# Simple, *ad-hoc* methods for coping with missing data and their shortcomings

James Carpenter & Mike Kenward

Department of Medical Statistics
London School of Hygiene & Tropical Medicine

James.Carpenter@lshtm.ac.uk
https://missingdata.lshtm.ac.uk

June 2005

# Table of Contents



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# Introudction

In contrast to principled methods, these usually create a single 'complete' dataset, which is analysed as if it were the fully observed data.
Unless certain, fairly strong, assumptions are true, the answers are invalid.
We briefly review the following methods:

- Analysis of completers only
- Imputation of simple mean
- Imputation of regression mean
- Last observation carried forward

# Completers analysis

The data on the left below has one missing observation on variable 2, unit 10.

| | 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 | ? |

# Completers analysis

- Completers analysis deletes all units with incomplete data from the analysis (here unit 10).
- It is inefficient.
- It is problematic in regression when covariate values are missing and models with several sets of explanatory variables need to be compared. Either we keep changing the size of the data set, as we add/remove explanatory variables with missing observations, or we use the (potentially very small, and unrepresentative) subset of the data with no missing values.
- When the missing observations are not a completely random selection of the data, a completers analysis will give biased estimates and invalid inferences.

# Simple mean imputation

The data on the left below has one missing observation on variable 2, unit 10.
We replace this with the arithmetic average of the observed data *for that variable*. This value is shown in red in the table below.

| | 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 |

# Simple mean imputation

- This approach is clearly inappropriate for categorical variables.
- It does not lead to proper estimates of measures of association or regression coefficients. Rather, associations tend to be diluted.
- In addition, variances will be wrongly estimated (typically under estimated) if the imputed values are treated as real. Thus inferences will be wrong too.

# Regression mean imputation

Here, we use the completers to calculate the regression of the incomplete variable on the other complete variables. Then, we substitute the *predicted mean* for each unit with a missing value. In this way we use information from the joint distribution of the variables to make the imputation.

# Regression mean imputation

Example

Consider again our dataset with two variables, which is missing variable 2 on unit 10:

| | 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 | ? |

# Regression mean imputation

To perform regression imputation, we first regress variable 2 on variable 1 (note, it doesn't matter which of these is the 'response' in the model of interest). In our example, we use simple linear regression:

$$V_2 = \alpha + \beta V_1 + e.$$

Using units 1-9, we find that $\hat{\alpha} = 6.56$ and $\hat{\beta} = -0.366$, so the regression relationship is

$$\text{Expected value of } V_2 = 6.56 - 0.366 V_1.$$

For unit 10, this gives

$$6.56 - 0.366 \times 3.6 = 5.24.$$

# Regression mean imputation

This value is shown in red below:

| Unit | Variables 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 |

# Regression mean imputation

- Regression mean imputation can generate unbiased estimates of means, associations ad regression coefficients in a much wider range of settings than simple mean imputation.
- However, one important problem remains. The variability of the imputations is too small, so the estimated precision of regression coefficients will be wrong and inferences will be misleading.

# Creating an extra category

When a categorical variable has missing values it is common practice to add an extra 'missing value' category. In the example below, the missing values, denoted '?' have been given the category 3.

| | Variables | | | | | |
|---|---|---|---|---|---|---|
| Unit | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 1 | 3 | 4.3 | 3.5 | 1 | 4.6 |
| 2 | 1 | 3 | ? | 3.5 | ? | ? |

# Creating an extra category

This is bad practice because:

- the impact of this strategy depends on how missing values are divided among the real categories, and how the probability of a value being missing depends on other variables;
- very dissimilar classes can be lumped into one group;
- sever bias can arise, in any direction, and
- when used to stratify for adjustment (or correct for confounding) the completed categorical variable will not do its job properly.

# Last observation carried forward (LOCF)

This method is specific to longitudinal data problems.
For each individual, missing values are replaced by the last observed value of that variable. For example:

| | Observation time | | | | | | |
|---|---|---|---|---|---|---|---|
| Unit | 1 | 2 | 3 | 4 | 5 | 6 | ... |
| 1 | 3.8 | 3.1 | 2.0 | ? –> 2.0 | ? –> 2.0 | ? –> 2.0 | |
| 2 | 4.1 | 3.5 | 3.8 | 2.4 | 2.8 | 3.0 | |
| 3 | 2.7 | 2.4 | 2.9 | 3.5 | ? –> 3.5 | ? –> 3.5 | |

# Last observation carried forward (LOCF)

Here the three missing values for unit 1, at times 4, 5 and 6 are replaced by the value at time 3, namely 2.0. Likewise the two missing values for unit 3, at times 5 and 6, are replaced by the value at time 4, which is 3.5.

Using LOCF, once the data set has been completed in this way it is analysed as if it were fully observed.

# Last observation carried forward (LOCF)

For full longitudinal data analyses this is clearly disastrous: means and covariance structure are seriously distorted. For single time point analyses the means are still likely to be distorted, measures of precision are wrong and hence inferences are wrong.

Note this is true even if the mechanism that causes the data to be missing is completely random. For a full discussion download the talk 'LOCF - time to stop carrying it forward' from the preprints page of this site.

# Conclusions

Unless the proportion missing is so small as to be unlikely to affect inferences, these simple *ad-hoc* methods should be avoided. However, note that 'small' is hard to define: estimates of the chances of rare events can be very sensitive to just a few missing observations; likewise, a sample mean can be sensitive to missing observations which are in the tails of the distribution.

# Conclusions

They usually conflict with the statistical model that underpins the analysis (however simple and implicit this might be) So they introduce *bias*.

As the assumptions about the reason for the data being missing that they implicitly make are often difficult to describe (e.g. with LOCF), they can make it very hard to know what assumptions are being made in the analysis.

They do not properly reflect statistical uncertainty: data are effectively 'made up' and no subsequent account is taken of this.