Principled methods

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

Principled methods

Wholly model based methods

Simple stochastic imputation

Multiple (stochastic) imputation

Weighting methods

Conclusion

Principled methods

These all have the following in common:

- No attempt is made to replace a missing value directly. i.e. we do not pretend to 'know' the missing values.
- Rather: available information (from the observed data and other contextual considerations) is combined with *assumptions* not dependent on the observed data.

Principled methods

This is used to

1. *either* generate *statistical* information about each missing value,
   e.g. *distributional* information: given what we have observed, the missing observation has a normal distribution with mean μ and variance σ^2, where the parameters can be estimated from the data.
2. *and/or* generate information about the missing value mechanism

Principled methods

The great range of ways in which these can be done leads to the plethora of approaches to missing values. Here are some broad classes of approach:

- Wholly model based methods.
- Simple stochastic imputation.
- Multiple stochastic imputation.
- Weighting methods

Wholly model based methods

A full statistical model is written down for the *complete data*.

Analysis (whether frequentist or Bayesian) is based on the likelihood.

Assumptions must be made about the missing data mechanism:

- If it is assumed MCAR or MAR, no explicit model is needed for it.
- Otherwise this model must be included in the overall formulation.

Wholly model based methods

Such likelihood analyses requires some form of integration (averaging) over the missing data. Depending on the setting this can be done implicitly or explicitly, directly or indirectly, analytically or numerically.

The statistical information on the missing data is contained in the model.

Examples of this would be the use of linear mixed models under MAR in SAS PROC MIXED or *MLwiN*.

Simple stochastic imputation

- Instead of replacing a value with a *mean*, a *random draw* is made from some suitable distribution.
- Provided the distribution is chosen appropriately, consistent estimators can be obtained from methods that would work with the whole data set.
- Very important in the large survey setting where draws are made from units with complete data that are 'similar' to the one with missing values (donors).
- There are many variations on this *hot-deck* approach.
- Implicitly they use non-parametric estimates of the distribution of the missing data: typically need very large samples.

Simple stochastic imputation

Although the resulting estimators can behave well, for precision (and inference) account must be taken of the source of the imputations (i.e. there is no 'extra' data). This implies that the usual complete data estimators of precision can't be used.

Thus, for each particular class of estimator (e.g. mean, ratio, percentile) each type of imputation has an associated variance estimator that may be design based (i.e. using the sampling structure of the survey) or model based, or model assisted (i.e. using some additional modelling assumptions). These variance estimators can be *very* complicated and are not convenient for generalization.

Multiple (stochastic) imputation

This is very similar to the single stochastic imputation method, except there are many ways in which draws can be made (e.g. hot-deck non-parametric, model based).

The crucial difference is that, instead of completing the data *once*, the imputation process is repeated a small number of times (typically 5-10). Provided the draws are done properly, variance estimation (and hence constructing valid inferences) is *much* more straightforward.

As is discussed more in the 'introduction to multiple imputation' document, the observed variability among the estimates from each imputed data set is used in modifying the complete data estimates of precision. In this way, valid inferences are obtained under missing at random.

Weighting methods

We give a simple illustration of weighting methods and contrast them with likelihood-based methods.

Example: simple continuous problem

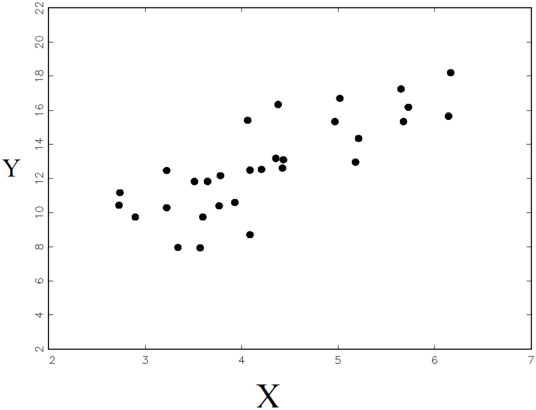
Consider a *simple linear regression* setting:

$$\mathbf{E}(Y_i) = \theta_0 + \theta_1 x_i = x_i^T \theta, i = 1, ..., n, \qquad (1)$$

where Y_i are independent and identically distributed as $N(0, \sigma^2)$.

Weighting methods

A typical data set might look like this:



Weighting methods

The ordinary least squares regression line (in this case maximum likelihood) is obtained by solving the *normal equations* for β:

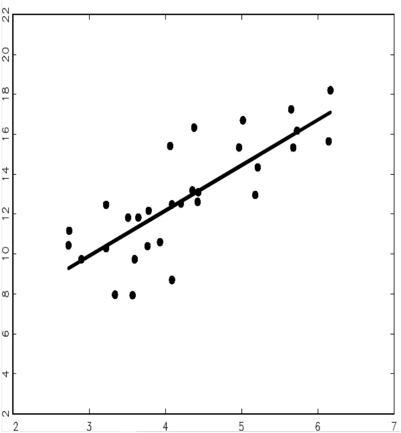
$$\sum_{i=1}^{n} \mathsf{x}_i (y_i - \mathsf{x}_i^T \beta) = 0.$$

More generally we can get parameter estimates by solving estimating equations:

$$\mathsf{U}(\mathsf{Y}; \hat{\theta}) = \sum_{i=1}^{n} \mathsf{U}_i(\mathsf{y}_i; \hat{\theta}) = 0.$$

Weighting methods

In this example, the estimates of the slope and intercept give the following line:

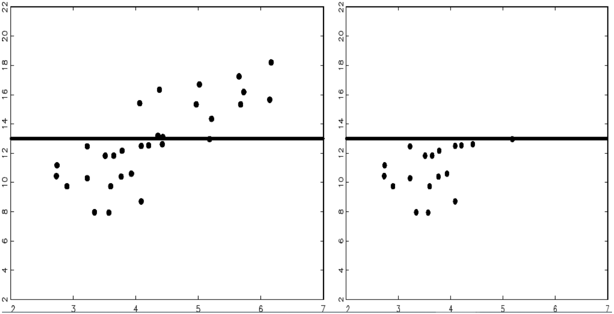


Weighting methods

Suppose now that some *response* (i.e. Y) observations are missing. The implications are (i) possible bias in the estimate of the intercept and slope and (ii) loss of precision in the estimate of the intercept and slope. Suppose in particular that the responses are MNAR; specifically that all observations greater than y=13 are unobserved.

Weighting methods

In other words we lose all observations above the horizontal line in the left-hand picture, leaving the observed data in the right hand picture:



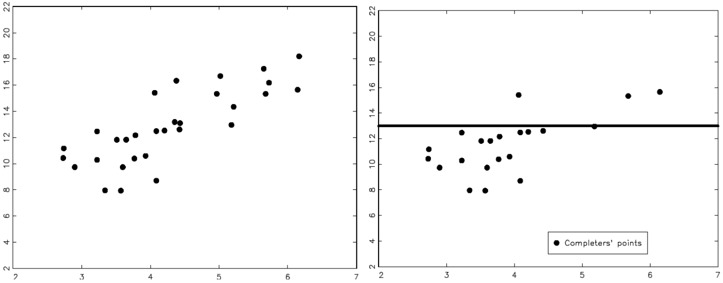
The 'completers' regression line is now biased (and inconsistent).

Weighting methods

However, because in this case we know the missing value mechanism and the distribution involved (which is unlikely in real applications) we can do a valid analysis using *likelihood* methods. In this special case the likelihood method is known as *Tobit regression*.

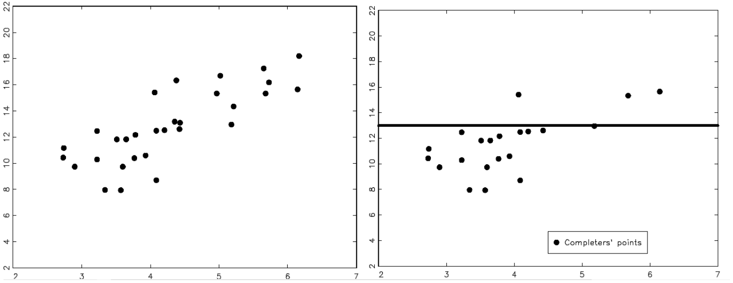
Weighting methods

Both the 'completers' and Tobit regression line are shown in the figure below, where the completers line is bottom line at the right hand end:



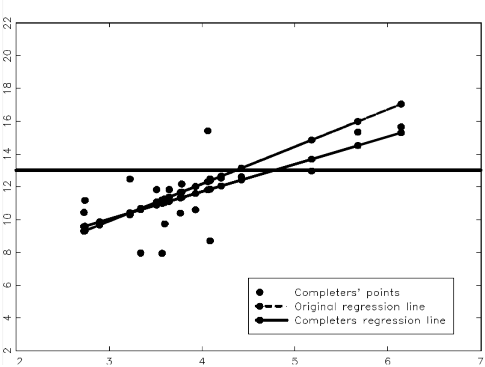
Weighting methods

To make it a little more realistic, suppose now that an observation greater than 13 has a probability of 0.25 of being observed; in other words instead of seeing the left hand plot below, we see the right hand plot.



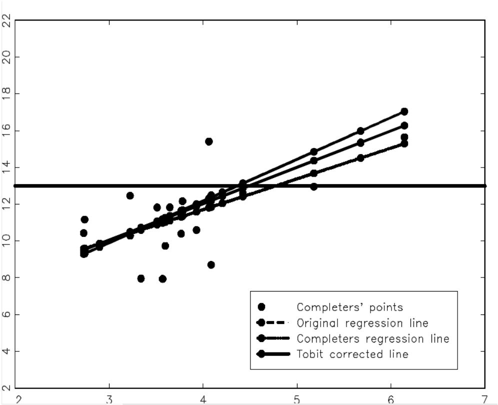
Weighting methods

The completers line is still inconsistent (lower line at right hand end):



Weighting methods

We could use Tobit regression to correct for this (top line at right hand end: original regression line; middle line at right hand end: Tobit regression line; bottom line at right hand end: 'completers' regression line).



Weighting methods

But there now exists an alternative correction, which requires only that we know the probability of Y_i being missing *given its value*. In other words, we don't need to know the distribution of the observations as we do for the Tobit regression.

Let R_i be a random variable indicating whether Y_i is missing or not, so $R_i = 0$ implies Y_i missing, and $R_i = 1$ implies Y_i is observed.

The following *weighted* estimating equation is unbiased for the regression parameters:

$$\sum_{i=1}^{n} \frac{R_i U_i(y_i; \hat{\theta})}{\Pr(R_i = 1 \mid Y_i)} = 0.$$

Weighting methods

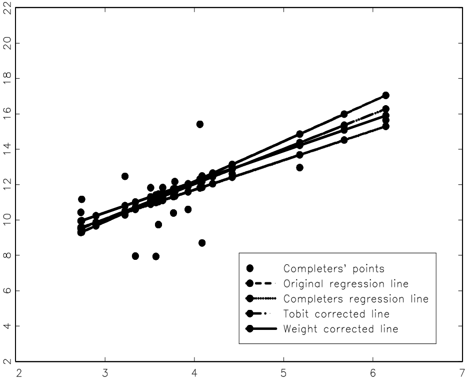
In this (artificial) example

$$\Pr(R_i = 1 \mid Y_i > 13) = \frac{1}{4}$$

and 1 otherwise, so we can use simple weighted least squares to make the correction.

Weighting methods

Comparison of weighting with other methods. At right hand end, top line is from weighted regression; second line is original regression line; third line is tobit regression and fourth line is completers analysis.



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Weighting methods

We now look at the performance of these two methods in this simple regression setting where the probability of observations greater than 13 being seen is 0.25. For sample sizes of 20, 100 and 1000, the table below shows the mean and standard deviation of the slope estimators (true value 2) over 10,000 simulations.

Weighting methods

Estimator	Expected value	SE
$n = 30$		
Completers only	1.73	0.39
Tobit	1.99	0.33
Weighted	1.95	0.45
$n = 100$		
Completers only	1.75	0.20
Tobit	1.98	0.18
Weighted	1.99	0.23
$n = 1000$		
Completers only	1.74	0.063
Tobit	1.98	0.055
Weighted	2.00	0.070

Weighting methods

We see that both tobit and weighted regression are unbiased, but that estimates from a weighted analysis are far more variable.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Conclusion

Our simple examples have illustrated that there are broadly two forms of principled analysis:

1. likelihood methods, which make distributional assumptions about the unseen data, and assumptions about the form dropout mechanism.
2. weighting methods, which use the inverse of

   $$\Pr(R_i = 1 \mid Y_i)$$

   as weights.

   In its simple form, weighting is much less precise. However, in the session on weighting, we will see that this can be addressed, albeit with difficulty.

Conclusion

In summary, in contrast to *ad-hoc* methods, principled methods are:

- based on a well-defined statistical model for the complete data (assumptions), and explicit assumptions about the missing value mechanism.
- the subsequent analysis, inferences and conclusions are valid under these assumptions.
- this doesn't mean the assumptions are necessarily *true* but it does allow the dependence of the conclusions on these assumptions to be investigated.