1

Introduction

Collecting, analysing and drawing inferences from data are central to research in the medical and social sciences. Unfortunately, for any number of reasons, it is rarely possible to collect all the intended data. The ubiquity of missing data, and the problems this poses for both analysis and inference, has spawned a substantial statistical literature dating from 1950s. At that time, when statistical computing was in its infancy, many analyses were only feasible because of the carefully planned balance in the dataset (for example, the same number of observations on each unit). Missing data meant the available data for analysis were unbalanced, thus complicating the planned analysis and in some instances rendering it unfeasible. Early work on the problem was therefore largely computational (e.g. Healy and Westmacott, 1956; Afifi and Elashoff, 1966; Orchard and Woodbury, 1972; Dempster *et al*., 1977).

The wider question of the consequences of nontrivial proportions of missing data for inference was neglected until a seminal paper by Rubin (1976). This set out a typology for assumptions about the reasons for missing data, and sketched their implications for analysis and inference. It marked the beginning of a broad stream of research about the analysis of partially observed data. The literature is now huge, and continues to grow, both as methods are developed for large and complex data structures, and as increasing computer power and suitable software enable researchers to apply these methods.

For a broad overview of the literature, a good place to start is one of the recent excellent textbooks. Little and Rubin (2002) write for applied statisticians. They give a good overview of likelihood methods, and give an introduction to multiple imputation. Allison (2002) presents a less technical overview. Schafer (1997) is more algorithmic, focusing on the EM algorithm and imputation using the multivatiate normal and general location model. Molenberghs and Kenward (2007)

Multiple Imputation and its Application, First Edition. James R. Carpenter and Michael G. Kenward.

focus on clinical studies, while Daniels and Hogan (2008) focus on longitudinal studies with a Bayesian emphasis.

The above books concentrate on parametric approaches. However, there is also a growing literature based around using inverse probability weighting, in the spirit of Horvitz and Thompson (1952), and associated doubly robust methods. In particular, we refer to the work of Robins and colleagues (e.g. Robins *et al*., 1995; Scharfstein *et al*., 1999). Vansteelandt *et al*. (2009) give an accessible introduction to these developments. A comparison with multiple imputation in a simple setting is given by Carpenter *et al*. (2006). The pros and cons are debated in Kang and Schafer (2007) and the theory is brought together by Tsiatis (2006).

This book is concerned with a particular statistical method for analysing and drawing inferences from incomplete data, called *Multiple Imputation (MI)*. Initially proposed by Rubin (1987) in the context of surveys, increasing awareness among researchers about the possible effects of missing data (e.g. Klebanoff and Cole, 2008) has led to an upsurge of interest (e.g. Sterne *et al*., 2009; Kenward and Carpenter, 2007; Schafer, 1999a; Rubin, 1996).

Multiple imputation (MI) is attractive because it is both practical and widely applicable. Recently developed statistical software (see, for example, issue 45 of the *Journal of Statistical Software*) has placed it within the reach of most researchers in the medical and social sciences, whether or not they have undertaken advanced training in statistics. However, the increasing use of MI in a range of settings beyond that originally envisaged has led to a bewildering proliferation of algorithms and software. Further, the implication of the underlying assumptions in the context of the data at hand is often unclear.

We are writing for researchers in the medical and social sciences with the aim of clarifying the issues raised by missing data, outlining the rationale for MI, explaining the motivation and relationship between the various imputation algorithms, and describing and illustrating its application to increasingly complex data structures.

Central to the analysis of partially observed data is an understanding of why the data are missing and the implications of this for the analysis. This is the focus of the remainder of this chapter. Introducing some of the examples that run through the book, we show how Rubin's typology (Rubin, 1976) provides the foundational framework for understanding the implications of missing data.

1.1 Reasons for missing data

In this section we consider possible reasons for missing data, illustrate these with examples, and draw some preliminary implications for inference. We use the word 'possible' advisedly, since with partially observed data we can rarely be sure of the mechanism giving rise to missing data. Instead, a range of possible mechanisms are consistent with the observed data. In practice, we therefore wish to analyse the data under different mechanisms, to establish the robustness of our inference in the face of uncertainty about the missingness mechanism.

All datasets consist of a series of *units* each of which provides information on a series of *items*. For example, in a cross-sectional questionnaire survey, the units would be individuals and the items their answers to the questions. In a household survey, the units would be households, and the items information about the household and members of the household. In longitudinal studies, units would typically be individuals while items would be longitudinal data from those individuals. In this book, units therefore correspond to the highest level in multilevel (i.e., hierarchical) data, and unless stated otherwise data from different units are statistically independent.

Within this framework, it is useful to distinguish between units where all the information is missing, termed *unit nonresponse* and units who contribute partial information, termed *item nonresponse*. The statistical issues are the same in both cases, and both can in principle be handled by MI. However, the main focus of this book is the latter.

Example 1.1 Mandarin tableau

Figure 1.1, which is also shown on the cover, shows part of the frontage of a senior mandarin's house in the New Territories, Hong Kong. We suppose interest focuses on characteristics of the figurines, for example their number, height, facial characteristics and dress. Unit nonresponse then corresponds to missing figurines, and item nonresponse to damaged – hence partially observed – figurines. □

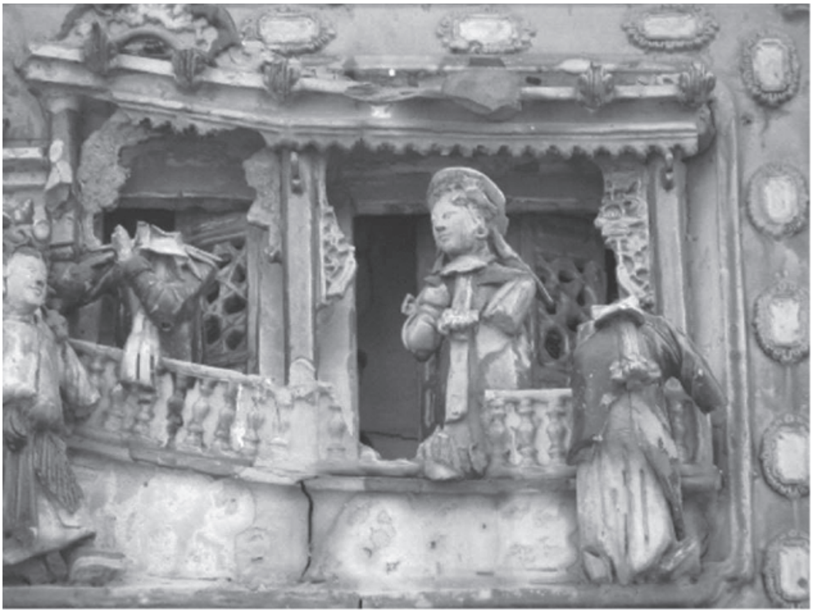


Figure 1.1 Detail from a senior mandarin's house front in New Territories, Hong Kong. Photograph by H. Goldstein.

1.2 Examples

We now introduce two key examples, which we return to throughout the book.

Example 1.2 Youth Cohort Study (YCS)

The Youth Cohort Study of England and Wales (YCS) is an ongoing UK government funded representative survey of pupils in England and Wales at school-leaving age (School year 11, age 16–17) (UK Data Archive, 2007). Each year that a new cohort is surveyed, detailed information is collected on each young person's experience of education and their qualifications as well as information on employment and training. A limited amount of information is collected on their personal characteristics, family, home circumstances, and aspirations.

Over the life-cycle of the YCS, different organisations have had responsibility for the structure and timings of data collection. Unfortunately, the documentation of older cohorts is poor. Croxford *et al.* (2007) have recently deposited a harmonised dataset that comprises YCS cohorts from 1984 to 2002 (UK Data Archive Study Number 5765). We consider data from pupils attending comprehensive schools from five YCS cohorts; these pupils reached the end of Year 11 in 1990, 1993, 1995, 1997 and 1999.

We explore relationships between Year 11 educational attainment (the General Certificate of Secondary Education) and key measures of social stratification. The units are pupils and the items are measurements on these pupils, and a nontrivial number of items are partially observed. □

Example 1.3 Randomised controlled trial of patients with chronic asthma

We consider data from a 5-arm asthma clinical trial to assess the efficacy and safety of budesonide, a second-generation glucocorticosteroid, on patients with chronic asthma. 473 patients with chronic asthma were enrolled in the 12-week randomised, double-blind, multi-centre parallel-group trial, which compared the effect of a daily dose of 200, 400, 800 or 1600 mcg of budesonide with placebo.

Key outcomes of clinical interest include patients' peak expiratory flow rate (their maximum speed of expiration in litres/minute) and their Forced Expiratory Volume, FEV_1, (the volume of air, in litres, the patient with fully inflated lungs can breathe out in one second). In summary, the trial found a statistically significant dose-response effect for the mean change from baseline over the study for both morning peak expiratory flow, evening peak expiratory flow and FEV_1, at the 5% level.

Budesonide treated patients also showed reduced asthma symptoms and bronchodilator use compared with placebo, while there were no clinically significant differences in treatment related adverse experiences between the treatment groups. Further details about the conduct of the trial, its conclusions and the variables collected can be found elsewhere (Busse *et al.*, 1998). Here, we focus on FEV_1 and confine our attention to the placebo and lowest active dose arms. FEV_1 was collected at baseline, then 2, 4, 8 and 12 weeks after randomisation. The intention was to compare FEV_1 across treatment arms at 12 weeks. However, excluding 3 patients whose participation in the study was intermittent, only 37 out of 90 patients in the placebo arm, and 71 out of 90 patients in the lowest active dose arm, still remained in the trial at twelve weeks. □

1.3 Patterns of missing data

It is very important to investigate the patterns of missing data before embarking on a formal analysis. This can throw up vital information that might otherwise be overlooked, and may even allow the missing data to be traced. For example, when analysing the new wave of a longitudinal survey, a colleague's careful examination of missing data patterns established that many of the missing questionnaires could be traced to a set of cardboard boxes. These turned out to have been left behind in a move. They were recovered and the data entered.

Most statistical software now has tools for describing the pattern of missing data. Key questions concern the extent and patterns of missing values, and whether the pattern is *monotone* (as described in the next paragraph), as if it is, this can considerably speed up and simplify the analysis.

Missing data in a set of p variables are said to follow a *monotone missingness pattern* if the variables can be re-ordered such that, for every unit i and variable j,

1. if unit i is observed on variable j, where $j = 2, \ldots, p$, it is observed on all variables $j' < j$, and
2. if unit i is missing on variable j, where $j = 2, \ldots, p$, it is missing on all variables $j' > j$.

A natural setting for the occurrence of monotone missing data is a longitudinal study, where units are observed either until they are lost to follow-up, or the study concludes. A monotone pattern is thus inconsistent with interim missing data, where units are observed for a period, missing for the subsequent period, but then observed. Questionnaires may also give rise to monotone missing data patterns when individuals systematically answer each question in turn from the beginning till they either stop or complete the questionnaire. In other settings it may be possible to re-order items to achieve a monotone pattern.

Example 1.2 Youth Cohort Study *(ctd)*

Table 1.1 shows the covariates we consider from the YCS. There are no missing data in the variables *cohort* and *boy*. The missingness pattern for GCSE score and the remaining two variables is shown in Table 1.2. In this example it is not possible to re-order the variables (items) to obtain a monotone pattern, due for example, to pattern 3 (N = 697). □

Table 1.1 YCS variables for exploring the relationship between Year 11 attainment and social stratification.

Variable name	Description
cohort	year of data collection: 1990, 93, 95, 97, 99
boy	indicator variable for boys
occupation	parental occupation, categorised as managerial, intermediate or working
ethnicity	categorised as Bangladeshi, Black, Indian, other Asian, Other, Pakistani or White

Table 1.2 Pattern of missing values in the YCS data.

Pattern	GCSE score	Occupation	Ethnicity	No.	% of total
1	✓	✓	✓	55145	87%
2	✓	·	✓	6821	11%
3	·	✓	✓	697	1%
4	✓	·	·	592	1%

Example 1.3 Asthma study *(ctd)*

Table 1.3 shows the withdrawal pattern for the placebo and lowest active dose arms (all the patients are receiving their randomised medication). We have removed three patients with unusual interim missing data from Table 1.3 and all our analyses. The remaining missingness pattern is monotone in both treatment arms. □

Table 1.3 Asthma study: withdrawal pattern by treatment arm.

Dropout pattern	Placebo arm						
	Mean FEV$_1$ (litres) measured at week					Number	Percent
	0	2	4	8	12		
1	✓	✓	✓	✓	✓	37	41
2	✓	✓	✓	✓	·	15	17
3	✓	✓	✓	·	·	22	24
4	✓	✓	·	·	·	16	18
	Lowest Active arm						
1	✓	✓	✓	✓	✓	71	79
2	✓	✓	✓	✓	·	8	9
3	✓	✓	✓	·	·	8	9
4	✓	✓	·	·	·	3	3

1.3.1 Consequences of missing data

Our focus is the practical implications of missing data for both parameter estimation and inference. Unfortunately, the two are often conflated, so that a computational method for parameter estimation when data are missing is said to have ‘solved’ or ‘handled’ the missing data issue. Since, with missing data, computational methods only lead to valid inference under specific assumptions, this attitude is likely to lead to misleading inferences.

In this context, it may be helpful to draw an analogy with the sampling process used to collect the data. If an analyst is presented with a spreadsheet containing columns of numerical data, they can analyse the data (calculate means of variables, regress variables on each other and so forth). However, they cannot draw any inferences unless they are told how and from whom the data were collected. This information is external to the numerical values of the variables.

We may think of the missing data mechanism as a second stage in the sampling process, but one that is not under our control. It acts on the data we intended to collect and leaves us with a partially observed dataset. Once again, the missing data mechanism cannot usually be definitively identified from the observed data, although the observed data may indicate plausible mechanisms (e.g. response may be negatively correlated with age). Thus we will need to make an assumption about the missingness mechanism in order to draw inference. The process of making this assumption is quite separate from the statistical methods we use for parameter estimation etc. Further, to the extent that the missing data mechanism cannot be definitively identified from the data, we will often wish to check the robustness of our inferences to a range of missingness mechanisms that are consistent with the observed data. The reason this book focuses on the statistical method of MI is that it provides a computationally feasible approach to the analysis for a wide range of problems under a range of missingness mechanisms.

We therefore begin with a typology for the mechanisms causing, or generating, the missing data. Later in this chapter we will see that consideration of these mechanisms in the context of the analysis at hand clarifies the assumptions under which a simple analysis, such as restriction to complete records, will be valid. It also clarifies when more sophisticated computational approaches such as MI will be valid and informs the way they are conducted. We stress again that the mechanism causing the missing data can rarely be definitively established. Thus we will often wish to explore the robustness of our inferences to a range of plausible missingness mechanisms – a process we call *sensitivity analysis*.

From a general standpoint, missing data may cause two problems: loss of efficiency and bias.

First, loss of efficiency, or information, is an inevitable consequence of missing data. Unfortunately, the extent of information loss is not directly linked to the proportion of incomplete records. Instead it is intrinsically linked to the analysis question. When crossing the road, the rear of the oncoming traffic is hidden from view – the data are missing. However, these missing data do not bear on the question at hand – will I make it across the road safely? While the proportion of missing data about each oncoming vehicle is substantial, information loss is negligible. Conversely, when estimating the prevalence of a rare disease, a small proportion of missing observations could have a disproportionate impact on the resulting estimate.

Faced with an incomplete dataset, most software automatically restricts analysis to complete records. As we illustrate below, the consequence of this for loss of information is not always easy to predict. Nevertheless, in many settings it will be important to include the information from partially complete records. Not least of the reasons for this is the time and money it has taken to collect even the partially complete records. Under certain assumptions about the missingness mechanism, we shall see that MI provides a natural way to do this.

Second, and perhaps more fundamentally, the subset of complete records may not be representative of the population under study. Restricting analysis to complete records may then lead to biased inference. The extent of such bias depends on the statistical behaviour of the missing data. A formal framework to describe this behaviour is thus fundamental. Such a framework was first elucidated in a seminal paper by Rubin (1976). To describe this, we need some definitions.

1.4 Inferential framework and notation

For clarity we take a frequentist approach to inference. This is not essential or necessarily desirable; indeed we will see that MI is essentially a Bayesian method, with good frequentist properties. Often, as Chapter 2 shows, formally showing these frequentist properties is most difficult theoretically.

We suppose we have a sample of n units, which will often be individuals, from a population that for practical inferential purposes can be considered infinite. Let $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \dots, Y_{i,p})^T$ denote the p variables we intended to collect from the i^{th} unit, $i = 1 \dots, n$. We wish to use these data to make inferences about a set of p population parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$.

For each unit $i = 1, \dots, n$ let $\mathbf{Y}_{i,O}$ denote the subset of p variables that are observed, and $\mathbf{Y}_{i,M}$ denote the subset that are missing. Thus, for different individuals, $\mathbf{Y}_{i,O}$ and $\mathbf{Y}_{i,M}$ may well be different subsets of the p variables. If no data are missing, $\mathbf{Y}_{i,M}$ will be empty.

Next, again for each individual $i = 1, \dots, n$ and variable $j = 1, \dots, p$, let $R_{i,j} = 1$ if $Y_{i,j}$ is observed and $R_{i,j} = 0$ if $Y_{i,j}$ is missing. Let $\mathbf{R}_i = (R_{i,1}, \dots, R_{i,p})^T$. Consistent with the definition of monotone missingness patterns on p. 10, the pattern is monotone if the p variables can be re-ordered so that for each unit i,

$$R_{i,j} = 0 \implies R_{i,j'} = 0 \text{ for } j' = j + 1, \dots, p. \tag{1.1}$$

The missing value mechanism is then formally defined as

$$\Pr(\mathbf{R}_i | \mathbf{Y}_i), \tag{1.2}$$

that is to say the probability of observing unit i's data given their potentially unseen values $\mathbf{Y}_i$. It is important to note that, in what follows, we assume that unit i's data exist (or at least existed). In other words, if it had been possible for us to be in the right place at the right time, we would have been able to observe the complete data. What (1.2) describes therefore, is the probability that the data collection we were able to undertake on unit i yielded values of $Y_{i,0}$. Thus, (at least until we consider sensitivity analysis for clinical trials in Chapter 10) the missing data are not counter-factual, in the sense of what might have happened if a patient had taken a different drug from the one they actually took, or a child had gone to a different school from the one they actually attended.

Example 1.1 Mandarin tableau *(ctd)*

Here, $\mathbf{Y}_i$ take the form of observations on the $n = 4$ figurines, describing for example their size and dress. $R_{i,j}$ indicates those observations that are missing on figurine i because its head is missing. Originally, of course, all the heads were present, so we can refer to the underlying values of the unobserved variables. □

Example 1.2 Youth Cohort Study (YCS) *(ctd)*

Here, underlying values of missing GCSE score, parental occupation and ethnicity exist, and given sufficient time and money we would be able to discover many of them. □

Example 1.3 Asthma study *(ctd)*

Were resources not limited, researchers could have visited each patient in their home at each of the scheduled follow-up times to record their data. □

We now come to the three classes of missing data mechanism. These describe how the probability of seeing the data depends on the observed, and unobserved (but potentially observable, or underlying) values. In general, depending on the context, we will think of the same mechanism applying either to all $i = 1, \ldots, n$ units in the data set, or to an independent subset of them.

1.4.1 Missing Completely At Random (MCAR)

We say data are *Missing Completely At Random (MCAR)* if the probability of a value being missing is unrelated to the observed and unobserved data on that unit. Algebraically,

$$\Pr(\mathbf{R}_i|\mathbf{Y}_i) = \Pr(\mathbf{R}_i). \qquad (1.3)$$

Since, when data are MCAR, the chance of the data being missing is unrelated to the values, the observed data are therefore representative of the population. However, relative to the data we intended to collect, information has been lost.

Example 1.1 Mandarin tableau *(ctd)*

Suppose we wish to summarise facial characteristics of the figurines, e.g. average head circumference. If the missing heads are MCAR, a valid estimate is obtained from the observed heads. Although valid, it is imprecise relative to an estimate based on all the heads.

Before moving on, note that the MCAR assumption is made for a specific analysis. It is not a property of the tableau. It may be plausible to assume that headgear is MCAR, while heads may systematically be missing because of racial characteristics. Further, if we step back from the tableau, we may see that missing heads correspond to missing, or recently replaced, roof tiles. If so, the mechanism causing the missing data is clear: however the assumption of MCAR is still likely to be appropriate, because the mechanism causing the missing data is unlikely to bear on (i.e., is likely statistically independent of) the analysis question.

Similarly, in certain settings we may find that the variables predictive of missing data are independent of the substantive analysis at hand. This is consistent with the MCAR assumption: analysis of the complete records will be unbiased, but some precision is lost. □

Example 1.2 Youth Cohort Study *(ctd)*

If data are MCAR in the YCS study, valid inference would be obtained from the 55145 complete records (Table 1.2). However, omitting the 8110 individuals with partial information means inferences are less precise than they could be. □

Example 1.3 Asthma study *(ctd)*

Assuming data are MCAR, a valid estimate of the overall mean in each group at 12 weeks is obtained by simply averaging the 37 available observations in the placebo group and the 71 available observations in the active group. This gives, respectively 2.05 *l* (s.e. 0.09) and 2.23 *l* (s.e. 0.10). □

1.4.2 Missing At Random (MAR)

We say data are *Missing At Random (MAR)* if *given, or conditional on, the observed data* the probability distribution of $\mathbf{R}_i$ is independent of the unobserved data. Recalling that for individual i we can partition $\mathbf{Y}_i$ as $(\mathbf{Y}_{i,O}, \mathbf{Y}_{i,M})$ we can express this mathematically as

$$\Pr(\mathbf{R}_i|\mathbf{Y}_i) = \Pr(\mathbf{R}_i|\mathbf{Y}_{i,O}). \tag{1.4}$$

This does not mean – as is sometimes supposed – that the probability of observing a variable on an individual is independent of the value of that variable. Quite the contrary: under MAR the chance of observing a variable will depend on its value. Crucially though, given the observed data this dependence is broken. Consider the following example.

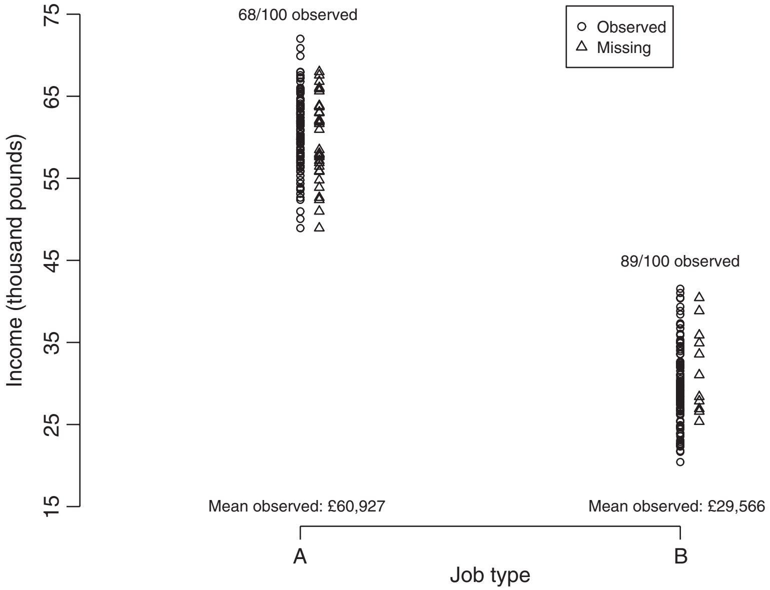


Figure 1.2 Plot of 200 hypothetical incomes against job type.

Example 1.4 Income and job type

Suppose we survey 100 employees of job type A and B for their income. Only 157 reveal their income, as shown in Figure 1.2. The figure shows that employees with higher incomes are less likely to divulge them: the probability of observing a variable depends on its value. However, if within job type A the probability of observing income does not depend on income, and within job type B the probability of observing income does not depend on income, then income is missing at random dependent on job type. □

The immediate consequence of this is that the mean of the observed incomes, marginal to (or aggregating over) job type is biased downwards. The data were generated with a mean income of £60,000 in job type A and £30,000 in job type B, so that the true mean income is £45,000. Contrast the observed mean income of

$$(68 \times 60927 + 89 \times 29566)/157 = £43{,}149.$$

We note three further points. First, if within job type the probability of observing income does not depend on income, it follows that:

1. To say 'income is MAR' is incomplete; we need instead to specify the variable which we assume makes income conditionally independent of job type. We could say

   'income is MAR, dependent on job type'

or, perhaps more explicitly,

> 'within categories of job type, income is MCAR.'

2. If income is MAR, dependent on job type,

   - in job type A the distribution of unobserved and observed incomes is the same, and
   - in job type B the distribution of unobserved and observed incomes is the same.

   Formally, let variable $Y_{i,1}$ be income and $Y_{i,2}$ be job type. Job type is always observed so $R_{i,2} = 1$ for all individuals i. The statement 'income is MAR, dependent on job type' is expressed

$$\Pr(R_{i,1} = 1|Y_{i,1}, Y_{i,2}) = \Pr(R_{i,1} = 1|Y_{i,2}). \tag{1.5}$$

   Now consider what this implies for the distribution of income given job type. By repeatedly using the definition of conditional probability

$$\begin{aligned}\Pr(Y_{i,1}|Y_{i,2}, R_{i,1} = 1) &= \frac{\Pr(Y_{i,1}, Y_{i,2}, R_{i,1} = 1)}{\Pr(Y_{i,2}, R_{i,1} = 1)} \\ &= \frac{\Pr(R_{i,1} = 1|Y_{i,1}, Y_{i,2}) \Pr(Y_{i,1}, Y_{i,2})}{\Pr(R_{i,1} = 1|Y_{i,2}) \Pr(Y_{i,2})} \\ &= \Pr(Y_{i,1}|Y_{i,2}), \end{aligned} \tag{1.6}$$

   where the last step follows from MAR, i.e., (1.5). The argument (1.6) holds if income is not observed, $R_{i,1} = 0$. Thus under MAR the distribution of income within job type is the same in the observed data, the unobserved data, and the population.

   In this case, to estimate the marginal income we average the observed income in each job type and then scale up:

$$(100 \times 60927 + 100 \times 29566)/200 = £45{,}247. \tag{1.7}$$

   Notice that to obtain this estimate we did not need to explicitly specify how the probability of observing income depends on job type; merely that given job type it does not depend on income.

3. The statement 'income is MAR, dependent on job type' is an untestable assumption. The data we would need to test it (represented by the triangles in Figure 1.2) is missing!

Of course, were it observed, we could 'test the MAR assumption' in two ways: first a logistic regression, for example:

$$\operatorname{logit}\Pr\{(R_{i,1}=1)\} = \alpha_0 + \alpha_1 Y_{i,1} + \alpha_2 Y_{i,2} + \alpha_3 Y_{i,1} Y_{i,2};$$

if MAR is true then the hypothesis $\alpha_1 = \alpha_3 = 0$ is true. Or, we could fit a corresponding regression:

$$\mathrm{E}(Y_{i,1}) = \beta_0 + \beta_1 Y_{i,2} + \beta_2 R_{i,1} + \beta_3 Y_{i,2} R_{i,1};$$

If MAR is true then the hypothesis $\beta_2 = \beta_3 = 0$ is true.

This simple example draws out the following general points:

1. statements relating the probability of observing data to the values of data have direct consequences for conditional distributions of the data, and
2. under the MAR assumption, the precise missing data mechanism need not be specified; indeed the precise form can be different for different individuals.

These two points together mean that the MAR mechanism is much more subtle than might at first appear; these subtleties can manifest themselves unexpectedly.

Example 1.4 Income and job type *(ctd)*

Suppose the mechanism causing the missing income differed for each of the 200 individuals, that is

$$\operatorname{logit}\Pr(R_{i,1}=1) = \alpha_{0,i} + \alpha_{1,i} Y_{i,2}.$$

Then missing data are still MAR, and (1.7) is still a valid estimate. □

Of course, it may be as contrived to think each individual has their own MAR mechanism as to think that the same mechanism holds for all. In a simple example this is not important, but in real applications a blanket assumption of MAR may be very contrived.

Example 1.5 Subtlety of MAR assumption

Suppose we have three variables, $Y_{i,1}$, $Y_{i,2}$, $Y_{i,3}$, and we are unfortunate, so that our dataset contains nontrivial numbers of all possible missingness patterns, as shown in Table 1.4.

If the same missingness mechanism applies to all the units, and it is either MAR or MCAR, then it must be MCAR. If we wish to assume data are MAR, we are forced to split the data into groups among which different MAR mechanisms are operating. These groups need not necessarily be defined by the missing data

Table 1.4 Three variables: all possible missing value patterns.

Pattern	Y_1	Y_2	Y_3
1	✓	✓	✓
2	✓	✓	·
3	✓	·	✓
4	·	✓	✓
5	✓	·	·
6	·	✓	·
7	·	·	✓

patterns; they could be defined by characteristics of the units. Settings like this are considered by Harel and Schafer (2009). To illustrate, though, we define groups by the missing data patterns.

For a MAR mechanism, we might assume the following:

- in patterns (1, 2) $Y_{i,3}$ is MAR given $Y_{i,1}$, $Y_{i,2}$;
- in patterns (3, 4, 7) $Y_{i,1}$ and/or $Y_{i,2}$ is MAR given $Y_{i,3}$, and
- in patterns (5, 6) data are MCAR.

In practice, often a relatively small number of the possible missingness patterns predominate, and it is assumptions about these that are important for any analysis. The remaining – relatively infrequent – patterns can often be assumed MCAR, with little risk to the final inference if this assumption is in fact wrong. □

Faced with complex data, there is a temptation to invoke the MAR assumption too readily, especially as this simplifies any analysis using MI. To guard against this, analysts need to be satisfied that any associations assumed to justify the MAR assumption are at least consistent with the observed data. Since consideration of selection mechanisms may not be as straightforward as might first appear, it can also be worth considering the plausibility of MAR from the point of view of the joint and conditional distribution of the data. As (1.6) illustrates, for MAR we need to be satisfied

1. that conditional distributions of partially observed variables given fully observed variables do not differ depending on whether the data are observed, and
2. in consequence the joint distribution of the data can be validly estimated by piecing together the marginal distributions of the observed patterns.

The above discussion explains why we do not regard the MAR assumption as a panacea, but nevertheless often both a plausible and practical starting point for the analysis of partially observed data. In particular, the points drawn out of Example 1.4 are not specific to either the number or type of variables (categorical or quantitative).

Example 1.1 Mandarin tableau *(ctd)*

Here the MAR assumption says that the distribution of head characteristics given body characteristics (i.e., dress, height, etc.) does not depend on whether the head is present. Thus, under MAR we can estimate the distribution of characteristics of figurines with missing heads from figurines with similar body characteristics.

Notice the two rightmost figurines in Figure 1.1 share the same necktie. Assuming headdress is MAR given necktie, the missing headdress on the rightmost figurine is similar to that on the second rightmost figurine.

Clearly this assumption cannot be checked from the tableau (data) at hand. However it might be possible to explore it using other tableaux (i.e., other datasets). If MAR is plausible for headdress given necktie, it does not mean it is plausible for skin colour given necktie. In other words MAR is an assumption we make for the analysis, not a characteristic of the dataset. For some analyses of partially observed data it may be plausible; for others not. □

1.4.3 Missing Not At Random (MNAR)

If the mechanism causing missing data is neither MCAR nor MAR, we say it is Missing Not At Random (MNAR). Under a MNAR mechanism, the probability of an observation being missing depends on the underlying value, and this dependence remains even given the observed data. Mathematically,

$$\Pr(\mathbf{R}_i|\mathbf{Y}_i) \neq \Pr(\mathbf{R}_i|\mathbf{Y}_{i,O}). \tag{1.8}$$

While in some settings MNAR may be more plausible than MAR, analysis under MNAR is considerably harder. This is because under MAR, equation (1.6) showed that conditional distributions of partially observed variables given fully observed variables are the same in units who do, and do not, have the data observed. However (1.6) does not hold if (1.8) holds.

It follows that inference under MNAR involves an explicit specification of either the selection mechanism, or how conditional distributions of partially observed variables given fully observed variables differ between units who do, and do not, have the data observed.

Formally, we can write the joint distribution of unit i's variables, $\mathbf{Y}_i$, and the indicator for observing those variables, $\mathbf{R}_i$ as

$$\Pr(\mathbf{R}_i|\mathbf{Y}_i)\Pr(\mathbf{Y}_i) = \Pr(\mathbf{R}_i, \mathbf{Y}_i) = \Pr(\mathbf{Y}_i|\mathbf{R}_i)\Pr(\mathbf{R}_i). \tag{1.9}$$

In the centre is the joint distribution, and this can be written either as

1. a *selection model* – the LHS of (1.9), i.e., a product of (i) the conditional probability of observing the variables, given their values and (ii) the marginal distribution of the data, OR

2. a pattern mixture model – the RHS of (1.9), i.e., a product of (i) the probability distribution of the data within each missingness pattern and (ii) the marginal probability of the missingness pattern.

Thus we can specify a MNAR mechanism either by specifying the selection model (which implies the pattern mixture model) or by specifying a pattern mixture model (which implies a selection model). Depending on the context, both approaches may be helpful. Unfortunately, even in apparently simple settings, explicitly calculating the selection implication of a pattern mixture model, or vice versa, can be awkward. We shall see in Chapter 10 that an advantage of multiple imputation is that, given a pattern mixture model, we can estimate the selection model implications quite easily.

Once again, as the example below shows, MNAR is an assumption for the analysis, not a characteristic of the data.

Example 1.1 Mandarin tableau *(ctd)*

It may be that the figurines with missing heads were wearing a head dress that identified them as a member of a class, or group, that subsequently became very unpopular – causing the heads to be smashed. This MNAR selection mechanism means that we cannot say anything about the typical characteristics of head dress without making untestable assumptions about the characteristics of the missing head dresses. Further, the MNAR assumption implies that the distribution of head dress given body dress is different for figurines with missing and observed heads.

We reiterate, under MNAR any summary statistics, or analyses, require *either* explicit assumptions about the form of the distribution of the missing data given the observed *or* explicit specification of the selection mechanism and the marginal distribution of the full (including unobserved) data. Contrast this with analyses assuming MAR, where these assumptions are made implicitly.

We repeat a point from the tableau: if head dress was the trigger for missing heads, but the type of head dress worn is not related to physical characteristics of the heads, analyses concerning their physical characteristics could be validly performed under MAR. Just because the heads are MNAR does not mean all analyses require the MNAR assumption. This underlines that, in applications, it is crucial to think carefully about the selection mechanism, and how it affects the analysis question. □

Example 1.6 Income MNAR

To illustrate (1.9), consider a simplified version of the income example above. Suppose that of the 100 people surveyed, 50 have the same income θ_L, and 50 have the same higher income, θ_U. Suppose further that all those with income θ_L disclose it, but only a fraction π of those with income θ_U disclose it.

This is an example of pattern mixture model, i.e., the RHS of (1.9). Let $1[\,.\,]$ be 1 if the statement in brackets is true and 0 otherwise. Then, in this simple example, it is clear what the the selection counterpart is:

$$\Pr(\text{income observed}) = 1 + (\pi - 1) \times 1[\text{income} = \theta_U], \quad \text{and}$$

$$\text{mean income} = (\theta_L + \theta_U)/2.$$

The pattern mixture model implies a selection model.

We now illustrate the same point with a bivariate normal model. Let Y denote income; to keep the algebra simple suppose $Y \sim N(0, 1)$, and we drop the index i. Let $R = 1$ if Y is observed, but now let $X \sim N(\mu_x, 1)$ be a normally distributed variable, correlated with Y, which is positive if Y is observed, that is when $R = 1$. We specify the selection model, and derive the pattern mixture model.

Let $\Phi(\,.\,)$ be the cumulative distribution of the standard normal, and suppose we choose the selection model as

$$\Pr(R = 1|Y) = \Pr(X > 0|Y) = \Phi(\alpha_0 + \alpha_1 Y). \tag{1.10}$$

Equation (1.10) thus assumes a specific MNAR mechanism, for α_0 and α_1 cannot be estimated from the observed values of Y.

Given (1.10) and the marginal standard normal distribution of Y, the joint distribution of (Y, X) is bivariate normal:

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left[\begin{pmatrix} 1 \\ \mu_x \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right], \tag{1.11}$$

where $\rho = \text{corr}(Y, X)$. Thus we have the central term in (1.9). It follows that

$$\Pr(X_i|y_i) \sim N\{\mu_x + \rho Y_i, (1 - \rho^2)\}.$$

Thus

$$\Pr(X > 0|Y) = \Phi\left(\frac{\mu_x + \rho Y}{\sqrt{1 - \rho^2}}\right) = \Phi\left(\frac{\mu_x}{\sqrt{1 - \rho^2}} + \frac{\rho}{\sqrt{1 - \rho^2}} Y\right).$$

Comparing with (1.10) we see $\rho = g(\alpha_1)$ and $\mu_x = h(\alpha_0, \alpha_1)$. Hence (α_0, α_1) define μ_x, which in turn defines the marginal probability, $\Pr(X > 0)$, of observing Y.

From the bivariate normal (1.11) the distribution of observed income, Y given $R = 1$ is $Y|x > 0$ which is

$$\frac{\phi(Y)}{\Phi(\mu_x)} \Phi\left(\frac{\mu_x + \rho Y}{\sqrt{1 - \rho^2}}\right) = \frac{\phi(Y)}{\Phi\{h(\alpha_0, \alpha_1)\}} \Phi(\alpha_0 + \alpha_1 Y).$$

A similar result follows for the distribution of unobserved income. Putting this together, we have arrived at the pattern mixture model, the RHS of (1.9). Specification of the selection mechanism, through α_0, α_1, together with the marginal distribution of income, fixes both the marginal probability of observing income and the distribution of the two ‘patterns’ of data: the seen and unseen incomes.

This is a simple example of the Heckman selection model, which is further discussed in Little and Rubin (1987), Ch. 11. More recently, it has also been used as a model for publication bias in meta analysis (Copas and Shi, 2000a). □

The example above illustrates that when data are MNAR, instead of thinking about the selection mechanism, it is equally appropriate to consider differences between conditional distributions of partially observed given fully observed variables. Under MAR such distributions do not differ depending on whether data is missing or not; under MNAR they do. Considering the conditional distribution of the observed data, and then exploring the robustness of inference as it is allowed to differ in the unobserved data, is therefore a natural way to explore the robustness of inference to an assumption of MAR. From our perspective it has two further advantages: (i) the differences can be expressed simply and pictorially, and (ii) MI provides a natural route for inference. Unfortunately, the selection counterparts, or implications, of pattern mixture models are rarely easy to calculate directly, but again MI can help: after imputing missing data under a pattern mixture model, it is straightforward to explore implications for the implied selection model.

Example 1.3 Asthma study *(ctd)*

We illustrate the above using the 12 week data from the asthma study. Suppose first that 12 week response is MAR given treatment group. Then, in each treatment group the mean of unobserved and observed data are the same, so the treatment effect is $2.23 - 2.05 = 0.18$ litres. Suppose we have a MNAR mechanism and we express this as a pattern mixture model. Let μ_P, μ_A be the mean response under placebo and active treatment. Then

$$\mu_P = 37 \times 2.05 + (90 - 37) \times (2.05 + \Delta_P), \text{ and}$$
$$\mu_A = 71 \times 2.23 + (90 - 71) \times (2.23 + \Delta_A),$$

where Δ_P, Δ_A are respectively the mean differences between observed and unobserved response in the placebo and active group.

Figure 1.3 shows how the estimated treatment effect varies as we move away from the assumption of MAR, i.e., that $\Delta_p = \Delta_A = 0$. Since many more patients are missing in the placebo group, the treatment estimate is much more sensitive to departures from MAR in this group.

Notice the inherently arbitrary nature of MNAR: because we cannot estimate Δ_A, Δ_P from the data at hand, all possible values are – in general – equally plausible. This issue is the motivation for our proposed approach to sensitivity analysis in clinical trials of this type in Section 10.4. □

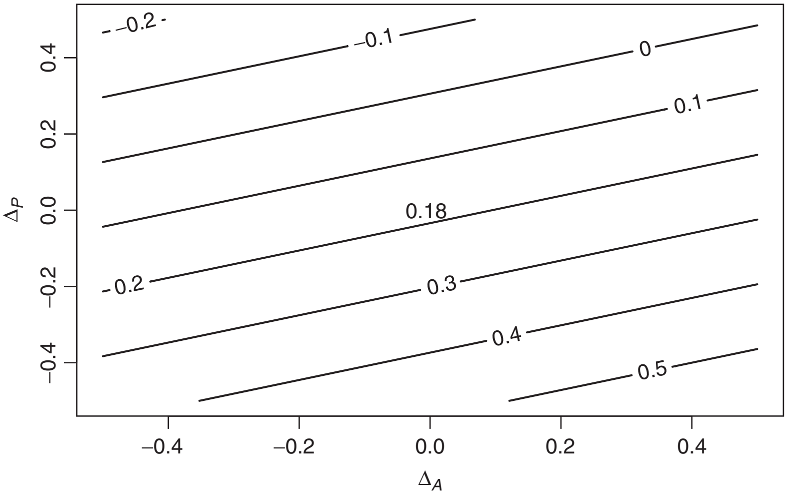


Figure 1.3 Contour plot of the difference in average FEV$_1$ (litres) between active and placebo groups, as Δ_P, Δ_A, vary. Under MAR, $\Delta_P = \Delta_A = 0$, and the difference is 0.18 litres.

1.4.4 Ignorability

If, under a specific assumption about the missingness mechanism, we can construct a valid analysis that does not require us to explicitly include the model for that missing value mechanism, we term the mechanism, in the context of this analysis, *ignorable*.

A common example of this is a likelihood based analysis assuming MAR.

However, as we see below there are other settings, where we do not assume MAR, that do not require us to explicitly include the model for the missingness mechanism yet still result in valid inference. For example, as discussed in Section 1.6.2, a complete records regression analysis is valid if data are MNAR dependent only on the covariates.

1.5 Using observed data to inform assumptions about the missingness mechanism

We have already noted that, given the observed data, we cannot definitively identify the missingness mechanism. Nevertheless, the observed data can help frame plausible assumptions about this – in other words assumptions which are consistent with the observed data. Exploratory analyses of this nature are important for (i) assessing whether a complete records analysis is likely to be biased and (ii) framing appropriate imputation models. Two key tools for this are summaries (tabular or graphical) of fully observed, or near-fully observed variables by missingness pattern and logistic regression of missingness indicators on observed, or near-fully observed variables.

Example 1.3 Asthma study *(ctd)*

Table 1.5 shows the mean FEV_1 by dropout pattern. In the placebo arm, patterns 3 and 4 have lower FEV_1 at baseline, and for patterns 2–4 FEV_1 declines from baseline to last visit. In the active arm, patterns 1, 2 show a similar increase of about 0.20 ml, while pattern 3 starts higher and shows little change, while pattern 4 shows marked decline. Notice also the increase in variance in the active arm over time which is different from the treatment arm. This is a common feature of such data, and should be reflected in the analysis.

MAR mechanisms that are dependent on treatment and response are consistent with these data. However, there is a suspicion that further decline between the last observed and first missing visit triggered withdrawal, probably followed in the placebo arm by switching to an active treatment. Thus it would be useful to explore sensitivity of treatment inferences to MNAR, which we do in Chapter 10. □

Table 1.5 Asthma study: mean FEV_1 (litres) at each visit, by dropout pattern and intervention arm.

Dropout pattern	Placebo arm						
	Mean FEV_1 (litres) measured at week					No.	%
	0	2	4	8	12		
1	2.11	2.14	2.07	2.01	2.06	37	40
2	2.31	2.18	1.95	2.13	–	15	16
3	1.96	1.73	1.84	–	–	22	24
4	1.84	1.72	–	–	–	16	17
All patients (Mean)	2.06	1.97	1.98	2.04	2.06	90	100
All patients (Std.)	0.57	0.67	0.56	0.58	0.55		
			Lowest Active arm				
1	2.03	2.22	2.23	2.24	2.23	71	78
2	1.93	1.91	2.01	2.14	–	8	9
3	2.28	2.10	2.29	–	–	8	9
4	2.24	1.84	–	–	–	3	3
All patients (Mean)	2.03	2.17	2.22	2.23	2.23	90	100
All patients (Std.)	0.65	0.75	0.80	0.85	0.81		

Example 1.2 Youth Cohort Study *(ctd)*

In Table 1.2, we saw that the principal missing data pattern has missing parental occupation. Let $R_i = 1$ if parental occupation is observed, and zero otherwise. Table 1.6 shows the results of various logistic regressions of R on the remaining fully, or near fully observed, variables: GCSE score, ethnicity, gender and cohort. The Receiver Operating Characteristic (ROC) is an assessment of how well a model discriminates between the missing and observed parental occupation, with a minimum value of 0.5 (no discrimination) and a maximum of 1. Of course, even if the model discriminated perfectly, this would say nothing about differences between observed and unobserved data, that is, whether the data are MNAR.

We see that GCSE score is the strongest predictor of missing parental occupation (ROC of 0.68), followed by ethnicity (here simplified to white/non-white) and cohort. Gender is a relatively weak predictor. Nevertheless, due to the size of the cohort, all are significant at the 5% level in model 4, which has reasonable discrimination (ROC = 0.74).

Figure 1.4 confirms that GCSE score is substantially higher among those whose parental occupation is observed (mean of 39 vs 28 points respectively). Further, 10% of children with missing parental occupation have no GCSEs (score 0) compared with 3% who have parental occupation observed.

We conclude the data are consistent with parental occupation missing at random, dependent strongly on GCSE score and ethnic group, but also associated

Table 1.6 Coefficients (standard errors), and receiver operating characteristic (ROC), from logistic models for the probability of observing parental occupation.

Variable	Models				
	1	2	3	4	5
cohort ’93	−0.085				−0.168
	(0.036)				(0.039)
cohort ’95	0.044				−0.212
	(0.038)				(0.042)
cohort ’97	0.178				−0.032
	(0.040)				(0.043)
cohort ’99	0.135				−0.165
	(0.040)				(0.046)
boy		−0.053			0.079
		(0.024)			(0.026)
GCSE score			0.037		0.038
			(0.001)		(0.001)
non-white				−1.723	−1.698
				(0.0288)	(0.031)
ROC	0.53	0.51	0.68	0.62	0.74

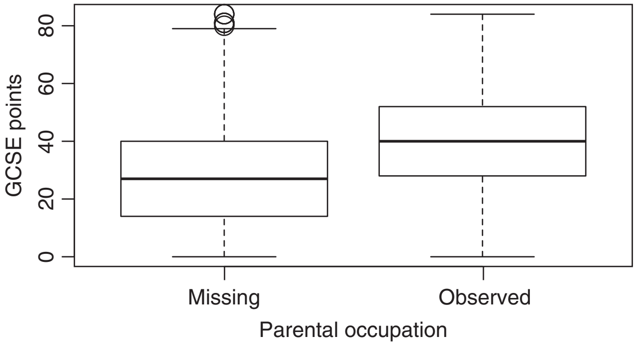


Figure 1.4 *Boxplot of GCSE points by whether parental occupation is observed.*

with cohort and weakly with gender. A relatively small number of values are missing for the other variables. It is plausible to assume these are either MCAR or perhaps MAR given observations on other variables; unless they are strongly MNAR this will have a negligible impact on subsequent inferences. □

1.6 Implications of missing data mechanisms for regression analyses

Usually, we will wish to fit some form of regression model to address our substantive questions. Here, we look at the implications, in terms of bias and loss of information, of missing data in the response and/or covariates under different missingness mechanisms. We first focus on linear regression; our findings there hold for most other regression models, including relative risk regression and survival analysis. Logistic regression is more subtle; we discuss this in Section 1.6.4.

1.6.1 Partially observed response

Suppose we wish to fit the model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad e_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n, \tag{1.12}$$

but Y is partially observed. Let R_i indicate whether Y_i is observed. For now assume that the x_i are known without error; for example it may be a design variable. Then the contribution to the likelihood for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ from unit i, conditional on x_i, is

$$L_i = \Pr(R_i, Y_i|x_i) = \Pr(R_i|Y_i, x_i)\Pr(Y_i|x_i). \tag{1.13}$$

Assume, as will typically be the case, that the parameters of $\Pr(Y_i|x_i)$, $\boldsymbol{\beta}$, are distinct from the parameters of $\Pr(R_i|Y_i, x_i)$.

Figure 1.2 suggests that, provided Y is MAR given the covariates in the model, units with missing response have no information about $\boldsymbol{\beta}$. To see this formally, first observe that as Y_i is MAR given x_i, only the second term on the RHS of (1.13) involves Y.

The contribution to the likelihood for an individual with missing response is obtained by integrating (for discrete variables summing) over all possible values of the missing response variable Y_i, given x_i. This is

$$\int \Pr(Y_i|x_i)dY_i = 1,$$

because we are integrating (summing) over all possible values of Y_i given $\boldsymbol{\beta}, x_i$ so the total probability is 1. Conditional on x, all individuals with missing Y thus contribute 1 to likelihood for $\boldsymbol{\beta}$, and so have no effect on, or information about, the maximum likelihood estimate of $\boldsymbol{\beta}$.

This may feel counterintuitive, especially if we have a large number of units with Y missing but X observed. Do they really have no information on the regression?

For linear regression, the answer is yes (there is no information), because the parameter space of the conditional distribution of Y given X is separate from that of the marginal distribution for X. In other words, the mean and variance of X have no information on, and place no restriction on, the parameters of the distribution of $Y|X$. Equivalently, the conditional distribution of $\Pr(Y|X)$ has no information on, and places no restriction on, the marginal distribution of X.

Example 1.3 Asthma study *(ctd)*

To illustrate the above, consider estimating the effect of treatment on the 12 week response, adjusting for baseline, setting aside the measurements at 2, 4 and 8 weeks. If we assume that the 12 week response is MAR given treatment, then from the above argument it follows that fitting the regression model,

$$Y_i = \beta_0 + \beta_1 1[\text{treatment = active}_i] + e_i, \quad e_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \tag{1.14}$$

using the complete records gives a valid estimate of the treatment effect (Table 1.7).

Now suppose, as Table 1.5 suggests, that baseline is also predictive of missing 12 week FEV_1; it is also strongly predictive of the actual 12 week FEV_1. Assuming 12 week FEV_1 is MAR given baseline and treatment, we can include baseline in the regression model (1.14). Again, the argument above shows that fitting this model to the observed data is valid and efficient; the results are in Table 1.7. Note that

Table 1.7 Asthma study: estimated treatment effect fitting treatment, and treatment and baseline. Inference is valid and fully efficient if assumption that data are MAR, dependent on the covariates in the model, is correct.

Covariates	n	Treatment estimate (s.e.)	p-value
Treat	108	0.172 L (0.149)	0.251
Treat & baseline	108	0.247 L (0.100)	0.016

(a) if baseline were predictive of underlying 12 week response, but given treatment not predictive of observing that response, we would still wish to include it, and

(b) in the unlikely case that baseline were predictive of missing 12 week response, but not related to the actual 12 week response value, there would be no benefit of including it.

We explore this study further, taking into account the longitudinal observations, in Chapters 3 and 7. □

The above argument extends naturally to partially observed multivariate responses. Suppose we have up to J observations on individual i, denoted $\mathbf{Y}_i = (Y_{i,1}, \dots, Y_{i,j})$. Suppose they are MAR given X_i, and – whatever the pattern of missing data – we partition $\mathbf{Y}_i$ into $\mathbf{Y}_{i,O}$ and $\mathbf{Y}_{i,M}$. Then the contribution of individual i to the likelihood for the regression of $\mathbf{Y}$ on X is

$$\int \Pr(\mathbf{Y}_i|\boldsymbol{\beta}, X_i)\, d\mathbf{Y}_{i,M},$$

in other words, the marginal likelihood of the observed data. For the multivariate normal distribution, this is readily calculated; in fact most software fits the model to the observed pattern of data by default. Once again, in this setting there is no advantage to, or gain from, using multiple imputation.

The last setting we consider in this subsection is when we have missing response data, but these data are MNAR given the variables we wish to include in the model of interest. For a direct exposition we return to univariate Y_i; the extension to multivariate $\mathbf{Y}_i$ is immediate.

Consider (1.13) and let the parameters of $\Pr(R_i|Y_i, X_i)$ be η and distinct from those of $\Pr(Y_i|X_i)$, i.e., $\boldsymbol{\beta}$. The contribution to the likelihood from individual i is

$$\int \Pr(R_i, Y_i, X_i)\, dY_i = \Pr(X_i) \int \left\{\Pr(R_i|\eta, Y_i, X_i) \Pr(Y_i|\boldsymbol{\beta}, X_i)\right\} dY_i. \quad (1.15)$$

We see the likelihood contribution for $\boldsymbol{\beta}$ is now caught up with the selection mechanism; we have to evaluate the integral on the RHS of (1.15) to obtain the contribution of individual i to the likelihood. Failure to do this leads to biased inference for $\boldsymbol{\beta}$.

Example 1.7 Linear regression

To illustrate this, we generate a sample of 200 observations from the regression model

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad e_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n \tag{1.16}$$

with $(\alpha, \beta, \sigma^2) = (5, 1, 4^2)$. These data, together with the fitted regression, are shown in Figure 1.5(a), together with the least squares fitted line, which has estimated parameters $(\hat{\alpha}, \hat{\beta}) = (5.14, 1.01)$.

Now suppose that some of the Y values are MNAR, and $R_i = 1$ if Y_i is observed and 0 otherwise. Suppose

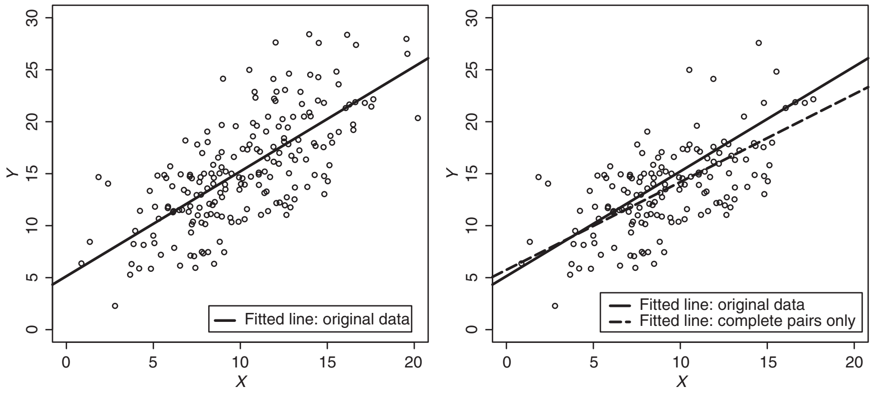
$$\Pr(R_i = 1) = \begin{cases} 0.8 & \text{if } Y_i > 18 \text{ and} \\ 0 & \text{otherwise} \end{cases} \tag{1.17}$$

Starting from the 200 observations shown in the left panel of Figure 1.5, the right panel plots a typical example of the complete records that remain under this mechanism. Fitting a regression line to the observed data gives $(\hat{\beta}_0, \hat{\beta}_1) = (5.75, 0.85)$. Because high values of Y, which correspond to high values of X, are likely to be missing, the intercept is biased slightly up and the slope down. □

Next, suppose that in addition to X, Y we have the fully observed variable Z. We suppose that Y is partially observed, we are interested in the regression of Y on X, that

$$\operatorname{logit} \Pr(R_i = 1) = \alpha_0 + \alpha_1 X_i + \alpha_2 Z_i, \tag{1.18}$$

and that Z is correlated with Y. Then, following from the discussion above, the regression of complete records Y_i on X_i will be biased, because setting Z aside



(a) Fitted regression line from the full data. $(\hat{\alpha}, \hat{\beta}) = (5.14, 1.01)$

(b) Y missing with probability 0.8 when $Y > 18$. $(\hat{\alpha}, \hat{\beta}) = (5.75, 0.85)$

Figure 1.5 Regression lines for synthetic data. Left panel: fitted regression line to the full data, $n = 200$. Right panel, original and fitted line when Y is MNAR.

Y_i is MNAR. However, the regression of complete records Y_i on X_i, Z_i will be unbiased, and efficient, because given X and Z, Y is MAR.

1.6.2 Missing covariates

We now consider the regression of Y on X, when Y is fully observed and X is partially observed.

Let $R_i = 1$ if X_i is observed and $R_i = 0$ otherwise. Consider the regression of Y on X estimated from the complete records, i.e., given $R_i = 1$. Following (1.6), for each individual pair,

$$\begin{aligned}\Pr(Y_i|X_i, R_i = 1) &= \frac{\Pr(Y_i, X_i, R_i = 1)}{\Pr(X_i, R_i = 1)}\\ &\frac{\Pr(R_i = 1|Y_i, X_i)\Pr(Y_i, X_i)}{\Pr(R_i = 1|X_i)\Pr(X_i)}\\ &\left\{\frac{\Pr(R_i = 1|Y_i, X_i)}{\Pr(R_i = 1|X_i)}\right\}\Pr(Y_i|X_i). \qquad (1.19)\end{aligned}$$

Thus, when the missingness mechanism for X, $\Pr(R_i = 1|Y_i, X_i)$, involves the response Y, restricting the analysis to the complete records gives biased point estimators and invalid inference. This holds whether the missingness mechanism only depends on Y, i.e., MAR, or whether it includes X as well, i.e., MNAR.

Example 1.6 Income *(ctd)*

Consider again the income example, but now suppose we wish to estimate the probability of job type given income, i.e., $\Pr(Y_{2,i}|Y_{1,i})$. As this is artificial data, we know the data generating mechanism:

$$Y_1 \sim N(60, 5) \text{ for job type A }, \qquad \text{and } Y_1 \sim N(30, 5) \text{ for job type B},$$

with Pr(job type A) $= \Pr(Z_2 = A) = 0.5$.

Thus

$$\Pr(Y_2 = A|y_1) = \frac{\Pr(Y_1 = y_1|Y_2 = A)}{\Pr(Y_1 = y_1|Y_2 = A) + \Pr(Y_1 = y_1|Y_2 = B)}.$$

Thus if $Y_1 = 45$, the probability $Y_2 = A$ is 0.5. In the original data (Figure 1.2) there is no overlap between the groups so again $\Pr(Y_2 = A|Y_1 = 45) = 0.5$. However, from the observed data, we estimate this as $68/(89 + 68) = 0.43$. This illustrates the general point above: in regression when covariates are MAR and the mechanism includes the response, complete records analysis is biased. □

From (1.19), we see that when the missingness mechanism for the covariate does not depend on the response Y_i the probability of $Y|X$ among the complete records is the same as that in the population. In other words, although the covariate is MNAR, estimating the regression using complete records is unbiased and gives valid inference, although a full likelihood analysis with the correctly specified selection mechanism would be more efficient. Again, we note that the precise form of selection mechanism can vary between units or individuals; its precise form is not relevant to the argument.

Example 1.7 Linear regression *(ctd)*

Continuing with this example, suppose we take the original 100 pairs and set all X values greater than 12 to missing. This is a strong MNAR mechanism, but given the (possibly unobserved) X value, the probability of X being missing does not depend on Y. Figure 1.6(a) shows the regression of Y on X fitted to the remaining points and the fitted line to the original data. They are virtually indistinguishable. Indeed using the observed points, $(\hat{\beta}_0, \hat{\beta}_1) = (5.16, 1.00)$.

Thus, as (1.19) implies, there is no bias, but some information is lost. It is also important to note that (i) in this situation an analysis under MAR would be biased but (ii) given the observed data, we cannot conclude that X is MNAR dependent only on X; indeed it would be plausible to have X MAR, or MNAR dependent on Y and X. □

Now consider the setting where the covariate, X, is MNAR depending on both X and Y. In this setting, (1.19) implies regression using the complete records will be biased.

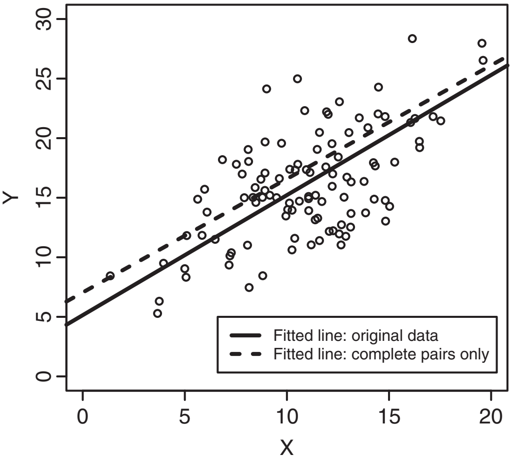


Figure 1.6 Missing covariates: effect of different mechanisms.

Example 1.7 Linear regression *(ctd)*

Continuing this example, suppose that

$$\Pr(X \text{ missing}) = \begin{cases} 0.8 \text{ if } Y < 15 \;\&\; X \;< 10 \\ 0.4 \text{ if } Y \geq 15 \;\&\; X \;\geq 10 \end{cases}$$

Figure 1.6(b) shows the results; the bias is clear. □

Lastly, consider the case where we have three variables (or sets of variables) X, Y, Z and we are interested in the regression of Y on X. Suppose X is MNAR given X, Z but that if we omit Z there is residual dependence of the missing mechanism on Y so that X is MNAR given X, Y.

In this setting, (1.19) shows us that using the complete records to regress Y on X will be biased; however, using the complete records to regress Y on X, Z will be unbiased for the latter, adjusted relationship. Unfortunately, unless Z is independent of X, so that including Z in the regression does not change the coefficient for X, it is not possible to use this to obtain a valid estimate of the regression of Y on X alone without making additional assumptions. Indeed, even if X is truly independent of Z, under a MNAR mechanism in the observed data they will typically be correlated.

1.6.3 Missing covariates and response

In our final, setting first suppose we have three variables, X, Y, Z and that Y and X are MAR given Z. Consider the linear regression of Y on X, Z. Units with X, Y missing contribute

$$\int \Pr(Y|\boldsymbol{\beta}; X, Z)\, dY = 1$$

to the likelihood $\Pr(Y|\boldsymbol{\beta}; X, Z)$. Thus, (1.19) implies the complete records analysis will be unbiased.

When we have additional variables predictive of Y and/or X then these may be used to recover information on the missing values and hence $\boldsymbol{\beta}$.

1.6.4 Subtle issues I: The odds ratio

This and the next two subsections consider some more subtle implications of the missingness mechanism for complete records analysis; some readers may prefer to skip to the summary on p. 35.

Harel and Carpenter (2012) consider the further complication that arises because some estimators possess a symmetry, which means they can be validly estimated from the complete records under a greater range of missing value mechanisms. The principal example is the odds ratio. Consider Table 1.8. We can either model A, C as binomial random variables with denominator $a + b$,

Table 1.8 Typical two-by-two table of counts relating outcome to exposure.

	Unexposed	Exposed
Good outcome	a	b
Poor outcome	c	d

$c + d$, or we can model A, B as binomial random variables with denominator $a + c$, $b + d$. In both cases estimates and inference for the odds ratio are identical. The first case corresponds to a case-control study, the latter to a cohort study.

Now suppose that the probability of outcome is MNAR dependent on only outcome. Consider the model

$$\operatorname{logit}\Pr(\text{good outcome}) = \beta_0 + \beta_1 \times 1[\text{exposed}].$$

The preceding discussion would lead us to suppose that both β_0 and β_1 will be biased. In fact, β_1 will be unbiased. Symmetry of the odds ratio means inference for this is the same as if we performed a logistic regression of exposure on outcome where outcome was MNAR dependent only on outcome. However this is an example of a covariate MNAR, and (1.19) shows that inference using the complete records is valid in this case. The same argument applies if exposure is MAR given outcome. Bias will only occur when estimating β_1 if data are MNAR dependent on both the outcome and covariate.

More generally, we will wish to estimate the log-odds ratio relating outcome, Y, to X for various possible confounders, say Z. Applying the above argument, Y may be MNAR dependent on itself and Z, yet the OR relating X to Y will still be validly estimated from the complete records. Or, Y may be MNAR dependent on itself and X, and then the OR relating Z to Y estimated using the complete records is still valid. However, if the MNAR mechanism depends on Y, X, Z, inference from the complete records is generally biased. This argument extends naturally to log-linear models for multi-category, rather than just binary, classifications.

Example 1.8 Odds ratio

Consider synthetic data relating binary outcome, Y, to binary X and a continuous Z. We generate $1 = 1, \ldots, 20,000$ observations as follows:

$$x_i = \begin{cases} 1 \text{ for } i = 1, \ldots, 10000 \\ 0 \text{ for } i = 10001, \ldots, 20000 \end{cases},$$

$Z \sim N(0.5 \times (x_i - 0.5), 1)$ and

$$\operatorname{logit}\Pr(Y_i = 1) = \beta_0 + \beta_1 X_i + \beta_2 Z_i. \tag{1.20}$$

where $(\beta_0, \beta_1, \beta_2) = (0, 1, 1)$.

Table 1.9 Missing data mechanisms, and bias of coefficient estimates with typical regression and logistic regression.

Mechanism depends on	Biased estimation of parameters using complete records					
	Typical regression			Logistic regression		
	constant	coeff. of X	coeff. of Z	constant	coeff. of X	coeff. of Z
Y	Yes	Yes	Yes	Yes	No	No
X	No	No	No	No	No	No
Z	No	No	No	No	No	No
X, Z	No	No	No	No	No	No
Y, X	Yes	Yes	Yes	Yes	Yes	No
Y, Z	Yes	Yes	Yes	Yes	No	Yes
Y, X, Z	Yes	Yes	Yes	Yes	Yes	Yes

We may consider either Z or X as the exposure. The relationship is confounded, so the unadjusted odds ratios are both biased.

Table 1.9 shows the mechanisms we consider, the bias we expect from a complete records analysis in a typical regression setting, and what we expect when using logistic regression (i.e., when we estimate log-odds ratios).

Notice that the bias does not depend on which variable has missing data, but instead on the mechanism that differentiates, or selects, the complete records from the rest of the sample. However, the appropriate approach for handling the bias (e.g. multiple imputation) will depend on the variable that is actually missing. For example, if the mechanism depends on Y and X is partially observed, data are MAR.

The results of fitting the logistic regression (1.20) for the seven scenarios in Table 1.9 are shown in Table 1.10. We see that when missing data depends on Y, odds ratios for coefficients are only biased if, in addition, the missingness mechanism depends on the covariates associated with those coefficients. This is a consequence of the symmetry of the logistic link. □

1.6.5 Implication for linear regression

Harel and Carpenter (2012) further show that, to first order approximation, the results for the odds ratio hold for linear and probit regression. In the case of linear regression of Y on X, Z, if missingness depends on Y and Z and the correlation between Y and X is moderate ($|\rho| < 0.75$), then when we estimate the regression using the complete records (i) the largest bias occurs for the coefficient for Z, but (ii) the coefficient for X is markedly less (but not completely) unbiased. As above, this applies even if the actual missing values occur in the variable Z.

Table 1.10 Empirical illustration of Table 1.9 using logistic regression.

Mechanism depends on	Probability of complete record	Estimated coefficients of		
		constant	X	Z
–	1	−0.03	1.03	1.03
Y	$[1+\exp(-y)]^{-1}$	0.34	0.99	1.01
X	$[1+\exp(-x)]^{-1}$	−0.04	1.02	1.00
Z	$[1+\exp(-z)]^{-1}$	−0.03	0.96	1.03
X, Z	$[1+\exp\{-(0.5(x-0.5)+z)\}]^{-1}$	−0.04	0.98	1.03
Y, X	$[1+\exp\{-(y+2(x-0.5))\}]^{-1}$	0.58	0.58	0.99
Y, Z	$[1+\exp\{-(y+z)\}]^{-1}$	0.38	0.96	0.82
Y, X, Z	$[1+\exp\{-(y+2(x-0.5)+z)\}]^{-1}$	0.63	0.58	0.81

This gives an informal guide to the difference between the coefficient estimates we might expect from a complete records analysis and those from an MAR analysis (typically obtained using MI). Because analysis under MAR, whether by MI or another route, is relatively complex – and thus relatively more prone to error – this provides a useful check on the plausibility of the results.

Related to this, Daniel *et al*. (2012) show how causal diagrams can be used to explore where bias due to missing data may arise. This can be a useful practical guide, both to whether it is worth using MI and to whether the results are consistent with the assumed missingness mechanisms.

Example 1.2 Youth Cohort Study *(ctd)*

Table 1.6 suggests that missing parental occupation depends on GCSE score and ethnicity. The above argument suggests that it is the coefficient for ethnicity that is most likely to be biased in the complete records analysis. After we have described MI for a range of data types, we return to this example at the end of Chapter 5. □

1.6.6 Subtle issues II: Subsample ignorability

Little and Zhang (2011) describe the related idea of subsample ignorable likelihood. Suppose we have four (sets of) variables, and the pattern of missing data shown in Table 1.11. We now make the *subsample ignorability*, that is:

1. within pattern 2, missing values of X and Y are MAR, and
2. within pattern 3, W is MNAR, with a mechanism that does not depend on Y.

Table 1.11 Missing data patterns for subsample ignorable likelihood. As before, '✓' denotes observed, '·' missing, and now '✓/·' denotes some observed and some missing.

Pattern	Variables				Number of observations
	Z	W	X	Y	
1	✓	✓	✓	✓	n_1
2	✓	✓	✓/·	✓/·	n_2
3	✓	·	✓/·	✓/·	n_3

Consider regression of Y on X, W, Z. Using the arguments developed earlier in this chapter, we see a complete records analysis will be invalid, because for observations in pattern 2 the missingness mechanism includes the response. Also, an analysis assuming MAR using observations from all three patterns will also be invalid, because data are MNAR in pattern 3. However, using only data from patterns 1 and 2, the missingness mechanism is MAR; therefore an appropriate analysis (e.g. using multiple imputation) in this setting gives valid inference. In essence this is a partial likelihood analysis, where the MNAR component is set aside.

Thus, by careful consideration of the reasons for missing data, we may be able to get valid inference via MI without recourse to a full MNAR analysis, even if a portion of the data are MNAR. A more formal justification of this approach is given by Little and Zhang (2011), who also present some simulations confirming the validity of inference when the subsample ignorability assumption holds, together with an example.

1.6.7 Summary: When restricting to complete records is valid

We have considered above the impact of various missing data mechanisms on regression analyses restricted to complete records. We note that restricting the regression analyses to complete records is generally invalid when the missingness mechanism includes the response. In establishing this, notice that what is important is the variables in the missingness mechanism, rather than variables with the missing data.

Consideration of the variables with missing data is important when deciding how to proceed beyond a complete records analysis. For instance, suppose the missingness mechanism depends on a covariate X and response Y but not on a third covariate, Z. Two possibilities are

1. Y partially observed, and

2. Z partially observed

In case (1) data are MNAR, so an analysis under the MAR assumption (e.g. using multiple imputation) will not be strictly valid. In case (2) we have a covariate MAR, so an analysis under MAR (e.g. using multiple imputation) will be valid.

In case (1), analysis under MAR may nevertheless be less biased, and the sensitivity to MNAR can be readily explored using multiple imputation, as we discuss in Chapter 10.

1.7 Summary

This chapter has introduced the central concepts involved in the analysis of partially observed data. These revolve around the 'reason for the data being missing' – more formally the missingness mechanism, and how this relates to the inferential question at hand. We have described Rubin's typology of missing data mechanisms (Rubin, 1976) and discussed these in the context of regression analysis.

We have stressed the importance of preliminary analysis of the data to identify the principal missingness patterns and elucidate plausible missingness mechanisms. Under particular missingness mechanisms, we have further explored when a regression restricted to complete records analysis is likely to give valid (if inefficient) inference.

The remainder of this book is concerned with using MI to obtain valid inference from partially observed data, predominantly the under assumption of MAR but also under the assumption of MNAR. However, there are a number of other methods that could be used to do this, for instance, the EM algorithm or a full Bayesian analysis (Clayton *et al.*, 1998). Why MI? The answer is because it is practical for applied researchers in a wide range of settings. The EM algorithm for parameter estimates is not computationally straightforward in general. Further it does not yield standard errors; a further step is required for this. A full Bayesian analysis usually requires specialist programming and will often be computationally demanding, particularly if a range of models have to be fitted.

By contrast, using multiple imputation, the researcher has to specify an appropriate imputation model. Robust software exists in many packages to fit (or approximately fit) this model, from which a series of say K imputed data sets are created. Assuming this has been done properly, the researcher can then fit their model of interest to each of the K imputed data sets in turn, obtaining K point estimates and standard errors. These are combined for final inference using Rubin's rules (Little and Rubin, 1987). These rules are relatively straightforward and perform remarkably well in a wide range of settings.

Thus, once the imputations have been created, inference proceeds using the usual software for fitting the model of interest to the complete records. It is therefore rapid. Further, analysis is not restricted to a single model: a range of models compatible with the imputation model can be explored. In addition, variables that the researcher does not wish to include in the model of interest (e.g. because they are on the causal path) can be included in the imputation model, improving both the plausibility of the MAR assumption and the imputation of the missing values.

The next chapter therefore introduces MI and sketches out its theoretical basis, illustrating this using linear regression. Subsequent chapters describe both algorithms for and application of MI to a broad range of social and medical data.