RT 2: Missing Data

Tessa Wilkie

Supervisor: Robin Mitra

1 Introduction

When we collect data, we are very unlikely to be able to collect all the data we need. So dealing with missing data is a common problem in statistics — getting it right is important.

Missing data is common in surveys, and, although applications run wider than this, this report will structure examples in the context of surveys, for ease of understanding.

We have responses, with observations that are either observed or missing for each question. We will refer to questions as variables.

The dataset is made up of all the responses (and recorded non-responses) in the survey.

Let's say a restaurant wants to re-work its menu to appeal better to the local population. They have conducted a survey of people who live nearby, to try to work out what they like and dislike. However, some respondents have not answered all of the questions in the survey. What should the restaurant do about this?

It might seem obvious that the only fair way to deal with missing data would be to delete any responses that are missing. They cannot just make up what people would have said, could they? That would surely be unfair and would bias the dataset.

But what if the questions with missing responses were about favourite meat dishes, and the respondents who did not answer the questions are vegetarians? If they deleted all of the vegetarians' responses to the survey, aren't they then biasing the analysis towards meat-eaters?

Another issue with deleting data could come up if a very large proportion of the respondents are vegetarians — so a lot of respondents had not completely answered the survey.

If the restaurant got rid of all of those respondents then any inference that they might make from the analysis of responses will lack power. They can say that 'the average person said X', but they will be statistically less sure about it. Their confidence intervals will be wide. This is not a good outcome either.

This report will explore methods to deal with missing data. These largely centre around ways to impute — fill in — the missing data with a plausible value.

This may seem a disingenous way to behave — we are making up the data!

But, as will be shown, it can result in more robust statistical analyses than if we didn't use imputation.

Analysts must be careful of how their method of dealing with missing data affects their analysis and what the risks are if they misuse it. To do this well, they should also be aware of what underlies the pattern of missing data. Is it, as above, a certain section of the population that is reluctant to answer a particular question? Is it totally random? If it is a certain sector of the population, can we identify them through other questions in the survey?

Of course, in data collection we may collect some data that is useless: it doesn't inform our analysis. For example, if we are interested in what might affect children's educational attainment, we would likely be interested in data on their location, nutrition and parental occupation, but we can probably live without knowing what their favourite colour was when they were six years old.

This report will first discuss the underlying driver of the missing data — these are often known as the missingness mechanism (Little and Rubin, 2020). They determine the relationship between whether or not data is missing and other variables in the dataset.

It then goes on to describe some simple ways to deal with missing data, including that of deleting any incomplete responses as described above.

A small simulation study, based on an imaginary survey of racehorses' heights and weights, follows this section to illustrate some of the simpler methods of imputation.

It then tackles more advanced techniques, that attempt to, as well as impute data wisely, give us a measure of uncertainty for the estimates that is due to the missing data.

Finally, the report will touch on the most difficult area of all: methods to deal with what is known as Missing Not at Random data. This section will discuss some open research questions in this area.

2 Types of Missing Data

As mentioned above, it is important not just to understand that data is missing, but why it is missing. The importance of considering the mechanism that drives missing data was first discussed by Rubin (1976).

Little and Rubin (2020) set up the problem in the following way.

We set $Y = (y_{ij})$ as a data set without missing values. Y is an $n \times K$ matrix, where K is the number of variables for which we collect data for each entry (so, the number of questions in a survey), and n the number of respondents.

For example, a survey of 1,000 graduates of working age that asks them their age, degree subject, career sector,and job satisfaction level, would have $K = 4$ and $n = 1000$.

The value of Y_j, a variable , for respondent i is defined as y_{ij}. We also have a "missingness indicator matrix", denoted by M. Entry $m_{ij} = 0$ if y_{ij} is not missing and 1 if it is.

We assume that the rows of Y and M are independently and identically distributed over i. The conditional distribution of m_i given y_i is denoted $f_{M|Y}(m_i|y_i, \phi)$. ϕ represents some unknown parameters. Observed entries of Y for unit i are denoted as $y_{(obs)i}$ and missing elements as $y_{(mis)i}$.

2.1 Missing Completely at Random

Data that is Missing Completely at Random (MCAR) is the easiest to deal with. This is because whether or not data is missing does not depend on the data.

If we imagine an online survey that had no missing values, but then a helpful computer virus has deleted some cell entries totally at random, then the nature of our data is Missing Completely at Random.

Little and Rubin (2020) define the model for this as:

$$f_{M|Y}(m_i|y_i, \phi) = f_{M|Y}(m_i|y_i^*, \phi).$$

This is for distinct values y_i, y_i^* and all i Little and Rubin (2020).

2.2 Missing at Random

Data that is Missing at Random (MAR) is where the mechanism that governs whether data is missing or not depends on the data, but that this data is observed.

This is where:

$$f_{M|Y}(m_i|y_{i(obs)}, y_{i(mis)}\phi) = f_{M|Y}(m_i|y_{i(obs)}, y^*_{i(mis)}, \phi),$$

for all responses i and all distinct $(y_{i(mis)}, y^*_{i(mis)})$, (Little and Rubin, 2020).

2.3 Missing Not at Random

If the above does not hold, the missingness mechanism depends on some of the unobserved values: then the data are Missing Not at Random, according to Little and Rubin (2020). That means that the missingness mechanism depends on data that are missing themselves. It is a very difficult type missing data to deal with. It is also, unfortunately, probably a common one.

3 Complete Case Analysis

A simple way to deal with missing data is Complete Case Analysis. This basically involves deleting an entry for where you have a missing data cell. So, in our matrix of survey responses, if line i for respondent i contains a missing observation, the whole line is deleted and not considered for analysis.

Little and Rubin (2020) consider Complete Case Analysis to only be suitable where the amount of missing data is very limited. Its main advantage is simplicity.

However, they point out that it has some major disadvantages: throwing away whole entries because they are missing data means we are throwing away potentially useful information. This means our estimates may lack precision, and they are likely to be biased (how likely depends on the mechanism that drives missingness).

Little and Rubin (2020) show that the bias of a sample mean under Complete Case Analysis is as follows:

$$\mu_{(obs)C} - \mu = (1 - \pi_{(obs)C})(\mu_{(obs)C} - \mu_{(mis)C}).$$

In the above equation, μ is the true mean; $\mu_{(obs)C}$ is the Complete Case mean — the entries that have no missing values; $\mu_{(mis)C}$ is the mean of the entries that have missing values; $\pi_{(obs)C}$ is the proportion of complete entries from the overall survey — in other words what proportion of the total we are able to keep.

Little and Rubin (2020) note that if the data are MCAR then the bias is zero.

Ways to improve Complete Case Analysis, according to Little and Rubin (2020) include examining the removed entries to see if we can legitimately proceed as if the data are missing completely at random.

3.1 Weighting

Little and Rubin (2020) also note that if we do not think the data is MCAR, then we can weight the data to attempt to adjust for bias. If we think that respondents to a survey that fit into a certain category are more or less likely to respond to certain questions in the survey — and we can estimate this probability from information they have supplied to the survey — then we can re-weight to make sure these respondents are not under-represented.

Little and Rubin (2020) formulate an estimator of the mean using weighting classes as follows: a sample is made up of J weighting classes; r_j is the number of respondents in weighting class j; the sample size for class j is n_j. The ratio $\frac{r_j}{n_j}$ is used to estimate the probability of response for observations in weighting class j.

Little and Rubin (2020) set that responding observations i in class j should be given weight:

$$w_i = \frac{r(\pi_i \hat{\phi}_i)^{-1}}{\sum_{k=1}^{r} (\pi_k \hat{\phi}_k)^{-1}}.$$

With $r = \sum_{j=1}^{J} r_j$ and $\hat{\phi}_i = \frac{r_j}{n_j}$ for respondents i in class j.

Little and Rubin (2020) recommend that weighting methods are used in situations where the analyst is more concerned with bias than a large sampling variance, as weighting methods can help with mitigating bias but they do not necessarily reduce sampling variance.

3.2 Available Case Analysis

Available Case Analysis involves throwing away less data than Complete Case Analysis, but it creates its own problems. With Available Case Analysis, an analyst would use all the observations for a category where the observation is recorded for that category, even if that response has missing values for other categories.

However, this means that the sample size could be different for each category being analysed, which would complicate analysis of the dataset as a whole.

Other problems noted by Little and Rubin (2020) are that Available Case analysis does not necessarily yield more efficient estimators than Complete Case analysis, despite using more data; it also can yield correlation estimates that are greater than |1|.

4 Single Imputation

This section introduces several techniques designed for a single round of imputation.

4.1 Unconditional Mean Imputation

Unconditional Mean Imputation is where missing data is replaced with the mean of the other entries for that category.

In Figure 1 we see an illustration of Unconditional Mean Imputation on a simulated data set of racehorses' heights and weights (the same example as used in the Simulation Study later in this report). Here, only weights are missing. They are missing according to an MAR pattern — in this case the taller the horse, the more likely its weight will not have been recorded.

Little and Rubin (2020) define unconditional mean imputation as estimating a missing value of Y_j for response i— y_{ij} — with the mean of the recorded values for category j being $\bar{y}_j^{(j)}$.

The sample mean (including observed and imputed observations for category j) is $\bar{y}_j^{(j)}$ and the sample variance is:

$$s_{jj}^{(j)} \frac{(n^{(j)} - 1)}{(n - 1)}.$$

The number of observations with recorded values for Y_j is $n^{(j)}$ and $s_{jj}^{(j)}$ is the estimated variance from the $n^{(j)}$ entries for Y_j that are not missing.

Even when the data are Missing Completely at Random, there are issues. In this case, $s_{jj}^{(j)}$ estimates the true variance (Little and Rubin, 2020), so the imputed data set will underestimate true variance.

Little and Rubin (2020) also do not approve of Unconditional Mean Imputation because it can produce bias.

We can see in Figure 1 that this method is unlikely to produce a satisfactory estimator for average racehorse weight.

There is positive correlation between the two variables, so the taller the horse, the more likely it is to weigh more. Yet because higher weights are more likely to be missing, the estimator looks bound to underestimate the true average weight of the sample of racehorses.

Conditional Mean Imputation is intended to improve over unconditional mean imputation. Regression imputation, considered in the next section, is an example of conditional mean imputation.

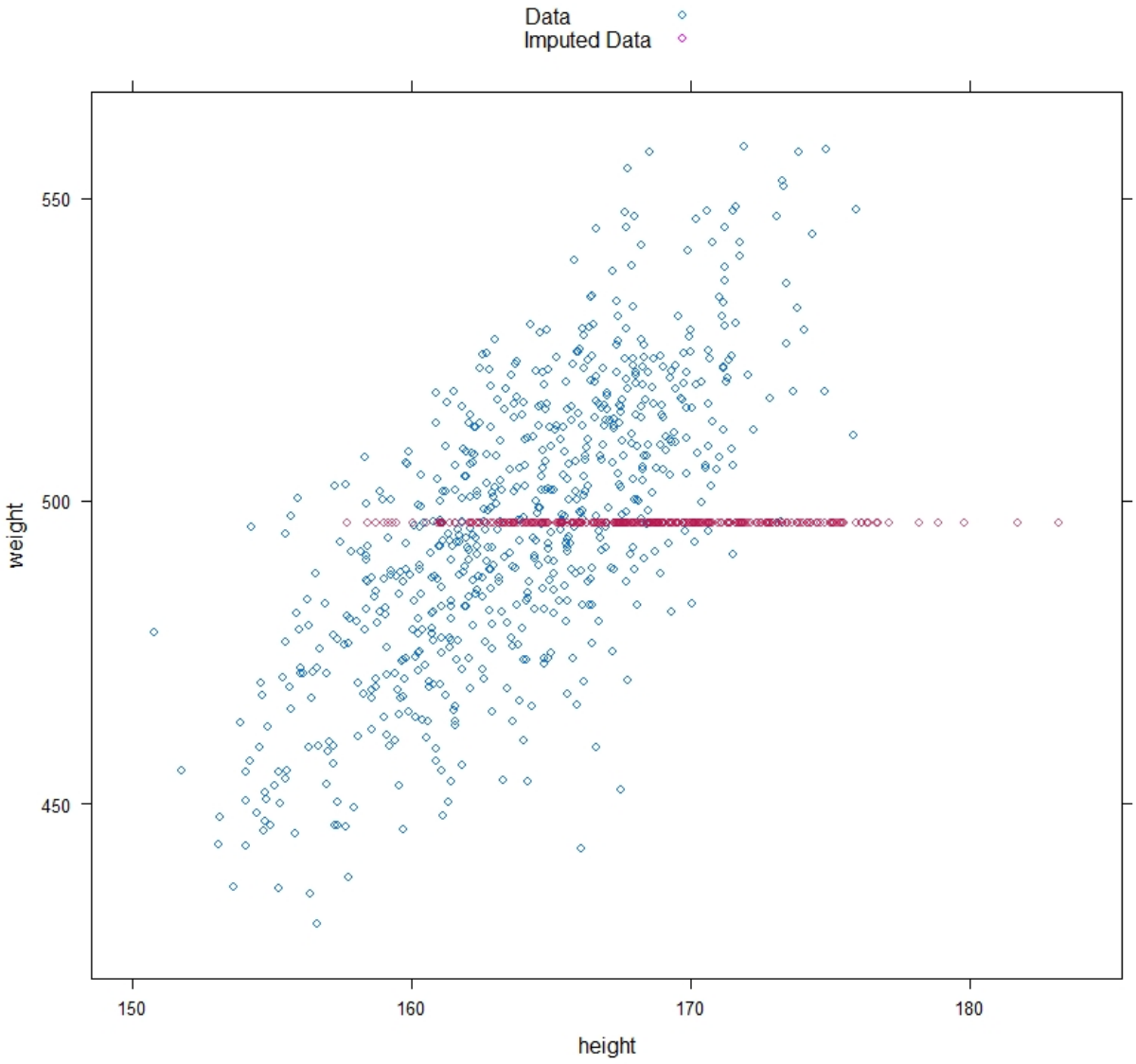

Figure 1: Imputation on MAR data using Unconditional Mean Imputation

4.2 Regression Imputation

This uses regression to predict what a missing value would be, using the observed values to inform the regression equation.

Little and Rubin (2020) give an example of univariate missing data. Suppose variables $Y_1, ..., Y_{K-1}$ have no missing response data, while variable Y_K has recorded data for the first r responses and after that data for Y_K is missing. So there are r recorded responses, and $n - r$ missing responses for Y_K.

If responses i has a missing value for variable K, but responses recorded for variables $1, ..., K-1$. We can impute the conditional mean with the following formula from (Little and Rubin (2020)):

$$\hat{y}_{iK} = \tilde{\beta}_{0\cdot 1,2,\ldots,(K-1)} + \sum_{j=1}^{K-1} \tilde{\beta}_{Kj\cdot 1,2,\ldots,(K-1)} \hat{y}_{ij}.$$

Little and Rubin (2020) define $\tilde{\beta}_{0.1,2\ldots,(K-1)}$ as the intercept in the regression. $\tilde{\beta}_{Kj\cdot 1,2,\ldots,(K-1)}$ is defined as the coefficient of Y_j from the regression of Y_K on $Y_1, ..., Y_{(K-1)}$.

As with Unconditional Mean Imputation, this method tends to underestimate the variance of the data, because it fills in the missing data by placing them on the regression line.

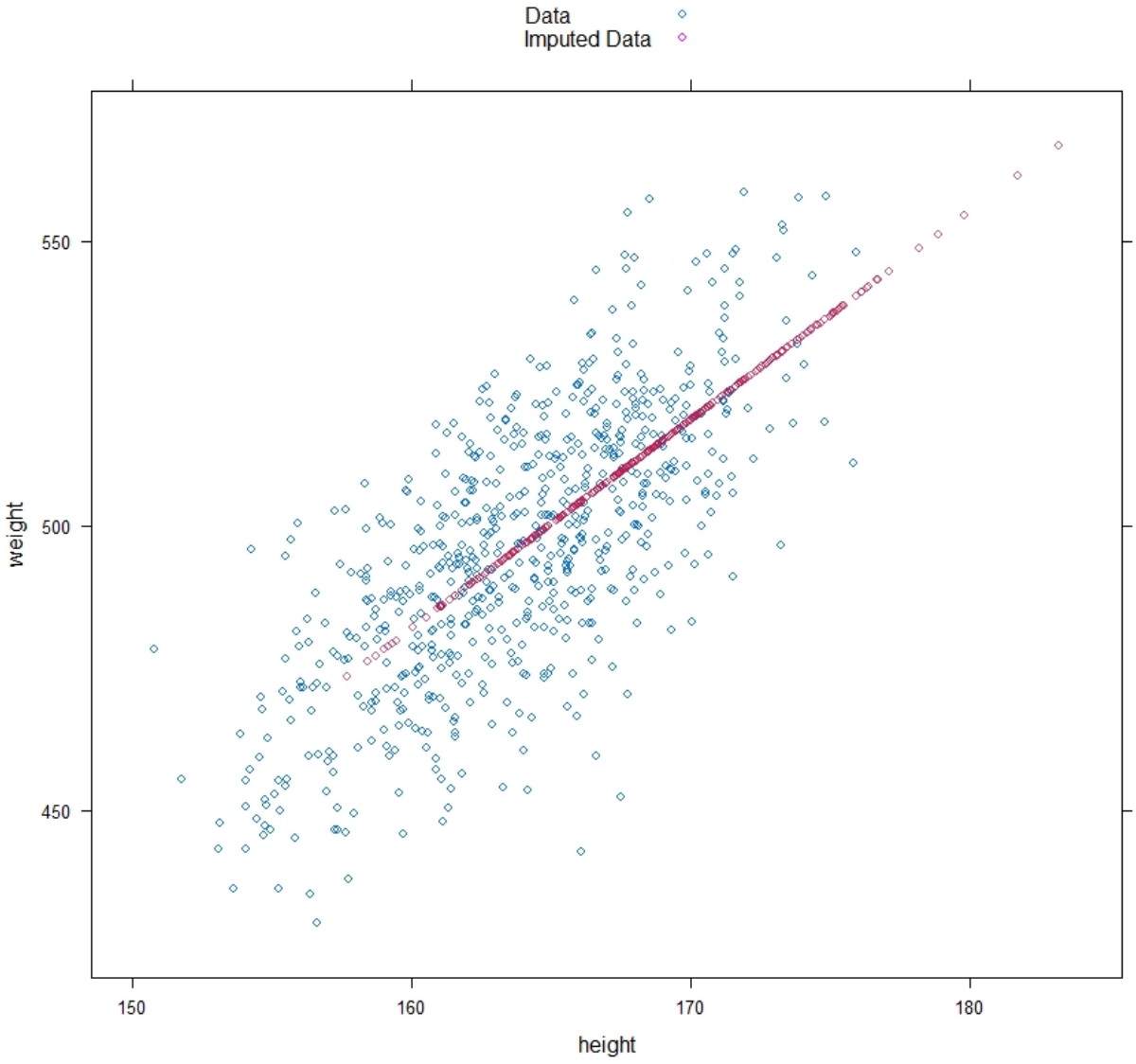

Figure 2: Imputation on MAR data using Regression Mean Imputation

In Figure 2 we can see Regression Imputation carried out on the simulated dataset of racehorse weights and heights.

The missing points are all imputed on the regression line. In reality, the missing points would

be scattered about the regression line.

The next technique described in this report aims to impute variables in a way that means the variance of the imputed dataset should be close to the true variance.

4.3 Stochastic Regression Imputation

Stochastic Regression Imputation builds on Regression Imputation by introducing a random element — so instead of imputing missing values as sitting on the regression line, missing values should deviate from it in a random manner.

We see it illustrated in Figure 3. Here, the imputed points are scattered, rather than sitting on the regression line.

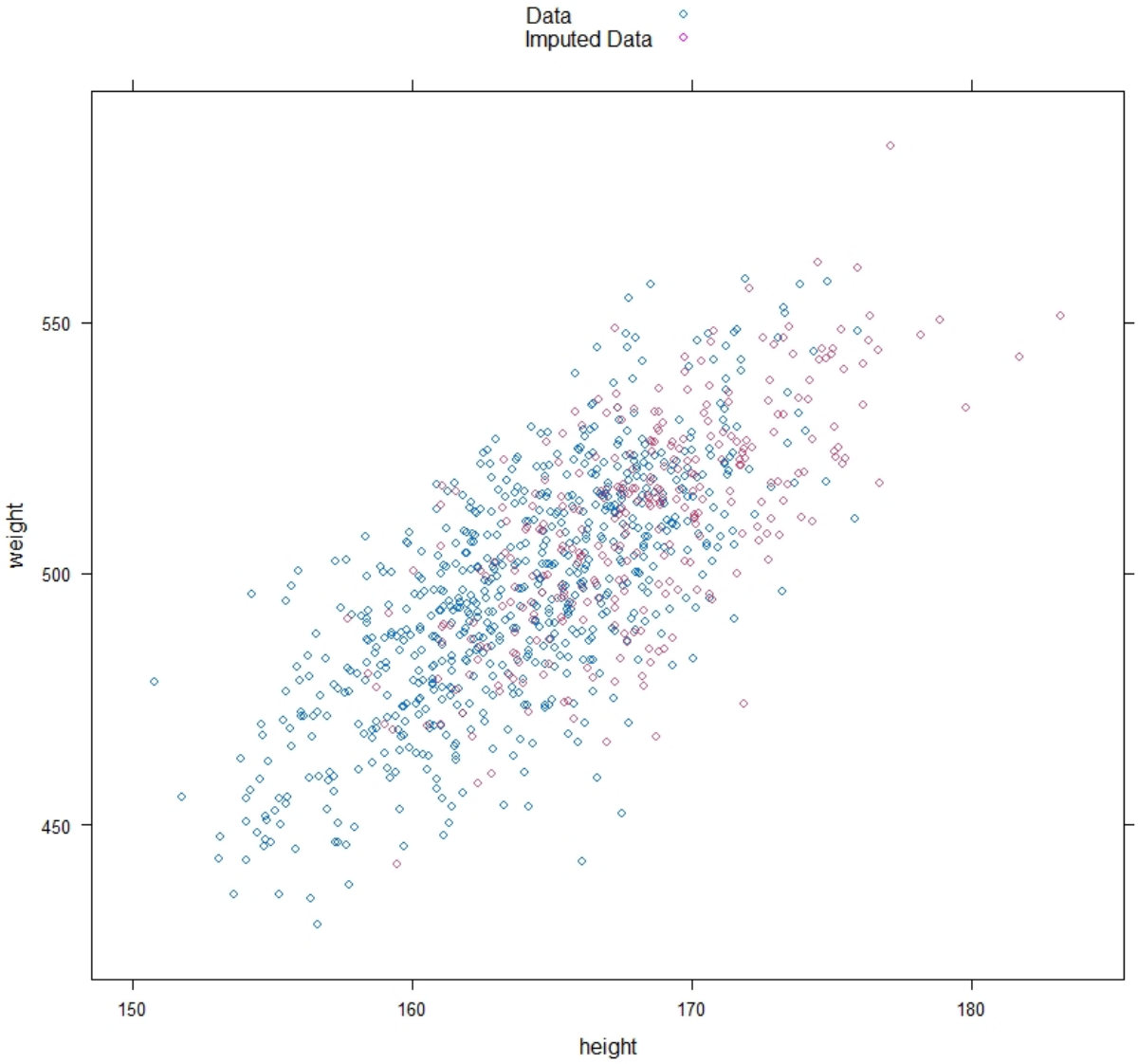

Figure 3: Imputation on MAR data using Stochastic Regression Imputation

Little and Rubin (2020) give the following formula for Stochastic Regression Imputation, applied to the same situation described in the above section on Regression Imputation.

$$\hat{y}_{iK} = \tilde{\beta}_{0.1,2...,(K-1)} + \sum_{j=1}^{K-1} \tilde{\beta}_{Kj.1,2...,(K-1)} \hat{y}_{ij} + z_{iK}.$$

The intercept, as before, is $\tilde{\beta}_{0.1,2...,(K-1)}$ and the coefficient of Y_j from the regression is, again, $\tilde{\beta}_{Kj.1,2...,(K-1)}\hat{y}_{ij}$. Here, z_iK is normally distributed with mean 0.

Unlike Unconditional Mean Imputation and Regression Imputation, Little and Rubin (2020) find that Stochastic Regression Imputation does not underestimate the variance, as the two previously discussed methods do.

They find that the method is not without its flaws, however. Firstly, it is less efficient than Unconditional Mean Imputation. Little and Rubin (2020) show that, for Stochastic Regression Imputation, the large sample variance of the estimator for the mean is larger than that derived by Unconditional Mean Imputation.

The second issue mentioned by Little and Rubin (2020) is that the standard errors of the parameter estimates derived from the imputed data set are too small, as they don't address estimation uncertainty.

4.4 Hot Deck Imputation

This method fills in missing values by using recorded responses from similar respondents in a survey or dataset.

In a survey setting (Andridge and Little, 2010), the respondent that has missing data to be imputed is known as the recipient. The similar respondents whose responses are used to impute the missing data are called donors.

A donor pool, or adjustment cell, is a pool of similar respondents whose responses may be used to impute the missing data of a respondent that is deemed similar enough to belong to that pool, according to Andridge and Little (2010). Analysts use covariate information — information from other answers to the survey — to attempt to categorise respondents into different donor pools.

Analysts need to take care in deciding which variables from covariate information are used to determine membership of a donor pool.

They need to examine whether a covariate is associated with the missing variable that is to be estimated, and whether it is associated with the probability of non-response (Andridge and Little, 2010). These relationships can have an impact on the bias and variance of estimators for the variable with missing data. For example, according to Andridge and Little (2010), if the covariates are associated with the probability of non-response but not with the outcome of the missing variable, then variance is upped and bias is not reduced.

4.4.1 Matching Metrics

The technique of choosing donors based on whether they are in the same donor pool as a recipient can be described more formally in the following way (Andridge and Little, 2010).

We let x_i denote values of covariates used to create donor pools for respondent i.

If respondent i has $x_i = (x_{i1}, ..., x_{iq})$ recorded as the values for the q covariates that are used to create donor pools. The donor pool (or adjustment cell) in which respondent i sits is denoted as $C(x_i)$. Then the matching metric is:

$$d(i,j) = \begin{cases} 0 & j \in C(x_i) \\ 1 & j \notin C(x,i) \end{cases} .$$

We want $d(i,j)$ to be less than a defined value, d_0, (Little and Rubin, 2020).

Other ways to find donors for recipients include using a distance metric — such as the Mahalanobis distance or the predictive mean— to try to identify respondents that are closest to the recipient.

The metric for predictive means is defined below (Andridge and Little, 2010):

$$d(i,j) = (\hat{Y}(x_i) - \hat{Y}(x_j))^2.$$

$\hat{Y}(x_i) = x_i^T \hat{\beta}$ is the predicted value of Y for recipient i, based on a regression of Y using data that is complete.

Little and Rubin (2020) prefer the predictive means matching method as it gives a higher weight to predictors that are better at predicting the missing variable.

Once an analyst has determined their donor pools or method of finding closest donors, they have to select which donor to use to impute missing data for a particular respondent.

Those working on distance metrics can use several methods (Andridge and Little, 2010). Among the methods described are deterministic (also known as nearest neighbour) methods where an analyst can pick a donor by using the distance metric to select the nearest one. An analyst could also randomly draw from a pool of donors that are defined as being within a certain distance of the recipient.

As with the other methods described so far, Hot Deck Imputation does not, of itself, account for uncertainty due to missing data — imputation uncertainty (Little and Rubin, 2020). Multiple Imputation is one way of dealing with this. It is described in Section 7. Others are resampling techniques such as Bootstrap and Jackknife. The Bootstrap procedure is detailed in Section 6.

5 Simulation Study

To illustrate some of the issues encountered with missing data, this report includes a short simulation study.

The aim is to illustrate some of the defects of the simplest single imputation methods even when data is MCAR or MAR.

Finally, it will show that widely used techniques can run into big problems when data is Missing Not at Random.

5.1 Method

We first simulate a bivariate normal distribution representing the heights (in cm) and weights (in kg) of racehorses. Variable x_1, the height, is normally distributed with mean 165 and standard deviation 5. Variable x_2, the weight, has mean 500 with standard deviation 25. The correlation coefficient, ρ, is 0.75.

We then, for each of Cases 1-3, remove approximately 30% of the weight data according to the three mechanisms:

- Case 1: Data is Missing Completely at Random - each weight entry has a 30% probability of being missing.
- Case 2: Data is Missing at Random. Weight records are more likely to be missing the greater the horse's height.
- Case 3: Data is Missing Not at Random. Weight data are more likely to be missing for horses that weigh more.

We then apply the following techniques to estimate the parameters of the distribution:

- Complete Case Analysis
- Unconditional Mean Imputation
- Regression Imputation
- Stochastic Regression Imputation

The process is replicated $1,000$ times for each case. The simulation is carried out using the `mice` package in R (van Buuren and Groothuis-Oudshoorn, 2011).

We record the sample mean, the bias of the sample mean, and the variance to examine which technique is most effective.

5.2 Results

In Table 1 we see that, as might be expected, all techniques do well at estimating the mean when the data is MCAR. However, we can see how much Unconditional Mean Imputation and Regression Imputation affect the variance (the true variance is 625). It is also worth noting that confidence intervals for Complete Case Analysis will be wider than for the other techniques, as they are based on a smaller sample.

Technique	Estimated mean	Variance	Bias of mean
Complete Case Analysis	500.01	623.45	0.01
Unconditional Mean Imputation	500.01	436.32	0.01
Regression Imputation	500.02	541.78	0.02
Stochastic Regression Imputation	500.02	624.31	0.02

Table 1: Table examining imputation techniques when data is MCAR

Technique	Estimated mean	Variance	Bias of mean
Complete Case Analysis	495.31	574.86	4.69
Unconditional Mean Imputation	495.31	402.38	4.69
Regression Imputation	500.03	541.83	0.03
Stochastic Regression Imputation	500.04	624.03	0.04

Table 2: Table examining imputation techniques when data is MAR

In Table 2 we can see that Complete Case Analysis and Unconditional Mean Imputation estimates of the mean begin to be biased when data is MAR. The pattern of missingness means taller horses are less likely to have their weights recorded, which is not properly reflected with these methods.

Only Stochastic Regression Imputation reflects variance well.

Technique	Estimated mean	Variance	Bias of mean
Complete Case Analysis	493.70	536.47	6.30
Unconditional Mean Imputation	493.70	375.51	6.30
Regression Imputation	497.01	484.08	2.99
Stochastic Regression Imputation	497.02	560.56	2.98

Table 3: Table examining imputation techniques when data MNAR

In Table 3 we can see that all techniques perform poorly — even Stochastic Regression Imputation has estimated the mean as too low. All of the techniques underestimate the variance — Unconditional Mean Imputation is the furthest wrong in this respect.

5.3 Discussion

In a simulation study we have access to the true parameters of the data, so we can see easily which techniques perform best according to our measures.

In reality, an analyst will be presented with a data set that has missing values. They will have to do their best to understand which missingness mechanism is driving the non-response pattern, and find an appropriate technique to deal with the missing data.

However, none of the estimates that the single imputation methods deliver give an analyst an idea of imputation uncertainty: how sure can they be about their estimates given they are partly based on imputed data?

The next sections explore techniques designed to deal with this problem.

6 Resampling techniques

One way to deal with imputation uncertainty is to repeatedly resample (with replacement) the observed data and perform imputation on each of these datasets, according to Little and Rubin (2020). Resampling with replacement is known as the Bootstrap process.

A Bootstrap procedure for missing data would go as follows (Little and Rubin, 2020).

Taking a random sample $S = \{i : i = 1, ..., n\}$ of responses, where some of these contain missing data, we can find B bootstrap estimates of a parameter of interest, θ, with the following process, which is described in Little and Rubin (2020).

Let the estimate $\hat{\theta}$ be the estimator for the overall sample S before bootstrap is carried out.

Let a bootstrap estimate be denoted by $\hat{\theta}^b$ for $b = 1, ..., B$.

1. Create a bootstrap sample, denoted $S^{(b)}$ from the original sample, by resampling with replacement.

2. Use an imputation process to impute missing values in $S^{(b)}$, to create the imputed bootstrap sample, $\hat{S}^{(b)}$.

3. Find the bootstrap estimate for $\hat{S}^{(b)}$. The bootstrap estimate of θ is:

$$\hat{\theta}_{boot} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}.$$

A consistent estimator of the sampling variance of $\hat{\theta}$ is:

$$\hat{V}_{boot} = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \hat{\theta}_{boot})^2.$$

Little and Rubin (2020) note that a less computationally intensive way of proceeding would be to create just one imputed data set and resample with replacement on this several times. However, they note this would not give a valid estimate of estimator variance as it doesn't appropriately reflect the imputation uncertainty.

7 Multiple Imputation

Multiple imputation helps users deal with uncertainty around imputation by carrying out several imputations of the same dataset.

There are generally three stages to analysis by multiple imputation, according to Wood et al. (2008). The description below follows theirs.

- The first is creating M multiply imputed data sets. This is done by replacing missing data with values drawn from a posterior predictive distribution: $p(Y_{mis}|Y_{obs})$, (Little and Rubin, 2020).

  In other words, we predict what the missing values would be given the observed ones, and use those predictions as our estimates of what the missing data should be. The draws from the posterior predictive distribution are independent samples (Wood et al., 2008).

- Secondly, the imputed data-sets can be analysed separately (and treated as complete data sets without missing values) to find estimators for each data set.

- Thirdly, these estimators can be combined.

  Wood et al. (2008) set out how this can be done. Let $\hat{\theta}_k$ be an estimator from the k_{th} imputed dataset, where $k = 1, ..., M$. The overall estimator is:

$$\bar{\theta} = \frac{1}{M} \sum_{k=1}^{M} \hat{\theta}_k.$$

  The within-imputation variance is:

$$\bar{W} = \frac{1}{M} \sum_{k=1}^{M} V_k.$$

  Here, V_k represents the estimated variance of $\hat{\theta}_k$. The between-imputation variance is:

$$B = \frac{1}{M-1} \sum_{k=1}^{M} (\hat{\theta}_k - \bar{\theta})(\hat{\theta}_k - \bar{\theta})^T.$$

  The overall variance is then:

$$Var(\bar{\theta}) = \bar{W} + (1 + M^{-1})B$$

Which single imputation methods are adaptable to multiple imputation? According to Little and Rubin (2020), any that involve draws from a predictive distribution of missing values.

7.1 Variable selection

An analyst who wants to impute missing data will generally want to then fit a model to the data to perform analysis. For example, in order to perform a multiple linear regression on a data set will then need to perform model selection. That is, which variables in the data set should be used as predictor variables in the linear regression and which can be discarded? Including all variables often means the model is unneccesarily complex and is against the principle of having a parsimonious model.

Under Multiple Imputation the problem of variable selection — which variables to use in your final model — can become a difficult one. Different iterations of imputations will yield slightly different data sets and so model selection processes may recommend different models for each one. How should an analyst pick which model to use?

Wood et al. (2008) analyse different methods of doing so. They note that Rubin's Rules (introduced in the 1987 edition of Rubin (2004)), prescribe calculating Wald tests on the overall estimators described above. This would mean fitting a proposed model to each imputed data set and then calculating the overall estimators.

They note that this could be computationally very expensive if a lot of imputed datasets are created or if the datasets themselves are big.

Wood et al. (2008) investigate simpler options. These are:

1. Model selection performed on the complete case dataset.

2. Model selection performed on just one of the data sets — Wood et al. (2008) use the first of the data sets from the Multiple Imputation process.

3. Imputation on each dataset resulting from Multiple Imputation. These might recommend different models, so they suggest an analyst could pick predictors that appear only in all the models, in any of the models, or in at least half of the models.

4. Stack all of the imputed datasets, weight them to adjust for the impact stacking has on standard errors, and then proceed with a variable selection process via weighted regression.

They find that the stacking technique is the best alternative, and should be used in situations where Rubin's Rules are impracticable.

7.2 Chained Equation Multiple Imputation (MICE)

This method is for multivariate missing data. Many multiple imputation models require specifying a joint distribution for the variables that the analyst wants to investigate. But, these do not always represent the data as well as we would like them to. This is where this method can help.

Chained Equation Multiple Imputation (MICE) imputes missing data based on conditional distributions. It is implemented in the `mice` package in R, (van Buuren and Groothuis-Oudshoorn, 2011).

Little and Rubin (2020) describe the method as follows.

Firstly defining $X_1, ..., X_k$ as a set of variables with missing values, and Z a vector of variables that are fully observed.

Then x_{obs} is the set of observed data in $X_1, ..., X_k$ and $x_{j(mis)}$ are the missing values for variable j, where $j = 1, ..., K$.

Little and Rubin (2020) recommend we find an apposite model for the conditional distribution $X_j|X_1,...X_{j-1},...,X_{j+1},...X_K,Z$ which has density

$$p_j(x_j|x_1,...,x_{j-1},x_{j+1},...,x_K|x,\theta_j).$$

Parameters for the conditional distribution are θ_j and the prior distribution is $\pi_j(\theta_j)$.

Then we can follow the procedure as follows (Little and Rubin, 2020):

1. Find initial imputations of the missing values in $X_1,...,X_K$. These are denoted $x^{(0)}_{1(mis)},...,x^{(0)}_{k(mis)}$.
2. At each iteration t, update the imputed values as draws from a sequence of predictive distributions, given the imputed values in iteration $(t-1)$.
   For $j=1,...,K$ draw:
   θ^t_j from
   $$\theta_j|x_{(obs)},z,x^t_{1(mis)},...,x^t_{(j-1)(mis)},x^{(t-1)}_{(j(mis)},...,x^{(t-1)}_{(K-1)(mis)}.$$
   Then draw:
   $$x^t_{j(mis)} \sim p(x_{j(mis)}|x_{(obs)},z,x^t_{1(mis)},...,x^t_{(j-1)(mis)},x^{(t-1)}_{(j(mis)},...,x^{(t-1)}_{K-1(mis)}).$$

Azur et al. (2011) note that the process is repeated for several cycles — the precise number to be determined by the analyst.

Little and Rubin (2020) note that unless the conditional distributions form a joint distribution then convergence may not occur, but that does not mean that the process won't yield useful imputations.

8 The Expectation-Maximisation Algorithm

The major alternative to Multiple Imputation to deal with missing data is the Expectation-Maximisation algorithm (Dempster et al., 1977). Dempster et al. (1977) specify as follows: we have complete data Y (which may include some parameters), and observed data $Y_{(obs)}$. We can only observe Y indirectly through $Y_{(obs)}$.

The algorithm consists of an Expectation step (E-step) and a Maximisation step (M-step). This report will first give a description of the simplest case detailed by Dempster et al. (1977): that where the data we want to estimate, $f(Y|\theta)$ are of the exponential family

This means that we can write:

$$f(Y|\theta) = \frac{b(Y)exp(\theta t(Y)^T)}{a(\theta)}.$$

Dempster et al. (1977) define θ as a vector of parameters and $t(Y)$ as a vector of sufficient statistics of the complete data.

The algorithm can run as follows (Dempster et al., 1977):

1. Initialise values for $t(Y)$ and θ.
2. The E-step: estimate $t(Y)$ at the p^{th} iteration with the equation:

$$t^{(p)} = E[t(Y)|Y_{(obs)}, \theta^{(p)}].$$

3. The M-step: estimate $\theta^{(p+1)}$ by solving:

$$E[t(Y)|\theta)] = t^{(p)}.$$

   This should find the maximum likelihood estimator for θ.

4. Repeat until convergence.

More generally, Dempster et al. (1977) introduce the function:

$$Q(\theta'|\theta) = E[logf(Y|\theta')|Y_{(obs)}, \theta]$$

They assume that this functions exists for all (θ', θ) and that $f(Y|\theta) > 0$ almost everywhere in the sample space of Y.

Then the p^{th} step of the algorithm can be described as below (Dempster et al., 1977):

- E-step: Find $Q(\theta|\theta^{(p)})$
- M-Step: Find the value of $\theta^{(p+1)}$ that maximises the likelihood, $Q(\theta|\theta^{(p)})$.

The EM Algorithm would produce a single estimate for each piece of missing data, so it does not of itself give information on uncertainty due to missingness (referred to earlier as imputation uncertainty).

One way to deal with this is the Expectation-Maximisation Algorithm with Bootstrapping (Honaker and King, 2010). This is compared with methods of Multiple Imputation in Takahashi (2017).

The algorithm involves drawing Bootstrap samples — resampling with replacement — on the original data set, then using the EM algorithm on each of the Bootstrap samples. This provides understanding of variation due to missing data.

9 Further Research: Missing Not at Random Data

A major area of further research lies in tackling missing data when it is Missing Not at Random.

This is a particularly difficult problem because not only do we not observe the missing data, but because the missingness depends on the data that is missing, we cannot observe much about what drives the missingness.

On top of that, it is impossible to test whether missingness is MAR or MNAR on real data Tompsett et al. (2018).

For example, to return to our simulated example of racehorse heights and weights. If we can imagine there was low correlation between height and weight, and that heavier horses are less likely to record weight, it could be very difficult to impute missing data because we have little information to tell us what the missing data might be.

We might imagine that weigh bridges for horses only go up to 600kg, so horses over 600kg are not recorded. If we do not have this information, it is difficult to impute the missing data.

Little and Rubin (2020) divide MNAR problems into two main types:

1. The missingness mechanism is known, albeit MNAR. So there are no unknown parameters Φ that describe the missingness mechanism. We might have this situation if we know that our horse weigh bridges do not work consistently when recording weights of more than 600kg, so horses that weigh more than 600kg are less likely to have their weight recorded.

2. The missingness mechanism is not known and there are unknown parameters Φ in the model. This would be if, for some unknown reason, heavier racehorses were more reluctant to step on the scales.

Little and Rubin (2020) discuss examples where the EM Algorithm can be applied to problems that fall into the first category.

The second category is more difficult. One simple way to deal with MNAR responses of the second type is to follow up on some non-respondents, to gain more information (Little and Rubin, 2020). Some others mentioned by Little and Rubin (2020) involve making assumptions about the type of MNAR mechanism showing in a dataset — and it is often difficult to tell whether it is reasonable to do so from the data.

Little and Rubin (2020) recommend Sensitivity Analysis as a way to assess sensitivity of parameter estimates to getting the MNAR mechanism wrong.

Analysts determine an MNAR model with unidentified parameters, the sensitivity parameters, and allow them to vary (Tompsett et al., 2018). This should show how sensitive the MNAR model is to variation in the parameters.

Sensitivity parameters have been adapted into a procedure related to MICE — Not at Random Fully Conditional Specification — which is discussed in Tompsett et al. (2018). The authors note that exploring real life applications of this method is an area for future work.

Another approach is to consider experiment design when data is MNAR. This is explored in Lee et al. (2018).

For example, Lee et al. (2018) note that estimates are bound to be biased when responses are MNAR, so they look to modify a design which might have only looked at minimising the variance of an estimator, to one minimising mean squared error (which includes bias and variance). Even this is not straightforward, however, because they will not know the precise nature of the bias.

To the authors' knowledge, this is the first exploration of adapting experiment design to tackle the problems that MNAR response data can throw up.

10 Conclusion

There are many methods out there to deal with missing data, and many are much more appropriate than simple methods like Unconditional Mean Imputation and Complete Case Analysis.

It is clear that methods developed to tackle missing data require understanding of the data set and what drives the missingness.

Doing what might at first appear to be fair — getting rid of incomplete data — can result in biased estimates and wider confidence intervals around them.

There is a lot of work to be done in the area of data that is Missing Not at Random.

Unfortunately, it is likely that a lot of data collected through surveys is Missing Not at Random, and yet it is often treated as Missing at Random due to a lack of consistent techniques that can be applied.

Current methods centre around attempting to lessen the impact of having Missing Not at Random response data, or of using Sensitivity Analysis to effectively try to tell us how big of a deal it is that we have MNAR data.

So far, the body of techniques to inform analysis of MNAR data is much smaller, and much more specialised than that of MCAR or MAR data.

References

Andridge, R. R. and Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Honaker, J. and King, G. (2010). What to do about missing values in time-series cross-section data. *American journal of political science*, 54(2):561–581.

Lee, K. M., Mitra, R., and Biedermann, S. (2018). Optimal design when outcome values are not missing at random. *Statistica Sinica*, 28(4):1821–1838.

Little, R. J. A. and Rubin, D. B. (2020). *Statistical analysis with missing data*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, third edition.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Takahashi, M. (2017). Statistical inference in missing data by mcmc and non-mcmc multiple imputation algorithms: Assessing the effects of between-imputation iterations. *Data Science Journal*, 16.

Tompsett, D. M., Leacy, F., Moreno-Betancur, M., Heron, J., and White, I. R. (2018). On the use of the not-at-random fully conditional specification (narfcs) procedure in practice. *Statistics in medicine*, 37(15):2338–2353.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in medicine*, 27(17):3227–3246.