & y_{ij} \text{ observed} \\ 0, & y_{ij} \text{ missing} \end{cases} \]

Model for Y and R

\[ f(Y, R | \theta, \psi) = f(Y | \theta) f(R | Y, \psi) \]

Complete-data model model for mechanism

Example: bivariate normal monotone data complete-data model:

\[ (y_{i1}, y_{i2}) \sim_{\text{ind}} N_2(\mu, \Sigma) \]

model for mechanism:

\[ (r_{i1}, y_{i2}) \sim_{\text{ind}} \text{Bern} \left( \Phi\left( \psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2} \right) \right) \]

\[ \Phi = \text{Normal cumulative distribution function} \]

Two likelihoods

- **Full likelihood** - involves model for R

  \[ f(Y_{\text{obs}}, R | \theta, \psi) = \int f(Y_{\text{obs}}, Y_{\text{mis}} | \theta) f(R | Y_{\text{obs}}, Y_{\text{mis}}, \psi) dY_{\text{mis}} \]

  \[ \Rightarrow L_{\text{full}}(\theta, \psi | Y_{\text{obs}}, R) = \text{const} \times f(Y_{\text{obs}}, R | \theta, \psi) \]

Likelihood ignoring the missing-data mechanism R

simpler since it does not involve model for R

\[ f(Y_{\text{obs}} | \theta) = \int f(Y_{\text{obs}}, Y_{\text{mis}} | \theta) dY_{\text{mis}} \]

\[ \Rightarrow L_{\text{ign}}(\theta | Y_{\text{obs}}) = \text{const} \times f(Y_{\text{obs}} | \theta) \]

Ignoring the missing-data mechanism

- **Note that if:**

  \[ L_{\text{full}}(\theta, \psi | Y_{\text{obs}}, R) = L(\psi | R, Y_{\text{obs}}) \times L_{\text{ign}}(\theta | Y_{\text{obs}}) \]

  where \[ L(\psi | R, Y_{\text{obs}}) \] does not depend on then inference about \[ \theta \] can be based on \[ L_{\text{ign}}(\theta | Y_{\text{obs}}) \]

The missing-data mechanism is then called **ignorable** for likelihood inference
Ignoring the md mechanism continued

- It is easy to show that sufficient conditions for ignoring the missing-data mechanism are:
  
  (A) Missing at Random (MAR):
  
  \[ f(R \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = f(R \mid Y_{\text{obs}}, \psi) \]
  
  for all \( Y_{\text{mis}} \)
  
  (B) Distinctness:
  
  \( \theta \) and \( \psi \) have distinct parameter spaces
  
  (Bayes: prior distributions of \( \theta \) and \( \psi \) are independent)

---

\textbf{Bivariate Normal Monotone Data}

\[
L_{\text{full}}(\theta, \psi \mid Y_{\text{obs}}, R) = \prod_{i=1}^{r} N_2(y_{i1}, y_{i2} \mid \mu, \Sigma) \times \Phi(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}) \\
\times \prod_{j=1}^{n} \int N_2(y_{j1}, y_{j2} \mid \mu, \Sigma) \times (1 - \Phi(\psi_0 + \psi_1 y_{j1} + \psi_2 y_{j2})) dy_{j2}
\]

Under MAR: \( \Phi(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}) = \Phi(\psi_0 + \psi_1 y_{i1}) \)

\[
L_{\text{full}}(\theta, \psi \mid Y_{\text{obs}}, R) = \prod_{i=1}^{r} N_2(y_{i1}, y_{i2} \mid \mu, \Sigma) \times \Phi(\psi_0 + \psi_1 y_{i1}) \\
\times \prod_{j=1}^{n} \int N_2(y_{j1} \mid \mu, \Sigma) \times (1 - \Phi(\psi_0 + \psi_1 y_{j1})) dy_{j1}
\]

\[
= L_{\text{ign}}(\theta \mid Y_{\text{obs}}, R) \times \prod_{i=1}^{r} \Phi(\psi_0 + \psi_1 y_{i1}) \times \prod_{j=1}^{n} (1 - \Phi(\psi_0 + \psi_1 y_{j1}))
\]

---

\textbf{Bayes: add prior distributions}

\[
p_{\text{complete}}(\theta, \psi \mid Y, R) = \pi(\theta, \psi) \times f(Y \mid \theta) \times f(R \mid Y, \psi)
\]

Prior dn: Complete-data model model for mechanism

\[
p_{\text{full}}(\theta, \psi \mid Y_{\text{obs}}, R) \propto \pi(\theta, \psi) \times f(Y_{\text{obs}}, R \mid \theta, \psi)
\]

- **Full posterior dn** - involves model for \( R \)

\[
f(Y_{\text{obs}}, R \mid \theta, \psi) = \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) f(R \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) dY_{\text{mis}}
\]

Posterior dn \textit{ignoring the missing-data mechanism} \( M \)

(simpler since it does not involve model for \( M \))

\[
p_{\text{ign}}(\theta \mid Y_{\text{obs}}) \propto \pi(\theta) \times f(Y_{\text{obs}} \mid \theta)
\]

\[
f(Y_{\text{obs}} \mid \theta) = \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) dY_{\text{mis}}
\]
Computational tools

- Factored likelihood for monotone patterns: e.g., for bivariate monotone data, factorize joint distribution of $Y_1, Y_2$ into marginal of $Y_2$ and conditional of $Y_2$ given $Y_1$. If parameters distinct, results in two complete data problem (Anderson, 1957). Works for both ML and Bayes
- ML for general patterns: EM algorithm (Dempster, Laird and Rubin 1977) and extensions (ECM, ECME, etc.)
- Bayes for general patterns: data augmentation, Gibbs’ sampler (Tanner and Wong, 1994)
- See e.g. Little and Rubin (2002)

Summary of Theory

- ML and Bayes are the main forms of likelihood-based inference
  - Parametric multiple imputation is also justified as a simulation approximation to posterior distribution
- Likelihood Inference Ignoring the Missing Data Mechanism is valid if
  - Model for $Y$ is correctly specified
  - Data are MAR
  - Fully efficient if distinctness condition holds
  - In contrast, many ad-hoc methods require the stronger MCAR assumption

Non-linear effects and interactions

We will focus on non-linear effects, since many of the issues with interactions are similar.

Specifically we consider two situations:
- when variables involved fully observed
- when variables involved missing
Example

The figure shows a plot of brain atrophy and levels of a cerebrospinal fluid peptide (C&K, p129).

94/199 individuals are missing the covariate.

How should we impute?

Option 1: passive imputation

For patient $i$, let $Y_{i,1}$ denote brain atrophy and $Y_{i,2}$ denote peptide level.

1. Impute, ignoring non-linear relationship:

   $Y_{i,1} = \beta_{0,1} + \epsilon_{i,1}$
   $Y_{i,2} = \beta_{0,2} + \epsilon_{i,2}$
   \($\epsilon_{i,1}, \epsilon_{i,2}$) $\sim N(0, \Omega)$.

2. Square imputed peptide values in each imputed dataset.

3. Fit the substantive model to each imputed dataset:

   $Y_{i,1} = \beta_0 + \beta_1 Y_{i,2} + \beta_2 Y_{i,2}^2 + \epsilon_i$
   and apply Rubin’s rules.

Results

Neural degeneration data: parameter estimates (standard errors) from fitting quadratic regression to the 104 complete records and after MI using the passive approach.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Complete records</th>
<th>Passive imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>34.2 (7.0)</td>
<td>20.7 (6.5)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$-0.254$ (0.075)</td>
<td>$-0.106$ (0.067)</td>
</tr>
<tr>
<td>$\beta_2(\times 10^4)$</td>
<td>5.33 (1.91)</td>
<td>1.61 (1.65)</td>
</tr>
</tbody>
</table>

Passive imputation dilutes the quadratic association.

No missing data in nonlinear relationships

Suppose we have three continuous variables, and substantive model

$Y_{i,1} = \beta_0 + \beta_1 Y_{i,2} + \beta_2 Y_{i,3} + \beta_3 Y_{i,3}^2 + \epsilon_i$, \(\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Suppose that $Y_2$ is partially observed, but that the other two variables are fully observed.

If $(Y_{i,1}, Y_{i,2})$ are approximately bivariate normal given $Y_{i,3}$, with conditional mean a linear function of $Y_{i,3}$, $Y_{i,3}^2$, then

$Y_{i,2} = \gamma_0 + \gamma_1 Y_{i,1} + \gamma_2 Y_{i,3} + \gamma_3 Y_{i,3}^2 + \tilde{\epsilon}_i$, \(\tilde{\epsilon}_i \overset{iid}{\sim} N(0, \tilde{\sigma}^2)$ (1)

is a congenial imputation model.

The key is that we need the non-linear effect of $Y_{i,3}$ in (1).

If $(Y_{i,1}, Y_{i,2})$ are not plausibly bivariate normal, see C&K §6.3.3 and p. 131.
Missing data in non-linear relationships
We consider two approaches
  - Just Another Variable (JAV)
  - Rejection sampling

Just Another Variable
Suppose we have two variables, both partially observed, and the model of interest is
\[ Y_{i,1} = \beta_0 + \beta_1 Y_{i,2} + \beta_2 Y_{i,2}^2 + e_i, \quad e_i \sim N(0, \sigma^2), \quad (2) \]
with \( Y_2 \) marginally normally distributed.

JAV proposes imputing \( Y_2 \) and \( Z = Y_2^2 \) treating them as if they were unrelated variables; hence the name ‘just another variable.’ Thus the joint imputation model is
\[ Y_{i,2} = \beta_{0,2} + \beta_{1,2} Y_{i,1} + e_{i,2} \]
\[ Z_{i,3} = \beta_{0,3} + \beta_{1,3} Y_{i,1} + e_{i,3} \]
\[ (e_{i,2}, e_{i,3})^T \sim \text{iid} N_2(0, \Omega). \]

Seaman, Bartlett and White (2012) criticize this approach, and show it only gives valid point estimates for linear regression models when data are MCAR. See also C&K p.134.

Full MCMC via rejection sampling
Consider again (2), and write the joint distribution of \( Y_1, Y_2 \) as
\[ f_{1,2}(Y_1, Y_2) = f_{1|2}(Y_1|Y_2)f_2(Y_2), \]
where \( f_{1|2}(Y_1|Y_2) \) is (2), and \( f(Y_2) \) is the marginal distribution of \( Y_2 \) which we assume is normal.

Suppose, given current draws for missing \( Y_2 \), we have used standard MCMC methods to draw
  - \( \beta, \sigma^2 \) (parameters of substantive model)
  - \( \tilde{\mu}, \tilde{\sigma}^2 \) (parameters of marginal distribution of \( Y_2 \))

We now use a rejection step to draw new values of missing \( Y_2 \).

Details I
For unit \( i \), suppose we draw missing \( Y_{i,2} \) from \( g \), and that
\[ \max_{Y_2} \left( \frac{f_{2|1}(Y_2 \mid Y_{i,1})}{g(Y_2)} \right) \leq K. \]

Then, if we accept the proposed value with probability
\[ \frac{f_{2|1}(Y_2 \mid Y_{i,1})}{g(Y_2) \times K}, \]
it is a valid draw from \( f_{2|1}(Y_2 \mid Y_{i,1}) \), i.e. the distribution of the missing given observed data.
Details II

If we choose \( g = f(Y_{i,2}; \tilde{\beta}, \tilde{\sigma}^2) \), i.e. as the marginal distribution of \( Y_2 \), then it follows that for a draw \( Y_{i,2}^* \), the acceptance probability is

\[
\frac{f_{21}(Y_{i,2}^* | Y_{i,1})}{g(Y_{i,2})} \times K = \exp\left\{ -\frac{1}{2\sigma^2}(Y_{i,1} - \beta_0 - \beta_1 Y_{i,2}^* - \beta_2 Y_{i,2}^*)^2 \right\}.
\]

For corresponding details for logistic regression see C\&K p.144, and for survival analysis p.171–2.

Bartlett et al (2014) show how this approach can be incorporated in the FCS framework.

Simulation study

Simulate data for \( n = 1000 \) units from the substantive model

\[
Y_{i,1} = \beta_0 + \beta_1 Y_{i,2} + \beta_2 Y_{i,2}^2 + e_i, \quad e_i \sim N(0, \sigma^2), \quad i \in (1, \ldots, n)
\]

and marginally, \( Y_{i,2} \sim N(0, 1) \).

Make values of \( Y_{i,2} \) either
- MCAR with probability 0.5, or
- MAR under the mechanism

\[
\text{logit} \ Pr(\text{observe } Y_{i,2}) = 0.37 - 0.41 Y_{i,1}
\]

Compare Passive, JAV and FCS (with rejection sampling).

Results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Complete records</th>
<th>Multiple imputation using passive</th>
<th>Multiple imputation using JAV</th>
<th>Rejection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data MCAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_0 = 0 )</td>
<td>0.003 (0.094)</td>
<td>0.435 (0.080)</td>
<td>0.012 (0.082)</td>
<td>0.002 (0.080)</td>
</tr>
<tr>
<td>( \beta_1 = 1 )</td>
<td>1.002 (0.079)</td>
<td>0.977 (0.102)</td>
<td>1.012 (0.082)</td>
<td>1.002 (0.079)</td>
</tr>
<tr>
<td>Data MAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_0 = 0 )</td>
<td>-0.461 (0.092)</td>
<td>0.649 (0.089)</td>
<td>-0.076 (0.111)</td>
<td>0.000 (0.085)</td>
</tr>
<tr>
<td>( \beta_1 = 1 )</td>
<td>0.900 (0.096)</td>
<td>0.844 (0.140)</td>
<td>1.267 (0.132)</td>
<td>0.993 (0.099)</td>
</tr>
<tr>
<td>( \beta_2 = 1 )</td>
<td>0.507 (0.079)</td>
<td>0.527 (0.056)</td>
<td>1.271 (0.103)</td>
<td>1.005 (0.069)</td>
</tr>
</tbody>
</table>

We see that only rejection sampling performs adequately. See also neural degeneration example, C\&K p.145.

Interactions: interaction variable fully observed

Consider a substantive model with an interaction involving a binary variable, for example if \( Y_{1,3} \) is binary:

\[
Y_{i,1} = \beta_0 + \beta_1 Y_{i,2} + \beta_2 Y_{i,2}^2 + \beta_3 Y_{i,2} Y_{i,3} + e_i, \quad e_i \sim N(0, \sigma^2)
\]

Suppose that there may be missing data on \( Y_1, Y_2 \) but that \( Y_3 \) is fully observed. Then we can proceed as follows:

- split the data into two sets by \( Y_3 \);
- perform MI (with \( Y_1, Y_2 \)) separately in each group, creating \( K \) imputed data sets for each group;
- For \( k = 1, \ldots, K \), append the \( k^{th} \) imputed dataset for \( Y_3 = 1 \) to that of the \( k^{th} \) imputed dataset for \( Y_3 = 0 \);
- Fit the substantive model to each imputed dataset and apply Rubin’s rules.

This approach can also be useful in clinical trials, where we may impute separately in each treatment group.
Interaction variable partially observed

In this setting we have three broad options:

1. Omit records with the interaction variable missing;
2. Impute with all interactions, and
3. Adopt the analogue of the rejection sampling approach discussed above for non-linear relationships

Option 1 is attractive if it removes relatively few records. We can then apply MI and impute separately in groups.

If the chance of the interaction variable being missing, given the other variables in the substantive model and any auxiliary variables, does not depend on the response, this will not result in any bias. See Carpenter & Plewis (2011) for an example.

Summary

- If non-linear effects and/or interactions are present in the substantive model, the imputation model must reflect this.
- If the variables involved are observed, this is relatively straightforward: we include them as covariates in our imputation model.
- If not, then the approach proposed above is preferable to both JAV and passive imputation.
- Stata software from www.missingdata.org.uk; a version of REALCOM with this functionality is forthcoming.
SESSION 2, Part I

Overview

- Inverse probability weighting (IPW) and extensions
- Little and Rubin (2002, chapter 3)
Complete cases

Imputation

Complete cases

Analysis

Incomplete

Imputations

Weights

Imputations

Weights

Imputations

Weights

Imputations

Weights

Imputations

Weights

Imputations

Weights

Imputations

Weights

Weighted CC analysis

• Weight respondents differentially to reduce nonresponse bias – e.g.
• Mean $$\bar{y} = \frac{\sum y_i}{n}$$ becomes weighted mean $$\bar{y}_w = \frac{\sum w_i y_i}{\sum w_i}$$
• Common for unit nonresponse in surveys
• “Quasi-randomization inference” : extends ideas of randomization inference in surveys

Unit nonresponse weights

• If probability of response was known, could obtain weight for units that are sampled and respond:
  
  \[
  w_i = \frac{1}{pr(\text{unit i is sampled and responds})} = \frac{1}{pr(\text{i sampled}) \times pr(\text{i responds|sampled})} = (\text{sampling weight}) \times (\text{response weight})
  \]

Since prob of response is not known, we need to estimate it.

Adjustment cell weighting

• Group respondents and nonrespondents into adjustment cells with similar values on variables recorded for both:
  
  • e.g. white females aged 25-35 living in SW
  
  100 in sample  <  80 respondents
  
  20 nonrespondents

  pr(response in cell) = 0.8
  
  weight for respondents in this cell = 1.25
Choice of adjustment cells

- To reduce bias, adjustment cell variable needs to be related to nonresponse $R$ and outcome $Y$
- To reduce variance, adjustment cell variables need to be related to outcome (otherwise weighting *increases* variance)
- Two methods for creating adjustment cells with multiple $X$’s are
  - Response propensity stratification, based on predicted propensities from regression of $R$ on $X$ estimated on all the data
  - Predictive mean stratification, based on predicted means from regression of $Y$ on $X$ estimated on the complete cases

Augmented IPW methods

Augmented IPW creates predictions from a model, and adds IPW residuals for robustness to model misspecification.

- **Double-robustness property**: Consistent estimates under MAR if either (a) the model relating the mean of the outcome variables to the design variables is incorrect or (b) if the missingness model is correct.
- Model predictions based on good auxiliary data increase efficiency of IPW

Weighted estimating equations

- More generally, estimator is solution of weighted estimating equations (WEE):
  \[ \text{If } E[w_i g(y_i, \theta)] = 0, \text{ then } \]
  \[ \text{Choose } \hat{\theta}_{\text{WEE}} \text{ to solve WEE: } \sum_{i \in \mathcal{I}} r_i w_i g(y_i, \theta) = 0 \]
- A particular case is to define $g$ as contribution to likelihood for a particular model
  - “quasi” or “pseudo” likelihood
  - Not in my view model-based, since weighting the likelihood contributions is ad-hoc

Augmented IPW estimate of mean

- Calibrate the predictions from a parametric model by adding means of the weighted residuals (Robins, Rotnitzky and Zhao, 1994; Scharfstein, Rotnitzky and Robins, 1999).
  \[ \hat{\mu} = n^{-1} \left( \sum_{i \in \mathcal{I}} \hat{y}_i \right) + n^{-1} \left( \sum_{i \in \mathcal{I}} w_i (y_i - \hat{y}_i) \right) \]
  \[ w_i = \pi^{-1}(X; \hat{\alpha}) \text{ is an estimate of the inverse of probability of being observed} \]
Double Robustness Property

- The augmented IPW estimate is doubly robust, meaning that it yields a consistent estimator for the marginal mean of $Y$, if EITHER:
  - (a) the mean of $Y$ given $X_1,...,X_p$ is correctly specified, OR
  - (b) the propensity is correctly specified
- That is, the calibration using the weighted residuals reduces bias if the model for $Y$ is not correctly specified
- BUT -- Does not necessarily help when both models are misspecification (Kang and Schafer, 2007)

An alternative approach: Penalized Spline of Propensity Prediction (PSPP)

- Uses the (estimated) response propensity as a covariate rather than as a weight
- Important to get relationship between $Y$ and response propensity correct, since misspecification of this leads to bias
- A property of propensity is that other covariates are balanced over respondents and nonrespondents, conditional on propensity scores (Rosenbaum and Rubin 1983); so misspecification of regression of these is less important (loss of precision, not bias).

The PSPP method

Define:  $Y^* = \text{logit} (\Pr(R = 1|X_1,...,X_p))$ (Need to estimate)

$$(Y | Y^*, X_1,..., X_p; \beta) \sim N(s(Y^*) + g(Y^*, X_2,..., X_p; \beta), \sigma^2)$$

- Nonparametric part
- Need to be correctly specified
- We choose penalized spline
- Parametric part
- Misspecification does not lead to bias
- Increase precision
- $X_i$ excluded to prevent multicollinearity

Double Robustness Property

- The PSPP method yields a consistent estimator for the marginal mean of $Y$, if:
  - (a) the mean of $Y$ given $X_1,...,X_p$ is correctly specified,
  - (b1) the propensity is correctly specified, and
  - (b2) relationship of outcome with propensity is correctly specified

Note: in (b), parametric part of model does not have to be correctly specified!
Comments

• Augmented IPW approach yields double robustness
• However, relatively easy to achieve double robustness in the direct prediction approach, using methods like PSPP (see Firth & Bennett 1998)
• With an appropriate robust model, GREG calibration is unnecessary and adds noise
• PSPP does well in frequentist simulations (Zhang and Little 2011; Yang and Little 2014)

Item nonresponse

• Item nonresponse generally has complex “swiss-cheese” pattern
• Weighting methods are possible when the data have a monotone pattern, but do not work for a general pattern
• Model-based multiple imputation methods are more suited to this situation
  – By conditioning fully on all observed data, these methods weaken MAR assumption

Inference from weighted data

• Role of survey weights in analytical inference (regression, factor analysis, ...) is controversial
• Can use packages for computing standard errors for complex sample designs -- but often these do not take into account sampling uncertainty in weights
• Bootstrap/Jackknife of weighting procedure propagates uncertainty in weights -- but weights need to be recalculated on each BS/JK sample

Summary of weighting methods

• Weighting is a relatively simple device for reducing bias from complete-case analysis
• Same weight for all variables -- simple, but better methods tune adjustment according to outcome
• No built in control of variance
  – ad-hoc trimming is common in surveys
• Less useful when:
  – Covariate information is extensive
  – Pattern of missing-data is non-monotone
Multilevel multiple imputation
Many datasets are hierarchical, and this should be respected when performing imputation, especially if the data are unbalanced (C&K, Ch.9).

Further, many datasets have genuine level-2 variables; treating these as level-1 observations is not generally appropriate.

Thus we now describe a multilevel latent normal model for multiple imputation of multilevel mixed response data (Goldstein et al, 2009).

We begin by describing a model for multilevel responses, and then extend to mixed responses and multiple imputation.

Multilevel responses: childhood and adult height
As an illustration of multilevel responses, consider modelling childhood and adult heights.

The dataset consists of 108 children with height measured on up to six occasions around the age of 13 together with their adult heights, altogether 436 growth measurements and 108 adult height measurements.

We have a two level model
- Level 1 is the repeated measures of childhood height
- Level 2 is the adult height

Such a model could be used to predict adult height from childhood height measurements.
Model
Let the superscript (1) denote level 1 (childhood heights) and (2) level 2 (adult height).

Let \( j = 1, ..., J \) denote people; \( i = 1, ..., I_j \) childhood height measurements.

Model:
\[
\begin{align*}
y_{i}^{(2)} &= \gamma_0 + u_{ij}^{(2)} \\
y_{ij}^{(1)} &= (\beta_0 + u_{ij}^{(1)}) + (\beta_1 + u_{ij}^{(1)})t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + e_{ij}
\end{align*}
\]
\[
\begin{pmatrix}
u_{ij}^{(2)} \\
u_{ij}^{(1)} \\
e_{ij}
\end{pmatrix} \sim MVN(0, \Omega_u)
\]
An MCMC algorithm for fitting this model is given in C&K, §9.2

Parameter estimates

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1 model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept (cm)</td>
<td>153.05</td>
<td>0.69</td>
</tr>
<tr>
<td>Age (centred 13 years)</td>
<td>7.07</td>
<td>0.16</td>
</tr>
<tr>
<td>Age-squared</td>
<td>0.294</td>
<td>0.054</td>
</tr>
<tr>
<td>Age-cubed</td>
<td>-0.208</td>
<td>0.029</td>
</tr>
<tr>
<td>Level 2 model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept (cm)</td>
<td>174.7</td>
<td>0.80</td>
</tr>
<tr>
<td>Level 2 covariance matrix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>57.77</td>
<td>1.30</td>
<td>50.01</td>
</tr>
<tr>
<td>1.30</td>
<td>0.53</td>
<td>1.24</td>
</tr>
<tr>
<td>50.01</td>
<td>1.24</td>
<td>69.42</td>
</tr>
<tr>
<td>Level 1 variance</td>
<td>3.21</td>
<td></td>
</tr>
</tbody>
</table>

Including binary data

Our imputation model will typically be a multivariate response multilevel model. Some of the responses may be binary. We accommodate these through a latent normal structure.

Let \( y_{i,j} \) be a binary variable. Define a latent normal variable \( z_{i,j} \) by
\[
z_{ij} > 0 \iff y_{i,j} = 1,
\]
and model
\[
z_{ij} = x_{ij}^T \beta + e_{ij}, \quad e_{ij} \sim N(0, 1)
\]
Then
\[
Pr(y_{i,j} = 1) = Pr\{e_{ij} > -(x_{ij}^T \beta)\} = \int_{-x_{ij}^T \beta}^{\infty} \phi(t) dt
\]
\[
= \Phi(x_{ij}^T \beta)
\]

Fitting: MCMC steps

- Given current draws of all parameters, and binary variable \( y_{i,j} \), update latent normal \( z_{i,j} \)'s. This can be readily done using rejection sampling.
- Once the latent normal values are drawn, the MCMC algorithm proceeds in the usual way.
- The exception is updating the level-1 covariance matrix, where the latent normal for the binary response is constrained to have variance 1.

To accommodate this, update the covariance matrix element-wise, using Metropolis-Hastings steps (C&K, p.97).

Binary data at level 2
We can include level 2 binary variables using the same approach. Now the level 2 covariance matrix needs to be updated element-wise too.
Suppose, \( y_{i,j} \) is a categorical response with \( M \) categories.

- We only need to model \( M - 1 \) probabilities. We do this by creating \( m = 1, \ldots, (M - 1) \) independent latent normal variables, \( z_{i,j,m} \), with variance 1. For each, we have a separate regression coefficient \( \beta_m \) relating covariates \( x_{i,j} \):

\[
z_{i,j,m} = x_{i,j}^T \beta_m + e_{i,j,m}, \quad m = 1, \ldots, (M - 1), \quad e_{i,j,m} \sim N_{M-1}(0, I_{K-1}).
\]

- Then, for \( m = 1, \ldots, (M - 1) \),

\[
Pr(y_{i,j} = m) = Pr(z_{i,j,m} > z_{i,j,m'}, \text{ all } m' \neq m) = Pr(e_{i,j,m} > x_{i,j}^T(\beta_m - \beta_{m'})), \text{ all } m' \neq m
\]

and

\[
Pr(y_{i,j} = M) = Pr(\text{all } z_{i,j,m'} < 0), \text{ all } m' = 1, \ldots, (M - 1).
\]

- We update the latent normal \( z \)'s, conditional on other responses and parameters, by rejection sampling from the appropriate conditional normal distribution (C&K, §5.3).

---

### Missing data

<table>
<thead>
<tr>
<th>%</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>miss</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adm</td>
<td>1197</td>
<td>1210</td>
<td>1188</td>
<td>1007</td>
<td>798</td>
<td>798</td>
<td>1116</td>
<td>1035</td>
<td></td>
</tr>
<tr>
<td>Mean comp (%)</td>
<td>89</td>
<td>72</td>
<td>88</td>
<td>83</td>
<td>32</td>
<td>49</td>
<td>63</td>
<td>75</td>
<td>0</td>
</tr>
<tr>
<td>PAR use (%)</td>
<td>97</td>
<td>82</td>
<td>96</td>
<td>92</td>
<td>22</td>
<td>52</td>
<td>60</td>
<td>97</td>
<td>0</td>
</tr>
<tr>
<td>Fem. ch (%)</td>
<td>42</td>
<td>39</td>
<td>41</td>
<td>39</td>
<td>44</td>
<td>38</td>
<td>44</td>
<td>41</td>
<td>6</td>
</tr>
<tr>
<td>Fem. clin (%)</td>
<td>76</td>
<td>60</td>
<td>52</td>
<td>70</td>
<td>64</td>
<td>92</td>
<td>58</td>
<td>54</td>
<td>14</td>
</tr>
<tr>
<td>Med yrs ex.</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>CME (%)</td>
<td>46</td>
<td>20</td>
<td>35</td>
<td>59</td>
<td>Not Offered</td>
<td></td>
<td></td>
<td></td>
<td>9</td>
</tr>
</tbody>
</table>

Most missing data is at the clinician level; this results in a complete records analysis of 6775/8349 (81%) records.

---

### Example: Kenyan paediatric hospital records

Ayieko et al (2011) report the results of a cluster randomised trial to evaluate an intervention to improve admission paediatric care in Kenyan hospitals.

We take a selection of the data, and our substantive model relates the percentage of admissions tasks completed (on any admission) to child and admitting clinician characteristics:

\[
\begin{align*}
\text{complete}_{i,j,k} & = \beta_0 + u_{j,k} + \nu_k + \beta_1[\text{PAR use}_{i,j,k}] + \beta_2[\text{female child}_{i,j,k}] + \\
& \quad + \beta_3[\text{female clinician}_{j,k}] + \beta_4[\text{years experience}_{j,k}] + \\
& \quad + \beta_5[\text{intervention hospital}_{i}] + \beta_6[\text{clinician takes CME}_{j,k}] + \\
& \quad + \beta_7[\text{PAR use}_{i,j,k}] + 1[\text{clinician takes CME}_{j,k}] + e_{i,j,k}
\end{align*}
\]

\[
\nu_k \sim N(0, \sigma^2_k), \quad u_{j,k} \sim N(0, \sigma^2_u), \quad e_{i,j,k} \sim N(0, \sigma^2_e).
\]

Here \( k \) indexes hospital, \( j \) clinician and \( k \) child.

---

### Multilevel Multiple Imputation

We split the data into 4 subsets defined by: H1–4 vs H5–8 and PAR use. We impute separately in each, then combine before fitting the substantive model. Imputation proceeds as follows:

- fit, using MCMC, a multivariate response model to the observed data, with partially observed variables as responses, and fully observed ones as covariates;
- draw the missing data from the posterior distribution, given the observed data, to ‘complete’ the data set;
- continue updating the sampler, and form a second ‘complete’ data set in the same way;
- repeat, till \( K \) ‘complete’ data sets have been obtained;
- fit the substantive model to each ‘completed’ dataset, obtaining parameter estimates and their variance/covariance matrix, and combine for final inference using Rubin’s rules.
Results (10 imputations)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete records</th>
<th>Multi-level MI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>SE</td>
</tr>
<tr>
<td>PAR</td>
<td>50.20</td>
<td>0.43</td>
</tr>
<tr>
<td>female child</td>
<td>0.26</td>
<td>0.23</td>
</tr>
<tr>
<td>female clin.</td>
<td>0.15</td>
<td>0.88</td>
</tr>
<tr>
<td>years exp.</td>
<td>-0.33</td>
<td>0.08</td>
</tr>
<tr>
<td>int. hosp.</td>
<td>9.99</td>
<td>3.10</td>
</tr>
<tr>
<td>CME</td>
<td>0.71</td>
<td>1.71</td>
</tr>
<tr>
<td>PAR use × CME</td>
<td>-0.46</td>
<td>1.37</td>
</tr>
<tr>
<td>constant</td>
<td>26.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Variance comp</td>
<td>95% Conf. Int.</td>
<td>95% Conf. Int.</td>
</tr>
<tr>
<td>hospital, $\sigma^2_u$</td>
<td>17.10</td>
<td>(5.02–58.25)</td>
</tr>
<tr>
<td>clinician, $\sigma^2_v$</td>
<td>41.06</td>
<td>(33.41–50.46)</td>
</tr>
<tr>
<td>child, $\sigma^2_e$</td>
<td>88.95</td>
<td>(85.93–92.07)</td>
</tr>
</tbody>
</table>

Bias corrected and inference improved.

Summary: multilevel imputation

- Building on similar models in the literature, we have developed a multilevel multivariate response model (Goldstein et al. 2009).
- The latent normal model is an attractive way of allowing mixed responses; for the extension to include Poisson data (Goldstein & Kounali, 2009).
- We have also explored a Box-Cox transformation to take account of non-normal responses.
- This approach has been successfully applied to cost-effectiveness data (Diaz et al, 2013).

Likelihood methods for non-ignorable models

- Statistical model + incomplete data => Likelihood
- For non-ignorable missing data, models needed for:
  - data without missing values
  - missing-data mechanism
- ML, Bayes and MI can be applied to these models
  - tools like factored likelihood, EM, Gibbs sampler apply to these models too
- But beware of unidentified or poorly-identified parameters, leading to numerical problems (e.g. lack of convergence). (Little and Rubin 2002, chapt. 15)
Models for Y and R

- Let \((y_i, r_i)\) denote the complete-data vector and missing-data indicator for the \(i\)th unit, and assume independence across units. Two generic modeling approaches are:
  
  **Selection models**, which factor:
  \[
  f(y_i, r_i \mid \theta, \psi) = f(y_i \mid \theta) \times f(r_i \mid y_i, \psi)
  \]

  - complete-data model
  - model for md mechanism

  **Pattern-mixture models**, which factor:
  \[
  f(y_i, r_i \mid \phi, \gamma) = f(y_i \mid r_i, \phi) \times f(r_i \mid \gamma)
  \]

  - model for \(y\)'s within pattern \(r_i\)
  - probability of pattern \(r_i\)

Selection or Pattern-Mixture Models?

- Selection models are:
  - more natural substantive formulation of model, if inference concerns the entire population
  - more common approach in literature
  - sensitive to specification of the form of the missing-data mechanism, which is often not well understood

- Pattern-mixture models are:
  - More natural when interest is in population strata defined by missing-data pattern
  - closer to the form of the data, sometimes simpler to fit
  - can avoid specifying the form of the md mechanism, which is incorporated indirectly via parameter restrictions.

Monotone Bivariate Normal: Probit Selection Model

\[
(y_{1i}, y_{2i}) \sim_{iid} N_2(\mu, \Sigma)
\]

\[
(r_{1i} \mid y_{1i}, y_{2i}) \sim_{ind} \text{Bern} \left[ \Phi(\psi_0 + \psi_1 y_{1i} + \psi_2 y_{2i}) \right]
\]

\(\Phi = \text{normal cumulative distribution function}\)

Intuitively, the parameter \(\psi_2\) is not estimable, because there are no cases with \(R = 0\) and \(Y\) observed

Solutions:
1. Assume \(\psi_2 = 0\), that is, data are MAR
2. Assume value of some other parameter, e.g. \(\psi_1 = 0\)
3. Vary \(\psi_2\) in a sensitivity analysis

Monotone Bivariate Normal: pattern-mixture model

\[
(y_{1i}, y_{2i} \mid r_{1i} = r) \sim_{iid} N_2(\mu^{(r)}, \Sigma^{(r)})
\]

\[
(r_{1i}) \sim_{ind} \text{Bern}[\gamma]
\]

The marginal of \(Y\) is then a mixture of normals

Here without more assumptions, there is no information about the parameters of the regression of \(Y_2\) on \(Y_1\) when \(R_2=0\)
Missing Data in Clinical Trials

- Recent U.S. National Research Council (2010) report on missing data in clinical trials recommends sensitivity analysis
- Parameters of MNAR models cannot be reliably estimated – identifiability requires structural assumptions that are often questionable
- Varying certain parameters in a sensitivity analysis is the preferred approach
- In many (not all) situations, it would be reasonable to choose an MAR primary model, and look at MNAR models via a sensitivity analysis to assess plausible deviations from MAR

Simple example

For treatment $t$, let

$\mu_{t1} = E(Y \mid T = t, R = 1), \mu_{0t} = E(Y \mid T = t, R = 0)$

Goal is to estimate $\mu_t = \pi_t \mu_{t1} + (1 - \pi_t) \mu_{0t}$

There is information about respondent means

$\{\mu_{t1}\}$ and response rates $\{\pi_t\}$

No information about nonrespondent means $\{\mu_{0t}\}$

Simple example

Consider first two treatments $T = 1, 2$, single outcome $Y$, no auxiliary data

Let $R = 1$ if $Y$ is observed, $R = 0$ otherwise

Problem is to estimated mean in each treatment arm based on data from subjects in each arm, where some of the $Y$'s are missing.

MAR would assume that within each treatment arm, the distribution of $Y$ for respondents was the same as for nonrespondents.

Simple pattern-mixture model for sensitivity analysis

$[Y \mid R = 1, T = j] \sim N(\mu_{1j}, \sigma_j^2)$

$[Y \mid R = 0, T = j] \sim N(\mu_{0j} + \Delta_j, \sigma_j^2)$

$Pr(R = 1 \mid T = j) = \pi_j$

MAR: $\Delta_1 = \Delta_2 = 0$

MNAR: For prespecified plausible values of $\Delta_1, \Delta_2$:

generate inferences about $\mu_1, \mu_2$ and $\mu_2 - \mu_1$

by replacing $\mu_{1j}, \sigma_j^2$ and $\pi_j$ by sample estimates, with associated standard errors.

Examples of choices of $\Delta_1, \Delta_2$:

$\Delta_j = k\sigma_j, k = 0.2, 0.5$
Variants and extensions

(A) Assume $\mu_0 = g^{-1}\{g(\mu_i) + \Delta_i\}$, where $g$ is a link function suitable for the outcome, e.g. logit for binary outcome.

(B) Include covariates and auxiliary variables, $Z$:

$[Y_i | R_i = 1, T_i = j, z_i] \sim N(\beta_{0j} + \beta_{1j}z_i + \sigma^2_j)$

$[Y_i | R_i = 0, T_i = j, z_i] \sim N(\beta_{0j} + \beta_{1j}z_i + \Delta_j, \sigma^2_j)$

Multiple imputation is a convenient way of marginalizing over auxiliary $z_i$.

If $\Delta_j = k\sigma_j$, $k = 0.2, 0.5$, then good covariates are "rewarded" by a reduction in size of $\sigma_j, \Delta_j$.

Sensitivity analysis for selection models

For selection models, a sensitivity analysis can be conducted by making an assumption about how the odds of nonresponse change with the values of the outcome $Y$.

For example, one can assume that the log odds of nonresponse differs by $\alpha$ for those who differ by one unit on $Y$, that is

$logit \{P[R = 0 | V,Y = y]\} = h(V) + \alpha y$.

This is a selection model since it models the probability of nonresponse as a function of the outcome and auxiliary variables $V$.

A sensitivity analysis for a survival model (Jansen)

Adopting a value of $\alpha$ is equivalent to adopting a known link between the distribution of the respondents and the nonrespondents.

A sensitivity analysis consists of repeating the inference for $\mu$ at different plausible values of $\alpha$ to assess the sensitivity of inferences about $\mu$ to deviations from MAR.

Prefer the pattern-mixture approach since it is easier to implement and explain to clinicians, and $\alpha$ has a tricky interpretation.

TEST 1: mITT/All Strata/Combined Doses
“Tipping Point” = 2300% inflation

Percent Inflation of the Hazard in Treatment group Subjects with Missing Data
Summary

- Sensitivity analysis is a scientific way of attempting to reflect uncertainty arising from potentially MNAR missing data
- Deciding on how to implement and interpret a sensitivity analysis in the regulatory setting is challenging
- The need and importance of sensitivity analysis increases with the amount of potentially MNAR missing data
- This reinforces the need to limit missing data in the design and implementation stage
  – Avoiding substantial amounts of missing data is key!

Conclusions

- Nonignorable mechanisms can be included in a missing-data analysis, but this is a difficult modeling problem
- Often little is known about the missing-data mechanism, and results may be sensitive to formulation
- Parameters of missing-data are often unidentified or weakly-identified from the data ...
- As a result, it may be more appropriate to do a sensitivity analysis, fixing weakly identified parameters at different values.
- Software for fitting non-ignorable models is not widely available
- Design to avoid nonignorable missing data is preferable if possible

Examples

We conclude with two examples of pattern-mixture sensitivity analysis for randomized controlled trials:
  - using expert information to quantify the departure from MNAR
  - borrowing information from other groups

In both cases MI provides a flexible route for estimation and inference.
Peer review trial

Schroter et al (2004) report a single blind randomised controlled trial among reviewers for a general medical journal. The aim was to investigate whether training improved the quality of peer review.

The study compared two different types of training (face to face training, or a self-taught package) with no training.

We restrict ourselves to the comparison between those randomised to the self-training package and no-training. Each participating reviewer was pre-randomised into their intervention group. Prior to any training, each was sent a baseline article to review (termed paper 1). If this was returned, then according to their randomised group, the reviewer was either (i) mailed a self-training package or (ii) received no further intervention.

Withdrawal pattern

<table>
<thead>
<tr>
<th>Group</th>
<th>Control</th>
<th>Postal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Returned review of paper 2</td>
<td>n</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>2.65</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.81</td>
</tr>
<tr>
<td>Did not return review of paper 2</td>
<td>n</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>3.02</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Review Quality Index of paper 1 by whether or not paper 2 was reviewed.

...ctd

Two to three months later, participants who had completed their first review were sent a further article to review (paper 2).

The analysis excluded all participants who did not complete their first review: this was not expected to cause bias since these participants were unaware of their randomised allocation.

Pattern mixture model

We use the following approach:

1. Draw a parameter vector from the Bayesian posterior distribution under MAR,
   \[ f(\eta|Y_O) \]

2. Modify \( \eta \rightarrow \tilde{\eta} \) to reflect the departure from MAR;

3. Impute the missing data from \( f(Y_M|Y_O, \tilde{\eta}) \)

4. Fit the substantive model to the imputed data set.

We repeat steps 1–3 \( K \) times, and then summarise the results for inference using Rubin's rules.
Peer review study

Focusing on the baseline adjusted comparison of the self-taught training package with no training, the model of interest is

\[ Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \]  

(1)

where \( i \) indexes participant, \( Y_i, X_{1,i} \) are the mean review quality index for paper 2 and paper 1 respectively and \( X_{2,i} \) is an indicator for the self-training group.

Specifying the departure from MAR

White et al (2007) devised a questionnaire which was completed by 2 investigators and 20 editors and other staff at the *British Medical Journal*. The questionnaire was designed to elicit the experts’ prior belief about the de facto difference between the average missing and average observed review quality index.

They showed that it was reasonable to pool information from the experts.

The resulting distribution is negatively skewed, with mean \(-0.21\) and SD 0.46 (on the review quality index scale).

...ctd

Suppose we denote by \((\delta_0, \delta_1)\) draws from the distribution of the mean difference in review quality between observed and unobserved reviews, in respectively the control and self-training groups. We adopt a bivariate normal model approximation to the prior:

\[
\begin{pmatrix}
\delta_0 \\
\delta_1
\end{pmatrix}
\sim N
\begin{pmatrix}
-0.21 \\
-0.21
\end{pmatrix},
0.46^2
\begin{pmatrix}
1 & \rho \\
\rho & 1
\end{pmatrix}
\]

Unfortunately, it was not possible to elicit a prior on \( \rho \) from the experts; we therefore analyse the data with \( \rho = 0, 0.5, 1 \) below.

Algorithm

Given a draw \((\delta_0, \delta_1)\) from this distribution the model is

\[ Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i \quad \text{if } Y_i \text{ observed}, \]
\[ Y_i = (\beta_0 + \delta_0) + \beta_1 X_{1,i} + (\beta_2 + \delta_1 - \delta_0) X_{2,i} + \epsilon_i \quad \text{if } Y_i \text{ unobserved}, \]

Thus the mean review quality, relative to that in the observed data, is changed by \( \delta_0 \) in the control arm and \( \delta_1 \) in the self-taught arm.
Following the general approach for estimating pattern mixture models via MI described above, we proceed as follows, noting that the imputation model under MAR and the substantive model are the same in this example:

1. Fit the imputation model to the observed data and draw from the posterior distribution of the parameters \( \eta = (\beta_0, \beta_1, \beta_2, \sigma^2) \).
2. Draw \( (\delta_0, \delta_1) \).
3. Using the draws obtained in steps 1 and 2, impute the missing data.

Steps 1–3 are repeated to create \( K \) imputed data sets. Then we fit the model of interest (1) to each imputed data set and apply Rubin’s rules for inference.

**RCT of inhaled steroid in chronic asthma patients**

Placebo and lowest active dose data.

<table>
<thead>
<tr>
<th>Deviation pattern</th>
<th>Placebo arm</th>
<th>Mean FEV1 (litres) measured at week</th>
<th>No.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.11</td>
<td>2.14 2.07 2.01 2.06</td>
<td>37</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>2.31</td>
<td>2.18 1.95 2.13 —</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>1.96</td>
<td>1.73 1.84 — —</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>1.84</td>
<td>1.72 — — —</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>All patients (Mean)</td>
<td>2.11</td>
<td>1.97 1.98 2.04 2.06</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>All patients (Std.)</td>
<td>0.57</td>
<td>0.67 0.56 0.58 0.55</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Deviation pattern</th>
<th>Lowest Active arm</th>
<th>Mean FEV1 (litres) measured at week</th>
<th>No.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.03</td>
<td>2.22 2.23 2.24 2.23</td>
<td>71</td>
<td>78</td>
</tr>
<tr>
<td>2</td>
<td>1.93</td>
<td>1.91 2.01 2.14 —</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>2.28</td>
<td>2.10 2.29 — —</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>2.24</td>
<td>1.84 — — —</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>All patients (Mean)</td>
<td>2.03</td>
<td>2.17 2.22 2.23 2.23</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>All patients (Std.)</td>
<td>0.65</td>
<td>0.75 0.80 0.85 0.81</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Framing assumptions**

Information may be available post-deviation, but for simplicity in this example we set post-deviation data to missing.

In order to complete the analysis, we need to specify a joint distribution for each patient’s pre- and post-deviation data.

Given this, we can use MI for estimation and inference by:
1. estimating each patient’s conditional distribution of post- given pre-deviation data;
2. drawing from this to give an imputed data set;
3. fitting our substantive model to this data set;
4. repeating steps 2–3 \( K \) times, and then applying Rubin’s rules to summarise the results for final inference.

**Results**

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Est</th>
<th>SE</th>
<th>MI df</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete records, MAR</td>
<td>0.237</td>
<td>0.070</td>
<td>N/A</td>
<td>&lt;0.001</td>
<td>(0.099, 0.376)</td>
</tr>
<tr>
<td>MAR, ( K = 20 )</td>
<td>0.245</td>
<td>0.073</td>
<td>302</td>
<td>&lt;0.001</td>
<td>(0.102, 0.389)</td>
</tr>
<tr>
<td>MAR, ( K = 10,000 )</td>
<td>0.237</td>
<td>0.070</td>
<td>( \approx \infty )</td>
<td>&lt;0.001</td>
<td>(0.099, 0.375)</td>
</tr>
<tr>
<td>MNAR, ( \rho = 0, ; K = 20 )</td>
<td>0.209</td>
<td>0.178</td>
<td>27</td>
<td>0.25</td>
<td>(\ approximate value)</td>
</tr>
<tr>
<td>MNAR, ( \rho = 0.5, ; K = 20 )</td>
<td>0.205</td>
<td>0.167</td>
<td>27</td>
<td>0.23</td>
<td>(\ approximate value)</td>
</tr>
<tr>
<td>MNAR, ( \rho = 1, ; K = 20 )</td>
<td>0.213</td>
<td>0.134</td>
<td>34</td>
<td>0.12</td>
<td>(\ approximate value)</td>
</tr>
</tbody>
</table>

**Conclusion:**

Taking into account a-priori expert opinion on the difference between reviewers who do, and do not, return papers, there is no benefit of the intervention.
Assumptions
The approach we explore here is, for each deviation (or—more likely—group of similar deviations occurring for similar reasons) we build an appropriate post-deviation distribution taking account of:

1. the patient's pre-deviation observations;
2. pre-deviation data from other patients in the trial;
3. the nature of the deviation, and
4. the reason for the deviation.

Such distributions need to be pre-specified, in an accessible way, and discussed with all those involved.

Options for post-deviation distributions

We begin by making the MAR assumption and—within each randomised arm—using the pre-deviation data to estimate the joint mean and covariance structure over the whole follow-up.

Then we propose a number options for constructing the joint distribution of each patient's pre- and post-deviation outcome data.

Each option represents a difference between de jure and de facto behaviour post-deviation. They are:

- on-treatment MAR
- Principled Last Mean Carried Forward
- Copy Reference
- Jump to Reference

Here we only consider on-treatment MAR and 'jump to reference'.

Randomised-arm MAR (de-jure)

The joint distribution of the patient's observed and post-deviation outcome data is multivariate normal with mean and covariance matrix from their randomised treatment arm.

Randomised-arm MAR (de-jure)

The joint distribution of the patient's observed and post-deviation outcome data is multivariate normal with mean and covariance matrix from their randomised treatment arm.
Jump to reference (de-facto)

Post-deviation, the patient ceases their randomised treatment, and their response distribution is now that of a ‘reference’ group of patients (typically, but not necessarily, control patients).

Such a change may be seen as extreme, and choosing the reference group to be the control group might be used as a worst-case scenario in terms of reducing any treatment effect since withdrawn patients on active will lose the effect of their period on treatment.
**Randomised-arm MAR: results**

Weeks 0 2 4 6 8 10 12

**Jump-to-reference (active): results**

Lowest active dose

Time from randomisation (weeks)

**Results**

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Treatment estimate (litres)</th>
<th>Std. Err.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DJ1</td>
<td>ANCOVA² JV</td>
<td>0.247</td>
<td>0.101</td>
</tr>
<tr>
<td>DJ2</td>
<td>ANCOVA SV</td>
<td>0.247</td>
<td>0.101</td>
</tr>
<tr>
<td>DJ3</td>
<td>SRM model, JV</td>
<td>0.283</td>
<td>0.094</td>
</tr>
<tr>
<td>DJ4</td>
<td>SRM model, SV</td>
<td>0.346</td>
<td>0.104</td>
</tr>
<tr>
<td>DJ5</td>
<td>Rand-arm MAR</td>
<td>0.334</td>
<td>0.107</td>
</tr>
<tr>
<td>DF1</td>
<td>J2R (active)</td>
<td>0.141</td>
<td>0.119</td>
</tr>
<tr>
<td>DF2</td>
<td>J2R (placebo)</td>
<td>0.264</td>
<td>0.108</td>
</tr>
<tr>
<td>DF3</td>
<td>LMCF</td>
<td>0.296</td>
<td>0.102</td>
</tr>
</tbody>
</table>

**Practical implications**

This approach allows the analysis to be pre-specified and performed step-by-step in terms of:

- estimands
- assumptions
- estimation

helping analysts can focus on what's important.

It is extremely flexible, and applicable to all outcome models.

It embraces missing data and non-compliance.

It has already been used in a number of regulatory submissions, and is at the heart of a DIA working group on missing data (hosted on www.missingdata.org.uk)

Full details in Carpenter, Kenward and Roger (2013).
Outstanding methodological issues

More methodological work to be done on:
- capturing and conveying assumptions
- using post-deviation information
- evaluating frequentist behaviour of MI intervals
- 'exact' alternatives to estimation via MI
- issues relating to survival data

Final Summary

- Analysis of partially observed data requires additional assumptions (MCAR, MAR, NMAR) which cannot be definitively verified from the data.
- Complete records analysis is inefficient, but will be valid under the assumption the missingness mechanism does not depend on the response in the substantive model.
- Multiple imputation provides a flexible approach to for analyzing partially observed data, and can accommodate:
  - non-linear relationships
  - measurement error
  - survey weights
  - multilevel structure
- It also provides a practical, flexible approach for sensitivity analysis.
References


for chained equations imputation. *BMC Medical Research Methodology, 14,* 28.


